Learning to translate land-cover maps

Luc Baudoux

Université Gustave Eiffel

Learning to translate land-cover maps: several multi-dimensional context-wise solutions

Thèse de doctorat de l'Université Gustave Eiffel

École doctorale n° 532, Mathématiques, Science, et Technologie de l'Information et de la Communication (MSTIC)
Spécialité de doctorat : Signal, Image, et Automatique
Unité de recherche : Laboratoire des Sciences et Technologies de l'Information Geographique (LASTIG), IGN.

Thèse présentée et soutenue à l'Université Gustave Eiffel, le 10/12/2021, par :

LUC BAUDOUX

Composition du Jury

Begüm Demir Professeur, Université de technologie de Berlin, Allemagne	Présidente du jury
Yuliya Tarabalka Chargée de recherche, LuxCarta, France	Examinatrice
Alexis (Lex) Comber Professeur, Université de Leeds, Royaume-Uni	Rapporteur
Germain Forestier Professeur, Université de Haute-Alsace, France	Rapporteur

Encadrement de la thèse

Clément MALLET Docteur, Université Gustave Eiffel/IGN/ENSG, France	Directeur de thèse
Jordi INGLADA Docteur, CNES/CNRS/IRD/INRAe/UPS, France	Co-Directeur de thèse

"The best translations are always the ones in the language the author can't read." - Jorge Amado

Acknowledgments

The great Jacques Rouxel once said, "It is better to mobilize your intelligence on bullshit than to mobilize your bullshit on intelligent things". Although I have often found myself doing the opposite over the past three years, I have always been able to count on the wise comments and advice of my two PhD advisors to get me back on the right track. I, therefore, address my heartfelt gratitude to Clément Mallet and Jordi Inglada for their unwavering ability to mobilize their intelligence on my bullshit.

I am happy to have been able to share these three years with all of my young engineering and researcher colleagues in the lab: Charly, without whom my incessant tea breaks would have seemed very lonely, but also Anatol, Azelle, Damien, Emile, Ekaterina, Florent, Gregoire, Helen, Medhi, Melvin, Romain, Solen, Teng, Yanis and Yizi with whom I had the chance to share beautiful moments of daily life. Although I cannot name them all because there are too many of them, I would also like to thank all of the "old" fellow researchers, especially Loic, with whom I was lucky enough to share an office until COVID and these long months of remote-working and Laurent for being such a good loser in our competition for whoever arrives the earliest at work.

Of course, I would like to thank IGN, CESBIO and ANR for making this thesis possible. Nonetheless, I am also thankful to the ENSG and its teaching team for allowing me to experiment with my teaching ability on friendly, motivated and, above all, unable to escape students.

Finally, the greatest Professor of all time, Albus Dumbledore, once said, "Happiness can be found, even in the darkest of times, if one only remembers to turn on the light". Thanks to my family and friends for having been this light and, in particular, to Serge, who was my beacon in the darkest hours of this thesis.

Résumé

La description de la couverture biophysique des surfaces terrestres, appelée occupation du sol, est d'une importance capitale dans de nombreux domaines, allant de l'urbanisme aux études climatiques en passant par la sécurité alimentaire. Historiquement produites à la main, les cartes d'occupation du sol ont profité de l'essor de l'imagerie satellitaire et des méthodes avancées de vision par ordinateur pour gagner en précision et en fréquence de mise à jour. Elles souffrent toutefois de deux inconvénients limitant leur utilisation. D'une part, la résolution spatiale des cartes produites est fixe. Or une carte d'une résolution de 10 mètres ne conviendra pas à l'analyse de phénomènes à grande échelle, ni à l'étude d'objets de moins de 10 mètres. D'autre part, la nomenclature de la carte est choisie pour répondre à un besoin spécifique qui ne correspond pas nécessairement aux besoins d'un autre utilisateur. Ainsi, une carte peut regrouper sous le terme "bâti" un ensemble d'éléments tels que des "routes" et des "habitations", qui dans d'autres nomenclatures seront classés séparément.

Les approches actuelles de traduction de nomenclatures sont principalement fondées sur des méthodes de traduction sémantique (LCCS...) appliquées au niveau de la nomenclature en comparant les définitions de classes (la classe "blé" sera traduite en "herbacée"). Ce faisant, elles négligent le fait que deux objets de la même classe peuvent être traduits différemment en fonction, par exemple, de leur contexte spatial ou de leur évolution temporelle. En outre, la traduction de la résolution spatiale est généralement traitée distinctement de la traduction de nomenclature alors que ces deux notions sont intimement liées (un arbre seul ne peut pas être traduit en "forêt").

Cette thèse aborde ce problème en proposant des méthodes de traduction contextuelle augmentant les possibilités de réutilisation et de génération de nouvelles occupations des sols. Dans un premier temps, nous proposons différentes stratégies, principalement fondées sur des réseaux de neurones à convolution apprenant à traduire une carte source en une carte cible en fonction du contexte. Nous montrons l'importance cruciale du contexte spatial et géographique (une forêt en montagne est probablement constituée de conifères) sur de multiples exemples de traductions. Dans un deuxième temps, partant du constat que les modèles de traduction multi-langues donnent de meilleurs résultats que ceux entraînés à traduire d'une seule langue source vers une seule langue cible, nous proposons un modèle de traduction multi-cartes permettant d'obtenir plusieurs nomenclatures cibles à partir d'une carte source. Nous montrons que ce modèle permet d'obtenir des résultats plus robustes que les modèles entraînés sur une seule traduction, en particulier sur des cartes avec peu d'échantillons d'entraînement. Troisièmement, nous expérimentons différentes configurations de fusion multimodale fusionnant des images satellites (optiques et radar) et des données d'élévation du terrain avec des cartes d'occupation du sol. Enfin, nous définissons la notion et proposons une méthode pour construire un espace de représentation

sémantique commun à toutes les occupations du sol. Nous ne représentons alors plus la traduction comme le passage d'un espace de représentation discret à n classes (une nomenclature) vers un autre espace, mais comme un simple changement d'interprétation d'un espace de représentation sémantique continu commun à toutes les nomenclatures. Nous proposons une première application de la notion d'espace de représentation sémantique à la traduction, en nous concentrant sur la traduction de cartes sources non vues pendant l'entraînement du modèle de traduction. Les codes et jeux de données (France entière, six cartes d'occupation du sol, images satellite, vérité terrain) produits au cours de cette thèse sont rendus accessibles pour la reproductibilité et des comparaisons futures.

Abstract

The description of the bio-physical coverage of the Earth's surface, termed land-cover, is of utmost importance in recent decades in many areas, ranging from urban planning to climate studies and food security. Historically manually produced, land-cover maps now take advantage of the recent boom of satellite imagery and computer vision techniques to gain more accuracy and higher update frequency. However, they still suffer from two disadvantages limiting their use. On the one hand, the land cover map spatial resolution is fixed, while a map at 10-meter spatial resolution will not be suitable for analysing large-scale phenomena, nor for monitoring objects less than 10 meters. On the other hand, the map nomenclature is chosen to meet a specific need which does not necessarily suit another user's needs. For instance, a nomenclature may group under the term "built-up areas" a set of elements such as "roads" and "dwellings", which other nomenclatures may classify separately. Current approaches target to adapt these nomenclatures and spatial resolutions. They are mainly based on pure semantic translation methods (LCCS...) applied at the nomenclature level by comparing class definitions. In doing so, they neglect that two objects of the same class can be translated differently depending, for instance, on their spatial context or temporal evolution. This thesis addresses this interleaved problem by proposing context-wise translation methods to increase re-use possibilities and new land-cover map generation. First, we propose different strategies, mainly based on convolution neural networks, learning to translate a source map into a target map context-wisely. In particular, we show the crucial importance of taking into account spatial and geographical contexts (a forest in the mountains is probably occupied by conifers) on multiple translation cases. Secondly, based on the observation that multi-language translation models provide better results than those trained to translate from a single source language to a single target language, we propose a multi-map translation framework allowing us to obtain several target nomenclatures from a unique source map. We show that this model allows for more robust results than models trained on a single translation, especially on maps with limited training samples. Thirdly, we experiment with different multi-modal fusion configurations merging satellite images (optical and radar) and elevation data with land-cover maps. Finally, we define the concept of, and propose a method to build, a semantic representation space common for all land-cover maps, no longer representing the translation as the transformation from a discrete representation space with n classes (a nomenclature) to another but as a simple change of the interpretation of a continuous semantic representation space shared between all nomenclatures. We propose the first application of the concept of common semantic representation space to translation, focusing on the translation of source maps unseen during the translation model training. The codes and datasets (France-wide, six land-cover maps, satellite imagery, and hand-annotated ground truth) produced during this thesis are also accessible for reproducibility and potential comparison purposes.

Contents

CHAPTER 1

Introduction

Land-cover (LC) is the "(bio-)physical material over the Earth's surface" [105]. As such, the land-cover is both spatially continuous (can be observed at an infinite number of different scales) and semantically continuous (an infinity of biophysical variables, such as humidity, height, biomass or chemical composition, describes each object). Modelling land-cover information requires using a categorisation method to represent this "infinite information" in a finite set by defining one (or, more rarely, several) scales of observation and a finite number of biophysical categories [321], *e.g.* forest, water... Land-cover classification aims to obtain this categorisation while preserving geolocation information. The product resulting from this process is called a *land-cover map*. This selection and combination of a finite number of biophysical variables into a finite number of groups, called classes, is referred to as the land-cover map *nomenclature*. Similarly, we speak of spatial resolution to describe the scale of observation adopted.

The knowledge of the elements comprising the soils and the subsoils of the Earth's surface is fundamental in many fields of research [104, 171, 255]. Consequently, over the last 30 years, Land-cover maps have become a mandatory baseline for monitoring the Earth's surface status and dynamics. They are, for example, used in areas as varied as urban planning [188] to better understand the mechanisms of artificialization [356], for the study of the climate since they allows albedo estimations [308] or even food security by estimating the proportion and type of cultivated areas [3]. It is also a large-scale monitoring tool for agricultural and environmental protection policies [64]. For that purpose, despite being a notoriously time-consuming procedure [250], many land-cover map products have been generated, covering the entire Earth's surface multiple times, at several spatial scales, and with various nomenclatures [106]. Therefore, the production of land use maps represents an economic, environmental, and scientific challenge.

Land-cover maps are generally tailored to offer a specific nomenclature and spatial resolution that meet a given user needs [292]. In the following sections, we focus on the challenges associated with the design and use of these maps. In particular, we show the considerable impact the choice of spatial resolution and nomenclature can have on the the map's potential applications. From this observation, several approaches, termed land cover translation (or harmonisation), have been proposed during the last 40 years to relax the

constraints related to the choice of a nomenclature and a fixed resolution. The translation aims to transform a source land-cover map nomenclature and resolution into a target one.

Current approaches, illustrated, for instance, by the well-known LCCS framework [68], assume that all elements of a given source class have the same possible translation. By analogy with language translation, we argue that those approaches act precisely like a *word-by-word* translation, as a given source class is always translated into the same target. Keeping this analogy, this manuscript focuses on breaking the *word-by-word* translation paradigm by incorporating context information, *e.g.* a forest in a mountainous area should not necessarily be translated identically as a forest near the sea. In this perspective, this manuscript identifies different context elements that can be beneficial for translation based on local spatial context (*Herbaceous* near *Water* might be *Wetlands*), geographical location or temporal constraints and propose ways to incorporate them into an operational translation framework.

Performing a manual analysis of the tremendous number of contextual elements and determining how to use them to improve translation is a time-consuming procedure not usable under operational constraints as a new analysis is required for each new source or target map. Instead, this manuscript explores data-driven strategies in which the contextual elements and how to use them are directly learned from existing source and target map samples.

Over the last decade, deep learning has been at the heart of significant advances in various disciplines, such as computer vision and natural language processing, thanks to its good results on various data [176] such as images [168] and text [318]. In general terms, deep learning can be seen as one of the numerous machine learning methods, *i.e.* as a set of automatic algorithms determining input data characteristics leading to a targeted result. The main advantage of deep learning methods lies in their ability to self-extract groups of well-tailored features from the input data to answer the problem [101]. This allows them to respond effectively to the problem posed without requiring a "manual" definition and computation of those features. This manuscript leverages this automatic feature extraction to extract contextual information for translation without explicitly defining the contextual elements to take into account.

As this thesis is to the best of our knowledge, the first to explores the potential of modern deep learning methods to perform land-cover map nomenclature and resolution translation, we first describe in the following sections the precise framing of the land-cover map translation. In particular, we describe the main characteristics of the land-cover maps and identify the different contextual information that can be used to achieve accurate translations. We also formalise the problem of land-cover map translation into three increasingly difficult problems. Finally, we present an overview of the contributions of this thesis.

1.1 Key concepts in land-cover mapping

1.1.1 A brief history

Old maps, such as the 1665 Atlas of Joan Blaeu [25], have always included some land-cover information as geographical markers, *e.g.* after the third forest turn on the right. As current topographical maps, they were designed as an abstract representation of reality in which elements are amplified, simplified or even not represented depending on the intended use of the map. For instance, Joan Blaeu's [25] atlas focuses on the road network and considerably simplifies forest geometry representing only specific ones visible from the road.

It will be necessary to wait for the development of aviation and photography, alleviating the need for the tedious collection of in-situ observations, for land-cover to become a proper cartographic subject resulting in the first real land-cover maps [207], *i.e.* geometrically accurate and spatially continuous land-cover focused cartographic products [192]. Often driven by military needs, those first land-cover focused maps were designed manually by photo interpreters analysing the images collected with a nomenclature focusing on potential obstacles for soldiers [194]. Due to the tedious aspect of this operation, Land-cover maps were mainly produced solely on a small spatial extent and never updated. Note that photo interpretation is still used nowadays for its outstanding quality results, which are difficult to match even with the current state-of-the-art image automatic analysis [296]. It not only provides some of the land-cover maps considered as references [118] but is also used for the constitution of data sets used for the automation of land-cover production and validation[189].

In the 1970's, the rise of computing science and the first satellite images [73, 81] launched the era of automatic remote sensing, *i.e.* the production of these maps by algorithms analysing images. The first algorithms were either based on physical modelling [146], requiring to manually define explicit sets of rules to translate an image pixel into a class ("a pixel with a high blue value is probably water") or on statistical resemblance analysis [35]. In parallel, the field of machine learning, which we can define as algorithm learning to "automatically improve through experience" [217], emerged and progressively replaced explicit sets of physical rules with implicitly learnt ones. The machine learning algorithms used for remote sensing mainly belong to the supervised learning paradigm. They learn the features responsible for class assignment on pre-existing pairs of image/map examples. Once trained, the algorithm can be applied to produce maps on new images. The machine learning community referred to the transformation of the almost continuous image information into a set of discrete classes as classification [166]. Therefore land-cover mapping is often termed land-cover map classification. These automated methods exhibited the advantage of their processing speed. They allowed the analysis of ever-larger geographical areas announcing the arrival of the first global-scale maps [191].

In recent years, the exponential improvement of satellite image resolution fostered the

emergence of a new paradigm replacing the independent per-pixel classification with a neighbour-aware pixel classification. This new paradigm called semantic segmentation, combined with the arrival of ever more efficient learning algorithms enabled the emergence of highly accurate maps with rich nomenclatures, high spatial and temporal resolution [334]. However, these improvements are made at the cost of an ever greater need for training data, particularly by analysing large time series of multi-spectral images via ever more resource-intensive algorithms [239]. As a result, producing land-cover maps remains a tedious process, based on advanced technologies and implying significant investment [203].

Ensuring that each map can be used for the highest possible number of applications is essential. From this observation rose a new field of study between the '70s and the beginning of the century: Land-cover map standardisation [119]. Standardisation aims to codify the nomenclature conception by using a shared vocabulary to facilitate the interoperability between LC. Considerable research was conducted on standardisation resulting in various solutions such as the famous Land-Cover Classification System (LCCS) [67], its improvement Land Cover Meta Language adopted as (ISO: 19144-2), and more recently, the EAGLE project [9]. However, none of those frameworks has been universally adopted by the remote sensing community for many reasons detailed in [54, 145]. Amongst those reasons, one of them is that, although indisputably necessary, nomenclature standardisation mainly focuses on methods applicable when creating new maps. They often do not or only partially adapt to the pre-existing maps and does not address resolution translation.

This thesis focuses on increasing the potential use of land-cover maps by presenting translation methods adapting the resolution and nomenclature of existing maps. The following sections introduce the notions of nomenclature and resolution and explain all the related issues.

1.1.2 Nomenclature

General definition A map's nomenclature (often termed classification system/class set/classification/legend) describes the categorisation of the infinite set of existing objects into a finite number of classes. Each class is described by a textual definition, which explains its expected content. Let U be the universal set of all possible objects on the ground surface, and let $n_1, n_2, ..., n_m$ be a nomenclature with m classes. The formulas below summarise some essential properties of the nomenclature.

Equation 1.1, referred as the completeness formula, states that the union of all the classes inside a nomenclature gives all the possible land-cover map types. Conversely, it ensures that each existing object on the Earth's surface can be associated with at least one class [79]. In other words, there must be no element that does not fit into the nomenclature. It is essential to specify that this principle often only applies to objects included in the map's spatial extent, which makes it challenging to use the nomenclature of a land-cover map covering a spatial extent A to one covering another spatial extent B. Moreover, some land-cover maps like Corine land-cover [121] defined their nomenclature some decades

ago and can have difficulty mapping recent land-cover types such as agrivoltaism surfaces mixing solar panels and pastoral activities. This results in potential no-data areas that are not reflected by Equation 1.1.

$$\bigcup_{i=1}^{m} n_i = U. \tag{1.1}$$

Equation 1.2, referred to as the unicity formula, states that the intersection of two classes gives the empty set or, in other words, that there should be no overlap between the definitions of the different classes [68]. It guarantees the unambiguity of the attribution of a class or that an object can only belong to a single class. A nomenclature's content stems from a compromise between the user needs and the obtainable accuracy given the algorithms and data used.

$$n_i \bigcap_{\substack{i,j=1 \\ i \neq j}}^{m} n_j = \emptyset. \tag{1.2}$$

Nomenclatures are traditionally organised hierarchically. The first level of the nomenclature has a minimal number of classes that can be divided into other classes when considering further levels (see Figure 2.1).

Land-cover and Land-use From the 1920's, we began to build an actual formalisation of the semantic content expected in a land-cover map, among other things, via the work of [266]. Foreseeing the economic potential of this type of mapping, particularly in terms of agricultural taxation, the author proposed to distinguish land-cover (the biophysical categories mentioned above, *e.g.* "forest") from land use (the anthropic use of land-cover, *e.g.* "forestry") which he believed to have more potential use cases. This is the first of many works to raise this distinction which agitated the world of cartographers and then the remote sensing community throughout the century [192]. Even though the transition from one to the other may seem trivial at first glance, the complexity stems from the multitude of possible associations between the cover and use [193]. For instance, grass-occupied soil can have various associated uses, like agricultural or recreational use (stadium) or even no use (natural meadow). [80] note that while most maps produced between the early XX$^{\text{th}}$ century and the 1970's were land-use oriented, ulterior maps were more land-cover oriented. They link this phenomenon to the evolution of mapping methods and, in particular, the arrival of the first low-cost, large-scale satellite images (Landsat-1) and the rise of computing, making possible their automatic interpretation. Their low spatial resolution compared to aerial images complicated land-use inference, which generally requires an analysis of the textural aspect of the image. Consequently, a progressive transition between land-use and land-cover mapping is observed due to the impossibility of determining the corresponding use.

The beginning of the XXIst century is marked by considerable improvement in satellite imagery resolution (Ikonos 2001, Quickbird, Spot5 2002), making use and cover analysis possible. Although the mixing of land-cover map and land use terminologies is pointed out as weakening the use of one or the other [53], most maps from then on mixed the two notions. Thus, in its current acceptance, the land-cover map nomenclature includes both the cover and use. The choice of the proportion of incorporated land-use class stems mainly from a compromise between preserving a high accuracy and responding to end users' needs.

1.1.3 Resolution

In remote sensing, the term resolution can either designate the: (i) temporal resolution, (ii) spectral resolution (iii) radiometric resolution or (iv) spatial resolution [238]. By analogy, we describe below the notion of resolution for a land-cover map.

Temporal resolution Traditionally describes the time step between two image acquisitions over a given area. By analogy, temporal resolution describes the time-step between two versions of a given land-cover map over a given area in this land-cover map oriented PhD manuscript.

Spectral resolution Traditionally describes the width and number of bands in the electromagnetic spectrum acquired by the imaging sensor. In this land-cover oriented manuscript, an analogy with nomenclature could be made. However, the remainder of the manuscript never refer to spectral resolution to avoid confusion and improper use of this term.

Radiometric resolution Sensitivity to slight energy difference in the image. Not applicable to this land-cover oriented manuscript.

Spatial resolution Describes the smallest distinguishable object by the sensor. By extension, spatial resolution is also used to qualify the smallest object visible in an image while being potentially different from the sensor resolution. To clarify the presentation, the term resolution always designate the spatial resolution in the remainder of this PhD manuscript unless mentioned explicitly.

Land-cover map spatial resolution General definition Cartographic spatial resolution is commonly defined as the smallest element visible [86] on the map. Unlike printed maps, where spatial resolution is assessed through a cartographic scale, *e.g.*. 1cm on the map represents 1km in the real world, land cover maps are intended to be used in a digital format to preserve precise geometric information. Consequently, Land-cover map spatial resolution is directly computed on the digital data and traditionally expressed in square meters, *e.g.*. the smallest element visible in the data is 100m^2 on the ground.

For a map in raster format, it is regularly associated with the spatial extent covered by one of its pixels. This assumption makes sense on maps resulting from automatic remote sensing. The classification is generally performed at a pixel level fostering a spatial resolution identical to the imagery used. Conversely, it is typically false for maps resulting from photo interpretation as they usually process the classification at an object-level, *i.e.* on a group of pixels, by manually drawing land-cover areas on images. To ensure both the global consistency of the product and reasonable conception time, they represent land-cover only if it covers a sufficient area, that is, an area superior to a fixed threshold. This threshold is referred to as the minimum mapping unit (MMU) and might differ significantly from the pixel size of the rasterised map version provided [267]. MMU might vary depending on the considered class or the spatial context, *e.g.* it is often different inside urban areas [322]. MMU can also include more constraints than the simple area threshold, such as width threshold on linear structures [164] or threshold on class proportion. Geometric and semantic information is partially lost due to the MMU, as multiple pixels with different classes might be grouped to achieve a sufficient spatial extent. As this thesis processes rasterised version of maps, we clearly distinguish the pixel resolution from the minimum mapping unit. As a way to simplify future discussion, we use the term *object* to describe either a pixel or a group of pixels. For instance, we state that "the translation is conducted at an object level" to designate that we transform the characteristics of each pixel/group of pixels independently, depending on their specific features. In opposition, we state that a "translation is conducted at a map or nomenclature level" when all objects with the same class are translated the same way.

Those spatial resolutions are strongly correlated to the nomenclature as the observation of a given class can only make sense within a specific range of resolution. For instance, the notion of "individual tree" does not make sense at one km^2 resolution, while "forest" does. In addition, the resolution of a map strongly constrains its use. As the resolution of the map increases, more phenomena will be identifiable, *e.g.* observing an urban heat island at a $1km^2$ resolution prevents fine-grain analysis. Additionally, spatial resolution tends to be constrained to the user's needs. Indeed, working with highly resolved maps is difficult when studying large-scale phenomena (such as climate-related topics) since it represents a massive volume of data and can behave like noisy data by adding too much information compared to the actual need.

1.2 Problem statement

This section introduces the precise framing in which we address land-cover map translation. Specifically, we provide insight into current issues with land-cover map translation methods and show the different ways in which land-cover map translation can be addressed.

We define land-cover map translation as the procedure aiming to transform simultaneously an existing map resolution and nomenclature into a target one. We argue that translation

is of significant interest for many downstream tasks as suggested by the plethoric number of papers using it for tasks as diverse as land-cover map fusion[285] comparison [204] or change detection [135, 208]. We identified the main potential applications requiring translation in the Figure 1.1 described below:

A The translation of an old source map into the same nomenclature and resolution as a more recent one enables studying land-cover **change detection** on matching classes and resolution.

B Translating a recent source map to the same nomenclature and resolution as an old target provides an **update** version of the old map.

C Translating a high quality land-cover map into a target nomenclature and resolution can provide **validation** samples to evaluate the quality of a target map.

D Translation can be used to **harmonise** multiple source data into a single target nomenclature resolution to achieve a downstream task such as land-cover fusion.

E Translation can be used to **complete** a small extent target map by incorporating the results of the translation of a large spatial extent source land-cover map.

F Translation can be used to **simplify the spatial resolution** of a highly resolved source into a coarser resolved one.

G Conversely, translation can be used to **improve the source resolution**.

H Translation can be used to, **modify the nomenclature** of the source, add new classes, merge some classes.

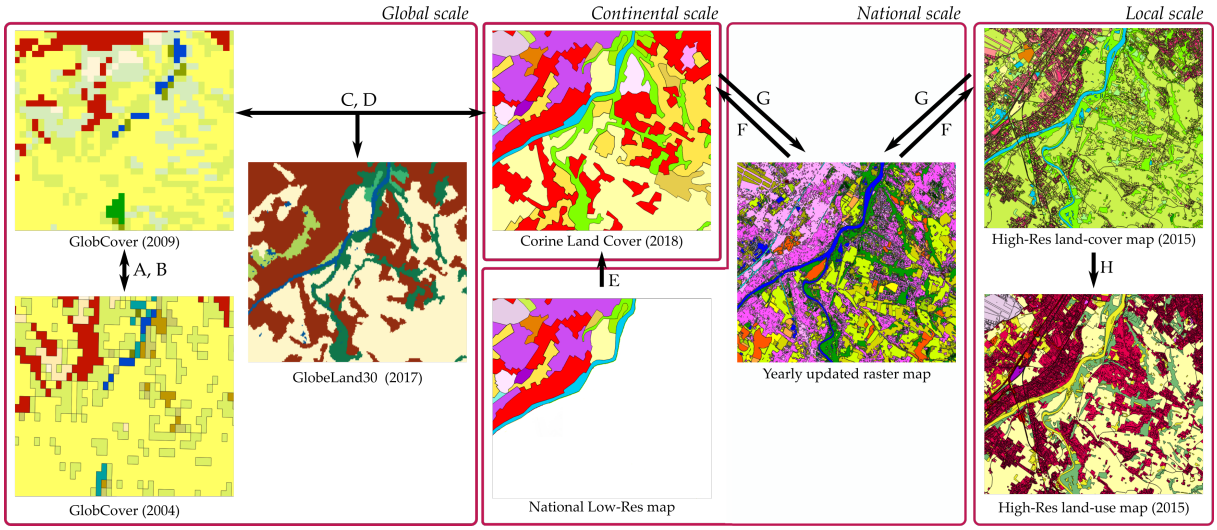

Figure 1.1: Main possibilities for land-cover map translation. (A) change detection, (B) updating, (C) validation, (D) harmonization, (E) completion, (F) spatial simplification, (G) spatial improvement, (H) semantic modification. See [106] for more details about the existing land-cover maps.

All those applications are, of course, also realisable using the traditional image-to-land-cover classification paradigm. However, we argue that the land-cover translation and the image classification paradigm are widely different and exhibit different strengths and weaknesses presented in Table 1.1. The main advantage of translation is to avoid processing vast stacks of temporal image acquisitions, which is both time and money-consuming, by leveraging the information already summarised in existing maps. Despite translation being a complex task (as exposed in the following sections), the state-of-the-art chapter shows that most works focus on the translation applications rather than the translation framework itself [344] resulting in low-quality translation. For instance, proposed translation approaches generally address the resolution and the nomenclature translation as two distinct problems. However, these notions are intimately intertwined, as pointed out in the previous section. For clarity, we address the main issues of the current state-of-the-art solutions separately in the two following paragraphs.

		Image Classification		Land-cover translation	
		Pros	Cons	Pros	Cons
Input data characteristics	Nature	Images = many geometric and semantic information	Requires huge temporal and multispectral images stacks and potentially multi modal data	A single source map resumes a temporal stacks of images	Discrete and compressed information representation = some information are enhanced but some are laking
	Features	Temporal, radiometric, texture = common features in computer vision	Spatial and temporal generalization issues = Domain adptation	Easy rule based predictions	Unusual features (Semantic, geometric, neigborhood) = lack of prior works
	Availability	High, many providers, high temporal resolution			Most maps are rarely updated
	Data volume & storage		High = multi modal,multi-temporal	light = one or few maps	
	Noise	Sensor noise = generally light and random			Heavy systemic noise (source map errors)
Output characteristics	Predictable classes	Potentially all classes defined by a distinctive temporal, radiometric or texture pattern are predictable	In practice, land-use classes are often badly predicted, most remote sensed map have less than 20 classes	Potentially, all classes provided	Widely depends on the source nomenclature and resolution
	Computation time		Long	Short	

Table 1.1: Comparison between the image to land-cover classification paradigm and land-cover map translation.

1.2.1 Nomenclature translation issues

Starting from the observation that the previously mentioned nomenclature standardisation methods failed to unite all remote sensing practitioners under a single standardisation tool and that those methods are often not applicable to already existing maps, many works have attempted to propose nomenclature translation approaches to transform a source nomenclature into a target one [119]. In this section, we briefly introduce their main limitations to explain the necessity of developing new approaches.

Existing translation methods are based on semantic analysis of class definitions and operate directly at the map nomenclature level: they seek to associate each source class with its potential corresponding target class, as shown in a simple version in Figure 1.2. In a nutshell, those methods generally rely either on an expert knowledge-based analysis, or on distances computed from an ontological representation of the classes, or on a ratio of shared features. A comprehensive presentation of how those methods define those semantic relations is given in Section 2.1.1.2. They often determine multiple possible associations per class, *e.g.* a *Forest* can either be translated as *Coniferous* or *Broad-leaved*. However,

they do not define object-level contextual conditions in which one translation is more likely than the other such as "Broad-leaved is a better translation for thin forest on a riverside". Thus, the only translation performed is the strongest one semantically, *i.e.* the closest in an ontology or the one with the highest number of shared features. This single final association is referred to as a *hard-association* by [180]. By analogy with the field of text translation, we consider these methods comparable to word-for-word translation of dictionaries of two languages. All possible word translations are known, but only the closest semantically is used. This causes multiple problems. First, the closest semantically can be ill-defined, *e.g. Forest* should theoretically be equidistant semantically from *Conifer* and *Broad-leaved*. Secondly, the closest semantically is not necessarily the most observed statistically. For instance, a class *Building* is semantically closer to *Dense urban* than *Sparse urban* as the second one might include many features other than buildings, *e.g.* gardens. However, as the two translations are meaningful and in order to reduce the error rate, *Building* should rather be translated into *Sparse urban* areas if *Sparse urban* are more frequently observed.

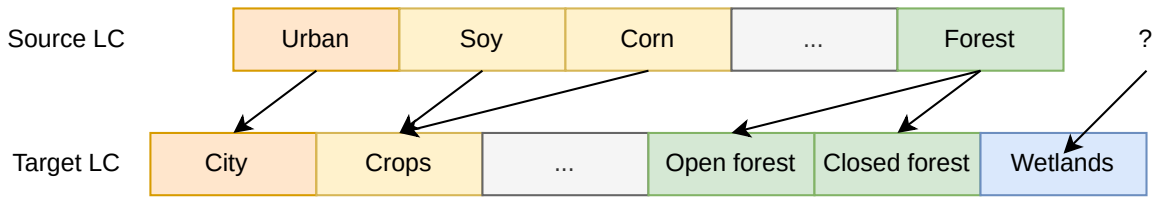

Figure 1.2: Example of semantic translation between two nomenclatures. Methods to determine the possible associations are presented in Section 2.1.1.2.

The resulting translated maps are of even lower quality as the number of possible associations for each source class is high. In order to overcome this problem, most of the studies using translation drastically reduce the number of classes in the target nomenclature to maintain good translation quality. To give an order of magnitude of the hard association's effect on the translation quality, we refer the reader to [225] results on the comparison between an expert-based hard-association translation output and the target LC, which resulted in a limited 57% agreement in their use-case.

Fusion of the translation of several source land-cover maps (land-cover fusion) is sometimes used to replace the nomenclature level translation by an object level one. For instance, if the first map translates one pixel into *Conifer* or *Broad-leaved* and the second map translate the same pixel into *Conifer* or *Shrub* the resulting translation should be *Conifer*. The several source map translations are merged at the pixel level. Using the same terminology as [180], we refer to this object-level translation as *soft-association* methods. However, this strategy supposes the availability of several sources with comparable resolutions.

To obtain good quality translations, linguists translate a word differently based on its context of use. By analogy, we propose to determine context elements that allow the translation of two objects of the same class differently. To our knowledge, no work has

been carried out on this subject. We propose below an initial analysis of the types of context that can be used when translating a land use at a pixel level. We distinguish four main types:

1. **Spatial context:** Local spatial pattern of the object to translate. It can further be decomposed into two aspects: i) the object's inner characteristics (shape, area, ...) and ii) the relations with close-by neighbouring objects (class, shape, ...) on the map. The notion of "closeby neighbouring" can not be defined precisely as it widely depends on the source map resolution and both the considered source and target/nomenclature. We present in Figure 1.3 some examples of a pixel of a given class.

2. **Geographic context:** Depends on the geographical location of the mapped object. In general terms, it accounts for spatial relationships at wide spatial extents. It can either involve i) the spatial context between the object of the source map for far-away structures, such as "an area of water far from the sea is not a saltwater marsh" or ii) spatial correlation with elements not inside the source map: geo-morphological or climatic features such as a tree in a mountainous area having a high probability of being a conifer.

3. **Temporal context:** Some land-cover classes are partly defined by their temporal patterns. For instance, Corine land-cover *Rice fields* include parcels which "As part of regular cultivation cycle, rice fields are occasionally left fallow for 1-3 years.". Incorporating temporal context implies using multiple source maps. In this example, three annual maps are needed to provide the required temporal context, as there may be no evidence of the land-cover class at a given epoch.

4. **Structured label noise:** Land-cover map often exhibits specific error spatial patterns [241] and specific error distributions. Figure 1.4 displays some of those specific patterns, such as isolated erroneous pixels or salt and pepper aspects on a small area. Instead of neglecting errors present in both the source and target map, one could leverage those specific error characteristics to increase translation quality. For instance, if a source pixel is classified *Water* while the target is *Forest*, the erroneous map might be identified based on the spatial patterns and error distribution of source and target maps. The ratio between the different error patterns can significantly change from one map to another due to changes in the conception method and data. For example, photo-interpreted maps do not exhibit isolated or "salt and pepper" pixels errors but are more prone to big area misclassification than automatically generated ones. For coherence with the three other terms we term Structured label noise as **Cartographic context**.

These different contexts have in common that they are challenging to set out in the form of explicit rules as they vary according to the intended translation and are plethoric.

However, it seems possible to learn these rules implicitly by training an algorithm to analyse existing maps.

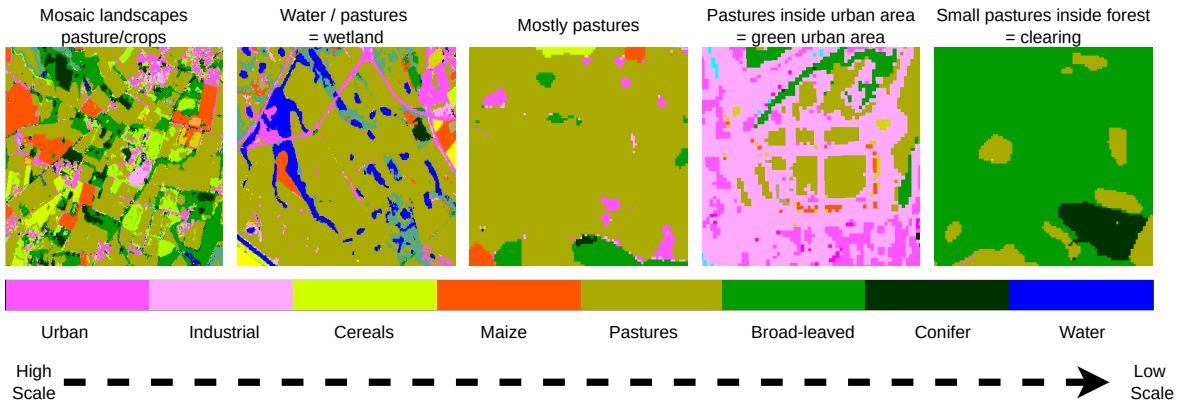

Figure 1.3: Example of various spatial contexts for a single class (Pasture) of the OSO map. Spatial context analysis can be used to determine potential land-cover map classes such as wetlands, or land-use classes such as green urban area.

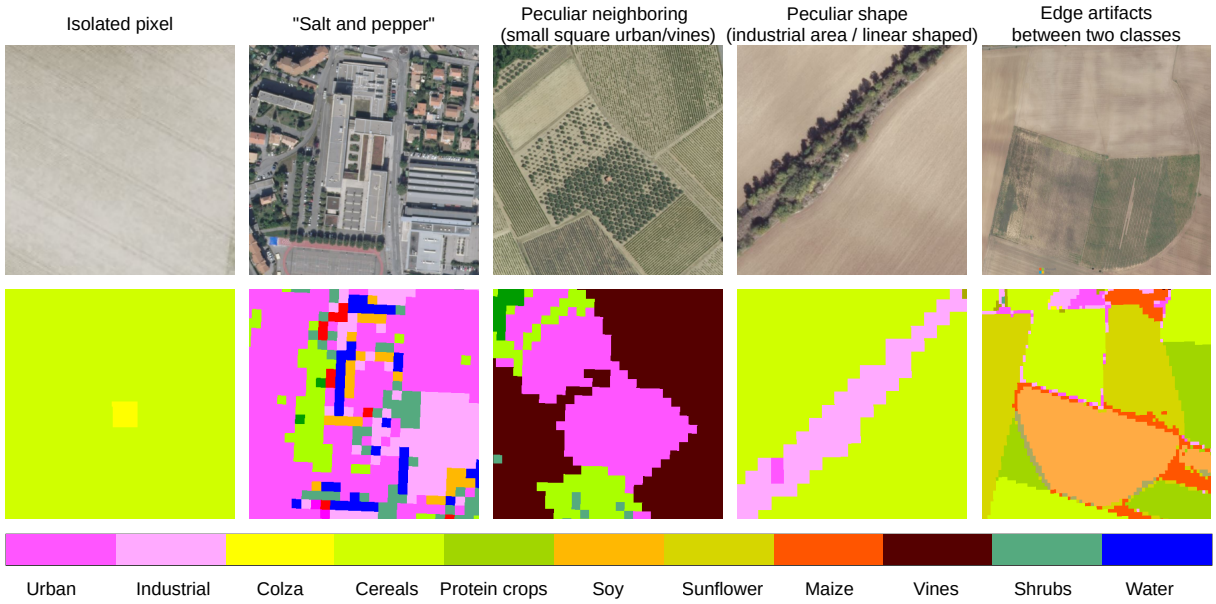

Figure 1.4: Various spatial patterns for errors in land-cover mapping.

1.2.2 Spatial resolution translation issues

The resolution translation can be either perceived as a super-resolution problem if one wishes to increase the map's resolution (by analogy with traditional image processing methods) or as a problem of cartographic generalisation if one wishes to decrease the resolution. We provide an in-depth analysis of those notions in Section 2.2, while this section only underline the main issues to tackle.

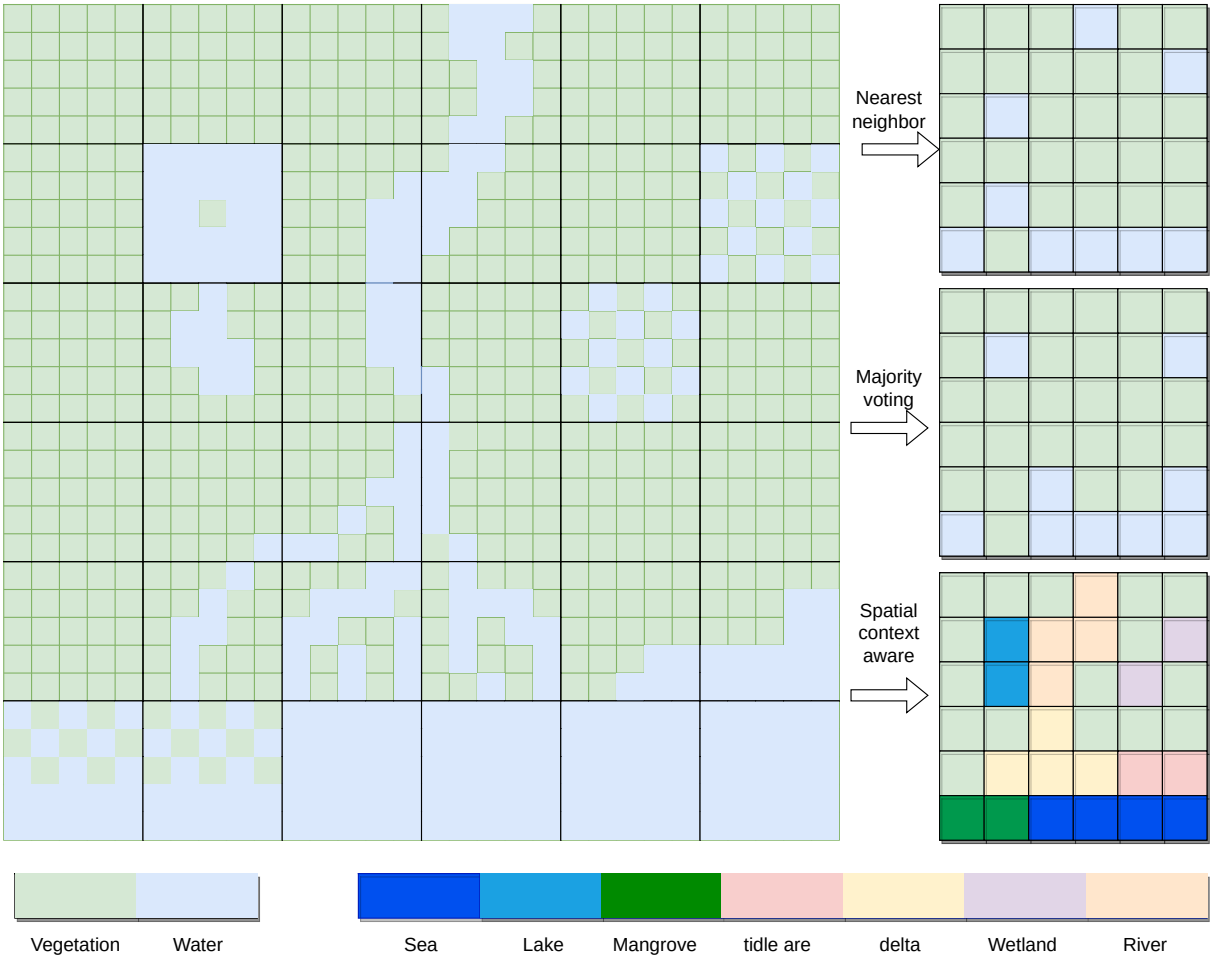

Figure 1.5: Example of a two-class land-cover map downsampling of factor 5. The result of the Nearest neighbour, a Majority voting and a hypothetical spatial context-aware resampling method are compared. Non-spatial context-aware methods i) perform poorly on cells with multiple land-cover map by erasing most of the information ii) do not provide potential new land-cover map classes

Historically carried out using a set of manually implemented rules, cartographic generalisation aims mainly at summarising the information in a map for better visualization [278]. It proceeds through different object operators such as collapsing, simplifying, displacement and exaggeration (*e.g.* simplifying the object shape) [82]. The traditional cartographic generalisation definition differs from the one expected for land-cover map generalisation by its objective: Land-cover map generalisation does not seek good visualisation or representation properties but to preserve, as much as possible, the geographical location and the semantic content of each object. This prevents the use of operators such as displacement or collapsing that do not preserve the geographical location. On the other hand, Land-cover map generalisation merges objects, including those belonging to different classes, even if it means creating new classes describing mixed land cover, *e.g.* a class for areas where small pastures and crop parcels are observed in a mosaic pattern. Multiple works have focused on this land-cover map generalisation issue using a tedious definition procedure of sets of

rule-based transformations that need to be redefined from scratch for each end-user needs and sometimes involve manual corrections. Therefore, no generic generalisation method can be applied to any land-cover map [143]. Moreover, most of the time, only the resulting generalised product is provided to users, not the generalisation framework. Consequently, most works involving generalisation tackle the resampling using the deletion operator through simple interpolation strategies such as majority voting or nearest neighbour. A comparison of those simple operators and a potential spatial context-aware resampling method is presented in Figure 1.5. The currently used nearest and majority voting method tend to neglect much land-cover information as the resulting map only accounts for one of the two classes initially present in the cells (*Vegetation* or *Water*). Conversely, by leveraging spatial context, one can achieve a resampling that combines the *Vegetation* and *Water* into new classes accounting for all the information initially included in the map.

Despite numerous works on a satellite image super-resolution for land-cover mapping [149], super-resolution is much less studied using maps as input due to two main reasons. First, defining explicitly rules to achieve super-resolution is a complex task which required significant advances in computer vision to train algorithms to perform it implicitly. Moreover, it involves adding more resolved data than the source map. Current most successful methods use convolutional neural networks to extract contextual information from images to learn to obtain high-resolved maps from coarse ones [202].

1.2.3 Towards continuous mapping

In [321], the authors advocate for the end of the land-cover classification paradigm. Making the same observation as the one we presented earlier on the land-cover map inherent continuous nature, they propose to replace classification, which inherently discretises the land-cover in a limited set of discrete classes by an object-based approach. In other words, instead of mapping an area with a single descriptor such as "forest", they propose to map it at an object level with a set of features which could, for instance, include information on the tree cover, tree height, tree essence, and eventual grassland presence. This new paradigm for land-cover mapping encountered considerable success during the past two decades, and multiple projects are now conducted to implement this paradigm in real operational cases such as the Spanish SIOSE [28], the EAGLE concept [8] and more recently, the CLC+ project [247]. This new paradigm gives LULC maps significantly richer information but also suffers a usability gap [353]. Moreover, as the manually defined features are often difficult to predict automatically [354] most of the current object-level approaches rely on photo interpretation.

Driven by recent advances in natural language processing on the semantic representation of words, sentences and concepts, a few works have recently chosen to learn to encode each pixel of images to output a per-pixel semantic representation of the object rather than classes [32]. They demonstrate that it enables to perform zero-shot land-cover map classification [246], *i.e.* to predict other classes than those used for training. We argue

that those learnt semantic features could be perfectly tailored to replace the difficult to automatically predict manually defined features mentioned above, especially in the land-cover translation setup. Indeed, unlike image classification, in which only the targeted classification can be semantically encoded, in the translation case, the semantic encoding can both be applied to the input (a land-cover map) and the output (another land-cover map). Theoretically, this could enable to train a model to translate a set of maps between each other and then apply this model to translate unseen during training source maps into unseen during training target maps. This is particularly interesting since it implies that if a model is trained on a sufficient number of translations, it would then be able to be applied to any new land-cover map without needing a sample of data for training. The obtained object-level semantic representation of land-cover (in this case, pixel level) could then exhibits the same properties as those expected for the object-based paradigm mentioned previously.

1.2.4 Challenges

Current solutions for land-cover map translation rest on manually defined sets of rules to perform both the nomenclature and the resolution translation. As underlined previously, we argue that contextual information can help perform significantly better translations. Since defining manually defined context rules for each kind of possible class is unfeasible, we propose to learn those rules directly from sets of examples using deep learning methods.

The main drawback of learning translation from data is that it requires the existence of target samples for the learning procedure. The first challenge is to propose methods with either high generalisation ability (perform well on vast spatial extent even when trained on small ones) or that can train to translate into a target nomenclature without needing explicit target examples.

As land-cover mapping is nowadays carried out on scales ranging from entire countries to continents and worldwide, the proposed solutions should lean toward identifying numerous kinds of context and class spatial repartition. In particular, all experiments are conducted France-wide (550,000km^2) to ensure the robustness of the learnt method.

As such, it involves defining ways to include far-off geographical contexts. This task is challenging as the convolution neural network used in this manuscript can only process small spatial extents due to memory limitations. For instance, studying a 25×25km area with a 10-meter resolved map implies using a 2500x2500 pixel image which is significant for CNN while small for geographical context.

The method should also be robust to errors in both the source and target maps, as most current land-cover map exhibits 10 to 30% errors. As seen in Figure 1.4, most errors tend to be systematic, with some objects in particular spatial contexts being more prone to errors. Thus the method could theoretically learn to reproduce target errors which is not the desired output. Independent validation data should be provided to evaluate whether

the method replicates target errors.

Since free Earth observation satellite programs such as the Sentinel mission started only recently (the sentinel mission started in 2015), most previous land-cover maps tend to be produced only a single time or with a significant time gap between each map due to high production cost. Translation methods should therefore be able to learn on pairs of source/target maps with long temporal gaps, increasing further the noise problem.

Translation also presents the problem of high-class imbalance that might be challenging for learning methods. According to Corine land-cover, the European reference for land-cover mapping, arable crops (the primary land-cover type over France) is around 16000 times more prevalent than the less common Corine land-cover classes agro-forestry. The rarer those training classes are in the training data, the harder they are to predict. Nonetheless, obtaining good results on rare classes is crucial for many applications; otherwise, they would not have been included in the source or targeted nomenclature.

Lastly, computational efficiency needs to be evaluated as land-cover mapping methods tend to be applied to sizeable spatial extents using large volumes of multi-temporal and multi-modal data. More specifically, translation methods should partly be assessed based on their performance/complexity trade-off against traditional image-based methods.

1.2.5 Problem statement

From the analysis of challenges conducted above, we derive the following problem statement:

Is machine-learnt context-aware land-cover translation viable to produce high-quality land-cover maps usable under operational constraints?

To answer this general question, we decompose the problems into four sub-questions:

1. How to translate from a source to a target using various context information?
2. Should land-cover translation be cast as an inherently multi-task problem in which one performs multiple source translations into multiple target ones? For which benefit?
3. In which conditions on the nomenclature and resolution of the source and target map should additional data, such as images, be used to improve translation? How shall the additional data and maps be merged?
4. Are class textual definitions an informative tool to enable using a model on nomenclature and resolution unseen during its training?

The following section presents in a condensed way the outline of this manuscript to answer these different sub-problems.

1.3 Contents

The following section reviews the content of this PhD manuscript. We assume that the reader is familiar with traditional machine learning algorithms, such as random forest, Principal Component Analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE) and deep learning. More specifically, we assume that the following concepts are familiar to the reader: supervised learning with a machine learning model, the main kinds of deep neural nets (perceptrons, convolutional neural networks, transformers) and training losses. We refer the reader to books [27, 101, 155] and available online material[1] if necessary. We provide for readers who just want a short explanation of key principles, a very concise (thus incomplete) presentation of deep-learning focusing on convolution neural network in Appendix G.

1.3.1 State of the art

In Chapter 2, we extensively review work related to translation. To clarify, we review nomenclature and resolution translation methods separately since most research does so even though, as mentioned earlier, those two aspects are highly intertwined. Section 2.1 reviews works on the nomenclature translation problem. We first present attempts to create universal nomenclatures applicable to all maps that are easily modulable to transform into other nomenclature. Then, we present some procedures to define links between two nomenclatures. Third section focuses on performing nomenclature-level or per-object translation once the class relations are characterised. Section 2.2 reviews work on resolution translation. We first present works focusing on down-resolving maps. Conversely, we then present works focusing on increasing map resolutions. Section 2.3 presents works focusing on how to learn to translate the four-levels of context we identified in section 1.2.1, namely : spatial, geographical, temporal, cartographic. Section 2.4 reviews works focusing on improving translation using multi-task and multi-modal data fusion that inspire the design of our methods for translation. Lastly, Section 2.5 reviews work focusing on how to learn to obtain a continuous semantic representation of a land-cover map to alleviate the need for a sample of existing targets based on semantic encoding.

1.3.2 Datasets and Evaluation metrics

In Chapter 3, we present the different datasets used for our experiments. As no works previously focused on learning the land-cover map translation, we needed to generate our training datasets. We publicly released those datasets which can hopefully foster more research on the translation problem or be used more widely for research on land-cover

[1] `http://introtodeeplearning.com/`

map fusion and multi-modal data fusion.

- OSO-to-CLC dataset was first released with our first paper on a single source to single target land-cover map translation. It holds 20k tiles of source/target pairs, covering all the 550,000km^2 of France's mainland territory. We included two dates for each land-cover map (2016/2018 for one and 2012/2018 for the other) and a more than 5000 point manually photo-interpreted ground truth focused on target classes. This dataset has been downloaded around 20 times since its publication.
- MLULC (Multi Land use land-cover map) was released later. While our first dataset could only be used to evaluate the translation from one land-cover map to another, MLULC can be used as a benchmark for multi-source to multi-target land-cover map translation, Land-cover map fusion, and image and map fusion. The dataset still covers the whole 550,000 km^2 of France's mainland territory. However, it includes six land-cover maps, 10 m Sentinel-2 cloud-free mosaic with visible and near-infrared spectral bands, the ALOS-WORLD3D digital elevation model and the corresponding aspect value, and 10 meter Sentinel-1 GRD with dual polarisation. MLULC was already downloaded more than 190 times at the time of writing.

In the last section, we present the metrics we use to assess all results. More specifically, we present two ways of assessing quality: i) comparison to target and ii) comparison to independent ground truth, the first offering estimation off per-class metrics while the second enables estimation of sensitivity to errors of the target data. We also provide formulas and short descriptions of used metrics.

1.3.3 Mono land-cover map translation

Chapter 4 explores the potential of machine-learning based context analysis for land-cover map translation. We work on the most straightforward setup: trying to obtain a single target map from a single source one, which we refer to as the mono-translation setup. We focus on pure land-cover map translation without using additional data such as images to identify the potential of context wise translation solutions. Section 4.1 presents a preliminary study comparing the non-contextual nomenclature level translation with simple context-aware techniques. We first present two traditional translation methods that perform nomenclature and resolution translation separately. Then we also introduce some simple statistical baselines based either on probability modelling or random forest. We show that such an approach performs poorly in many translation cases and presents few experiments on local spatial context. In Section 4.2, we introduce one of our key-contribution, the Asymmetrical-Unet (A-Unet) proposed for spatial context-aware translation. The network is designed to have the following characteristics: i) change the input resolution into the output ii) take into account large spatial context patterns iii) preserve spatial generalisation ability iv) ensure that rare classes are preserved. In Section 4.3, we answer the geographical context problem by adding a small geographical coordinate sub-module

to encode geographical knowledge and demonstrate the potential of this method to take into account far-off spatial contexts. In Section 4.4, we experiment with the temporal dimension of translation. We first focus on the impact of noise induced by the temporal gap between the source and target on the learning results by comparing three different scenarios. Secondly, we evaluate the potential of the temporal context presented earlier by using multiple dates on the source map. Lastly, in section 4.5, we present a strategy to mitigate the impact of target errors during the training procedure.

1.3.4 Multi-land-cover, Multi-modal translation

In Chapter 5, we successively broaden the translation problem to i) perform multi-source multi-target translation and ii) incorporate other types of data. Building a successful multi-land-cover map translation framework is crucial as obtaining multiple nomenclatures and resolutions from a single one is a step towards obtaining a continuum of nomenclatures and resolutions. Furthermore, designing ways to add complementary pieces of information such as optical images or DEM is not only interesting for the translation problem. It also offers an insight into the high complementarity of those products when trying to obtain high-quality land-cover maps. This task exhibits high importance as the fusion between images and land-cover map databases have always been a standard procedure in photo interpretation but is poorly addressed by current automatic methods. In section 5.1, we present the multi land-cover translation network MLCT-Net we design to address the multi-translation task. The network learns to project all maps into a shared representation space before translating to ensure high generalisation ability, even for maps with few training data. We demonstrate that learning a single multi-land-cover map translation model outperforms widely training multiple mono-translation models. Section 5.2 investigates methods to incorporate multi-modal data into the MLCT-net network to improve the result. We thoroughly present experiments for various kinds of data (optical, DEM, SAR) and demonstrate the interest in land-cover map translation for improving current land-cover mapping methods.

1.3.5 Building a continuous semantic land-cover map

Chapter 6 focuses on obtaining a continuum of representation nomenclature and resolution for land-cover maps. We propose to represent each land-cover map class in a continuous semantic space based on class definition, *e.g.* *Corn* is closer to *Soy* than *Water* in the semantic space. As mentioned earlier, no works have explored this subject from a land-cover map translation perspective. Thus, we take deep care in defining the suitable characteristics for such space and exploring some of the possible applications. In section 6.1, we introduce precisely the yet unexplored in the literature notion of continuous semantic space for land-cover translation. In section 6.2, we describe the expected properties of the semantic space and propose related quality evaluation metrics. In section 6.3, we investigate how

to build this semantic representation by comparing a pre-trained language model with a specially fine-tuned to land-cover translation one. In section 6.4, we propose the first application of semantic space to land-cover res-sampling and zero-shot translation.

1.3.6 Publications

Significant parts of the work presented in this PhD manuscript were published in international journals and conferences during the completion of the doctorate.

International journals

- Baudoux, L.; Inglada, J.; Mallet, C. "Toward a Yearly Country-Scale CORINE land-cover map without Using Images: A Map Translation Approach." Remote Sens. 2021, 13, 1060., https://doi.org/10.3390/rs13061060
- Baudoux, L.; Inglada, J.; Mallet, C. "Multi-nomenclature, multi-resolution joint translation: an application to land-cover mapping", International Journal of Geographical Information Science, 2022, https://doi.org/10.1080/13658816.2022.2120996

International conferences

- L. Baudoux, J. Inglada and C. Mallet, "Contextual land-cover map Translation with Semantic Segmentation," 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021, pp. 2488-2491, https://doi.org/10.1109/IGARSS47720.2021.9553693
- L. Baudoux, J. Inglada and C. Mallet, "Deep-Learning Based Multiple Land-Cover Map Translation," IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 2022, pp. 1260-1263, https://doi.org/10.1109/IGARSS46834.2022.9883056

Code In a commitment to reproducible research, we provide our main research code under free software license in the following repositories:

- https://github.com/LBaudoux/Unet_LandCoverTranslator
- https://github.com/LBaudoux/MLULC

CHAPTER 2

Literature review

This chapter reviews valuable works to understand the remainder of the manuscript. Land-cover map translation being an understudied subject, some of the hereafter mentioned works belong to diverse scientific fields such as cartography, remote sensing, natural language processing, computer vision and machine learning. Since this PhD thesis is most likely to be read by members of the remote sensing field, we assume good knowledge of related topics from the readers. We already provide the readers' computer vision and machine learning materials in Section 1.3. Additional material is provided each time a concept not commonly found in remote sensing is addressed.

2.1 Nomenclature Translation

Nomenclature translation methods aim to transform all classes from a given source nomenclature to a target one. Nomenclature translation is as a two-step procedure which: i) define the relations between the classes of the two nomenclatures and ii) Use those relationships to translate each source class. Let S (respectively T) be the source (respectively target) nomenclature defining a a set of classes as described in Equation 2.1.

$$\begin{aligned} S &= \{\, S_i, \mid 1 < i < c_S \text{ with } c_S \text{ number of source classes} \,\} \\ T &= \{\, T_j, \mid 1 < j < c_T \text{ with } c_T \text{ number of target classes} \,\} \end{aligned} \tag{2.1}$$

Each class is described as a set holding all the objects (in this case, pixel p_k) belongings to that class; for instance, source class 1 and target class 3 are described in Equation 2.2.

$$\begin{aligned} S_1 &= \{\, p_k, \mid 1 < k < t_S(1) \text{ with } t_S(1) \text{ total number of pixels with source class 1} \,\} \\ T_3 &= \{\, p_k, \mid 1 < k < t_T(3) \text{ with } t_T(3) \text{ total number of pixels with target class 3} \,\} \end{aligned} \tag{2.2}$$

[335] and [145] distinguish four possible types of correspondence between a source class S_i and a target class T_j namely: the perfect match when the two classes describe the same objects (Equation 2.3), the inclusion of S_i in T_j (Equation 2.4), *e.g.* S_i is "corn field" and T_j is "cropland", and conversely the inclusion of T_j in S_i (Equation 2.5) or no relation at

all (Equation 2.6).

$$S_i = T_j. \quad (2.3) \qquad S_i \cap T_j = S_i. \quad (2.4) \qquad S_i \cap T_j = T_j. \quad (2.5) \qquad S_i \cap T_j = \emptyset. \quad (2.6)$$

From observation of real land-cover map translation scenarios, we derive one more relation, namely: the partial overlap, *i.e. Wetland* can partially overlap with *Water* or *Grassland*, but is not included them (Equation 2.7 where S_i^c denote objects that are not in class S_i).

$$S_i \cap T_j \neq \emptyset \quad and \quad S_i \cap T_j^c \neq \emptyset \quad and \quad S_i^c \cap T_j \neq \emptyset. \quad (2.7)$$

Translation from a source class S_i is straightforward in the case of Equation 2.3 and Equation 2.4 as all pixels belongings to S_i also belong to T_j. The two other cases are more challenging as some elements of S_i belong to T_j while some do not. The following sections first review how to define those relations between classes or, in other words, how to identify the Equation corresponding to each couple of source/target classes. We then review how to translate using those relations once they are defined.

2.1.1 Defining the relations between classes

The following sections review works in chronological order to highlight current nomenclature translation trends. The first section reviews nomenclature standardisation attempts starting in the '70s which proposes to define all land-cover maps with a single nomenclature system to avoid the partial overlap problem. The second section introduces works starting in the '90s on nomenclature harmonisation, *i.e.* which tries to establish the links between two nomenclatures.

2.1.1.1 Nomenclature standardisation

Nomenclature standardisation frameworks, also termed classification systems, aim to provide a single unified nomenclature in which all land-cover map can be expressed. Nomenclature standardisation requires providing an accurate definition of each class content through a set of descriptors (classifiers) [144]. For example, a class *Forest* is not simply described as "multiple high trees gathered on a wide spatial extent" because the highly subjective terms "multiple", "high", "gathered", and "wide" can be interpreted differently depending on readers field of work or the spatial extent they are mapping.

Table 2.1 reviews some of the most famous standardisation approaches to date. This table does not seek exhaustivity due to the high number of either country or thematic-specific standardisation system but tries to highlight the main trends. We propose to distinguish three main archetypes of standardisation approaches :

- The predefined classes, hierarchical classification system (PDCHCS) represented in Table 2.1 by USGS-LULC, CLC, GEOCOVER, NLCD and NLUD-C: This approach

Name	Abbrevation	Year	Link with other methods	Intended Spatial extent	Intended spatial resolution (min area - width)	Classification type	Example of class definitions to describe a wheat crop
USGS LULC Classification Systems [6]	USGS-LULC	1976	-	USA originally but applied worldwide	4ha/16ha-200m	hierachical: -level 1: 9 classes -level2: 37 classes	"Cropland and Pasture: include: cropland harvested, including bush fruits; cultivated summer-fallow and idle cropland; land on which crop failure occurs; cropland in soil-improvement grasses and legumes; cropland used only for pasture in rotation with crops; and pasture on land more or less permanently used for that purpose.
CORINE land-cover [121]	CLC	1990	-	Europe	25 ha - 100m	hierarchical: level 1 : 5 classes level 2 : 15 classes level 3 : 44 classes	"Non-irrigated arable land: Cultivated land parcels under rainfed agricultural use for annually harvested non-permanent crops, normally under a crop rotation system, including fallow lands within such crop rotation. Fields with sporadic sprinkler-irrigation with non-permanent devices to support dominant rainfed cultivation are included."
EarthSat GeoCover Land Cover Legend [61]	GEOCOVER	1990	USGS-LULC	Worldwide	900m²-30m	13 classes	"Agriculture, General: All non-rice agricultural fields, both with crop or fallow; highly managed pastures and hay lands (but not grasslands commonly referenced as "rangeland"); complex mosaics of natural vegetation and cropland. Some orchards and tree plantations, such as palm or date plantations, may be included in this category."
National land-cover Data Classification System [125]	NLCD	1992	USGS-LULC	USA	900 m² - 30m	hierarchical : level 1 : 6 classes level 2 : 25 classes	"Cultivated Crops: areas used for the production of annual crops, such as corn, soybeans, vegetables, tobacco, and cotton, and also perennial woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20% of total vegetation. This class also includes all land being actively tilled"
National land use/cover database of China [362]	NLUD-C	1995	-	China	900 m² - 30m	hierarchical : level 1 : 6 classes level 2 : 25 classes	"Dryland: Cropland for cultivation without water supply and irrigation facilities; cropland that has water supply and irrigation facilities and planting dry farming crops; croplandplanting vegetables; fallow land."
Land-cover Classification System [68] (AFRICOVER [303])	LCCS	1996	-	Worldwide	Scale independent	1 to n classes	See Figure 2.2
International Geosphere-Biosphere Programme-Data and Information System [190]	IGPB-DIS	1996	LCCS	Worldwide	1km²-1km	17 classes	"Croplands: At least 60% of area is cultivated croplands"
South African Standard land-cover Classification System [303]	SASLCCS	1996	LCCS	South Africa	900 m² - 30m	hierarchical : level 1 : 12 classes level 2 : 27 classes	"Temporary crops: Includes areas currently under crop, fallow land, and land being prepared for planting (excluding timber production). Land is harvested at the completion of the growing season and remains idle until replanted. Physical boundaries are broadly defined : may include small interfield cover types and farm infrastructure. "
Land-cover Working Group [297, 327] (of the Asian Association on Remote Sensing)	AARS	1997	IGBP-DIS	Asia	1 km² - 1km	hierachical: up to 7 levels up to 61 classes	"wheat: wheat"
UMd Global Land-cover Classification [111]	UMd	1998	IGBP-DIS	Worldwide	900 m² - 30m	13 classes	"Croplands: lands with 80% of the landscape covered in crop-producing fields. Note that perennial woody crops will be classified as the appropriate forest or shrubs land-cover type"
National Forest Inventory Land-cover Classification Scheme [333]	NFI	1999	-	Canada	900 m² - 30m	hierarchical: up to 64 classes	"Herb – Graminoids: A Herb polygon with graminoids greater than 50% of the herb cover. Where graminoids are defined as herbaceous plants with long, narrow leaves characterised by linear venation; including grasses, sedges, rushes, and other related species.
North American Land Change Monitoring System Legend [175]	NALCMS	2005	LCCS	North America	900m²/4ha - 30m	19 classes	"Cropland: Areas dominated by intensively managed crops. These areas typically require human activities for their maintenance. This includes areas used to produce annual crops, perennial grasses for grazing and woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20% of total vegetation. This class does not represent natural grasslands used for light to moderate grazing."
Sistema de Información de Ocupación del Suelo en España~[28]	SIOSE	2005	-	Spain	Object-oriented	Object based	See Figure 2.2
EIONET Action Group on Land monitoring in Europe~[8]	EAGLE	2013	-	Worldwide	Object-oriented	Object-oriented	See Figure 2.2

Table 2.1: Some of the main land-cover standardisation approaches are ordered by year of conception. Blue lines denote the most influential ones (those used for multiple projects). Intended spatial extent plays a crucial role in the nomenclature content, ie. AARS (Asia) include a "Tea" class that is not included in the Corine land-cover (Europe). The intended resolution is important to define the "purity" of the classification for example, the first version of NLCD only requires 20% of the vegetation being cropland to be classified as cropland.

first proposed by [6] with the USGS-LULC defines an ontological representation of n classes that can aggregate multiple classes into a single one as shown in Figure 2.1. This approach was later extended with IGPB-DIS [190] system and is still used in many currently produced land-cover maps. The main advantage of this strategy is its simplicity and the possibility to fuse multiple classes into one to reduce thematic errors.

- The hierarchically organised classifiers classification system (HOCCS) represented in Table 2.1 by LCCS, IGPB-DIS, SASLCCS, AARS, NFI and NALCMS: This approach first proposed by LCCS [68] avoid defining a fixed set of classes. It first consists of a three-level hierarchical dichotomous nomenclature structure comparable to the PDCHCS-based classification system resulting in 8 classes distinguishing vegetated and non-vegetated (level 1), terrestrial and aquatic (level 2), artificial and natural (level 3) land-cover map types. Then instead of keeping this hierarchical structure, LCCS proposes a modular phase which defines a set of hierarchically organised descriptors such as Life Form, Cover, Height, and Macro pattern (examples of descriptors are presented in Figure 2.2). Those descriptors are intended to be grouped to form various sets of mutually exclusive classes. This represents a significant paradigm shift. Instead of focusing on classes name, this approach focuses on their descriptors. The descriptor set has evolved over time and has been adopted as an ISO standard under Land-Cover Meta Language (LLCS 3.0).

- The object-oriented classification system (OOCS) represented in Table 2.1 by EAGLE and SIOSE: This recent approach is illustrated by the EAGLE [8] project. Like LCSS, this approach focuses on defining land-cover map descriptors. However, these descriptors are intended to be applied at an object level. This enables the creation of maps in which each object is described as a potentially unique set of descriptors instead of a predefined class. Some argue that it is a new paradigm that cannot be perceived as a classification system [321] since there is no class anymore.

Standardisation considerably simplifies the translation by ensuring that the source and target nomenclature are both a subset version of the standardised one. Indeed, in this configuration, the previously mentioned partial overlap between two classes is rarely observed. However, from Table 2.1, we observe that none of the three presented paradigms has been commonly adopted [41]. A detailed review of the potential reasons for the non-adoptions of a standardised system is presented by one of the LCCS creators in [145]. To summarise their observation, the main difficulty is that depending on the covered spatial extent and the thematic purpose, the descriptors of a class can vary profoundly. For instance, a forest definition often includes descriptors on tree height or coverage or even tree essence that varies profoundly depending on the geographical area and the intended usage of the map [42]. Even when using one of the PDCHCS or the HOCCS approaches, two forests can have widely different definitions, which considerably limits the interest in standardising in the first place. Moreover, some descriptors can be challenging to determine

for given pixels either due to a lack of information in the raw data or the algorithm used to make the map [178]. Lastly, to the best of our knowledge, the OOCS approach is currently only used in operations set up in Spain through the SIOSE project [28] (the EAGLE concept is not applied yet but is expected to be used for the Corine land-cover map product CLC+ in 2024). This is mainly explained by the fact that mapping descriptors represent a tremendous amount of work at an object level (many descriptors per pixel instead of a single class) when photo-interpreted and the difficulty through automatic image analysis. The most closely related automatic strategy found in literature focuses on unmixing the content of each pixel into fractions of the target classes [24, 31, 83, 84]. Those approaches are comparable to OOCS, for which the descriptors are the target class percentage inside a given object instead of physiologic/biotic descriptors. Using the same terminology as [75], we refer to those maps expressed at pixel levels with target class percentage as "fuzzy land-cover map classification". Conversely, the traditional one class per pixel approach is termed "crisp land-cover map classification".

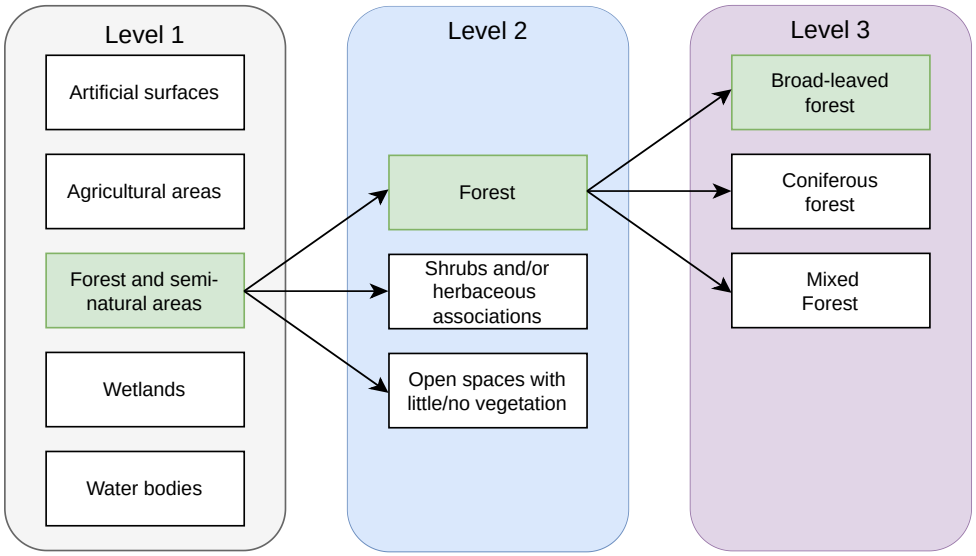

Figure 2.1: A simplified example of hierarchical classification based on the Corine land-cover 3 level nomenclature. Green illustrates the per-level classification of an oak forest.

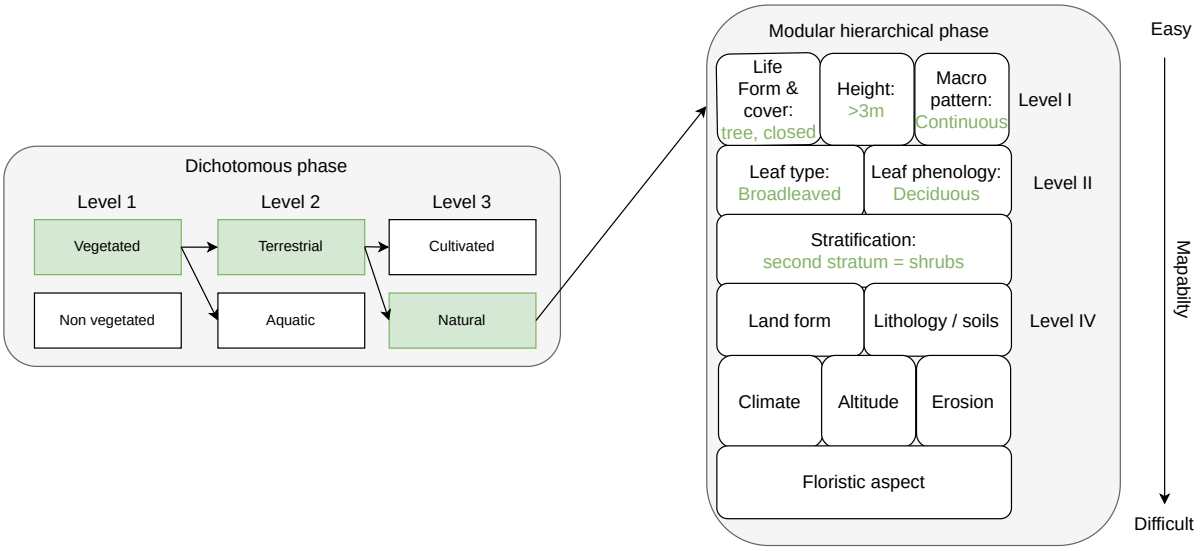

Figure 2.2: LCCS standardisation approach with the example of the classification of an oak tree forest. The first phase (dichotomous phase) is similar to the hierarchical classification and enables an 8-class nomenclature at its third level. The second phase describes the class with a custom set of descriptors. Oak forest definition could incorporate a height descriptor stating that an oak tree is between 3 and 20 meters tall. The descriptors are hierarchically organised according to their expected mappability (the difficulty of determination).

2.1.1.2 Nomenclature harmonisation

The previous section point out that most land-cover fallow different standardisation method making standardisation based translation difficult. This section investigates nomenclature harmonisation methods that enable associations between classes of different nomenclatures even when they do not fallow the same standardisation method. Many methods have been proposed to find the relations between two classes, most relying on a semantic analysis of the class definition [339].

The simplest solution is to assess the possible association through expert knowledge [234]. When challenging to obtain from the source map classes are targeted, remote sensing data are often integrated to bring additional information on the target class characteristics [1]. This strategy has the advantage of simplicity, albeit not allowing a detailed understanding of the quality of the translation. Moreover it is hardly generalisable as two different experts can provide different associations, especially on nomenclatures with a high number of classes [78] (more than 10).

Numerous more sophisticated solutions have been proposed to automatically estimate the similarity between classes based on semantic similarities [55, 147]. Semantic similarity is a family of metrics assessing quantitatively two class closeness based on their definitions. It can be computed in numerous ways. For instance, one can represent the two nomenclatures into a shared hierarchical nomenclature tree like the one presented in Figure 2.1 and then count the number of links that separate the two classes [220]. The main advantage

of this method is that it ensures symmetric properties for the semantic similarity, *i.e.* $d(S_i, T_j) = d(T_j, S_i)$. It can also be perceived as a distance measurement used to embed classes in a euclidean representation space [93]. However, it assumes that the source and target nomenclature can both be represented into a single unified hierarchical tree with for each class of Level n+1 only a single super-class at Level n. Based on the observation that this assumption is often not met [123], *i.e.* a *herbaceous* class might be simultaneously in the super-class natural *grassland*, *crops* or *pasture*, other kinds of shared nomenclature representation structures have been proposed such as semi-lattices [159]. The main advantage of a semantic similarity-based approach is that it defines a correspondence value between each source and target classes [65].

Some recent references in semantic similarity computation are based on using the LCCS standardisation framework [305]. All nomenclatures are first expressed into the LCCS framework in a standardisation step which is challenging in multiple cases [54]. The class association is then determined by computing the percentage of shared descriptors between classes using a simple version of the Tversky loss [313] defined in Equation 2.8 where s_i and t_j denote the descriptors of the class S_i and T_j and $|s_i \cap t_j|$ the number of descriptors shared by the two classes.

$$Similarity = \frac{|s_i \cap t_j|}{|s_i|}. \tag{2.8}$$

Multiple improvements to this formula have been proposed in the literature. For instance, [76] propose to take into account the descriptors of S_i that are not included in T_j (the dissimilarities) since the original equation only focuses on the shared ones (the similarities) while both information is equally important. Furthermore, [252] proposes assigning different weights depending on the descriptors as they do not necessarily all revert the same importance. Lastly, in the case of hierarchically organised nomenclature, [253, 254] assigns different weights to each feature based on the hierarchical level of the source and target class. An important observation is that those attribute-based semantic similarities measurement does not preserve the symmetric property of ontological-based measurement, *i.e.* $d(S_i, T_j) \neq d(T_j, S_i)$; thus, they do not correspond to the mathematical definition of a distance metric. From a semantic point of view, this is meaningful, especially in hierarchically organised nomenclatures. For instance, *wheat crops* is very close to *cropland*, semantically speaking, as they all belongs to a cropland. Conversely, *cropland* is not so close to *wheat crops* as all are not *wheat crops*. For a complete review of the different semantic similarity measurements, refer to [271].

Another way of defining a relationship between two nomenclatures is data-driven, with a direct comparison of pairs of maps. Statistical links between the nomenclatures are obtained by analysing spatially co-occurring sources and target classes. A first naive approach uses a confusion matrix between the source and target maps and assigns the most likely class in the target nomenclature to each source class. [56] propose to translate a source map into a coarser resolved target map by using the number and class of the source

pixels inside each targeted segment as information to train a discriminant analysis. More recently, Latent Dirichlet Allocation has been adopted in [180] as an unsupervised way to merge statistical analysis with semantic distances between classes. One of the method's main advantages is that there may be a significant difference between the theoretical semantic content of a class and its actual content, limiting the real meaning of the measures of semantic similarities. For instance, suppose that the class *crops* of a source map has an accuracy of 70% (because of confusion between the two classes *natural* and *cultivated grasslands*). The semantic definition of the class only accounts for 70 % of the actual content of the class. Determining these associations on the real content of the target classes through a data-driven rather than definition-driven approach tackles this issue. However, it exhibits two main limitations: it is not flexible and result in poor quality similarity metrics if a huge resolution gap exists between the source and target.

An important observation is that most real case nomenclature harmonisation results in low semantic similarity scores between a source class and its ideal target [65]. For instance, [78] computed the semantic similarity between two famous land-cover maps (CLC, the reference for Europe land-cover and NLCD, the referenced for USA). They conclude that amongst the 44 classes of CLC and the 21 classes of NLCD, only three semantic similarities are equal to 1. Thus, most classes exhibit the partial overlap mentioned at the beginning of the section.

2.1.2 Translation based on the semantic relation(s) between classes

Once the relations between all the source and the target classes are obtained, the translation from the source nomenclature to the target is performed.

This task is straightforward when class S_i has a unique association in the target nomenclature (Equations 2.3 and 2.4). Other cases are far more challenging and are poorly addressed by semantic methods [120, 145, 165]. Two different strategies can be adopted: i) The hard association method, which consists in assigning to each source label its most likely correspondence, and ii) The soft-association method, when one tries to act at the object level by assigning a different label to a source label depending on some object characteristics.

The hard association method, by far the most observed in literature [120, 142, 272, 309, 311], translates each source class into the most similar target class. This approach has a significant limitation: when a class has more than one non-zero semantic counterpart, translating it to the semantically closest class *de facto* ignores all other possible associations. By analogy with the natural language processing field, we compare this approach to a "word by word" translation in which one knows all the possible translations for a word but only assigns the most frequently observed one. Those issues are illustrated by [225] who translated GLC2000 [17] into CORINE land-cover map [121] only using a semantic hard association approach. They obtained a low 57% agreement with the observed correspondence between the two maps.

The soft association method tries to translate each source class into several target classes depending on some contextual information. They either rely on: i) local spatial context summarization [56] or ii) on merging multiple independent land-cover map translation contexts [317]. For example, local context summarisation is illustrated by [56], which performs the translation at an object level (the object is a segment based on the target segmentation). Based on multiple source pixels class distributions inside each segment, one can train a model (a discriminant analysis) to predict the target class. With this strategy, two source pixels with the same class in the source nomenclature can be translated into two different target classes, provided those two pixels belong to two distinct segments. However, this method has many limitations. First, the targeted map must have a coarser resolution than the source one. Moreover, the method assumes that the target map segmentation is available as source input, which might not be the case in an operational scenario. Last but not least, semantic distances are evaluated for a pair of source/target classes and in no way for a fusion of multiple source classes into a single target [65]. Thus it requires relying data-driven strategy to proceed to translation.

The other strategy merges several source maps to obtain a single target. A semantic harmonisation method determines the associations between source and target classes for all maps. Then, a vote is cast at the pixel level to determine the target label according to the pixel composition in the source classes. Multiple methods have been proposed for such a decision: sum [157] or weighted sum [55, 317] of the semantic similarities of source maps. Recently, [180] proposed a hybrid approach, combining the semantic similarities with statistical correspondences between source and target classes. They trained a Latent Dirichlet Allocation model to obtain a discriminant embedding of the spatial co-occurrence between the classes of the two source maps for 300×300 pixel tiles. Once this spatial co-occurrence embedding is obtained, they rely on a simple probabilistic model to predict the target. They assume that the spatial co-occurrence of the classes observed in the 300×300 pixel tiles is statistically informative on the per-pixel translation. For instance, a tile where 80% of pixels labelled S_i^1 in source map one and S_k^2 in source map two corresponds to the target class T_j, should translate all pixels of the labelled simultaneously S_i^1 and S_k^2 in T_j. The authors pointed out that this assumption is not always met, especially when there is a resolution gap between the two source maps, making the method usable only with maps with the same resolutions. Moreover, this soft association model requires the availability of multiple source maps. Therefore, [202] replaced the multiple maps approach with satellite images to perform soft association without using several source maps. The method relies on the idea that one can learn to classify an image into a target land-cover map using only the source map and prior knowledge of the spatial co-occurrence between the source and target classes. Each image's expected target class distribution is estimated using the source map and the co-occurrence matrix. Then a network is trained to predict this expected target class distribution using a Kullback–Leibler divergence loss function. This method suffers two principal limitations. First, it only uses the source map to estimate the target class distribution instead of using it jointly with the image, losing all the location information. Moreover, it assumes that feeding the network with

the expected class proportion for a given image result in a consistent classification. In an informal and simplified way, this can be conceptualised by the idea that if one knows that 20% of pixels should be water and that the remaining are trees, if inside the data 20% of the pixels shares similar features, it seems plausible to classify all blue pixels as water. However, we argue that the higher the number of target classes, the less consistent these assumptions get. Consequently, [202] only presented their results for a target map with four classes (Water, forest, field, and impervious surfaces) with a source map with 20 labels.

To summarise, current nomenclature translation techniques are conducted as a two-step procedure. They first define the relations between each source and target class through semantic or statistical analysis. This is almost straightforward when all maps are standardised from the beginning, but it is otherwise more challenging. Then, they perform either a complex association that gives poor quality results or a soft association that either requires downgrading resolution or merging multiple maps.

These methods all disregard the different contexts mentioned in the section 1.2.1 (spatial, temporal, geographic, cartographic). For example, the class *Grass* of GlobeLand30 [43] can be translated into several CORINE land-cover map classes: *Green urban areas*, *Sports and leisure facilities*, *Pastures*, *Natural grassland* or *Sparsely vegetated areas*. In such cases, an analysis of the local neighbourhood is required—this advocates for integrating the spatial context in the translation task. As in natural language processing (NLP), land-cover map translation involves too many possible contextual configurations and cannot be manually defined. In NLP, this issue is tackled using machine learning procedures on text corpora [330]. Surprisingly, no attempt at land-cover map translation using machine learning-based contextual translation frameworks has been proposed in the literature so far.

2.2 Spatial resolution translation

This section reviews current work on changing the resolution of land-cover maps. Changing the resolution of a map is a complex problem that can not be tackled using the same technique as those used for images. Changing an image's resolution is tackled using interpolation techniques that compute local neighbourhood interpolation, *i.e.* weighted means. This is not possible with maps, as the pixel values denote categorical variables. If a forest is labelled 1, water is labelled 2, and a cropland is labelled 3, the mean between a forest pixel and cropland ($\frac{1+3}{2} = 2$) should not be considered as water. To avoid this problem, most of the land-cover focused works used discrete interpolation methods, either nearest neighbour or majority voting, that account only for a part of the information, as illustrated by Figure 1.5 on page 13. A clear distinction must be made between the terms up/down-resolving and up/down-scaling. Up/down-resolving is the operation to either increase (up) or decrease (down) the number of pixels of a given image. On maps,

this consist of obtaining more/less pixels respectively on the same spatial extent, *i.e.* an up-resolve map has a finer resolution than the down-resolve one. Conversely, up/down-scaling is a cartographic-specific term which consists of either increasing (up) or decreasing the scale of the map. As large-scale values describe coarsely resolved maps, an up-scale map has a coarser resolution than the down-scale one (opposite of the up-resolution). For clarity, we only use the terms up-resolve (obtain a fine resolved map) and down-resolve (obtain a coarse resolved one). This section review works on down and up-resolving maps separately as the involved methods are significantly different. We put to the reader's attention that far more literature on the down-resolving of maps can be found than on up-resolving, which is the exact opposite of what is observed in images. Consequently, the up-resolving section mostly mention works on images rather than maps that might need some adaptations to work on categorical data.

2.2.1 Down-resolving land-cover maps

Down-resolving land cover maps consist in resuming multiple p_S source categorical pixels into p_T target pixels (with $p_T < p_S$). The different methods proposed in the literature can be categorised according to i) how they aggregate p_S pixels into M groups and ii) How they resume the information inside the M groups.

Two aggregation strategies are commonly found in the literature: the grid one and the target segment aware. The grid aggregation strategy consists in packing the p_S pixels according to a regular grid with p_T cells, as illustrated in Figure 1.5. It is the most commonly adopted strategy as it does not involve any extra knowledge. Conversely, the target segment aware aggregation strategy assumes one already has access to the M intended groups. For instance, cadastral parcels can be used to aggregate multiple source pixels into a single value per parcel to which they belong. In the land-cover translation scenario, we do not assume that the target segmentation can be used as an input as it would drastically reduce the use case. A third and less studied strategy, the characteristics-based aggregation, groups the p_S pixels according to their characteristics, either semantic [293] (put adjacent objects with the same class into one group) or geometric [329] (put adjacent objects with the geometric characteristics like shape or orientation into one group). However, those methods are usually defined for working on vector databases and are thus ill-defined when working on images as they induce strong spatial deformations. Conversely, we observe that context-wise semantic segmentation methods are known for their ability to learn to predict the target segmentation. We argue that in the specific case of learnt land-cover translation, one could theoretically achieve a third grouping strategy (neither grid nor target segment aware) that we term context-aware grouping.

Multiple strategies have been proposed to resume the information on land-cover maps. The most straightforward strategy adopted in most studies dealing with multiple land-cover maps with various resolutions is to rely on the nearest neighbour interpolation [286] as this method is simple and does not require additional information. However, only one

source pixel inside each of the M groups is responsible for the final classification; most of the source information is lost. The second most adopted strategy is the majority voting rule, which resumes the content of the M groups using the most frequently observed class, better reflecting the actual content of each group. An interesting observation is made by [345], which showed that the map obtained from either nearest or majority rule techniques applied to a fine resolved source map and the classification from a coarse image give significantly different results. They underlined that those techniques do not give satisfactory results as they do not realistically resume the information. They link this problem to the fact that, as the resolution gap between the source and target increases, the majority class tends to represent only a tiny fraction of all the pixels inside one group. For instance, if in every group, the majority class is S_i and represents 25% of pixels inside the group, the resulting majority voting resume all the groups to S_i despite S_i being only 25% of the land-cover observed in the source map. [153] proposed an interpolation method that preserves the per-class areas observed in the source map to partially alleviate this problem. However, this interpolation technique presents the reverse problem. For instance, the group with the highest proportion of S_i will be classified S_i, even if it represents only 5% of the group, if there are a hundred target groups and a class S_i representing 1% of the area in the source map. [291] observes that all those methods completely neglect the noise in the land-cover maps, which can reach 50% depending on the considered land cover and class. They propose to correct the majority voting using the confidence in each source pixel classification (they assume the availability of this confidence matrix).

However, those strategies neglect that when resuming information, one might want to prioritise some information by giving it more weight. For instance, classes such as *Closed forest* are often defined with a statement such as "more than 85% of the area should be covered by trees". Consequently, if the group have 51% of pixels labelled *Closed forest*, the resulting label should not be *Closed forest* even if it is the majority class. This idea of down-resolving a map by applying selective rules depending on some intended characteristics of the down-resolved map closely relates to Cartographic generalization [261]. Those rules detail how to combine a set of operations to transform the resolution of the maps [233]. Many attempts at classifying those operators can be found in the literature [39, 82, 211, 347]. They usually distinguish the operators that modify the attributes (the class) of the object, such as a reclassification operator, from those that modify its geometry, such as a displacement operator, which represents the object slightly out of place. The set of operators used for land-cover map generalisation might vary profoundly depending on the objective of the generalization [143]. In particular, attribute operators are usable in the land-cover translation case as we seek to reclassify data. Conversely, most geometric operators are unusable as we need to preserve the correct geo-location of objects. A recent trend in land-cover map generalisation tends to make the generalisation model more and more aware of the spatial context of an object before generalizing [329], through machine leaned strategies [301], in order to improve the classification of heterogeneous areas. Current attempts mainly focus on training deep learning models [276], usually in a self-supervised manner using generative adversarial networks due to the lack of a

clear definition of a correct generalization when one wants to achieve a map with high readability [306]. However, in the specific land cover translation case, this consideration is irrelevant as we do not seek the land cover's readability but to represent the information semantically and geometrically accurately. Moreover, those cartographic generalisation methods are for now only applied to spatially discontinuous objects of a single class, such as roads or buildings [77, 275, 343]. In the land-cover translation, we must necessarily perform transformation on spatially continuous space with multiple classes. For instance, if a forest is adjacent to a crop field, augmenting the size of the forest necessarily reduce the crop field and eventually require reclassification.

2.2.2 Up-scaling: land-cover super resolution

Increasing the resolution of land-cover maps is a hot topic that involves numerous studies. However, most focus on how to transform coarse images into finely resolved maps [158, 304, 360] rather than on how to transform coarse maps into fine resolved ones. They either rely on: i) the use of multi-temporal acquisition of images and make use of the slight spatial shift occurring between each acquisition to estimate a super-resolved version of the images and then proceed to classification [184] or ii) on a two-step procedure which first estimates the fraction of each class inside each coarse image pixel using a spectral unmixing method and then use this fuzzy classification to determine a high resolved map [183]. We point out that the current state-of-the-art methods in super-resolution of land-cover map using images rely on deep convolutional neural networks to analyse the spatial context of each pixels [149, 182].

Conversely, very few works directly address how to transform coarse maps (instead of an image) into finely resolved maps. The limited work on the subject mainly comes from the previously mentioned two-step procedure of creating a fuzzy classification from a coarse image and then super-resolving it. This second step, *i.e.* super resolving a coarse fuzzy classification into a highly resolved crisp classification, closely relates to this PhD manuscript objective of super resolving a coarse crisp classification into a higher resolved one. Indeed, a crisp classification can be seen as a particular case of fuzzy classification where each pixel has a 100% fraction of a single class and 0 for all others. [29] assimilate this problem as an inverse problem [23, 295]. They argue that the forward problem, *i.e.* determining coarse fractions from a fine resolved crisp land-cover map, is straightforward. Conversely, they insist that the inverse problem is inherently under-determined as multiple plausible solutions can be found. They insist on the need to use prior information that resolve the inherent ambiguity by constraining the number of spatial patterns of classes that can occur at that resolution. Literature can be divided either by the prior information or the algorithm used. Prior information relies most of the time on predefined assumptions on the spatial distribution of the labels, such as the supposition that the output high resolved classification should be the one with the higher maximum class auto-correlation with coarse maps [10, 213, 319] or a mix between ensuring a spatial continuity of same

class object, which is a common assumption [268], and preserving class proportions [298]. More rarely, this preliminary information is directly extracted from images [287]. The algorithm used for performing the super-resolution also varies profoundly from linear interpolation [319] to Hopfield networks [298], to genetic programming [213].

Nonetheless, the crisp and coarse classification to crisp and highly resolved classification task is significantly more challenging than using a fuzzy classification as input as it increases the inherent ambiguity of the problem, *i.e.* the problem is less constrained and thus has more solutions. Instead of giving exact information about the expected proportion of each class for a given coarse pixel location, only one class is provided. To the best of our knowledge, the only proposed method classifies a high-resolved image into a finely resolved land cover using a coarse one during training [202]. As explained in Section 2.1.2, this works assumes that the transition matrix M between the coarse and high resolved land-cover map is known. Each coarse label S_i probability of corresponding to a fine resolved target T_j is known. For a given source map, one can estimate D the expected number of pixels labelled T_j in the target map while being labelled S_i in the source map by simply multiplying the n number of pixels labelled S_i by the probability for a pixel label S_i to be T_j in the target map $M(S_i, T_j)$ as $D(T_j|S_i) = n * M(S_i, T_j)$. A deep convolutional neural network is trained to predict this expected target label distribution using a statistic-matching loss function. However, as pointed out previously, this method performs poorly as the number of target labels increases and uses the source map only as a guide for training an image classification procedure.

In a slightly different setup where the source and target maps are continuous (vegetation height), [198] proposed a solution that alleviates the need to assume prior knowledge of M by replacing the distribution-based loss with a per-pixel loss by assuming that if a coarse pixel has an X value then the global average of all super-resolved pixel inside the coarse one should be close to X. In practice, they process each source pixel value with a Multi-Layer Perceptron (MLP) and compute the sum of the mean absolute error (MAE) between multiple corresponding targets predicted values and the coarse source pixel. They argue that this per-pixels loss function results always in the same solution for one specific coarse source pixel value independently from its spatial context. Thus, they also propose to add information on the pixel location by processing the x and y pixel coordinates in a separate (MLP) and merging it with the pixel representation. However, their methods appear challenging in the land-cover setup, as the categorical representation of land-cover data prevents using their per-pixel MAE loss. Furthermore, it still neglects the geometric information of the source map by processing only the image. Lastly, choosing a per-pixel analysis with an MLP prevents the analysis of the spatial context, which appears not optimal for the super-resolution task. Indeed pixel super-resolution in a land-cover map should be highly dependent on the surrounding of the considered pixel, *i.e.* a pixel in a middle of a forest or on one of its edges should not be super-resolved the same way.

To sum up, this state-of-the-art on-resolution translation pledges for the use of spatial context information to obtain high-quality results. Moreover, Convolutional neural networks

(CNN) achieve the current state-of-the-art results both in the super and down resolution cases. However, as current literature only includes a limited number of works on CNN applied to maps, discussion on the potential and limitation of this technique on the data must be conducted.

2.3 Translation using the four levels of context

In the introduction, we identified four relevant context information levels that one could use to achieve high-quality translation (spatial, temporal, geographical, and cartographic). This section focuses on considering four different contexts using a data-driven strategy rather than the earlier rule-based approach. Influenced by the state-of-the-art results obtained by convolutional neural networks, we mainly concentrate on strategies adopted in this field.

2.3.1 Spatial context: analysing the spatial relations between objects

We assume that the reader already has knowledge of deep learning and convolution neural networks and do not detail the underlying methods in this manuscript. We suppose that the reader is familiar with supervised training, convolutions and their parameters (size, stride, dilation), convolution neural network and continuous optimization with loss functions. Readers unfamiliar with those topics are preferably referred to [27, 101, 155] and available online material[1]. However, for a concise (3 pages) introduction to the main principle of machine learning and deep learning concepts used in this manuscript, refer to Appendix G.

2.3.1.1 Definition

Spatial context has long been recognised as a significant element in computer vision [15] and, more specifically, in the characterisation of land-cover maps [288]. Two levels of spatial context for a given land-cover map object are identified. The inner spatial context, commonly used in the remote-sensing community through the term (GEographic) Object-Based Image Analysis (GEOBIA or OBIA), increases classification performance by describing the shape characteristics of the object to classify [18]. For instance, a linear-shaped element is probably a road or a river but is unlikely an ocean. Those methods, traditionally used to classify images, consist of a two-step procedure [130]. First, the images are segmented in homogeneous areas using a segmentation algorithm, the homogeneousness criterium being user defined. In the land-cover translation scenario, the first step of this procedure is by nature irrelevant as we already have access to a

[1] `http://introtodeeplearning.com/`

homogeneous segmentation in the form of contiguous pixels with the same class in the source map. However, one must notice that this source segmentation might significantly differ from the target one resulting in geometric inaccuracies. Secondly, the information on each segment characteristic is computed. Those different characteristics can be grouped into three distinct categories [71]: spectral information (e.g. the average pixel value inside one segment, its standard deviation...), textural information (e.g. spatial patterns of the pixels inside the segment), and shape characteristics (e.g. is the segment wide, linear, circular ...). As we assume in the land-cover translation scenario that the segment holds one homogeneous categorical variable (the same class for all pixels in the segment), the notion of spectral or textural information is irrelevant. Conversely, we present in the Table 2.2 a set of common shape indicators computable on the land-cover translation setup, mainly inspired from [151]. We acknowledge that many other indicators have been proposed but argue that, as underlined by [130] most of them are highly-correlated and, therefore, including all of them does not improve classification accuracy.

Name	Formula	Description	Helps to discriminate
Area	$A_p = \frac{1}{2}\sum_{i=1}^{n} x_i(y_{i+1} - y_{i-1})$	x_i: x coordinate of the ith point of the polygon.	all classes: discriminates small objects from larger ones
Elongation	$EL = A_p/P_p$	P_p: Perimeter of the polygon	elongated objects: roads, rivers, sand coastlines
Circularity	$CI = \frac{A_p}{A_{cp}} = \frac{4\pi A_p}{P_p^2}$	A_{cp}: Area of the circle with the same perimeter	compact objects: circular irrigated crops, some planted forest
NestedPoly	-	Count the number of polygons inside the polygon	Wide extent object in which other object are inserted: urban areas (parks, road, buidings) and croplands (60% of France cover)
Convexity	$CO = \frac{A_p}{A_{ch}}$	A_{ch}: Area of the convex hull of the polygon	regular/irregular shaped polygons: a rectangle cropland has $CO = 1$, conversely a star shaped building have a lower CO
MBRH	-	Minimum bounding rectangle height	Estimate the shape spread: building have usually small spatial spread compared to Forest or croplands
MBRW	-	Minimum bounding rectangle width	
MBRArea	$MBRHeight \times MBRWidth$	Minimum bounding rectangle area	
MBRFlatness	$\frac{MBRH}{MBRW}$	Flatness of the minimum bounding rectangle	Complementary with elongation: helps to distinguish ellongated from compact objects
MBRAngle		Angle between the width of the MBR and the North	Spatially oriented objects: vineyards are often South oriented

Table 2.2: Common shape features used for evaluating the shape of an object mainly inspired from [151].

Conversely, the outer spatial context, often termed as texture when the considered object is a pixel or as fragmentation when the considered object is a segment, describes the surrounding of the object [16], *i.e.* this green object is surrounded by green, blue ones. This outer spatial context has been widely studied in images by studying spectral variations between objects. However, to the best of our knowledge, no study has been conducted on describing the outer spatial context directly on a land cover map, *i.e.* a grassland near a river is probably a wetland.

The last ten years have demonstrated the high ability of Convolutional neural networks (CNN) to automatically directly learn custom context indicators (rather than manually defined ones). Therefore it appears attractive to explore those techniques to perform inner and outer spatial context-wise land-cover translation.

2.3.1.2 CNN based spatial context analysis

We highlight that as no previous work has been conducted on applying CNN using land-cover maps as input, we refer to literature on traditional images. Therefore some of the

conclusions obtained in the mentioned papers will not apply to land-cover maps. For instance, CNN are known to be far more sensible to texture than shape information on traditional images [96, 117]. However, they can also extract shape information in a low texture environment [12] such as the one met in land-cover translation.

[315] stated that the number of papers dealing with land-cover classification through deep learning has almost doubled yearly since 2015. More importantly, they pointed out that convolutional neural networks performed better than other methods to learn very different spatial contexts, as demonstrated by the various studies on specific land-cover types such as urban land-use [133], wetland [141], forest [110] or agriculture [169]. Most work focuses on learning spatial context and considers all the exploitable spatial contexts from the inner one to the farthest possible outer spatial context, which has been recognised as helpful in many domains [365]. The maximum distance influencing the classification of a pixel by the network is referred to as the receptive field. According to [197], it consists either of techniques that down-resolve the images before convolution [69], techniques that modify the convolution properties such as dilated convolutions [186] or increasing the depth of the network. Since reviewing all the network architectures would be impossible, we only review the two networks our work is built on.

The first one is the famous U-Net architecture [257] which offers a good compromise between having a large receptive field through downs-resolving while preserving high-resolution information through skip connections [341]. This network is commonly used in the remote sensing community for its lightweight aspect, which enables both rapid training and inference and avoids overfitting on small datasets. We present the U-Net architecture in Figure 2.3.

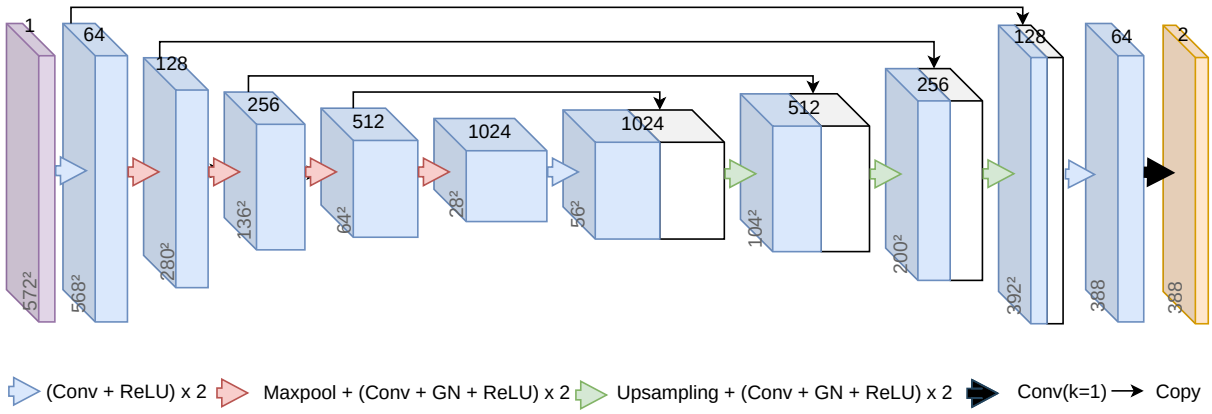

Figure 2.3: Original U-Net-architecture from [257]. The image is progressively down-resolved to ensure large receptieve field. Conversly, copy is used to preserve edges informations.

The second, called atrous spatial pyramidal pooling, comes from the deeplabV3 architecture [45] and is inspired by [115]. It proposes producing multiple down-resolved versions of the same feature maps using dilated convolutions. The core idea behind the dilated

convolution is illustrated in Figure 2.4. Instead of performing the convolution on the eight direct neighbours of a given central pixel (like the upper row of the Figure 2.4), the convolution filter is split to still take into account eight neighbours but at an increased distance from the centre pixel (as is in the second or third row of the Figure 2.4). All the outputs down-resolved images are then concatenated and given to another convolution layer.

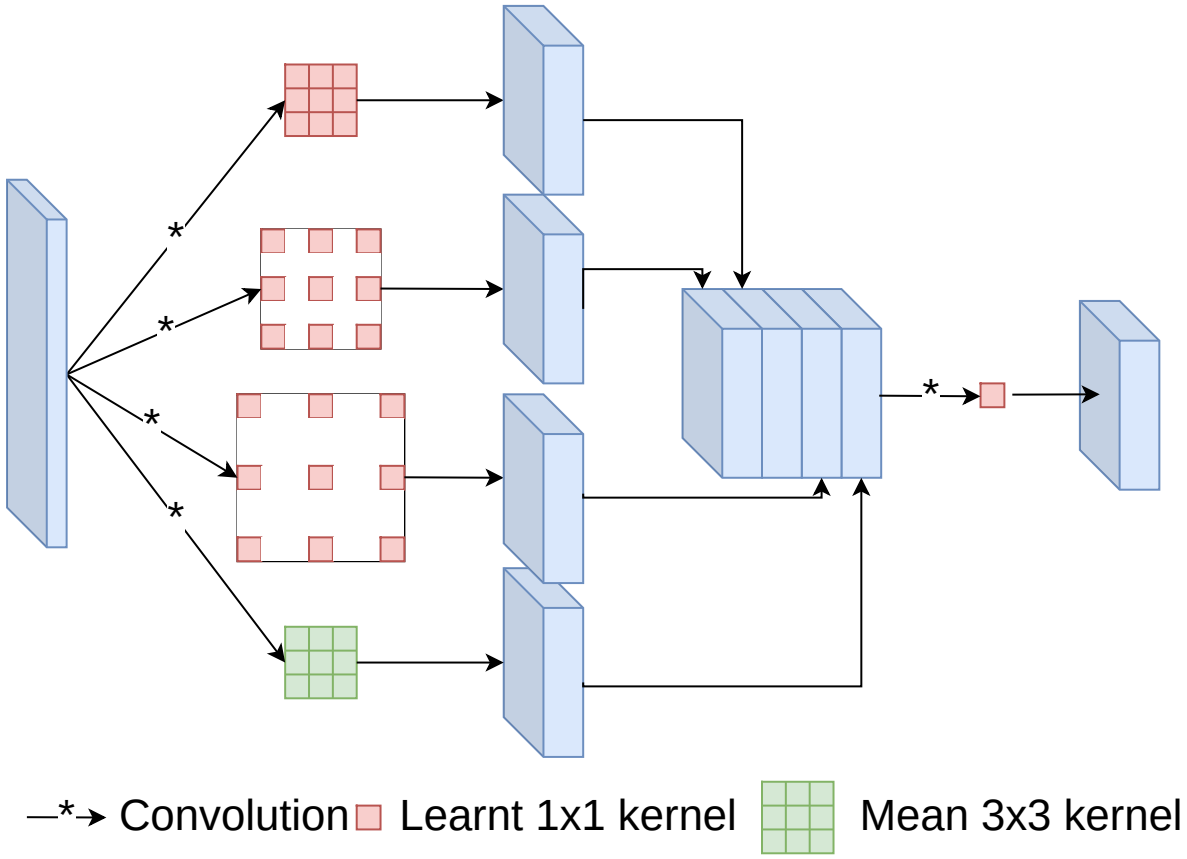

Figure 2.4: Illustration of spatial pyramidal pooling architecture [45]. The input feature map is convoluted with various independent dilated kernels and a mean average filter. The down resolution is achieved through a stride parameter on the convolution represented on this illustration. The different feature maps obtained are then concatenated and linearly combined by a 1x1 learnt kernel.

One primary specificity of learning spatial context for land-cover mapping is that a large network receptive field is insufficient. Indeed, there is no guarantee that all the relevant spatial context is present in the image in the first place. This stems from two phenomena. First, most studies on land-cover mapping focus on a specific spatial extent; information close to but outside the spatial extent is ignored. Secondly, neural networks can only process limited-size images due to memory constraints. The remote sensing community usually uses a patch-based approach to decompose the study area into small patches that are classified separately [277]. For instance, most of them process images from a few dozens of pixels [273] up to a thousand pixels [359] width, which for a 10-meter resolution

image correspond to 10 km. Consequently, the spatial context can not include far-range information, such as whether the pixel is in a mountainous area.

[7] demonstrated that increasing the receptive field of a network is always beneficial but that the increase of accuracy follows a logarithmic curve. They linked this behaviour to [196] works that demonstrate that more weight is always given to the centre of the receptive field. This appears problematic as far-range spatial context is lowly taken into account. Moreover, far-range context is not directly interesting by itself, *i.e.* knowing the class of an object 20km away from the object we want to classify is not really informative. The potential resides in the combination of all the object classes and spatial arrangement from the closest to the farthest ones, *i.e.* the general characteristics of a wide area around the object to classify, such as geomorphological information (the area is far from the sea and near a mountain) or landscapes (the area is fragmented into small parcels). For this specific context information, there is no real reason to give more weight to the centre of the area. The following section introduces methods to incorporate information on wide-scale spatial context, which we term Geographic context.

2.3.2 Geographic context: Tackling specific regional land-cover patterns

We define geographic context as the macro-scale spatial patterns of classes. The spatial distribution of land-cover classes is highly correlated to environmental, climatic and anthropogenic factors [171]. For instance, in France, trees in mountainous areas are likely to be conifers, the crops cultivated in the northern and southern part of the country are different (due to temperature gradient), and far from cities areas are more likely to hold shrubs than close one (fallow croplands). Consequently, a land-cover translation should be aware of this geographic spatial context.

We can distinguish two main strategies in the literature. The first one is defining homogeneous from a land cover point of view spatial units. For instance, [129] proposed to decompose the globe into 61 bio-realms representing unique combinations of biome and biogeographical realm [177], for which the land-cover macro-scale spatial patterns are considered homogenous. Those homogeneous areas are either used to train separate models on different zones like in [139] or to fine-tune multiple local models from a global one [150, 174]. This strategy has the main limitation that it is time-consuming as it requires training/fine-tuning multiple models. Moreover, its performance is highly dependent on how the local areas are defined, *i.e.* Eco-climatic areas used in [139] do not give the same information as the grid area used in [174]. Lastly, [129] observes that, in reality, macro-scale spatial patterns of land-cover do not exhibit sharp divisions but progressive ones, which this strategy does not deal with.

The second strategy is to feed the classification model directly with information on geographic localisation [201]. Very few works have been conducted so far and to the best of our knowledge, no attempts other than our work [19] experimented with this approach

for land-cover map mapping. Attempts in literature mainly concentrate on using the geographic coordinates to produce a coordinate embedding that could later be used for any classification task. For instance, [348] propose to build a vocabulary $V = t1, t2, ..., tn$ consisting of n words extracted from image tags collected worldwide on data sources such as Flickr, Twitter, and Foursquare. Then they proposed a spatial grid-based approach for each cell represented by the tags found in the cell. For instance, if the vocabulary includes three words("water", "Eiffel tower", "glaciers"), a grid cell on the Paris area is encoded as (1,1,0) as there is water and Eiffel tower in Paris but no glaciers. To avoid having two close-by cells with very different embedding, they also include a neighbourhood weighting on the embedding. As they argue that building this per grid cell embedding is tedious, they proposed to train an MLP to predict this embedding directly from the grid cell coordinate. This grid-level semantic embedding can provide helpful information when training a machine learning method. [49], which focuses on image classification based on geotags, established an interesting comparison between using geographic coordinates to either: i) establish a whitelist of possible target labels depending on the location; e.g. was a Swann already observed nearby those coordinates if yes then the Swann is a possible prediction otherwise its not ii) train an MLP to predict a per-label correction from the coordinates, to be multiplied with the per-class prediction obtained with a pre-trained CNN (with fixed weights) applied to the image with the corresponding (the same strategy is found in [200]) or iii) learn the image per-class prediction and coordinate per-label correction simultaneously. They demonstrate that training using the image and the coordinates simultaneously works better on small datasets. Conversely, combining a pre-trained image prediction with a learnt coordinate correction is more efficient on bigger ones. The main potential of [49] is that, unlike other methods, it does not process data according to a predefined grid but directly the coordinates. This considers that macro-scale spatial patterns of land-cover exhibit smooth variation instead of the sharp ones involved by predefined grid systems.

Interestingly multiple other research has been conducted to incorporate 1D location information (different from 2D geographical coordinates), such as in natural language processing to add information on word location inside sentences or in 3D protein structure analysis to add information on the sequential order of amino acid [363]. The current state of the art approach, relies on a strategy called positional encoding [318] which was developed to encode word position in sentences. The core idea is that a correct location encoding should i) be unique for each location ii) ensure that distance in the embedding space between any two locations (e.g. between the 2^{nd} and 4^{th} word) should be consistent across sentences with different lengths and iii) Should be bounded to generalize to any sentences length. They proposed to encode the word position using a set of sine and cosine functions with varying frequencies before adding them to the network. Let d be the total number of dimension of a word embedding and k be one of those dimension, then position encoding is given by Equation 2.9.

$$\overrightarrow{p_x}^{(i)} = f(x)^{(i)} := \begin{cases} sin(\omega_k x) & \text{if i=2k} \\ cos(\omega_k x) & \text{if i=2k+1} \end{cases} \quad \text{where} \quad \omega_k = \frac{1}{10000^{2k/d}}. \tag{2.9}$$

As the positional encoding provides a tremendous improvement in the natural language processing field, some works have been conducted to adapt it to work on 2D location information that could be used into CNN. We distinguish two kinds of coordinates that can be position encoded: the intra-image coordinates termed pixel coordinate (the pixel is in the ith row and jth column) and the geographical coordinates (this pixel is locates at X/Y longitude/latitude). Current work focus solely on pixel coordinates. They aims to remove the convolution translation invariance [187]. All pixels are submitted to the same convolution when using a CNN, which might be unsuitable in some specific cases, *i.e.* when working with rasterized wind data, a different part of the images is not exposed equally to wind [314]. [235] propose a strategy encoding image pixel coordinates using a simple adaptation of the positional encoding: rows and columns are independently encoded using Equation 2.9 and concatenated.

In the land-cover translation scenario, we aim to use geographical coordinates. Two key difference are observed with pixel ones. First, the translation invariance should be preserved as all the pixels of one patch belong to the same limited spatial extent; thus have the same geographic context. Consequently, each pixel of a patch should be summed with the same geographical context embedding. Secondly, while pixel coordinates are the same in train and test (there is always a first row/column pixel), the geographical coordinates used are different in training and testing. Thus the effectiveness of positional encoding is yet to be shown in this case as it requires generalisation ability.

2.3.3 Temporal context

2.3.3.1 Definition

We define the temporal context of a land-cover nomenclature as all the explicit and implicit temporal patterns exhibited by the classes of the nomenclature. Explicit temporal patterns include all information on temporality included in the class definition. For instance, in the *Rice Field* definition presented in Definition 1 we underlined typical explicit temporal patterns. We observe that inside a definition, multiple temporality levels might be included, "periodically" here describes a temporal pattern occurring multiple times across a year. At the same time, "regular cultivation cycle" and "occasionally left fallow for 1-3 years" describe a pattern with a lower temporality (every three years).

"Rice Fields: Cultivated land parcels prepared for rice production, consisting of periodically flooded flat surfaces with irrigation channels. As part of regular cultivation cycle, rice fields are occasionally left fallow for 1-3 years. These parcels are considered to be rice fields, too." - Corine Land-cover (1)

Implicit temporal patterns include all information on temporality that is not directly included in the class definition but characterises the class content evolution across the land-cover reference period. For instance, a *Wheat* class inside a map covering an annual period describes a temporal pattern with a period of bare soil followed by wheat growth and another bare-soil period.

Those temporal patterns are directly influenced by the reference period covered by a land-cover map. On the same given spatial extent, a map covering a monthly period distinguish *Bare soils* from *Vegetation*, an annual distinguish *Fallow* from *Annual crops* and *Perennial* ones by studying the variation of vegetation across the year, and a map covering several reference year distinguish *Natural areas* from *Cultivated* ones by studying the vegetation variation across multiple years [221]. Translating often involves matching classes with different temporal contexts, *e.g.* a *Rice crops* observed during a one-year period, one observed once during a three-year period or a flooded area observed during a one-month period.

Land-cover translation temporal context must be clearly distinguished from the operational temporal gap between the source and the target map encountered when training a machine learning model to perform the translation. We consider this temporal gap as noise-inducing and address it in the following subsection.

2.3.3.2 Methods

Historically, classification from multiple images with different acquisition dates is mainly proceeded by concatenating all images into a single one and directly processing the raw data [91, 126]. A common strategy to improve this multi-temporal classification was to complement it with additional spectral features computed separately for each date based on spectral indices such as NDVI [103, 323] or EVI [13, 26] and texture indices [88]. SpectroTemporal aware features were later proposed to reflect the temporal evolution of the data better. Standard spectrotemporal features mainly include statistical metrics (average, min/max, standard deviation of per pixel spectral values on the period) [51, 89, 95, 210], phenology metrics derived from statistical metrics (beginning/end, length, amplitude of the growing season) [148], or on reshaped raw data feature space through dimensionality reduction techniques such as PCA [100]. In the land-cover translation scenario, those spectrotemporal features are not applicable as they were designed to work on continuous feature space, *i.e.* the average value of a pixel for different reference periods has no real meaning on land-cover maps as it involves doing the mean of categorical variables. Therefore, those works should be adapted to be usable in the land-cover maps translation setup.

The obtained features are historically used to train various machine learning algorithm including decision tree [95, 126], random forest [51, 139], support vector machines [37, 100], Gaussian Maximum Likelihood [148, 249], or multi-layer perceptron [13, 103]. More recent works focusing on the integration of large temporal sequences of data into neural

architectures (RNN [137], LSTLM [273], transformer networks [94]) are considered out of scope of this manuscript as most land-cover maps are only available in very small temporal sequences. Indeed most are produced only once or on less than five reference periods.

2.3.4 Cartographic context: learning with noisy data

Translation raises questions about the impact of source and target data errors. The traditional rule-based translation approach *de facto* ignores this problem as (i) the target map is not used in a learning procedure; target map errors can not be learnt (ii) the nomenclature translation is based on source class description, which does not take errors into account. Thus errors in the source map are agnostically translated into target classes. Learning to translate from a source to target behave oppositely. First, some errors in the source maps can potentially be compensated by the translation procedure. For instance if all roads in the source maps are misclassified as urban areas, some of them can be correctly classified as roads in the target nomenclature by using this specific erroneous spatial pattern of linear-shaped urban areas to distinguish roads. Secondly the targeted map errors can influence the translation results. For instance, if all roads in the target map are misclassified as urban areas, the network might learn to replicate the same error by transforming all roads of the source nomenclature to urban areas. Fighting against the effect of label noise is mandatory as the method should be designed to be able to translate between source and target with numerous classes, which tend to be have significant error rates [302].

Learning with noisy data is a highly studied subject. The manuscript only focuses on label noise (in opposition to image noise). Label noise in land-cover mapping can be systemic or random depending on the noise source [87]. Most of the systemic noise in land-cover maps is induced by errors of the classifier producing the maps, which tend to repeat the same error on similar objects. Therefore it is instance dependent rather than class dependent. For instance, *Croplands* pixels on the field's edge can be classified as a *Forest*, with this confusion only happening in this specific spatial setup. Random label noise is more often observed on photo-interpreted maps, mainly due to temporal gaps between the data used and the target temporality. In the real case scenario, most errors tend to be observed between classes semantically close for photo-interpreted land-cover maps or on classes close in the feature space for automatically derived land-covers. Unlike image noise which is often considered additive to the true information, label noise behaves as correlated to original information noise.

[284] propose to classify the different deep learning approach to deal with label noise in four categories: i) architecture-based [336] ii) regularization based [280] iii) Loss based [205] iv) sample selection based [283], which can be combined [364]. Loss-based strategies, which aims to correct the loss computation by assuming some knowledge on the label noise appears particularly relevant in the land-cover translation case as unlike most problems a noise transition matrix is most of the time provided (also termed confusion matrix in

remote sensing). For instance, assuming that the noise transition matrix is known, one can correct the prediction of the network by multiplying it by its error probability [237].

However, current loss based solutions use transition matrix to correct prediction at a per-pixel level, e.g. if *Forest* is misclassified in *Water* 20% of the time in the target map, then noise transition aware loss encourages the prediction on a *Forest* pixel to be 80% *Forest* 20% *Water*. We believe it corresponds to misuse of the noise transition matrix, which are class level indicators rather than instance level [46], *i.e.* one object is 100% *Forest* or *Water*, and 20% of all objects labeled *Forest* should be *Water*. We underline that the main reason for the current misuse of the noise transition matrix is that they were mainly developed to address image classification. In this setup, one image has one class. Thus, in an iterative algorithm such as those used to train deep learning models, each iteration only processes a small number of objects per-class. This prevents computing per-class distribution estimates such as the previously mentioned 20% of all objects labeled *Forest* should be *Water* as there could be only two images labeled *Forest*. We argue that in the case of land-cover translation, we could alleviate this limitation as we have access to the vast number of annotations; land-covers are semantic segmentation of images in which each pixel has one class. We underline that only a very limited number of works have been conducted to adapt label noise correction methods to semantic segmentation maps and that none of them was loss based to the best of our knowledge.

2.4 Robust translation

This section focuses on obtaining high-level quality translation models, which we mainly define as: i) robust to change in the class distribution and spatial arrangement to enable generalisation on spatial extent not covered initially, ii) able to preserve a maximum of the target class diversity. Three main research fields have been proposed in the traditional image to land-cover classification.

First, domain adaptation [312] defines methods able to achieve high-quality results when the training and testing data exhibit statistically different distributions [167]. Radiometric values in two different eco-climatic areas can be widely different; the results of a model trained on one ecoclimatic area might not be transferable to another. In the case of land-cover map translation, spatial domain-adaptation does not deal with radiometric values but with class distribution and varying spatial patterns, *e.g.* fragmented-crop landscapes might be observed in the training data while open-field is observed in the testing data. The literature mainly focuses on extracting and projecting features into a representation space shared between the training and the testing data. In this space, training and testing data are expected to exhibit the same statistical properties without any observable shift between them. Traditional methods mainly rely on statistical matching strategies such as multidimensional histogram matching [138], or principal component analysis (PCA) [228]. More recent works adopted deep neural networks for their high

generalisation ability [226]. Two constraints are found in the literature to enforce the training and testing to be mapped into a shared space: (i) minimising the distance in the representation space between two identical elements through a loss term [231], (ii) adversarial training [342] in which a discriminator enforces source and target observations to be comparable. In the land-cover translation scenario, loss based strategy aims to ensure minimal distance in the representation space for comparable target classes independently from their nomenclature. For instance, *Forest*, *Coniferous*, *Broad-leaved* or *Shrubs* should all be represented closely in the representation space, to ensure that a model trained to translate a source map into a target on area where there is no *shrubs* is usable on an area with *shrubs*. This shared representation space should not only account for classes but also spatial context, *i.e.* two elongated forest should be represented closer from one another than a more circular one. Theoretically adversarial training should obtain the same result but is confronted with the well-known difficulties in optimising adversarial networks and are not be explored in this manuscript.

A second strategy is to rely on multi-modal data fusion, which focuses on defining methods to combine heterogeneous sources of information [98]. Fusing independent sources of information (such as image, text, video, or audio) describing the same object tends to enforce learning an object representation less sensitive to local variations and imprecision. Most current research focuses on merging optical and radar data in land-cover mapping as they bring complementary features [269]. In the land-cover translation setup, image and map fusion could enable to learn a more robust representation of the data, enabling the prediction of higher class diversity. Papers resuming the output of the GRSS data fusion contest (held each year since 2006 [232]) are a rich source of information on the current state-of-the-art method for performing a fusion of remote sensing data. Until 2018, proposed methods relied on extracting manually defined features from the various data, like NDVI for optical data and class proportion for land-cover data [349] and feeding it to an ensemble classifier (often random forest). [340] pointed out that the 2018 contest was the first one in which the best results were always observed using deep learning to extract features from raw data directly. From then on, even though some slight changes are observed in the deep learning architecture, the core remains the same. Each data is encoded using a separate network and merged at a level of abstraction, going from simply concatenating all raw data to just before a final classification layer. Even though fusing multiple data at the raw-data level should yield the best inference as it ensures that no information is lost, [170] observe that various complicating factors such as noise and contradicting data might benefit from fusing the different data at a higher level of abstraction [160]. Different works have concluded differently on the best location to merge data. Choosing the right location to perform data fusion remains highly empirical and depends on the task and the chosen architecture [134, 244]. Experiments are yet to be conducted for the land-cover translation setup. A recent trend in multi-modal data fusion, is to incorporate an attention mechanism [21]. First introduced for natural language processing, it learns to balance each feature's importance better by analysing all features simultaneously [48, 318]. Figure 2.5 presents one of the multiple adaptations of

the attention mechanism to images proposed by [358]. The main idea is to process a set of feature maps in three independent feature linear combination layers (f,g,h). The attention map is obtained as the scalar product between transposed f and g. It corresponds to a learnt per-pixel weight. Lastly, the scalar product between this attention map and h is computed to obtain the output of the attention mechanism. Following [199] observation, we underline that the attention mechanism can be used to give selectively more weights to some modalities when merging multiple data sources.

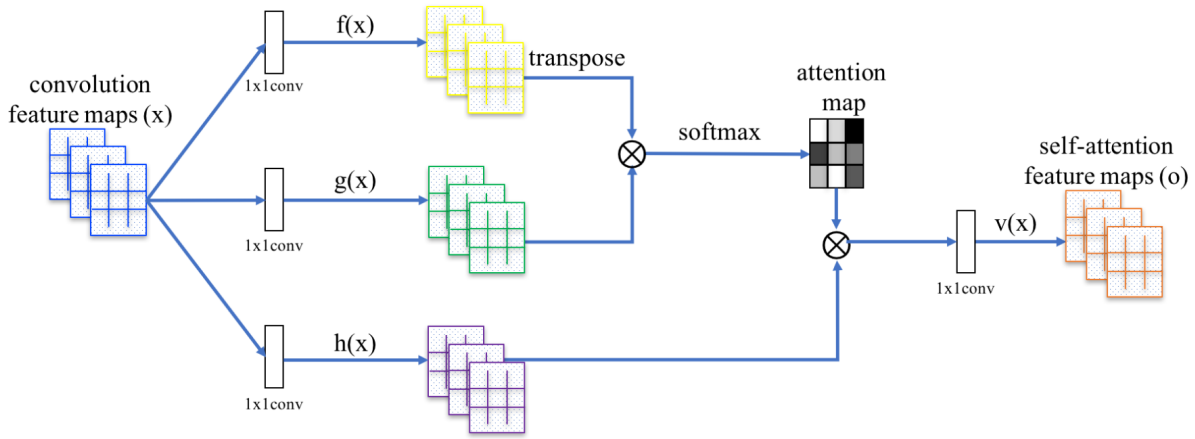

Figure 2.5: Per-pixel Self-Attention mecanism for CNN proposed by [358]. Figure reproduced from [358].

The last strategy consists in multi-task learning, which aims to improve inference accuracy on several tasks by training simultaneously on all of them [74], based on the assumption that at least a subset of the task is related [361]. In the image to land-cover classification, common multi-task strategies involve either predicting simultaneously land-cover and land-use [22], land-cover and change detection [63] or land-cover and height [38]. We argue that land-cover translation is an ideal setup for multi-task prediction. The translation could benefit from training a single model to transform one source land-cover into multiple targets. For instance, recent advances in natural language processing have shown that learning single multiple translation models achieved comparable results with learning multiple one-to-one cases. However, the former yields more robust results on languages with few samples and has better generalisation abilities [59]. Multi-language training seems to benefit from the obtained multi-language common representation space [245]. Multitask networks are also commonly trained with the dual objective of reconstructing the input data (self-reconstruction) and achieving the desired tasks [346] to enforce mapping to a shared representation while preserving the unique features of each input.

2.5 Towards learning a semantically consistent representation of land-cover maps

We mentioned in Section 1.2.3 that the current trend in land-cover map standardisation is to perform an object-based nomenclature, in which each object is assigned a set of descriptors depending on its characteristics [4]. A similar tendency is observed in the field of zero-shot learning. The core idea in zero-shot learning is to train a network to encode a subset of classes into a consistent semantic space. At inference, classes unseen during training are predicted based on auxiliary information on the links between the classes used for training and the unseen ones. [90] takes the example that with the information "a zebra is more-or-less like a horse but with black-and-white stripes", a child can recognise a zebra without seeing one before, provided he was first taught what a horse and a stripe pattern look like.

We believe that this notion of semantic representation at an object level should be a cornerstone of machine-learnt translation for two reasons: operational and information preservation. First of all, from an operational point of view, working on semantic representation on classes could alleviate the need to retrain the method at each new source or the target map, *i.e.* once the concept of "forest", "trees", "coniferous", "broad-leaved", "mixed forest", "dense vegetation" have been learnt by the model at various scales and on various spatial extent through the use of multiple sources and target maps it should not be necessary to retrain the model if one wants to translate the concept "woodland". Secondly, from an information preservation point of view, the whole idea of applying a machine learning strategy is to extract contextual information on each object of the source land-cover map to perform better translation than a non-context-wise one. However, while precious, this contextual information is lost once used to assign the target class.

[324] reviews current strategies used for zero-shot learning. They first propose to distinguish them according to the nature of the learnt semantic representation: manually engineered semantic spaces or learnt ones. From a land-cover translation point of view, the LCCS framework presented earlier might be seen as manually engineered semantic spaces in which each class is described by a set of descriptors. They also distinguish them by the nature of the method trained to map into the semantic space. Classifier-based approaches focus on how to directly learn a classifier for the unseen classes, while instance-based focuses on obtaining labelled instances belonging to the unseen classes and using them for classifier learning.

Most works for land-cover map focus on the method to train rather than on the nature of the semantic space. For instance, from a land-cover mapping point-of-view, one of the first works was conducted by [107, 246]. They used a pre-trained word2vec [214] model to translate each class name into a feature vector. They then trained a CNN to obtain these feature vectors from a source image, *i.e.* instead of training to assign to each pixel a class, such as a forest, they assign the corresponding feature vector. During the test phase, they

rely on a simple approach, such as a k-nearest neighbour, to assign unseen classes.

We believe that more importance should be given to the definition of the embedding space. Indeed the traditionally used word2vec representation of classes has many limitations, namely: i) it only takes into account the class name rather than class description, ii) it is not specialised in land-cover; natural grassland and cropland can be very close in the feature space while we traditionally make a distinction between the two in land-cover mapping.

An important observation is that significant works on label encoding have been conducted during the last decades, especially on how to encode label hierarchy [256], class co-occurrence statistic [212], semantic attributes [172] and subsets of those information [279]. Thus it appears interesting to see how to adapt those works to encode semantically and contextually land-cover maps.

2.6 Conclusion

In summary, current land-cover translation solutions address the nomenclature and resolution translation separately. Nomenclature translation solutions are mainly based on semantic associations evaluated at the nomenclature level (LCCS, EAGLE ...). Most of the time, all pixels of a given source class are only translated into a single target (the semantically closest one). Resolution translation is, on the contrary, tackled using either simple interpolation techniques such as the nearest neighbour or majority voting that are not able to infer new classes (a patchwork of water and trees is probably a wetland) or not addressed when it comes to land-cover super-resolution (exception made of [202]). We propose to investigate how to improve translation by performing the nomenclature and resolution translation context-wisely jointly. The currently best-performing methods to incorporate spatial, geographical and temporal contexts replaced manually defined feature-based solutions with machine learnt ones, principally convolution neural networks. As those solutions require training data, we present the dataset used for our experiments in the following section.

CHAPTER 3

Datasets

This chapter introduces the datasets used during this PhD thesis. Learning the translation requires selecting land-cover maps to train the models. As we concentrate on supervised learning strategies, the introduced datasets propose pairs of the corresponding source and target maps. Since no public dataset with multiple land-cover maps is available, the Multiple Land-use Land-cover (MLULC) dataset, including six land-cover maps, is introduced. A smaller version of this dataset, the OSO to CLC dataset, is also provided to enable specific experimentation. The slight variations between the two datasets are explicitly mentioned when appropriate. Section 3.1 introduces the different steps to create a land-cover translation dataset. Criteria influencing the choice of the study area and the choice of the used land-cover maps are emphasised. Section 3.2 introduces the study area and the used land-cover maps main characteristics. Section 3.3 presents the differences between the land-cover maps and underlines the main challenges of the dataset. Section 3.4 introduces additional data (optical images, synthetic aperture radar images and digital elevation models) provided with the dataset that can be used to improve the translation quality. Finally, Section 3.5 presents a manually built ground-truth dataset that enables fair assessment of the quality of the translation by avoiding quality evaluation based on a comparison to a noisy target reference. This last section additionally reviews the metrics used in this PhD manuscript to assess the quality of the translation.

3.1 Dataset creation protocol

This section introduces the different operations used for the dataset creation and highlights the elements that influence the choice of the study area and land-cover maps. The dataset creation procedure involves seven steps presented in the following subsections.

Download Land-cover maps can be distributed in two different formats: raster or vector. Raster format (*i.e.* image format) describes the land cover using a grid, in which each cell is represented by a pixel value. This format is usually the one used by land-cover maps stemming from automatic image classification. Conversely, vector format describes the land-cover using a set of spatially contiguous polygons, often obtained by manual

delineation by photo-interpreters. As this manuscript focuses on convolutional neural network application, raster format is the final delivery format of our dataset. However, when possible the land-cover are dowloaded in vector format as the dataset creation procedure involves re-projecting all land-cover maps in the same geographic coordinate system, which involves significant deformation on raster data [274]. Those vector maps are rasterised after the reprojection and aligning step.

Alignment and reprojection All maps are cropped and aligned according to France's borders and are re-projected to the French official projection system EPSG:2154. This step involves nearest neighbour resampling for maps only available in raster format to preserve the original resolution. This step produces a spatial shift for raster maps with a degradation of the geometric resolution that can reach half the size of a pixel in the input image x or y directions [260] and $\frac{\sqrt{2}}{2}$ pixels in the diagonal directions [274].

Optional reclassification Land-cover classes with unclear (such as *Other*) or mixed content can be reclassified at this stage by either merging them with other classes or labelling them as no-data.

Rasterization Vector maps are rasterised. The resolution is determined using the official target pixel resolution when provided or is arbitrarily considered as half the MMU value.

Patch decomposition CNN are originally built to analyse traditional photography, which exhibits a limited number of pixels (often less than $1,\ 000 \times 1,\ 000$) or can be resized. Conversely, this assumption does not hold for the obtained raster land-cover map. The number of pixels is often very high (120, 000×120, 000 for a ten-meter map covering all of France), and can not fit directly in memory when processed by a CNN. A common strategy is to rely on a grid-based approach to decompose the wide image into a subset of smaller images, as represented in Figure 3.1. Each of these small images, termed *patch*, is then processed by the CNN like it would have been for traditional photography. One of the main limitations of the patch-based approach is that areas close to the edge of the patch only have access to a limited amount of spatial information, as up to half of it might be outside of the patch (Figure 3.1). To circumvent this issue, a common strategy is to process spatially overlapping patches. However, this strategy is incompatible with some of the random train/test split procedure presented in the following paragraph as it would make it possible to identify neighbour train/test patches using the overlapping area. Of course, in an operational setup in which one wants to produce a France wide map, one should use spatially overlapping patches. However, for experimental purpose, experiments are carried out with non overlapping patches to enable unbiased measurements. We choose to perform translation on a 6×6 km^2 which ensures reasonable memory consumption (600×600 pixel for a 10m land-cover map and 60×60 for a 100m one), while limiting the proportion of pixel under edge influence.

Train/test/validation split Patches are separated into a train (65%), test (35%) and validation (5%) to ensure that the CNN is not evaluated on the data used for training it. This separation can either be random or geographically organised, as illustrated in Figure 3.2. The first strategy enables estimating accurate metrics over France. In contrast, the second is more appropriate for evaluating spatial generalisation ability and is more difficult to analyse as it highly depends on the chosen geographical repartition. We choose to rely on random sampling to design this dataset to simplify the analysis of the results and make the quantitative translation result comparable to those assessed by the original land-cover producer.

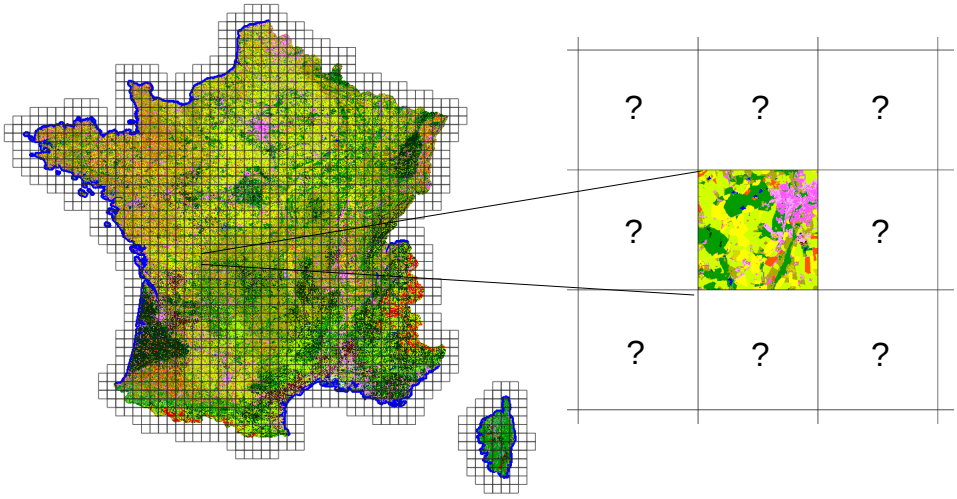

Figure 3.1: Patch based based approach. A France wide map is split into small tiles to enable processing by a neural network. This strategy creates a detrimental loss of information on the edge of the patch as each patch is analysed independently from others (right). For visibility issues 20×20 km^2 patches are represented while 6×6 km^2 are used.

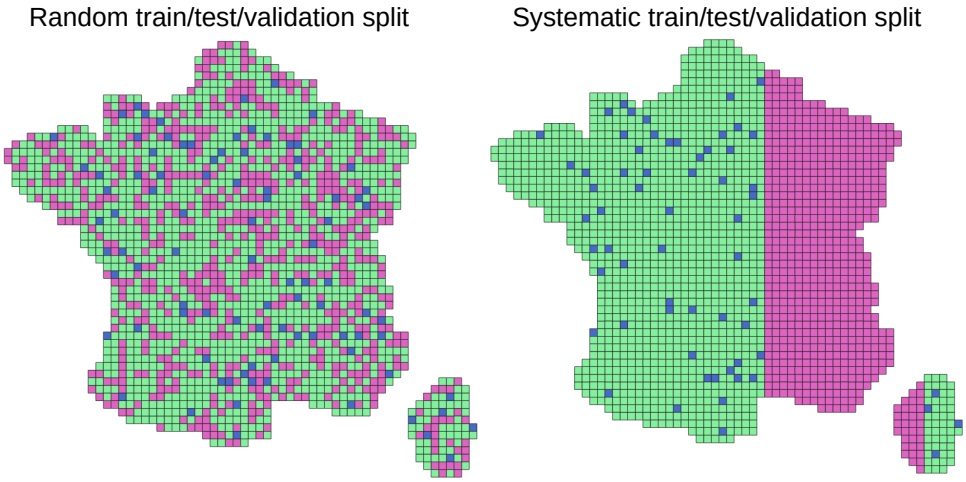

Figure 3.2: Separation of train (green), test (purple) and validation (blue) data performed randomly (left) or with a spatially systematic approach (right).

3.2 Study area and used land-cover maps

This section focuses on the land-cover maps used for the experimentation presented in this PhD manuscript. The maps shown here were selected from the numerous existing ones based on four criteria.

1. **Spatial extent**: To ensure spatial overlap between maps, we selected a spatial extent on which our study is conducted: the whole 550,000 km^2 of mainland France territory. Selected maps should be broad enough to ensure enough class and landscape diversity. We arbitrarily set up a 1% overlap with the territory selection limit. We underline that the smaller the spatial extent is, the less geographical context there is. Learning to analyse geographical context on a small extent land-cover map, even though possible, is less interesting as it provides results close to a non-context wise method.

2. **Nomenclature diversity**: Selected land-cover maps should at least include ten classes to ensure a minimum diversity of classes. Otherwise, the translation is either straightforward (as all the pixels of the source class should only be translated into a single target) or impossible (when the source class shall be translated into multiple target as the less source classes the less spatial context).

3. **Semantic accuracy**: Selected maps should exhibit at least a 70 % overall accuracy to avoid working on too noisy data. The 70% bound is chosen arbitrarily to limit the number of classes with a per-class accuracy below 50%. Indeed each class with an accuracy below this number holds more erroneous than reliable examples and is deemed to enforce the model to learn erroneous translations. We underline that removing all land-cover maps holding at least one class below 50% accuracy per class accuracy is, in practice, unfeasible as it would remove too many automatically classified maps.

4. **Open access license**: To be selected, a land-cover map must be published under an open access policy. This ensures that datasets can be shared to foster more research on land-cover map translation.

This first section gives some insightful details about France's territory's characteristics. We then review the characteristics of the six-land-cover maps fulfilling the previously mentioned criteria. Lastly, we provide a first analysis of the associated challenges in terms of compatibility for semantic translation and machine-learning-related challenges.

3.2.1 Study area

Metropolitan France, *i.e.* the part located on the European continent, covers an area of 550,000 km^2, making it the third largest country in Europe. This vast spatial extent enables building a large patch-based dataset, ensuring sufficient data to perform train/test

splits between patches with all classes and features in both datasets. Moreover, it avoids the need to restrain the patch size to very small spatial extents like the 32×32 pixels of [273], which necessarily limits the spatial context available.

France's varied landscapes make it challenging for land-cover map translation. As a coastal country, it is bordered by maritime facades to the North, West and South, with a total length of coast greater than 3,427 km. These coastlines offer varied landscapes, from fallout mountain ranges in the South-East, plateaus ending on cliffs in the North or vast sandy plains in the South-West. France's territory also offers various topographical units via multiple eroded massifs (Armorican, Central, Corsica ...) or higher ones (Alpine, Pyrenean) with a peak at 4,808 meters above sea level. These massifs delimit several sedimentary basins, especially the Aquitaine Basin to the South-West and the Paris Basin to the North. The latter exhibits particularly fertile soil resulting in an agricultural land spatial distribution asymmetry.

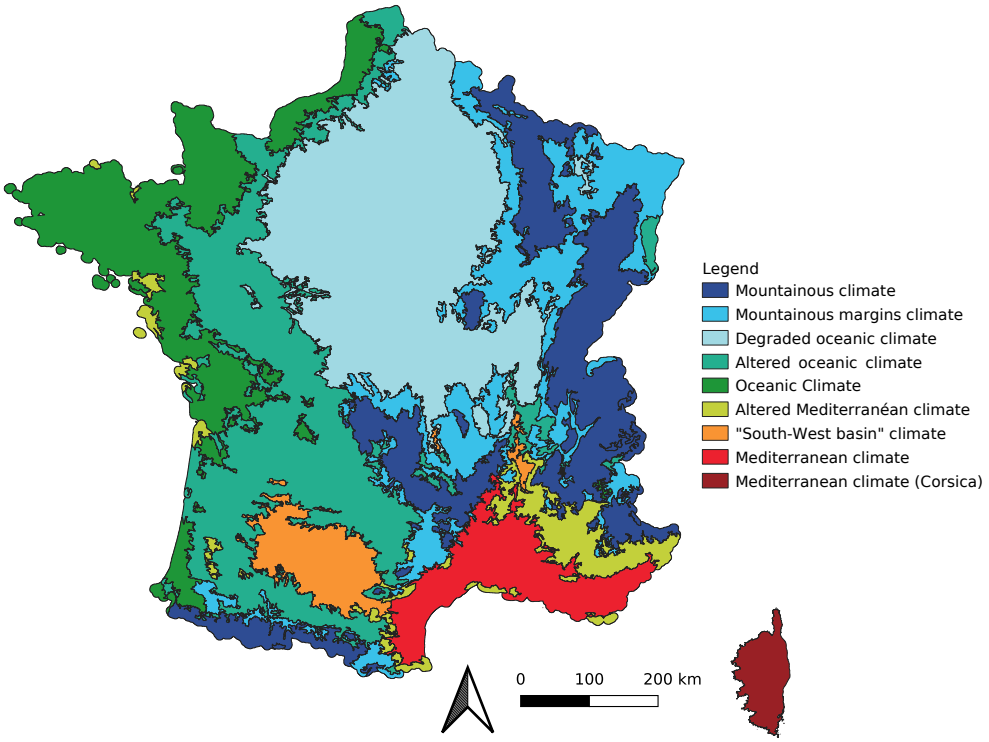

Figure 3.3: Eco-Climatic areas defined by [154]. For convenience, areas under 100km^2 are removed. We added a distinct Ecoclimatic area for Corsica, which was not represented in the original map (this climate is close to the Mediterranean one).

The fairly marked climate regional variations of metropolitan France are also interesting as they enforce intra-class diversity and spatial contexts, *i.e.* forest tree density or tree essence vary significantly depending on the climate. France's climate is organised in two gradients. The Atlantic Ocean on the West coast creates a West-East gradient with a higher wind exposure and annually more stable temperature on the West. Additionally,

the traditional North-South gradient, with higher temperatures in the South (Northern hemisphere), is reinforced by the presence of the Mediterranean sea in the South. These climates can all be declined in more local variants due to mountainous areas (East and West borderline). Figure 3.3, for instance, presents an Eco-climatic area delineation into 9 areas proposed by [154].

France also encompasses broadly different land-cover types heterogeneously distributed. About 5% of its territory is currently artificialised in scattered urban blocks, the Paris region alone representing nearly 20% of the population while accounting only for 2% of the country's total surface. This asymmetry is materialised by a "diagonal of low densities" from North-East to South-West, characterised by a low population compared to the rest of the country. Most of the surfaces are dedicated to agriculture (60%), with a prominent contrast depending on the region's climate and geological characteristics, *e.g.* a fragmented hedgerow landscape on the West side of the country and Open-field areas in the Paris region. France is also one of the most wooded countries in Western Europe, with forests occupying 34% of the metropolitan territory, mainly in the form of deciduous trees but also of conifers in the mountains and the planted forest of the moors to the West. Wetlands, which covered nearly a quarter of the country, have declined sharply since the 19^{th} century and represent less than 1% of the territory to date, as do lakes and rivers.

To sum up, the broad range of landscapes and climate makes the study area interesting for evaluating land-cover translation in a wide variety of situations. However, the prominence of two super-class, Forest and Agricultural land (94% of the territory), enforces the selection of maps with rich nomenclatures. For instance, translation into a simple 5-class nomenclature, including a Forest and a Cropland class, easily achieves high-quality results.

3.2.2 Selection of a restricted number of land-cover maps

Numerous land-cover maps respect the four criteria presented above. For instance, amongst the land-cover maps reviewed by [106], six global (MCD12Q1, GlobCover, GLCNMO, GLC SHARE, GeoWiki) and 3 Europe comprehensive (CORINE, LUCAS on 2km grid, GlobCorine) land-cover products respect the above criteria. Additionally, other global products proposed since 2015 (GlobLand30, ESRI LULC) and various France-wide (OSO) or local scale land-cover maps (OCSGE, MOS with one map per administrative area, CRIGE-PACA) all match previous criteria. As we can not analyse all possible translations, we arbitrarily choose to focus only on six maps. The selection of those six maps tries to maximise the complexity of translation by enforcing high diversity in production method, spatial resolution, nomenclature and spatial extent.

Amongst those four characteristics, the production method is the most important to diversify as it influences the three other ones. Photo-interpreted land-cover maps usually exhibit rich nomenclatures (more than 15 classes), often mixing land-use and land-cover definitions, coarse spatial resolution and are often coarser when they cover a wide spatial extent. Unlike most automatically obtained land-cover maps, they exhibit a minimum

mapping unit which correlates the classification of one pixel to its surroundings. They often exhibit fewer errors (around 10%) than automatically derived ones, with most errors being region shaped, *i.e.* multiple contiguous pixels are all misclassified identically and with the same magnitude per-class error ratio. Conversely, most automatically derived land-cover maps exhibit smaller nomenclatures (less than 15 classes), mainly land-cover oriented, as land-use is often challenging to obtain from images. Their spatial resolution is often significantly higher, and they rarely exhibit a minimum mapping unit, *i.e.* the content of one pixel is less dependent on its surrounding. They often exhibit significantly higher error rates (up to 35%), most errors exhibiting systematic spatial (edges between land-cover) and class (class close in the feature space are more likely confused) patterns. Nomenclature, resolution and error patterns of automatically obtained maps highly depend on the source data used (MODIS, Sentinel, PROBA-V...) and the algorithm characteristics (pixel classification vs OBIA classification, RF vs SVM vs Deep learning).

Lastly, the selected land-cover maps should be temporally close to ensure that the learnt translation does not suffer significantly from a temporal gap between the source and target land-cover. We arbitrarily set this closeness limit to 10 years.

3.2.3 Presentation of the input land-cover maps

Table 3.1 summarises the main characteristics of the six land-cover maps selected which exhibits a broad range of production methods (either photo-interpreted or automatically generated, different Source data), spatial resolutions (from 10 to 100 m), nomenclatures (from 11 to 44 classes, cover and use) and spatial extent (from 10,000 to 550,000 km^2): CGLS-LC100 [33], CORINE land-cover map [218], OSO [139], OCS-GE cover, OCS-GE use, and MOS.

	CGLS-LC100 [33]	CLC [218]	OSO [139]	OCS-GE cover [229]	OCS-GE use [229]	MOS
Extent	World	Europe	France	West and South France	West and South France	Paris area
Generation	Machine-Learning	Photo-interpreted	Machine-Learning	Photo-interpreted	Photo-interpreted	Photo-interpreted
Source data	PROBA-V	Landsat, Sentinel-2	Sentinel-2	Aerial imagery	Aerial imagery	Aerial imagery
Distribution format	raster	vector	raster	vector	vector	vector
Selected year	2018	2018	2018	2014-2015	2014-2015	2017
Number of classes	12	44	23	14	17	11
Pixel resolution (m)	100	100	10	10	10	20
Minimum mapping unit (m^2)	10,000	250,000	100	200-2,500	200-2,500	400
Official geometric accuracy (m)	100	100	10	5	5	5
Official semantic accuracy	73 % (Europe)	92 % (Europe)	87% (France)			
Accuracy on ground truth	80% (France)	88% (France)	86% (France)			

Table 3.1: Main characteristics of the six selected LULC maps.

In order to reduce the impact of changes occurring between two maps in the translation procedure, we carefully selected the year of the maps to make them the closest possible to each other (selected years are indicated in Table 3.1). Few maps are produced on a yearly basis which inevitably generates discrepancies between the six maps.

The Copernicus Global Land Service land-cover map (**CGLS-LC100**) map has global coverage and is released annually in raster format. Based on PROBA-V image time series

classification with a supervised Random Forest framework [33], each map covers a civil year reference period with five released versions so far (2015, 2016, 2017, 2018, 2019) [1]. Main map characteristics include a spatial resolution of 100 m, up to 22 classes (with a fine-grained separation into 12 forest labels), and hierarchically organised into a 3-level nomenclature. Level 1 merges all forest classes into one (leading to 11 classes), and level 2 distinguishes open from closed forests. We choose to rely on the level 2 nomenclature (see Table A.6), instead of the level 3 due to its higher accuracy (estimated overall accuracy over Europe of 80% at level 1, 73% at level 2 and not communicated at level 3 [310]). Indeed, our proposed solution relies on a supervised learning process: inserting a too significant noise level would be detrimental [224]. Moreover, working with level 3 labels would have also required dealing with complex classes such as *Unknown open forest types* that are poorly handled by translation systems.

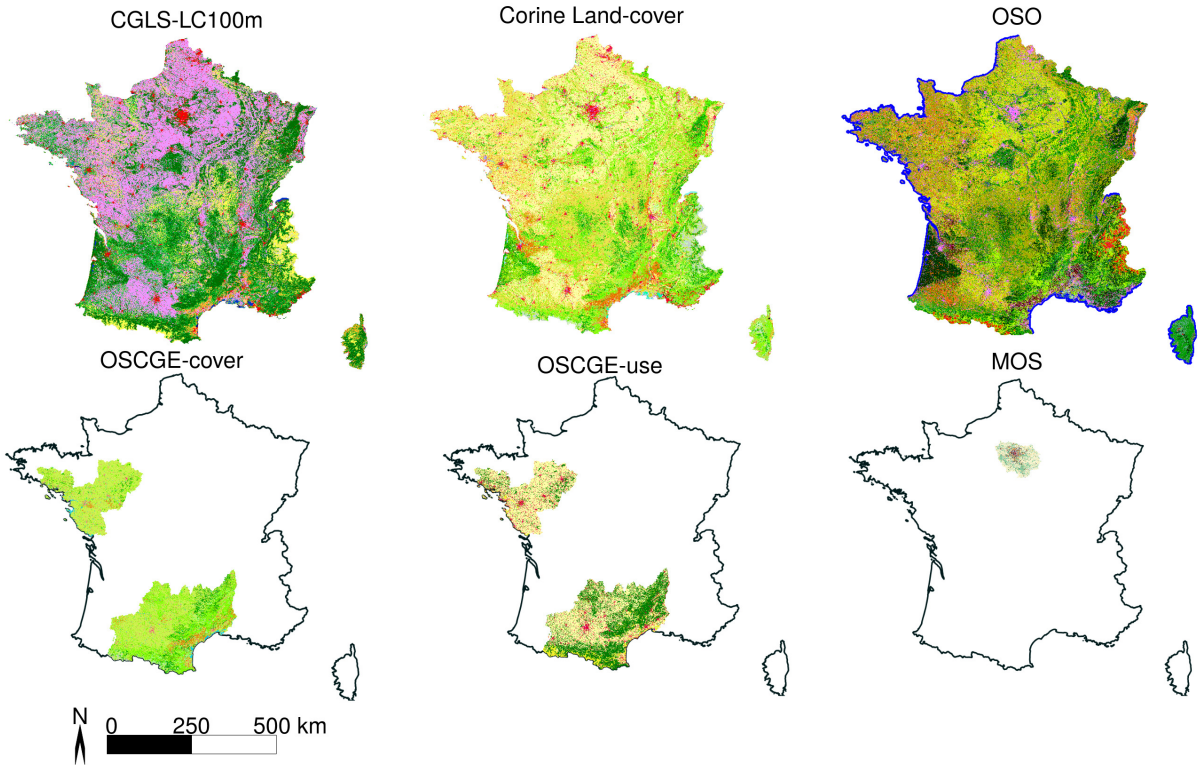

Figure 3.4: Spatial extent of the 6 land-cover maps used in this work.

The CORINE land-cover (**CLC**) database and its 92% thematic accuracy [218] has been the reference for land-use and land-cover map documentation at the European scale for the last three decades. As part of the European project Copernicus, five versions of the product have been released so far (1990, 2000, 2006, 2012, 2018)[2], covering up to 39 countries in 2018. CLC is mainly generated through visual inspection of both mono and multi-temporal high resolution/very high-resolution optical satellite images (Landsat, Sentinel-2, SPOT),

[1] `https://land.copernicus.eu/global/products/lc`
[2] `https://land.copernicus.eu/pan-european/corine-land-cover`

complemented with local databases. CLC is released dually in vector format with a 250,000 m^2 minimum mapping unit (MMU) for classes represented by polygonal objects and an additional 100 m width constraint for linear features and in raster format with a 100 $\times$ 100 m pixel spatial resolution. The nomenclature includes up to 44 classes (see Table A.2), hierarchically organised into a 3-level nomenclature. Most current translation methods relie on the first or second level of nomenclature as the translation accuracy highly depends on the semantic and spatial correspondences between the source and the desired nomenclatures, which increase for land-cover map with few classes [20]. Conversely, in the following, we target full CLC level 3 translation (44 classes) to understand better and assess which classes can be distinguished using contextual methods. Indeed, context-based translation solutions exhibit a significant potential for some challenging CLC level 3 classes (*e.g.*, *Mixed Forest*, or *Green urban areas*) that calls for fine assessment.

The Occupation des Sols Opérationnelle (**OSO**) covers Metropolitan France and is released annually in raster format. Based on Sentinel-2 image time series classification with a supervised Random Forest framework [139], each map covers a civil year reference period with six released versions so far (2016, 2017, 2018, 2019, 2020,2021). Main map characteristics include a spatial resolution of 10 m, 23 classes with a fine-grained 11 class agricultural discrimination (see Table A.1), and an overall accuracy higher than 85%. This product is valuable to this study for its high resolution coupled with a detailed crop nomenclature. The OSO product is freely distributed around April each year [3].

The Occupation des Sols à Grande Echelle (**OCS-GE**) map covers West and South-West France (125,000 km^2), and is expected to be updated at least on a 5-year basis. Based on photo-interpretation of aerial visible and near-infrared imagery at 20 cm, each administrative state is mapped independently, with the first campaign between 2014-2015 and one between 2019-2021([4]). Our work only includes 2014-2015 maps, the more recent one still being under review at the moment of this writing. Main map characteristics include a spatial class-dependent resolution between 5 and 10 m, a minimum mapping unit between 200 and 2,500 m^2 depending on the class and the location and two land-cover map/land-user nomenclatures: 14 labels for land-cover map (see Table A.4) and 17 for land-use. This joint LC/LU product is particularly interesting in studying automatic land-use prediction from land-cover map. So far, the two products are generated on the same spatial support: a territory segmentation is automatically performed using a database with road and rail network to obtain a global skeleton which is later subdivided in more refined units by photo-interpreters independently for land-use and land cover. In the remainder, we refer to those two nomenclatures as **OCS-GEc** for land-cover map and **OCS-GEu** for land-use. The choice has been made to remove the following three classes from OCS-GEu: *Other primary productions*,*Other transport networks* and *Unknown use*, due to their mixed and complex content.

The Mode d'Occupation des Sols (**MOS**) map covers the Paris region (12,000 km^2) and is

[3] `https://www.theia-land.fr/en/product/land-cover-map/`
[4] `https://geoservices.ign.fr/ocsge`

released approximately every four years in vector format. Based on the visual interpretation of 0.15 m aerial optical imagery, each map covers a civil year reference period with nine released versions so far (1982, 1987, 1990, 1994, 1999, 2003, 2008, 2012, 2017)([5]). Main map characteristics include a spatial resolution of around 20 m, up to 81 classes (with a fine-grained 68 built-up classes), hierarchically organised into a 4-level nomenclature. The choice to rely only on the 11 class level 1 nomenclature (see Table A.3) has been made since the other levels are not freely available.

3.3 MLULC challenging characteristics

This section briefly highlights some of the main challenges faced when translating between the six selected land-covers. We identify four principal challenges: nomenclature translation, resolution translation, translation of erroneous land-cover maps, and spatial generalization.

3.3.1 Nomenclature translation issues

3.3.1.1 Quantifying the nomenclature translation issues

We provide in Appendix B various tables describing independently for each source/target map the main semantic correspondences from one class to another based on a manual analysis of the main semantic links (see Section 6.2.2.4 for more detail on this procedure). In particular, we established that amongst the independent translation of the 118 accumulated classes of the dataset into each of the five other nomenclatures (590 possible source/target association), only 315 (53.3%) source classes have a single relatively close correspondence in the target nomenclature (denoted 1-to-1 translation), *i.e.* for which one translation appears more consistent than all the others. The translation of those 315 source classes results in 151 target classes; thus 25.6% of classes are semantically easily obtainable. Conversely, 11.6% of classes can never be obtained by semantic translation of a source map due to a lack of semantic correspondence (denoted 0-to-1 translation). Appendix C presents for each source to target map translation the proportion of target classes obtained from a 1-to-1 and 0-to-one translation (with two different techniques) to help identify the most challenging translation scenarios from a semantic point of view. The main observation is that the only source map usable to obtain more than 50% of target classes using a pure 1-to-1 semantic approach is CLC due to its high number of classes. All other source-to-target translation scenarios exhibit configurations in which the proportion of target classes obtained by translating one source class into multiple targets is above 70%. Obtaining high-quality semantic translation requires defining context or ancillary data-based criteria to translate each of those source classes into their multiple possible translations. Synthetically, the most challenging class to obtain are OSO fine grained

[5] `https://www.data.gouv.fr/fr/datasets/mode-doccupation-du-sol-mos-en-11-postes-en-2017/`

11 agricultural classes, CLC 11 urban classes and 5 wetland oriented classes and almost all OCGSE use classes due to a lack of semantic correspondences in the other land-cover maps.

3.3.1.2 Imperfect semantic matching

We underline that even source classes with a clear principal semantic correspondent in the target nomenclature (1-to-1) often corresponds to an imperfect translation due to: variations in class characteristics, temporal patterns and arbitrary rules.

Varying class characteristics Class definition often includes a threshold on specific characteristics expected for an object to be attributed to the corresponding classes. Those thresholds are often widely different depending on the considered land-cover map. For instance, we provide below the threshold definition to attribute the class *Broad leaved forest* for CLC (2), OSO (3), and OCSGE cover (4).

"The predominant classifying parameter for this class is a crown cover density of > 30 % or a minimum 500 subjects/ha density [...] The minimum tree height is 5 m." - CLC (2)

" Land with an absolute tree cover rate greater than or equal to 10 % with trees reaching or capable of reaching a height greater than 5 meters on-site" - OSO (3)

" Absolute tree cover rate greater than or equal to 25%" - OCSGE cover" (4)

Even though translating one of those classes into one of the other is the natural translation, it results in inaccurate translation for objects between the source and target thresholds. As many such inconsistencies are observed in the dataset, we can not present them all. However, we underline that this class characteristic variation mainly affects biotic land-cover types.

Varying temporalities We highlight that due to the production scheme of CLC, CLC class definition tends to exhibit a multi-year temporal pattern, especially for agricultural areas, shrubs and forest areas (see Section 2.3.3.1). Conversely, all other maps tend to define their classes on a year-based analysis, making the translation to CLC difficult when only one source map is used. The translation from CLC to one of those maps is ill-defined in those classes if no additional data with a date close to the target map is used.

Arbitrary rules Land-cover maps often include sets of arbitrary rules, mainly in forms of spatial constraints that separate different land-cover types. Most affected translation includes the distinction between different sorts of water objects, such as the separation between salted and non-salted waters in Figure 3.5 (*i.e.* CLC *Water courses, Water bodies, Coastal lagoons, Estuaries, Sea and Ocean*, CGLS, *Permanent waters bodies, Ocean*). Occasionally some other land-cover types can exhibit those spatially arbitrary delineations. For instance, OCSGE use *Secondary or tertiary production and residential usage* seems

to include all elements inside an arbitrary, non-constant across-territory buffer around built elements while not giving information on this buffer in the class definition. Therefore translating those classes implies to infers those arbitrary rules directly from the map by using additional data or inferring them from context.

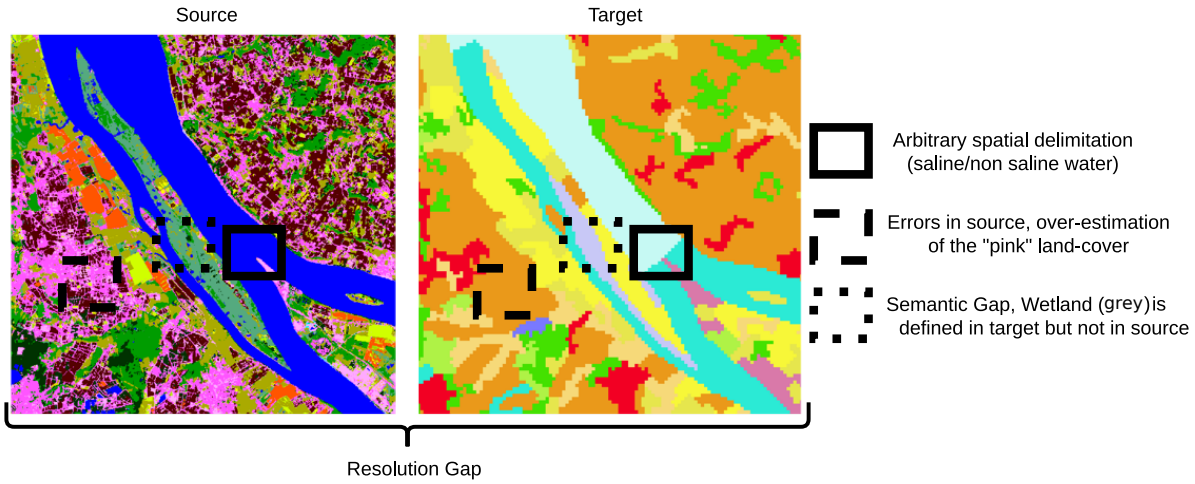

Figure 3.5: Illustration of the resolution and semantic discrepancies between OSO and CLC. See Tables A.2 and A.1 for the detailed nomenclatures.

3.3.1.3 Constraints specific to machine-learnt translation

The high over-representation of forest and agricultural lands in France (94%, see Section 3.2) compared to the rest of land cover fosters a significant inter-class imbalance, especially for maps with a high number of urban or wetland-oriented classes such as CLC and MOS or land-use oriented such as OCSGEu. For instance, in the CLC map, the ratio between the most and the least represented classes is higher than 200 in level 2 and higher than 7000 in level 3. When performing manual semantic translation, this does not represent a significant issue as long as the statistically most frequent classes are advantaged compared to statistically less observed one on 1-to-n translations (which is not the case in most current translation frameworks). However, from a machine-learning point of view, this is challenging as it requires a solution that is not too favourably biased towards the most statistically probable translation.

3.3.2 Resolution translation issues

In the dataset, the gap of pixel resolution is of a factor of 100 between the most resolved maps (OSO, OCSGE) and the least resolved ones (CGLS-LC100, CLC) and goes up to 2500 when the minimum mapping unit is taken into account. Figure 3.5 illustrates this resolution gap between OSO and CLC, *i.e.* 100 OSO pixels are resumed into one CLC pixel, and each CLC segment at least includes 25 pixels (one segment resume at least 2500 OSO pixels).

In the case of up-sampling (e.g. CLC translated to OSO), this very wide resolution gap necessarily implies using highly resolved additional data to obtain both a semantically and geometrically more refined segmentation of the territory, *e.g.* to predict 2500 pixels with potentially different target classes from a 25 pixels segment with a single source class. However, Section 2.4 underlined that very few works study how to merge additional data and maps in a context-aware fashion. Thus research must be conducted on this particular topic. Examples of usable additional data are provided with the dataset and presented in Section 3.4.

In the case of down-sampling (e.g. OSO translated to CLC), most current resampling techniques use a grid-based approach (e.g. a cell of 100 OSO pixels is resumed into one CLC pixel) to resample the land-cover map. Section 2.2.1 underlined that the technique used for resuming the information, either nearest neighbour or majority voting, neglects information for heterogeneous cells holding multiple classes. We argue that context and class proportion should be used to obtain specific classes, *e.g.* a cell mixing OSO *Water* and *Pastures* should be translated a CLC *Inland wetlands*.

From a machine-learning point of view, we underline that learning the MMU on the dataset is a difficult task that requires a spatial context-aware translation framework that learns rules such as: *'If two adjacent areas of discontinuous and continuous urban fabric occur, each of them <25 ha, but in total >25 ha, they should be mapped as one single polygon, and discontinuous urban fabric is privileged'*. The main difficulty in learning the minimum mapping unit is that errors in the source or target maps can be perceived as minimum mapping rules when they exhibit systematic patterns. For instance, if small discontinuous urban fabrics are often confused with industrial and commercial units, the method could transform the previous rule in 'If two adjacent areas of industrial and commercial unit and continuous urban fabric occur, each of them <25 ha, but in total >25 ha, they should be mapped as one single polygon, and discontinuous urban fabric is privileged'. The impact of errors on the dataset is addressed in the next Section.

3.3.3 Errors

Appendix E presents the confusion matrices of the France wide-land-cover maps (CLC, CGLS-LC100, OSO) computed on the ground-truth provided with the MLULC dataset (see Section 3.5.1) (official confusion matrices are available but produced using heterogeneous methods and does not cover the same spatial extent). From these confusion matrices we derive Figure 3.6, which presents the dispersion of per-class accuracies.

We observe that most of the land-cover maps exhibit a high precision variation depending on the considered class. For instance, almost half of CGLS-LC100 level 2 classes exhibit an accuracy of less than 70% and two classes of OSO exhibit a precision below 30%. From a translation point of view, this is challenging, as errors in the source map have a high risk of being mistranslated into the targeted nomenclature, *i.e.* translating from a map with 70% accuracy is unlikely to give a target map with higher accuracy.

Finally, unlike semantic-based translation methods that are very sensitive to errors in source data, machine learnt methods are more sensitive to errors in target data. For instance, a per-class analysis of the OSO product reveals that the OSO *Road surfaces* are more than 50% of the time misclassified as *Industrial and Commercial units*. Thus, the translation from MOS *Transports* to OSO is likely to learn to translate into *Industrial and Commercial units*.

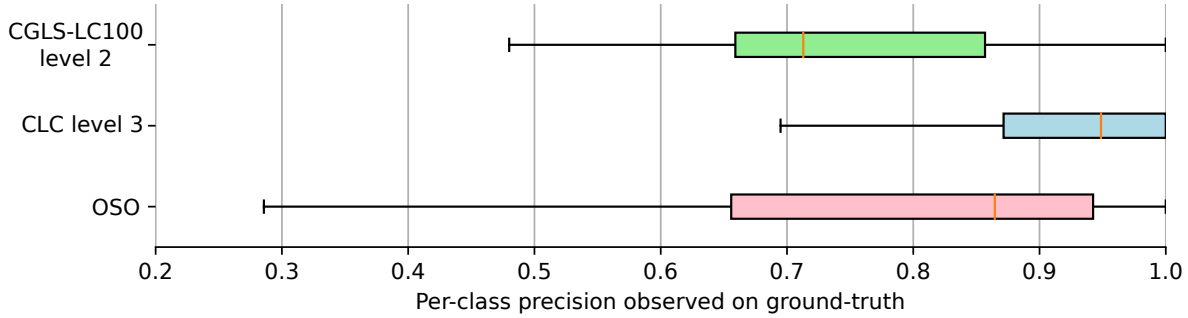

Figure 3.6: Box plot illustration of per-class accuracy dispersion for the three France-wide land-cover maps of the dataset. Per class accuracy are computed on the ground truth presented in Section 3.5.1.

3.3.4 Spatial generalization

Since several maps do not cover the whole French territory, the number of available maps varies depending on the considered location, as shown in Figure 3.4. Using this particularity, one can study spatial generalisation to unseen areas on a real operational setup, e.g. translate the OSO map into MOS (less than 2% of the territory) to obtain a France wide MOS. However, this raises two main concerns. First, some OSO classes such as *glaciers*, *natural grassland*, or *bare soils* are never seen in the MOS original extent (see Figure 3.4). The translation from those OSO classes to a MOS class is never learnt. In this configuration, a first solution is to rely on rule-based non-contextual translation for those specific classes. Developing a machine learning strategy to address this issue is far more challenging and is discussed in Section 5.1. Consequently, they are hard to fit in the MOS nomenclature. Secondly, from a machine learning point of view, this difference in spatial extent creates a significant imbalance between the number of patch of the different LULC, making approaches performing multiple translations at the same time or multi-data fusion more difficult.

3.3.5 Conclusion

As we aimed to study translation in complex cases, the six selected land-cover maps exhibit complicated nomenclature and resolution translation problems and include a significant number of errors.

From the observation that the spatial co-occurrence between source and target classes observed between two maps of the dataset directly depends on the nomenclature, the resolution and the errors, we propose to summarise the difficulty by comparing the semantic-based nomenclature translation matrices provided in Appendix B with the spatial co-occurrence. As an illustration of this idea, we provide Table 3.2 in which we compare the observed percentage of spatial co-occurrence between OSO 2018 pixels at 10 m and CLC 2018 (at level 2 for readability) with their possible semantic association, e.g. one pixel labelled as *Discontinuous urban fabric* in OSO should semantically be translated into *Urban fabric* in CLC. In contrast, this association is only observed 45% of the time in the dataset.

OSO Classes	**CLC Classes**	**Observed**
Dense urban	Urban fabric	87%
Sparse urban	Urban fabric	45%
Industrial and commercial	Industrial, comercial and transport units	14%
Roads	Industrial, comercial and transport units	16%
Rapeseeds	Arable land / Heterogeneous agricultural areas	91%
Cereals	Arable land / Heterogeneous agricultural areas	90%
Protein Crops	Arable land / Heterogeneous agricultural areas	91%
Soy	Arable land / Heterogeneous agricultural areas	91%
Sunflower	Arable land / Heterogeneous agricultural areas	89%
Maize	Arable land / Heterogeneous agricultural areas	83%
Rice	Arable land / Heterogeneous agricultural areas	96%
Tubers	Arable land / Heterogeneous agricultural areas	96%
Orchards	Permanent crops / Heterogeneous agricultural areas	62%
Vineyards	Permanent crops / Heterogeneous agricultural areas	82%
Pastures	Pastures / Heterogeneous agricultural areas / Artificial non-agricultural vegetated areas	69%
Lawn	Shrub and herbaceous associations	39%
Shrub	Shrub and herbaceous associations	41%
Broad leaved	Forest / Artificial non-agricultural vegetated areas / Heterogeneous agricultural areas	82%
Coniferous	Forest / Artificial non-agricultural vegetated areas / Heterogeneous agricultural areas	79%
Mineral surfaces	Open space with little or no vegetation / Mine, dump and construction sites	86%
Sand	Open space with little or no vegetation	65%
Glaciers and snow	Open space with little or no vegetation	100%
Water	Inland water / Marine Water	84%

Table 3.2: Semantic association between OSO and CLC level 2. The third column gives the observed percentage of spatial co-occurrence between one of the source class and all the proposed semantic translation(s) *i.e.* only 62% of OSO orchards pixels are translated into one of the semantically corresponding CLC class *Permanent crops* or *Heterogeneous agricultural areas* .

3.4 Ancillary data for enhancing the translation process

In the previous section, we explained that land-cover map translation is a difficult task that can even be impossible to achieve using just one source/target pair of maps due to a lack of semantic correspondence, a gap in resolution, and errors. In this section, we briefly present additional data that can be used to improve the translation quality when facing one of those cases. More specifically, we introduce data rich in semantic information to help with the lack of semantic information and errors and geometric information to help with the resolution gap. We investigate three sorts of data: Optical imagery, Synthetic aperture radar imagery, and Digital Elevation model. Mono-temporal data acquisitions are used to

increase the translation quality in those challenging cases. Multi-temporal acquisitions are deliberately ignored as we focus on the complementarity between additional data and land-cover maps for translation rather than producing the best quality maps. The choice of the data provider is made according to three main factors, the open access rights, which enable to share the dataset, the resolution of the data, which should be close to or higher than the most resolved map (10×10m), and the worldwide availability to enable comparable results on different spatial extents.

3.4.1 Optical imagery: Sentinel-2

Optical imagery, also termed optoelectronics, acquires information based on the reflection of the sunlight on the Earth's surface. The data is collected using mainly the wavelength between 400 and 1400 nanometers. Amongst the multiple optical data providers, we selected Sentinel-2 according to three criteria:

- A 10-meter spatial resolution, on par with highest resolved maps, to test the coarse to fine resolved translation in a non-ill-defined setup. As we do not seek to test the image potential for a given target land-cover, but it's complementarity with the source, more resolved images are irrelevant as the additional gain in translation accuracy would only be due to the image.
- The availability of cloud-free synthesis. As we proceed to a mono-temporal analysis at a France-wide scale, we require cloudless images to proceed with relevant data fusion.
- Free open access right enabling redistribution. This enables the distribution of the MLULC dataset.

The twin satellites Sentinel-2A and 2B are carrying multi-spectral optical sensors acquiring thirteen spectral bands at different spatial resolutions [72]. As we are mainly interested in gaining geometric information, we only keep the four 10-meter resolved bands (Red, Green, Blue and Near-infrared) in our analysis. The two satellites exhibit the same orbit but are phased at 180° to enable a temporal resolution of 5 days at the equator (slightly better at higher latitudes).

Two main levels of pre-processing Sentinel-2 images are distributed. Level 1 provides geometric correction to account for the main image distortions factor (satellite's motion, Earth's rotation, acquisition angle, orbit and terrain topography) to provide an accurate location for each pixel. Level 2 adds to level 1 the atmospheric properties, which are not constant in time and thus require corrections for applications performing a temporal analysis of an object, *i.e.* it removes exogenous factors such as atmospheric absorption or scattering.

As we do not plan to use multi-temporal acquisitions of Sentinel-2 but only a single one across all of France, those two processing levels are insufficient. Indeed, they do not

consider the cloud cover, which is more or less critical depending on the geographical areas and strongly degrades the data availability [289]. We relied on the Weighted Average Synthesis Processor (WASP) algorithm [109] implemented by the Theia French Land Data Center to produce a cloudless synthesis image. The core idea of the WASP processing chain is to perform a weighted average of all the Sentinel 2 images available during 45 days. The weight of each date is determined by taking into account both the time gap between this date and the mid of the 45-day date (to ensure a spectral coherence between adjacent pixels) and the likeliness of the pixel being cloud occulted. An almost cloudless single France wide-image with only a slight spectral shift between two tiles of different orbits is obtained. However, this processing chain still exhibits artefacts when all acquisitions during the 45-day periods (8 to 9 acquisitions) are cloudy on the same pixel. Artefacts are also observed on surfaces, such as snow and water, that change quickly over time.

Adding geometric information to the land-cover translation, such as the target segmentation, using an image, implies using an image from which target segmentation can be inferred. We observe that the potential of target segmentation inference from an image widely depends on the considered date, especially for agricultural land-scapes (60% of French territory) in which the visible segmentation varies across the year. The choice of the 45-day period is essential as land-cover map mapping accuracy widely depends on the date used due to seasonal variation [185]. A first constraint on selecting the date is that the usage of 45 days (9 images) does not ensure a cloudless synthesis, especially in the very cloudy months between October and March (the kept image must be outside of this period). To our knowledge, no studies focused on the best dates to perform land-cover map mapping over France. However, some thematic-specific studies on urban areas [248] have demonstrated to produce more accurate land-cover maps during the Spring or Autumn season. Since we previously mentioned that France's land-cover map mainly includes croplands (nearly 60%) and that most of them are already harvested in the autumn season, we selected a spring season cloudless synthesis centred on mid-April 2019. Ideally, we would have processed the 2018 image to be as close as possible to land-cover maps, but the synthesis of 2018 was too cloudy.

Last but not least, the image is aligned with the different land-cover maps in an alignment step. It mainly consists of a re-projection from EPSG:4326 to EPSG:2154 on the same grid as the one used for the land-cover map. This involves resampling, for which we relied on a bi-cubic interpolation. As such, the effective resolution of the obtained image is slightly under the original one.

3.4.2 Radar image: Sentinel-1

Radar imagery acquires information based on the reflection of a radar signal emitted by the satellite on the Earth's surface. The data is collected using a wavelength between a cm and a few meters. Since the reader might not be familiar with the SAR imaging system, this section provide critical relevant information for understanding the dataset. However,

for a real in-depth explanation of the radar imaging system, we refer the reader to [14, 62, 219]. We selected the Sentinel-1 data following the same criterion list as the one used to select the data provider for optical images.

The twin satellites Sentinel-1A and 1B carry a C-SAR (synthetic aperture radar) instrument acquiring in dual polarisation (HH+HV, VV+VH) with an incidence acquisition angle between 20-46°. This right-sided acquisition angle has many consequences, such as producing images with a pixel resolution varying across the satellite track. The Sentinel-1 data are either distributed with all the possible information (complex signal with phase and amplitude) and with the unequal pixel size in a mode called Single-Look Complex (SLC) or in a simpler form including only the modulus of the signal (with no phase information) in a resampled $10m \times 10m$ pixel-sized image in a mode called Ground Range Detect (GRD). As our work does not involve phase exploitation, traditionally used for SAR interferometry, we only discuss the characteristics of the GRD product.

Radar images' main characteristics stem from the used wavelength in the case of Sentinel 1, the C-band, which corresponds to 5.5 cm. As a diffraction-limited system [209], a radar imaging system cannot acquire information on objects too small compared to the wavelength used. Clouds mainly consisting of water droplets of a millimetre width are seen through, enabling image acquisition even in cloudy weather. More generally, the use of a significantly higher wavelength than optical imagery brings widely different information. For instance, the signal tends to penetrate the tree and soil cover a bit before being reflected. It provides additional geometric information that is not seen in an optical image.

However, because the radar signal is coherently emitted in space and time, the multiple backscattered signal of a single pixel content creates some constructive and destructive interference. This sum of signals inside each pixel results in a noise-like pattern referred to as speckle, which can be modelled as a multiplicative noise (assuming a set of realistic assumptions referred to as the Goodman hypothesis [102]). As speckle makes the radar images usage considerably harder, various denoising (despeckling) methods have been proposed. Local spatial filtering methods have been proposed and perform local weighted means to reduce the noise at the cost of down-resolving the image. Temporal filtering methods have also been proposed. They use the random aspect of the noise by averaging multiple noise-independent acquisitions, with the drawback of being inconsistent on surfaces changing over time. Lastly, mixes between those strategies have also been proposed to alleviate the problems of both methods.

Since we mainly want to preserve the geometric information rather than the radiometric one, we propose performing a temporal average of 3 years of sentinel-1 acquisition across France (2017-2019). As the SAR acquisition is realized with an angle, each object can be seen from two points of view depending on the orbit direction (ascending or descending). We randomly selected the descending. Furthermore, we only included images from the S1B satellite to avoid averaging with the S1A, which has very slightly yet different acquisition parameters. This results in averaging up to 120 GRD images on some locations producing a nearly noiseless SAR image.

Preprocessing steps include: calibration to consider the incidence angle and remove thermal noise, orthorectification and a temporal average. As the last step, we reprojected all the images in EPSG:4326 along the grid used for land-cover maps, which involves a bi-cubic resampling. All the processing steps except the temporal average are proceeded using the S1Tiling software[6].

3.4.3 DEM: Alos-World3D

Elevation and land-cover are highly intertwined, making elevation a good additional data to increase the quality of land-cover translation. For instance, France vineyards tend to be cultivated on lands with a slope (that ensures excellent drainage properties) south oriented (to ensure high solar exposition).

Ideally, the digital elevation model (DEM) should be around 10 meters resolved to exhibit the exact resolution as the most resolved maps. However, even though some corresponding digital elevation models are available in France, they do not provide the worldwide extent mentioned. We choose to rely on one of the two publicly available worldwide DEM at 30×30m resolution: Alos-World3D [2] (AW3D30) was chosen over the SRTM due to its higher height accuracy [264].

The AW3D30 provides a single image, with each pixel representing the height of the object inside with a 5-meter precision. The data is reprojected in EPSG:2154 using the same grid as the one used for land-cover maps and a bi-cubic resampling algorithm.

Commonly computed DEM-derived features (slope, aspect, topographic position and roughness index) are also provided to help to increase the quality of the translation. In particular, we underline that high slope areas are often mineral areas, that some crops, such as vineyards, exhibit a specific slope orientation towards the sun (aspect), and that wetlands are often observed downhill (topographic position index).

3.5 Ground truth and quality measurement

3.5.1 Ground-truth datasets

Comparison between translated and target maps is the simplest way to assess the quality of the translation. Comparison can be performed pixel-wise all over the test set, offering many samples per class to evaluate absolute and per-class metrics. However, those measurements are maximised only when the translation exhibits the same errors as the target data, *i.e.* a translation corresponding at 100% to CGLS is a translation replicating the 30% error rate of CGLS. We refer to this comparison as an *agreement measure* rather than an accuracy measure. Moreover, it is worth noting that the agreement measure can only be computed

[6] `https://gitlab.orfeo-toolbox.org/s1-tiling/s1tiling`

on the intersection between the source and target maps. For instance, the agreement measurement between the translation from MOS to OSO and OSO can only be computed on the Paris area, while the result of the CLC-to-OSO translation can be computed over entire France. Consequently, those two agreement measurements are not comparable.

Conversely, the comparison with an independent ground truth gives a better estimate of the accuracy. However, creating such a ground truth on each specific map spatial extent for all of the six maps with a large enough sample to compute significant per-class accuracies [85] is unrealistic for both time and lack of expertise reasons. This ground truth should be country-wide (to study generalisation to broader spatial extents) and with classes compliant with the specifications of each map. During this PhD, we released two ground truths datasets; one including ground truth annotation for the six land-cover maps of the MLULC dataset, including 2,300 points and one specifically focusing on CLC at level 2 but with a higher number of annotations (5,000) to ensure enough per-class sample to estimate per-class metrics. As the creation procedure of those two ground truths are identical, we only focus on the broad one (including the six land-cover maps) in the remaining to simplify the reading.

The choice of the ground truth sample size and sampling strategy directly influences the precision of the computed metrics. To enable a fair comparison between different land-cover maps, the sampling strategy consist in annotating the same sample for all land-cover maps, *i.e.* one ground truth point is annotated six times. As a direct consequence, stratified random sampling strategies aiming to provide enough samples per class to compute accurate per-class metrics are difficult to conduct. We choose to rely on pure random sampling, which provides accurate overall accuracy across the territory at the cost of poor per-class metrics estimation. For a suitable ground truth sample size n ensuring reasonable overall accuracy evaluation, we rely on Equation 3.1 [52, 230]:

$$n = \frac{z^2 \alpha(1-\alpha)}{m^2}. \tag{3.1}$$

where z is a percentile from the standard normal distribution, α is the real overall accuracy and m is the authorised margin of error. We arbitrarily choose $z = 1.96$ and $m = 2\%$, corresponding to a 95% confidence interval with a 2% margin on the overall accuracy prediction as those values are the most commonly found in land cover literature. We choose to set $\alpha = 50\%$ (the worst-case scenario leading to the largest available sample size), as we can not predict the real overall accuracy of the different translations. This results in an expected ground truth sample size $n = 2,300$.

These 2,300 points are randomly sampled from the test set; thus, they can not be used to compute per-class accuracy due to the low (or null) number of samples for rare classes. We also provide 400 additional points (non-randomly sampled), focusing on rare classes to ensure an arbitrary minimum number of 15 points per class to enable rough per-class metric estimation. The points in the ground truth are sampled with a minimum distance of 2.5 km to reduce spatial correlation. The MMU of CLC on linear elements does not

guarantee independence below this distance.

Ground truth labelling relies on photo-interpretation of Sentinel-2 (multiple dates available) and Spot 6 and 7 (High resolution) imagery, and two independent sources of information: (i) the French authoritative cartographic database (BD Topo), yearly updated at 2 m with more than a hundred classes, and (ii) the national Land Parcel Information System (RPG, Registre Parcellaire Graphique), a 10 m farmers declarative database for European Common Agricultural Policy (CAP) [36]. We consider the target data valid unless it disagrees with those sets, in which case photo interpretation is performed. Such ground truth exhibits multiple limitations:

- First, the BD Topo and RPG only cover about 75% of France since some structures are excluded (*e.g.*, sidewalks), and information is lacking (missing farmer declarations, especially for crops not included in CAP subsidies). Thus exhaustive ground truth representation of some classes is not guaranteed.
- The generated ground truth is a partially corrected version of the original data instead of a completely independent ground truth (*i.e.*, favourably biased toward the original data).
- The 400 additional points are mostly added on the rarest of the 44 CLC classes to increase their sample size. Thus they abide by the 25 ha MMU of CLC, which significantly affects statistics for other maps. For instance, most CLC *Sport and leisure facilities* additional points added are *golfs* since they cover large surfaces and subsequently artificially enriches the MOS *Artificial green urban areas* with numerous golfs.
- A France-wide ground truth can only be used to assess France-wide quality: evaluation of translation quality on a smaller extent (MOS, OCSGE) can not be performed.

The obtained ground truth enables fairly computing the original land-cover accuracy. The obtained accuracy for OSO (86%) and CGLS-LC100 (73%) (is on par with their official nomenclature (respectively 87% and 72% at the European level for CGLS-LC100) while the one of CLC strongly differs (89% on the ground truth compared to 94.2% in [218]. This stems from multiple variations with CLC's quality assessment protocol :

- Two different operators double-checked the CLC official validation dataset, while ours includes only one interpretation for each point.
- The official validation is achieved with respect to the CLC initial segmentation (to avoid taking into account geometric errors, separately evaluated), while ours correct CLC segmentation when needed. Therefore, our interpretation of the same point might differ, especially on the edges.
- The official validation is performed on the vector data, while ours is performed on a rasterised version, which tends to amplify CLC segmentation errors.

Evaluation data	Target map	Random ground truth	Enriched ground truth
Sample size	>100,000 for all LULC	2,300	2,300+400
Minimum sample per class	>1,000	0	10
Pros	- Big sample size = per-class metrics are computable	- France wide coverage for all maps	- France wide coverage for all maps
Cons	- Same errors as target data - Only covers the target extent	- Small minimum sample per class	- Partially biased to increase sample size of rare classes
Usage	- Overall accuracy - Per class accuracy	- Overall accuracy - Generalisation	- Per class accuracy - Generalisation

Table 3.3: Summary of the characteristics of the three data sets used for translation evaluation.

3.5.2 Quantitative indices

Quantitative evaluation is performed using traditional land-cover metrics to compare with other works. In particular, we provide the confusion matrix to discriminate the nature of errors. The Overall Accuracy computation is used to account for global quality. LULC data sets are highly class-imbalanced: high accuracy can be achieved by correctly predicting the most frequent classes (often not the most difficult to discriminate). Therefore macro f1-score and Kappa are computed to assess the quality of the translated classes more accurately. Standard per-class metrics (precision, recall, f1-score) are also computed. We underline that other metrics used in semantic segmentation could have been used but exhibit either a perfect correlation with one of those indicators (IoU and f1-score), or are not commonly found in the land-cover field (mean Average Precision) preventing comparison with other works. Formulas for per-class metrics and overall metrics are given in Equations 3.2 and 3.3, respectively.

$$p_i = \sum_{j=1}^{c} \frac{m_{ii}}{m_{ji}}, \qquad r_i = \sum_{j=1}^{c} \frac{m_{ii}}{m_{ij}}, \qquad F1_i = \sum_{j=1}^{c} 2\frac{p_i r_i}{p_i + r_i}. \tag{3.2}$$

$$OA = \frac{\sum_{i=1}^{c} m_{ii}}{\sum_{i,j=1}^{c} m_{ij}}, \qquad mF1 = \frac{1}{n}\sum_{j=1}^{c} F1_i, \qquad K = \frac{OA - p_e}{1 - p_e}. \tag{3.3}$$

p_i, r_i, and $F1_i$ are the precision, recall, and f-score for a given class i, respectively. OA is the overall accuracy, $mF1$ is the macro f1-score, c is the number of classes, and m_{ij} is the element in the j^{th} row of the i^{th} column of the confusion matrix, *i.e.*, the number of pixels of class j classified as i. p_e is the hypothetical overall accuracy obtained with a classifier replicating the class proportions randomly. These statistics are computed separately by comparing the translation with (a) the target LULC and (b) the ground truth. For clarity, we denote OA_{ag} and $mF1_{ag}$ the metrics computed by comparing the translation to the

target and OA_{gt} and $mF1_{gt}$ the metrics computed by comparing the translation to the ground truth. We underline that all the metrics displayed in this manuscript are rounded to the closest per cent, as we believe variations below a per cent are not interesting from an operational point of view. Moreover, to be statistically significant, it would require dozens of independent training for each experiment as the results between two different training of the models of this manuscript often vary up to 0.25%.

Evaluating the quality of the geometric preservation of a map is rarely done in land-cover map classification studies while being extensively studied in image reconstruction or denoising [262]. Thus, we lack predefined quantitative measures to assess it. We propose to use the Edge Preservation Index (EPI),also termed Edge correlation [199], traditionally used in [351]. It computes the correlation between edges in a pair of images. EPI is defined as [265]:

$$EPI = \frac{\Gamma(\Delta l - \overline{\Delta l}, \Delta t - \overline{\Delta t})}{\sqrt{\Gamma(\Delta l - \overline{\Delta l}, \Delta l - \overline{\Delta l})\Gamma(\Delta t - \overline{\Delta t}, \Delta t - \overline{\Delta t})}}. \tag{3.4}$$

where l is the map predicted by the network, t is the ground truth target, Δ denotes a high pass filter (we use a simple Laplacian kernel), $\overline{\Delta l}$ denotes the mean of the high pass filtered image, and Γ is defined as:

$$\Gamma(x, y) = \sum_{i,j \in (nr,nc)} x(i,j) \times y(i,j), \tag{3.5}$$

where nr and rc represent the number of pixels in rows and columns, respectively, of the given image.

Since we work on categorical data, edges are discretized ($0 = non\ edge$, $1 = edge$). Ideally, the EPI should be computed between the prediction and perfectly segmented ground truth. Since creating an object-oriented ground truth would be overly time-consuming, we compute the EPI between the prediction and the target. It assesses the agreement between the prediction and the target edges rather than the true correlation between the prediction and a perfect segmentation. The larger the EPI value is, the more edges are maintained. The EPI value is highly dependent on the expected proportion of edge in the target, which makes it quite comparable to the dependence of the proportion of each class of the *kappa* value. Unlike Kappa, for which interpretation key enabling distinction between low and high agreement have been provided [173, 281], no such key exists for EPI. We propose to build an interpretation key for each land-cover maps (see Table H.1) using a protocol described in Appendix H. We propose to use the average of each threshold resulting in the following key: poor (EPI$<$0), slight ($0 <$ EPI < 0.16), fair ($0.16 <$ EPI < 0.30), moderate ($0.30 <$ EPI < 0.45), substantial ($0.45 <$ EPI < 0.60), almost perfect (EPI > 0.60).

CHAPTER 4

Mono land-cover map translation

This chapter tries to answer the question: How to translate from a source to a target leveraging various context information? In this direction, we decompose the chapter in the following sections.

First, Section 4.1 introduces baseline land-cover map translation methods directly inspired by the literature. As current literature relies principally on a resampling followed by nomenclature-level rule-based translations, these non-contextual methods are assimilated as a reference group to evaluate the performance of the context-aware methods presented in other sections.

Secondly, Section 4.2 investigates the potential of spatial context for translation. First, a method based on standard manually defined shape indicators is used to perform context-wise translation using the shape of the object a pixel belongs to. Secondly, a convolutional neural network is used to replace manually defined features with learnt ones, hoping to extract spatial context information better.

Section 4.3 focuses on the geographical context. First, we identify and compare two data sources of geographical context, *i.e.* ecoclimatic areas and geographical coordinates. Two different strategies are proposed to leverage these data based on independent model training per area or direct incorporation as learnable features to train a single model. We then focus on incorporating geographical coordinates in the Convolutional neural network proposed in the previous section.

Section 4.4 addresses the temporal context. A first investigation of the potential of multi-temporal analysis is conducted. In particular, a distinction between improvements due to higher temporal information and those due to a higher ability to identify source errors is conducted. Secondly, the effect of a temporal gap between the source and target map used for learning the translation is assessed and linked with real operational situations.

Section 4.5 addresses the cartographic context. We propose a loss-based method to learn translation when the target data is heavily noisy (30% when the target is CGLS) using the noise confusion matrix provided by the land-cover map producers.

All experiments are conducted on the MLULC dataset introduced in Section 3, leading to 26 possible translations with various resolutions and nomenclatures. Therefore we

expect the results to be generalizable to other land-cover maps. As no additional data is considered in this section, we underline that the presented translations results from coarsely resolved sources (CGLS and CLC) to highly resolved target maps (OSO, OCSGE use, OCSGE cover, MOS) are ill-defined but still presented for reference.

4.1 Map translation without contextual information

4.1.1 Motivation

This section proposes a set of translation baselines, inspired by the state-of-the-art, that we used as references to estimate the potential of the context-wise methods introduced later on. The six proposed baselines, summarised in Table 4.1, reflects the diversity of translation approaches proposed in the current literature. We identify three key differences between translation methods: soft or hard association, semantic or statistics translation, and learnt or manually defined resampling. We review them briefly below.

Section 2.1.2 distinguished the hard association, *i.e.* one source class is translated into a single target class, from soft association, *i.e.* one source class have varying translations depending on the considered pixel. We underlined that current soft association techniques either require downsampling or merging multiple source. As this thesis mainly focuses on the translation framework rather than on merging hypothetically available maps, soft association baselines proposed in this section relies on source down-sampling.

We also distinguished baselines according to the nature of the translation performed: Semantic or Statistic. Semantic translation (**HardSem**, **SoftSem**) of source relies on a set of manually defined rules linking the source and target nomenclature with the main advantage of not requiring an existing target sample for training. Statistic based methods (**SoftStat**, **SoftMaxProba**, **SoftLearntFreq**, **SoftLearntGridPattern**) evaluate the correspondence between source and target classes by computing spatial co-occurence between source and target classes on pairs of spatially overlapping source and target.

Finally we distinguish machine-learnt statistic downsampling strategies (**SoftLearntFreq**, **SoftLearntGridPattern**) from manually defined ones (**SoftSem**, **SoftMaxProba**). Manually defined downsampling techniques combine translation conducted independently for multiple source pixels to obtain a single target class. This approach inherently ignores the potential translation synergies of the different source pixels. For instance, *Water* and *Grasslands* are not very close semantically and statistically from *Rice crops* when analysed separately. Conversely, when 2 pixels, one of *Water* one of *Grassland*, are to be translated together into a single target class, *Rice crops* is a semantically and statistically valid translation. The machine-learnt statistic downsampling strategies enable an analysis of those source class compatibilities instead of considering each translation independently.

In the next subsection, we present the implementation details of each corresponding baseline used in this manuscript.

	Method		Characteristics				Related works
	Nomenclature translation	Resolution translation	Translated class diversirty	Sensitivity to noise in target	Sensitivity to noise in source	Apply MMU rules	
HardSem	Semantic-based : Highest semantic affinity association	- After nomenclature translation - Majority voting	$-$	$+++$	$---$	$---$	[243, 305, 317]
HardStat	Statistic-based : Highest probability association	- After nomenclature translation - Majority voting	$-$	$--$	$--$	$---$	[225]
SoftSem	Semantic-based : Highest combined semantic affinity association	- After nomenclature translation - Majority voting	$+$	$+++$	$---$	$---$	
SoftMaxProba	Statistic-based : Highest combined probability association	- After nomenclature translation - Sum of probability	$+$	$--$	$--$	$---$	[60, 92]
SoftLearntFreq	Statistic-based : Learnt class frequency association	- Simulatenous with nomenclature - Frequency association	$++$	$---$	$-$	$+$	[56, 180]
SoftLearntGridPattern	Statistic-based : Learnt class spatial pattern association	- Simulatenous with nomenclature - Spatial pattern association	$++$	$---$	$-$	$+$	

Table 4.1: Presentation of the six baseline methods used for comparison. We represent positive/negative characteristics in the forms of $+/-$. The number of signs denotes the degree of positiveness/negativeness of the method on an arbitrary scale ranging from $---$ (the method is the worst possible for this characteristic) $+++$ the method is the best.

4.1.2 Baselines implementation

Three of the six baselines mentioned below (**HardSem**, **HardStat**, **SoftLearntFreq**) are found in the literature. The three others are natural adaptations of close-by methods and are part of our contribution to the translation problem.

HardSem (Figure 4.1) is the most commonly found approach in literature. It relies on a separate nomenclature and resolution translation. Most works perform the resolution translation before the nomenclature translation. However, we argue that the resolution translation should always be performed on the target nomenclature to keep the most relevant information and thus proceed in this order: **HardSem** first applies a semantic translation, then the resolution one. The semantic translation is performed in a hard association manner, *i.e.* one source class is associated with one unique target class. Since finding the main correspondent of each source class into the target nomenclature is straightforward for a human operator, the association is performed manually instead of using a semantic harmonisation tool such as LCCS. The results of this one source/one target class association is provided in Appendix A with the nomenclature of each of the 6 maps. For instance, all the pixels of class *Individual housing* of the MOS map are translated into *Built up* in the CGLS-LC100 map. Most current papers perform the resampling using the nearest neighbour method. This is not detrimental when the target resolution is equal to or finer than the source. However, this resampling method is the worst possible when the target resolution is coarser than the source. n source pixels are resumed into a single target one using only the content of one of them, thereby ignoring all the others. Instead, we propose to perform, in this case, the resampling using a majority voting rule, *i.e.* the class of the target pixel is the most frequently observed one amongst the n source pixels. This resampling system is rarely used because it is more computationally intensive and rarely implemented in the standard image processing software.

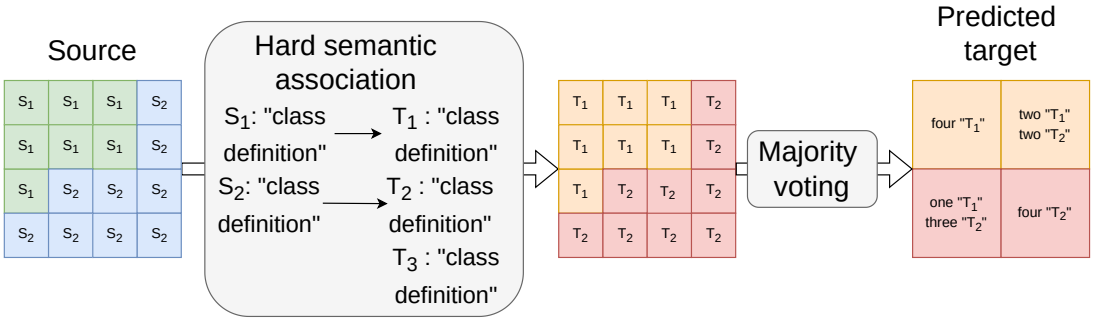

Figure 4.1: The **HardSem** translation: Source S_i and S_j are translated into their closest semantic correspondent, target class T_k and T_l, respectively. A down-resolution translation of factor 2 is then performed using majority voting.

HardStat relies on the exact same strategy: nomenclature translation followed by a resolution translation (see Figure 4.2). The only difference is that the nomenclature translation is based on statistical matching. An association matrix, giving the spatial

co-occurrence proportion of each source with each target class, is computed. Each source class is translated into its most frequently spatially co-occurring target class. For instance, if a source class S_i is observed at the same location than a target class T_k in 10% of cases, class T_l in 60% and class T_m in 30%, S_i is translated into class T_l. As some of the soft baselines presented use a machine-learning algorithm such as the Random forest, we underline that this method gives strictly equivalent results to training a random forest to perform the translation from source to target using only the source class as a feature.

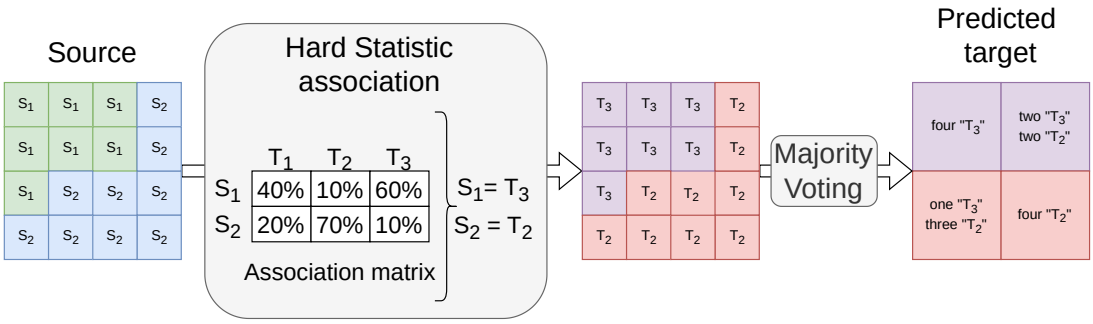

Figure 4.2: **HardStat**: Source S_i and S_j are translated into the most frequent corresponding target , target T_m and T_l, respectively. A down-sampling of factor 2 is then performed by majority voting.

SoftSem presented in Figure 4.3 can be used when the target resolution is coarser than the source one. It relies on a separate nomenclature and resolution translation. The nomenclature translation is achieved by using the soft association between the source and the target maps described in Appendix B. Each source class might be semantically linked with multiple target ones. Since we do not assume prior knowledge of the target segmentation, the resolution translation is performed with a grid-based approach. The possible target translations of the two source pixels are summed, and the target class with the most votes is assigned. Thus two pixels with the same source class can be translated differently depending on their neighbourhood.

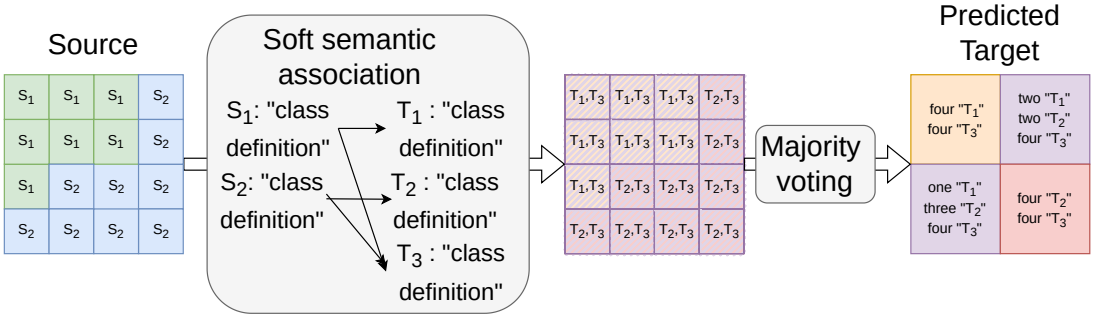

Figure 4.3: **SoftSem**: Source S_i and S_j are associated to all their possible semantic correspondent (T_k/T_l and T_l/T_m respectively). A down-sampling of a factor 2 is then performed by majority voting.

SoftMaxProba (Figure 4.4), based on [60], can be used when target's resolution is coarser than source's. It relies on a separate nomenclature and resolution translation. The nomenclature translation is achieved using the same association matrix as the **HardStat**. The average of all the observed proportions of co-occurrence for each of the n source pixels is performed, and the most probable one is then assigned to the target. This acts differently than the **HardStat** method, which only considers one possible translation per source pixel (the most frequently observed one). However, the statistical links between the source and target nomenclature do not consider multiple source class compatibility.

Figure 4.4: **SoftMaxProba**: the translated class is obtained by averaging the probability of each source class to correspond to each target class and choosing the most probable one.

SoftLearntFreq directly inspired from [56], can be used when the target resolution is coarser than the source one (Figure 4.5). It relies on a simultaneous nomenclature and resolution translation and tries to alleviate the **SoftMaxProba** main limitation by considering multiple source class compatibilities. The core idea is that compatibility can be inferred from data through machine learning algorithms. For instance, one can learn to translate differently depending on the proportions of each source class amongst n-pixels to translate into one. Additionally, it enables us to learn thresholds rules such as "*it is forest only if trees cover more than 70% of the surface*". Unlike [56], we do not assume prior knowledge of the target segmentation, relying instead on a grid-based segmentation at the target resolution. This approach appears more realistic in most land-cover translation problems, as the target segmentation is usually unknown. Moreover, [56] analysed the proportions of classes using a discriminant analysis while we replaced it with a random forest, showing better results in our case. **Implementation details**: A cross-validation experiment is conducted to estimate the best random forest parameters according to the $mF1_{ag}$ using a grid search approach on the number of trees (50 to 500, with a 50 step), maximum tree depth (from 5 to 40, with a 5 step) and the minimum samples per leaves (from to to 1 to 101, with a step 10). Even though slight variations are observed between different map 100 trees, with a 25 maximum depth and 10 minimum samples per leaf

appears sufficient to achieve the best mF1 score for all maps at ±1%.

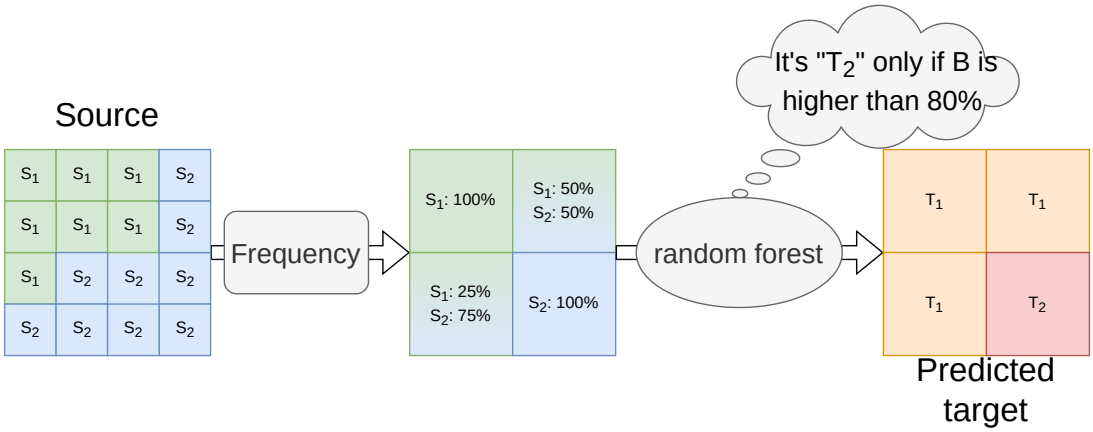

Figure 4.5: **SoftLearntFreq**: The frequency of each source class is used to train a random forest. Translates differently depending on the proportion of each source class.

SoftLearntGridPattern (Figure 4.6) is very close to **SoftLearntFreq**. However, instead of learning how to translate source class proportions, the method directly processes the n pixels preserving their local spatial arrangement. We replace the random forest with a convolutional neural network, the ResNet50, as a random forest would be unable to learn all the possible patterns due to the very high number of configurations (16 in the Figure 4.6 with only two source classes and a small downsampling factor of 4). **Implementation details**: We used the classical RestNet50 implementation [114] in its pytorch implementation [1].

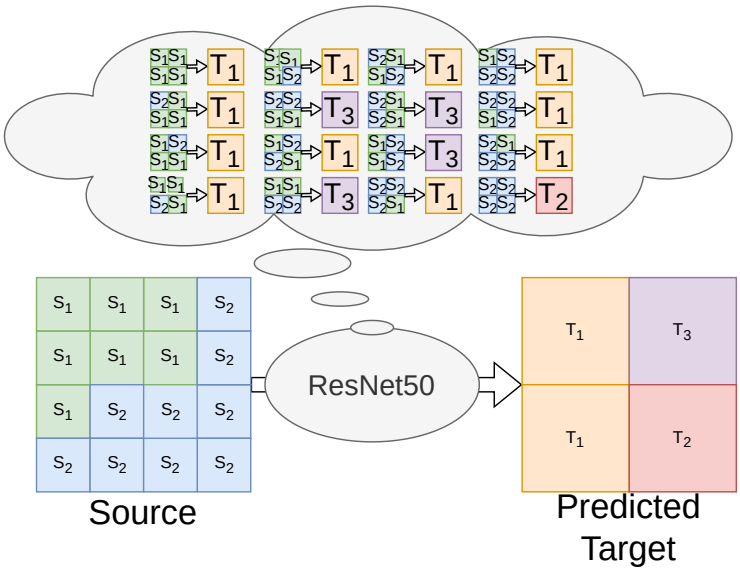

Figure 4.6: **SoftLearntGridPattern**: The translation of each source class is performed by a ResNet50 trained to translate differently based on the local class spatial arrangement.

[1] `https://pytorch.org/hub/pytorch_vision_resnet/`

4.1.3 Results

4.1.3.1 Qualitative analysis

Figure 4.7 presents an illustrative sample of patches of translation obtained on our France-wide test set. Only translations from high to coarse resolved maps are displayed to be able to compare the result of the four Soft baselines.

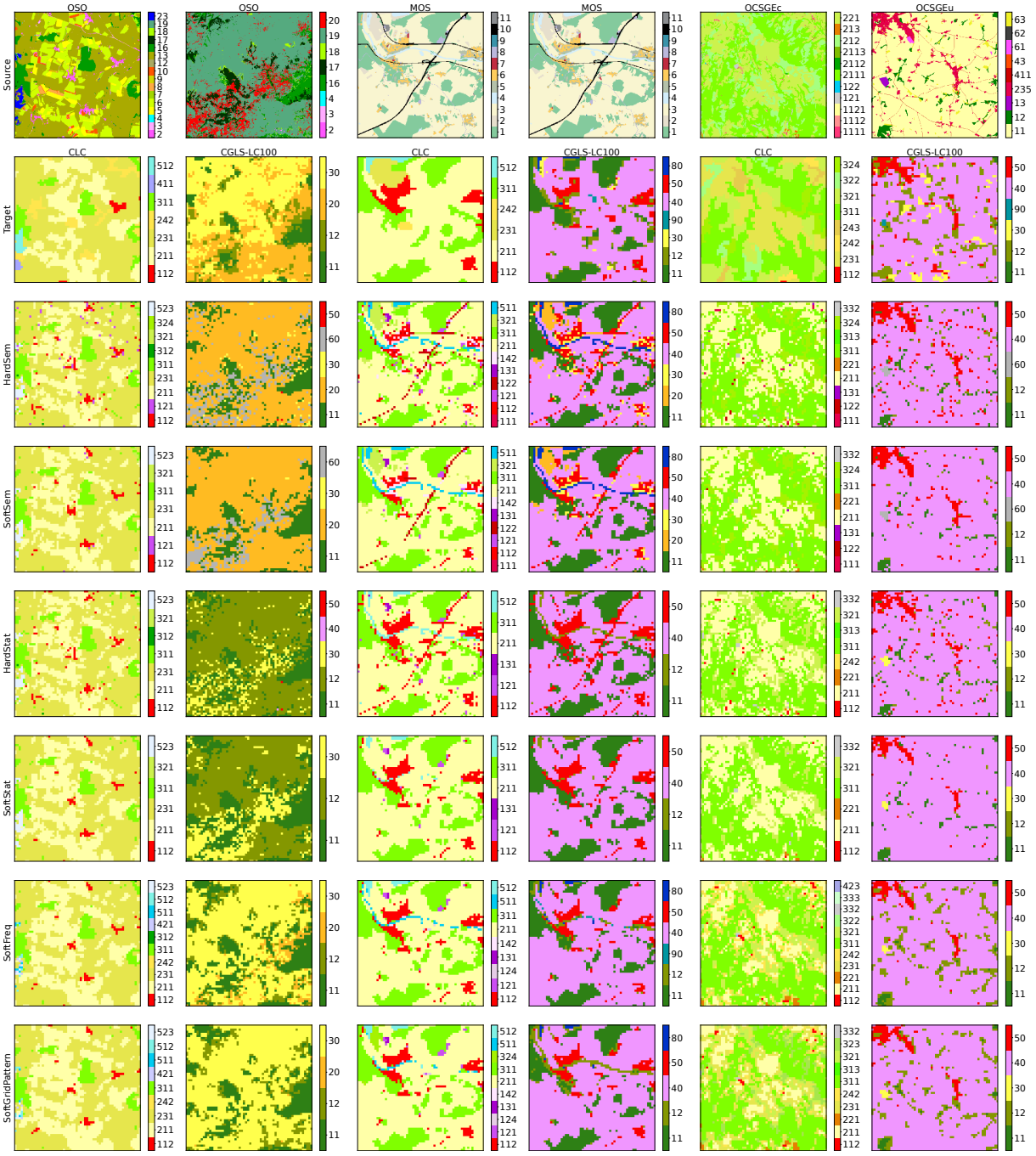

Figure 4.7: Sample of results for six different areas with each of the six baseline methods. The class/label correspondence tables of each LULC are available in Appendix A.

We categorize those results into three qualitatively homogeneous groups: the semantic group (**HardSem**, **SoftSem**), the simple statistic group (**HardStat**,**SoftMaxProba**), and the machine-leaning based downsampling group (**SoftLearntFreq**, **SoftLearntGridPattern**). The distinction between those three groups is especially visible in the second column for which OSO 19:*Woody moorlands* is translated as CGLS 20:*Shrubland* by semantic-based methods, as CGLS 12:*Open forest* by simple statistic methods or as a mix between CGLS 20:*Herbaceous vegetation* and *Shrubland* by more machine learning based methods. The difference between the semantic group and the other is mainly observed for classes for which the main semantic correspondent differs from the statistically most observed one. Such differences are often observed on erroneous source classes, as semantic translations assume that the description of a class and its real content is identical, *i.e.* no errors in the source map. For instance, in the second column (OSO to CGLS-LC100), the source OSO maps widely overestimates *Bare rock*areas (in red) which are wrongly translated into CGLS-LC100 *Bare / sparse vegetation* (in grey). Conversely, statistic methods determine the translation based on the real noisy content of the source and target class resulting in different translations, *e.g.* most OSO *Bare rock* corresponds to *Herbaceous vegetation* in CGLS. Consequently, statistic methods partially compensate for errors in the source maps but are prone to replicate errors of the target.

The simple and machine learnt statistic groups often give almost identical results. Most of the differences are located on the edges of objects, as illustrated in the two last columns. For instance, edges between OCS-GEc 1111:*Built-up areas* and 222:*Herbaceous formations* are translated as CLC 211:*Non-irrigated arable land* by simple statistic methods while being translated as 242:*Complex cultivation patterns* by the machine learnt ones. This distinction makes sense as the surrounding of small cities are very often classified as 242 in the CLC product and is only obtainable by learning to perform the dual translation of the two classes simultaneously. We highlight that those two groups are highly dependent on the noise in the target. For instance, CGLS-LC100 tends to misclassify *Water* as *Open Forest* on rivers, and both groups replicate the same error in the translation as shown in the fourth column (MOS to CGLS-LC100).

A general weakness shared between all those methods is that they fail to predict many classes resulting in a low diversity of translated classes. Moreover, none of them can apply the CLC MMU of 25ha (25 pixels), resulting in this noisy single isolated pixels pattern.

4.1.3.2 Quantitative analysis

As for the qualitative review, quantitative metrics for soft association methods can only be computed for translation in which the target map is coarser resolved than the source one. Tables presenting the quantitative translation results hold empty cells for translation in which the target is identically or higher resolved. In this case we arbitrarily consider that **SoftSem** is identical to the **HardSem** . Similarly, other soft methods are considered identical to the **HardStat**.

Source		CGLS (P)					CLC (C)					OSO (O)					OCS-GEc (G1)				OCS-GEu (G2)				MOS (M)			Total	Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	average	on soft
OA_{ag}	HardSem	52	42	56	**70**	**75**	65	49	67	77	**79**	57	54	69	76	80	53	39	34	87	54	37	31	75	77	72	59	61	58
	SoftSem											62	59			**82**	56	41			57	40			80	76		62	61
	HardStat	54	44	65	70	75	68	55	71	78	79	61	54	73	80	81	55	41	49	89	54	37	41	78	80	77	62	64	60
	SoftMaxProba											65	61			**82**	57	44			57	40			83	81		65	63
	SoftLearntFreq											**72**	**62**			**82**	**63**	**47**			**62**	**41**			**84**	81		**66**	66
	SoftLearntGridPattern											**72**	**62**			**82**	**63**	**47**			**62**	**41**			**84**	**82**		**66**	**67**
$mF1_{ag}$	HardSem	**13**	17	**22**	15	**24**	46	**32**	**36**	**31**	**42**	36	17	**36**	**20**	38	24	9	17	**27**	19	8	8	**29**	35	17	**19**	25	23
	SoftSem											38	19			38	27	10			20	8			38	19		25	24
	HardStat	13	18	19	16	24	47	32	33	30	42	33	16	34	20	38	26	10	20	27	19	8	10	27	31	15	18	24	22
	SoftMaxProba											36	18			39	27	10			20	9			32	17		25	23
	SoftLearntFreq											**48**	**26**			41	**38**	**18**			**28**	**12**			42	24		**31**	**27**
	SoftLearntGridPattern											47	24			**42**	37	15			**28**	**12**			**43**	**25**		30	**27**
EPI	HardSem	**28**	**5**	**6**	**5**	**17**	**30**	**6**	**6**	**7**	**20**	32	30	**28**	**32**	57	29	29	29	79	31	31	29	**77**	53	49	**38**	30	38
	SoftSem											36	36			58	28	32			32	36			54	54		**31**	41
	HardStat	27	**5**	3	**5**	16	**30**	**6**	5	**7**	**20**	30	27	25	31	56	30	29	**30**	**80**	31	31	**30**	**77**	47	46	**38**	29	37
	SoftMaxProba											36	36			58	28	32			32	36			54	54		**31**	41
	SoftLearntFreq											**42**	**36**			56	**35**	**31**			34	**33**			56	54		**31**	**42**
	SoftLearntGridPattern											**42**	35			**57**	**35**	**31**			**35**	32			**57**	**55**		**31**	**42**

Table 4.2: Agreement metrics between the prediction and the target map for all pairs of overlapping source/target maps. Cells are coloured in grey for methods based on statistical correspondences and in white for those based on semantics. Cells are kept empty for soft baselines when the target resolution is not coarser than the source. The average performance is computed (i) on the 26 translations, filling the empty cells with their corresponding **HardSem** or **HardStat** results (Total average) or (ii) on the nine translations when the target resolution is coarser (Average on soft).

Source		P					C					O					Total
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	average
OA_{gt}	HardSem	47	**46**	62	79	**81**	**72**	51	76	84	**85**	68	64	**86**	**86**	**92**	72
	SoftSem											71	65			**92**	72
	HardStat	**52**	45	**68**	**80**	**81**	68	**57**	**77**	**85**	**85**	64	63	**86**	**89**	91	73
	SoftMaxProba											67	66			**92**	73
	SoftLearntFreq											**73**	**68**			**92**	**74**
	SoftLearntGridPattern											**73**	**68**			**92**	**74**
$mF1_{gt}$	HardSem	12	**18**	**29**	22	**26**	**62**	**42**	**56**	**55**	**60**	45	20	**51**	30	41	38
	SoftSem											**47**	22			43	38
	HardStat	**13**	**18**	22	25	**26**	59	37	50	48	**57**	35	17	48	**31**	41	35
	SoftMaxProba											37	20			42	36
	SoftLearntFreq											41	26			44	36
	SoftLearntGridPattern											41	**28**			**45**	**37**

Table 4.3: Quality of the translation computed on ground truth. Cells are coloured in grey for methods based on statistical correspondences and white for those based on semantics. Cells are kept empty for soft baselines when the target resolution is not coarser than the source. The OA_{gt} is evaluated on the 2300 random points ground truth, while the $mF1_{gt}$ is computed on the 2700 points enriched with rare classes.

Table 4.2 resumes the agreement metrics of the various methods, Figure 4.8 resume the per-class agreement metrics, while Table 4.3 resumes the accuracy on the ground truth. The agreement metrics assesses the agreement between the target land-cover map and the translated map giving a detailed insight into predicted class diversity. However, it is maximised only when the translation method learns to replicate errors. Conversely, the ground truth enables us to determine the robustness to label noise but is not well suited for per-class metrics as the ground truth is too small. Lastly, agreement is computed on the target map extent, *i.e.* can not be computed between two maps with no overlap (MOS, OCSGE-cover and use). Conversely, accuracy on ground truth can only be computed on a

wide extent to avoid a too small sample size. Thus the translation results on the ground truth are only provided when the source map is CLC, OSO or CGLS-LC100.

Figure 4.8: Per-class F1 agreement computed on the sum of the translation confusion matrices of all the sources to one target. See Appendix A for class label.

The first observation is that textbfHardSem, which is by far the most used in literature, gives worse results than **HardStat**. This underlines that in the case of a source class with multiple possible translations, the closest semantically is often not the most probable statistically. This observation is of utmost importance as current nomenclature translation methods mainly focus on defining semantic similarity between class definitions rather than

defining a probabilistic matching between all the source and target classes. It pledges for more research on automatic statistical matching methods such as those presented in this manuscript.

A second observation is that when target is coarser than source, the four baselines performing soft associations (**SoftSem**, **SoftMaxProba**,**SoftLearntFreq**, **SoftLearntGridPattern**) give better results than **HardSem** and **HardStat**. This behaviour advocates for the need to perform the resolution and nomenclature translation jointly rather than separately, as most methods do.

The **SoftSem** and **SoftMaxProba** baseline comparison reveals that the latter exhibits a significantly higher OA_{ag} and OA_{gt} than the first while achieving almost the same $mF1_{ag}$ and $mF1_{gt}$. It appears that statistic-based methods are biased towards achieving the highest possible overall accuracy while neglecting the diversity of translated classes (see Figure 4.8). The methods proposed in this manuscript should thus focus on ensuring target class diversity.

The last observation is that the **SoftLearntFreq** and **SoftLearntGridPattern** methods outperform the others significantly (+6% OA_{ag} compared to HardStat). We link this behaviour to the fact that they learn the direct translation of a combination of source classes instead of combining individual class translations. We note that the knowledge of the local spatial arrangement also increases slightly all metrics.

4.2 Spatial Context

This section investigates the spatial context potential for translation. First, we propose a simple yet efficient way to evaluate spatial context based on manually defined shape features (area, elongation, compactness...) , *e.g.* is the water pixel in a linear shaped object or not. Once determined we train a random forest to use this shape information to translated differently pixels with the same class but exhibiting to different shape *e.g.* the random forest learns that a water pixel in a linear shape object should be translated into river while other should be translated into lakes. Secondly, we propose a more complex approach replacing those manually defined features with automatically learnt ones, hoping that it could increase the translation quality. The proposed approach relies on a convolutional neural network method that jointly performs nomenclature and resolution translation.

4.2.1 Manually defined shape indicators

4.2.1.1 Motivation

This section is built on the observation that two pixels with the same source class but a different target class often belong to segments with very different shapes. Figure 4.9

illustrates this idea by displaying segments with the same MOS classes but a different CLC translation. For instance, two pixels belonging to two different MOS *Forest* segments should be translated into (i) CLC *Non-irrigated arable land* when the Forest exhibits a thin linear-shaped structure, *i.e.* is a hedgerow or in (ii) CLC *Broad-leaved forest* when the Forest exhibits a more circular shape structure with a significantly larger area. We argue that the knowledge of the shape of source segments is crucial for translation.

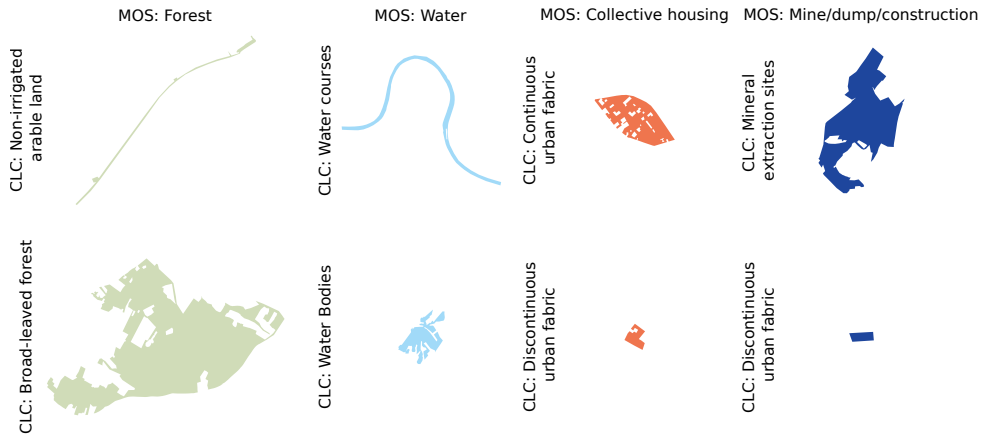

Figure 4.9: Illustration of the relevance of the shape information for translation. For each of the MOS classes, *Forest*, *Water*, *Collective housing* and *Mine/dump/construction*, two different objects with different CLC translations are provided. Shape analysis of those MOS objects appears relevant to determine which CLC class to assign.

Section 2.3.1 identified two approaches for extracting shape information in the literature: manually defined shape features (MDSF) or machine-learnt shape features (MLSF). As no works have focused on incorporating shape in the land-cover translation framework, this section studies MDSF as a simple tool to improve translation. An MLSF method dealing with the shape and class spatial correlation is explored separately in Section 4.2.2.

Manually defined shape features have partly been explored in Geographic-object based image analysis (GEOBIA). In land-cover mapping, GEOBIA segmentations are often obtained from automatic image analysis methods aiming to group pixels in homogenous segments. The automatic algorithms are parametrized to achieve a tradeoff between (a) ensuring that contiguous pixels belonging to different classes belong to different segments and (b) that pixels belonging to the same class belong to the same segment. For instance, all contiguous pixels of a *Forest* must belong to a single segment, and this segment must only include *Forest*. As most approaches favorites (a), an over-segmentation is observed in most applications, *e.g.* a single *Forest* is composed of multiple segments. Consequently, few experiments on the use of MDSF for land-cover mapping have been conducted as the used segmentation often poorly reflects the true shape. Conversely, in the land-cover translation scenario, obtaining the source segmentation is straightforward and directly reflects the geometric properties of the object to translate. As the few works conducted on using those MDSF for land-cover mapping focus principally on urban area change

detection, we lack a proper analysis of the best MDSF to use. The experimental setup section, first proposes a set of common shape indicators obtained from the literature that could provide helpful information for translation.

Additionally, shape information is rarely included in land-cover map nomenclature definitions to ensure that nothing is misclassified due to a peculiar shape. Establishing rule-based semantic translation based on shape information is unfeasible. A machine learning (ML) strategy is explored to learn to translate using both the source label and the MDSF. The choice of the ML method is discussed to both respects the characteristics of the MDSF and offer the possibility to obtain a general intuition on which shape elements are important for land-cover translation.

4.2.1.2 Experimental protocol

Manually defined shape features Ten shape features, commonly used in the GIS community [18] and easily analysable, are experimented with (refer to Table 2.2). We succinctly introduce them below and underline examples of classes for which they are sufficient alone to improve translation significantly. Combining those MDSF can help identify more classes not mentioned below. **Area** is used to evaluate the size of the segments and is particularly informative when the source resolution is finer than the target *e.g.* a 100m^2 grassland is probably a garden while a 10^6m^2 is probably a pasture or a natural grassland. **Elongation** is used to evaluate the "thickness" of a segment using the ratio between area and perimeter (close to 0 for elongated segments). It helps translating segments involving thickness constraints *e.g.* translating water in a river or lake, or with cartographic generalisation rules *e.g.* a thin forest is probably a hedgerow between crops classified as *arable crops* in CLC. **Circularity** compares the segment area to the area of a circle with the same perimeter (1 for a circle). Especially useful to translate circular irrigated crops. **NestedPoly** counts the number of the segment within the considered segment and is the only selected MDSF rarely found in the literature. It enables translating classes characterized by density constraints *e.g.* an urban area segment holding many other segments of different land-cover classes is less likely to be a dense urban area. **Convexity** is the ratio between the area and the convex hull area of a segment. It indirectly evaluates the shape complexity. It is mainly used in urban mapping to distinguish different building's usage. **MBRH, MBRW, MBRArea, MBRFlatness** Height, width, area and Elongation of the minimum bounding rectangle of the segment. Complementary to **Area**, it helps to evaluate the Height, Width disproportion and, indirectly, the global spread of the land cover. **MBRArea=Area** implies that the shape is rectangle *e.g.* pastures often exhibit rectangle patterns while natural grassland do not. **MBRAngle** gives a rough estimate of the principal orientation of the segment. Especially useful to translated land-cover classes correlated to sun position and topography *e.g.* hillsides vineyards, or wind constraints *e.g.* hedgerows planted to limit wind influence.

Figure 4.10 presents the correlation matrix between the different shape indicators of

CGLS-LC100, CLC and MOS. The diagonal delivers the auto-correlation of the indicators for one map. For instance, the left top cell presents the auto-correlation of CGLS-LC100 shape indicators. A first observation is that the shape indicators of one map tend to be highly correlated, especially the Area, **MBRH**, **MBRW**, **MBRArea** and **NestedPoly**. This is easily explainable as wide area polygons necessarily have a wide **MBRArea** (thus a wide **MBRH** and **MBRW**) and is more likely to have multiple inserted sub-polygons. Additionally, the correlation between two features of the same map is highly dependent on the resolution and nomenclature of the map. For instance, The MBRAngle is mostly positively correlated with the other features for CGLS, negatively for CLC, and almost unrelated to MOS.

Interestingly, inter-correlation between two map features exhibits the same sort of correlation as those observed inside one map. In particular, we observe that those inter-correlation coefficient are high (most of the time superior to 0.5), denoting that source map segmentation shape indicators provide information on target segmentation shape.

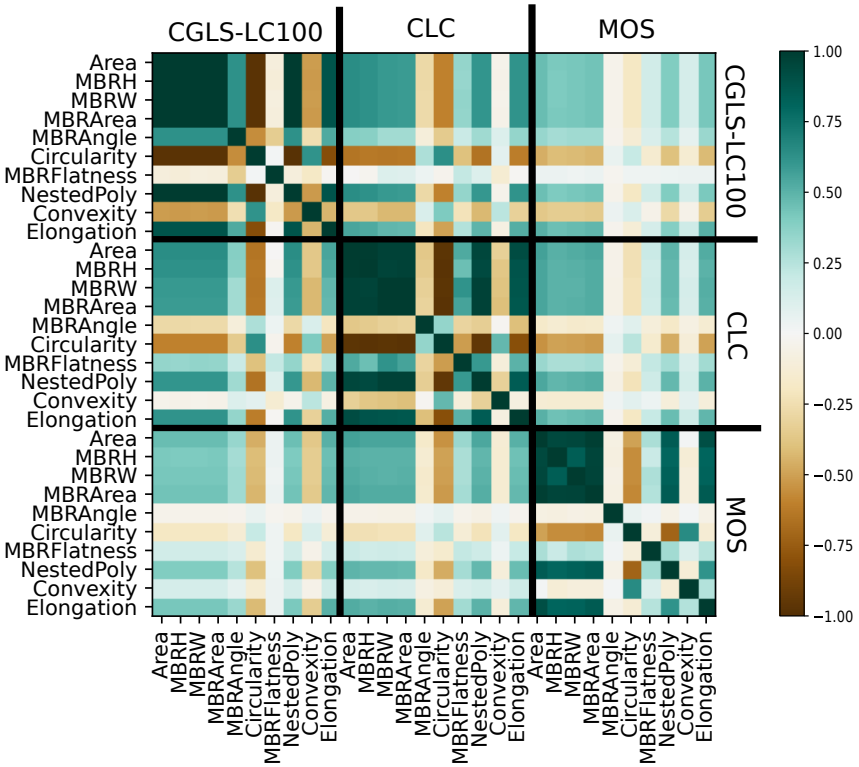

Figure 4.10: Spearman's correlation coefficient between the shape indicators of CGLS, CLC and MOS. The diagonal of the matrix denotes the autocorrelation between the indicators of the map. The rest of the matrix gives the correlation between the shape indicators of different maps.

Unfortunately, as MDSF computation is particularly time-consuming for high-resolved maps as it requires both i) to have access to a vectorised version of the map and ii) to compute the indices for each of them, experiments are only conducted for the three coarser of the six maps: CGLS-LC100, CLC, and MOS.

Machine Learning The selected machine learning algorithm must be robust to highly correlated features. Moreover, the model's decision should be interpretable to understand important shape characteristics. The random forest algorithm exhibiting these two characteristics [97] appears a natural choice. Source labels are one-hot encoded and combined with the MDSF features to train the random forest algorithm. This method is termed **SpaShape**. The **HardStat** method is used as a control to evaluate the role of shape indicators since its results are identical to a random forest trained using the source labels as features.

Implementation details: As pixels belonging to different patches might belong to the same segments, the independence between training and testing can not be ensured using the train/test based on patch separation presented in Section 3.1. We process the raw vector version of the map. Adjacent segments with the same class are merged to limit arbitrary segmentation artefacts, and MDSF are computed. Segments are randomly split in train/test using the same ratio as in the MLULC dataset (65,35) and rasterised to the source resolution. The random forest is trained pixel-wisely on truly independent train segments while preserving the fact that two pixels of a given segment have the same feature but can have a different corresponding target class. The confusion matrix and afferent metrics computed on the test are corrected by sample inclusion probability to be comparable with the results of the baselines, i.e. the random split in train/test/validation at the segment level might introduce a slight shift in the respective proportion of each class.

The same grid search cross-validation strategy as for **SoftLearntFreq** is used to get the best parameters for the random forest with almost identical results: 300 trees, five minimum samples per leaf and a max depth of 25.

4.2.1.3 Results

Quantitative aspect Qualitative aspect does not differ much visually as target classes benefiting from MDSF-aware translation are mostly rare classes. Therefore, we only present quantitative results in this section. Table 4.4 compares the results with the two hard-association baselines. A first observation is that the effect of adding shape information widely differs depending on the considered translation. When translating from a coarse map (CLC or CGLS-LC100) to a highly resolved one (MOS), adding spatial context does not improve the translation significantly ($+0.5\% OA_{ag}$ and $mF1_{ag}$). This is widely understandable because the translation is ill-defined. Additional high-resolved data is mandatory to obtain good results. Moreover, as very large objects usually encompass various classes due to the minimum mapping unit, their shapes give limited information on their actual content. On the contrary, when the source resolution is finer or comparable to the targeted one, the overall agreement and agreement f1-score are improved by a vast margin ($+4\% OA_{ag}$,$+6\% mF1_{ag}$) compared to the best baseline, HardStat). We underline that those results are even better than those observed with the soft-translation methods.

Source		CGLS (P)		CLC (C)		MOS (M)		Total
Target		C	M	P	M	P	C	average
OA_{ag}	HardSem	52	**75**	65	79	77	72	70
	HardStat	54	**75**	68	79	80	77	72
	SpaShape	**57**	**75**	**70**	**80**	**84**	**83**	**75**
$mF1_{ag}$	HardSem	13	24	46	**42**	35	17	30
	HardStat	13	24	47	**42**	31	15	29
	SpaShape	**19**	**25**	**52**	**42**	**39**	**24**	**34**

Table 4.4: Agreement metrics obtained using shape indicators compared to baselines

Feature Importance We study the importance of each MDSF indicator to provide insight into shape elements impacting land cover translation. Feature importance is assessed using a random feature permutation approach which is more robust to the highly correlated features than the more traditional mean decrease in impurity [30]. A random forest is first trained on the training set. Then, an iterative comparison is conducted on the test set between the obtained $mF1_{ag}$ and the $mF1_{ag}$ obtained with one randomly shuffled feature. This strategy for assessing feature importance tends to underestimate the importance of correlated features [30]. Consequently, the importance of Area, MBRh and MBRW features is probably underestimated. Figure 4.11 presents the results of the feature importance evaluation when MOS is used as the source map, but the same observation can be made using other source maps. The main observation is that the most important features tend to be either those assessing the spatial extent of the objects (Area, MBRarea, NestedPoly) or the shape compactness (Elongation). It appears important to ensure that any proposed spatial context-aware methods can assess the size of the object they are translating.

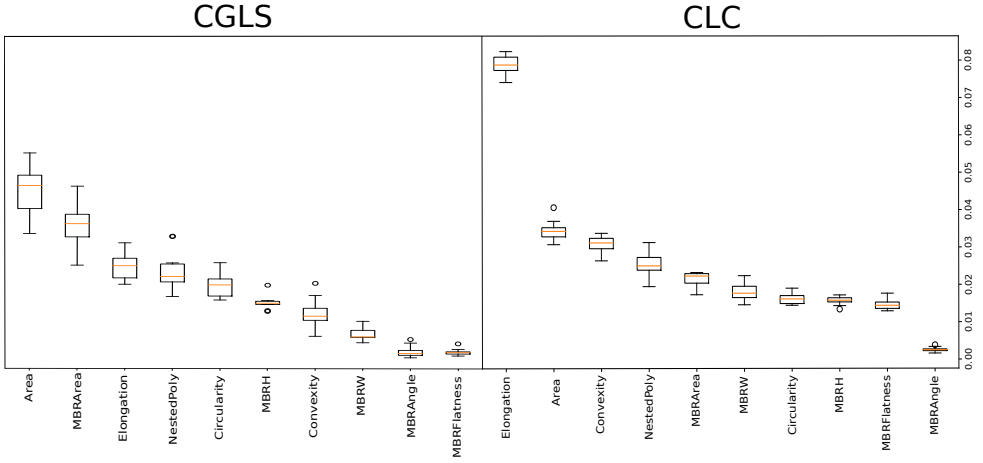

Figure 4.11: Feature importance for the $mF1_{ag}$ metric for the translation from MOS to either CLC or CGLS evaluated using a random feature permutation technique. The source label feature is removed for readibility due to its high importance score (0.55). Some features importance might be underevaluated due to high feature correlation (refer to Figure 4.10)

Discussion Our experiments demonstrate that the use of simple MDSF significantly improved the translation results when the source resolution is finer than the target one ($+4\%OA_{ag}$,$+6\%mF1_{ag}$). Conversely, when the source resolution coarse the observed improvement is low ($\pm$ 0.5%). We link this behaviour to the fact that MDSF usage relevance directly depends on the assumption that segment shape accurately reflects its semantic content. However, land-cover maps with coarse resolution often mix multiple land-cover inside a single segment, especially when the minimum mapping unit is high (CLC). This effect is further amplified by the fact that low-resolution land-cover maps often exhibit complex shaped segments, including holes (NestedPoly>1) for which MDSF provides unreliable information *e.g.* **Circularity** does not reflect accurately segment compactness if the shape has a big hole in it.

The experiment design, *i.e.* a limited number of manually defined shape indicators, probably leads to an underestimation of the shape information potential for translation. MLSF could partially alleviate this problem by automatically learning translation-tailored shape features instead of arbitrarily chosen ones.

4.2.2 Machine Learnt contextual features

4.2.2.1 Motivation

The previous section underlined that the shape of source segments might significantly improve the land-cover map translation. However, the method relies on manually defined shape features, which cannot precisely estimate all possible shapes. This section investigates how to learn translation-tailored contextual features automatically. Unlike the previous section, the proposed method also incorporate the previously neglected neighbouring spatial context *e.g.* sand is usually near the sea and wetlands near rivers. Inspired by the literature review conducted in Section 2.3.1, a Convolutional Neural Networks (CNN) approach is experimented as they are perfectly tailored to integrate the semantics of the pixel with its spatial context.

4.2.2.2 Method

Resolution gap Unlike land-cover map classification, which usually preserves the input-data spatial resolution, translation is frequently confronted with a difference between source and target resolution. A first simple strategy to deal with the resolution gap is to either (i) resample the source to the target resolution before network processing or (ii) resample the output to the target resolution. (i) is mainly used when the source resolution is coarser than the target, while (ii) is used when the source is finer resolved. Resolution and nomenclature being highly intertwined, it appears detrimental to process them separately, especially when translating from a fine resolved to a coarse target map (see Figure 1.5). This section primarily focuses on the case of translating a fine-resolved

map to a coarse target by adapting the network architecture to translate nomenclature and resolution simultaneously. The reverse translation from a coarse to a highly resolved map involves using additional data, complexifying the distinction between the spatial context importance and the additional data. We address this issue separately in Section 5.2.2.

A downsampling network Image to land-cover map classification is traditionally cast as a semantic segmentation task, in which a set of remote sensing images are transformed into a class map [136, 206, 236]. Similarly, we observe that land-cover map translation can also be seen as a semantic segmentation task where the input pixels are not physical values (namely, the optical spectral bands or SAR polarized channels), but semantic classes, *i.e.*, nominal categorical data with low cardinality. Therefore we base our network on a popular semantic segmentation network, the U-Net [257] introduced in Section 2.3.1 and presented in Figure 2.3. The main idea of U-Net is to encode the image input into a vector representation using successive down-sampling and convolution steps and then restore (decode) the image using successive up-sampling and deconvolution layers. Skip connections between the encoder and the decoder convolutions are used to avoid losing spatial information during the MaxPooling process. The standard U-Net skip connections impose aggregating features of the same scale in the encoder and decoder sub-networks resulting in output with the same resolution as the input. We propose a simple U-Net adaptation, termed Asymmetrical U-Net **A-Unet**, to achieve the desired simultaneous nomenclature and resolution translation. Figure 4.12 presents it implementation for a resolution gap of a factor 10. It consists in (i) removing some of the skip connections and (ii) choosing different MaxPooling ratios for the down-sampling and the up-sampling parts. Let r be the resizing factor between the input and the output of the network, and $D = (d_1, d_2, ..., d_h), d_i \in \mathbb{N}$ the downsizing factor of the different pooling layers in the encoder. We need to ensure that:

$$r = \prod_{i}^{h} d_i. \tag{4.1}$$

Pooling parameters must be as small as possible to reduce the loss of spatial detail.

$$D = argmin_{d_i} \prod_{i}^{h} d_i. \tag{4.2}$$

This problem has a unique solution obtainable by prime decomposition [355]. In the OSO to CLC translation case, the target map is ten times smaller than the source map. This leads to apply a five and a two pooling layers in the asymmetrical part of the encoder. To avoid information loss, we apply the two pooling layers first. In the case where the source and target resolutions are identical, this network is equivalent to a U-Net. In the case where the source resolution is coarser than the target, the A-Unet first resample the source to the target resolution using the nearest neighbour operator and provide the data to a classical U-Net. We remind the reader that this problem is ill-defined without additional data and that this solution is only proposed to provide reference results for Section 5.2.2.

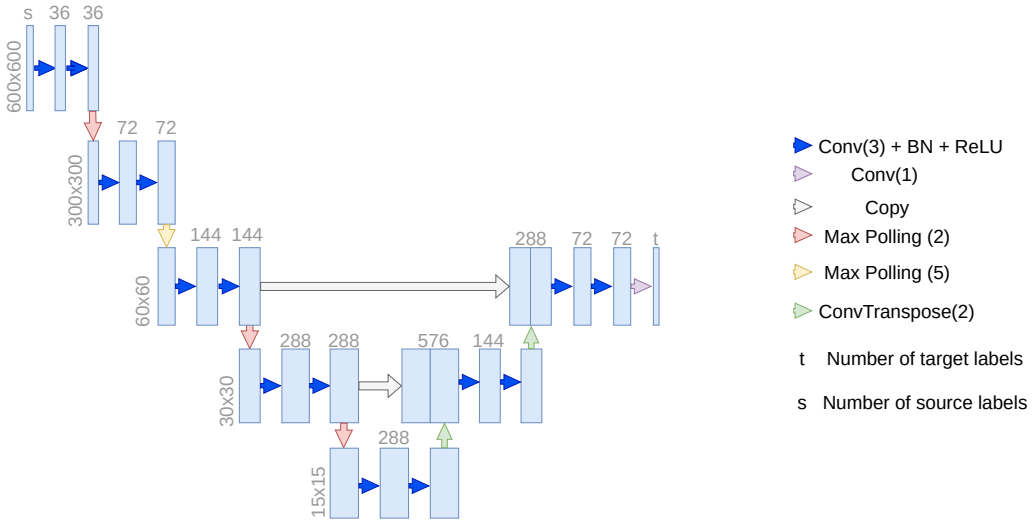

Figure 4.12: Our U-Net adaptation. An asymmetrical encoder-decoder is designed to handle the downsampling of fine-resolved sources to coarse-resolve targets efficiently.

Implementation details: As we rely on the basic U-Net architecture, the only hyper-parameters to tune are the: optimizer, learning rate, batch size and the number of feature maps of the first convolution block (the number of feature maps is simply increased by a two factor at each block in the U-Net architecture). We relied on Adam optimizer [163] as it offers a quick convergence with high robustness to learning rate value. Multiple loss are tested by reducing the loss of factor 0.5 starting from 0.02 using an "on plateau strategy". Experiments on the number of feature maps and batch size are conducted on the OSO to CLC translation as it involves the most resolved source map with the most target classes. 8, 16, 32, 64 feature maps were tested (the original implementation used 64). We highlight that 32 feature maps were sufficient to reach a performance plateau both in terms of OA_{ag} and $mF1_{ag}$. Various batch sizes ranged from 12 to 192 with a 2 factor step. The results from batch size 12 to 120 are almost identical, with less than 1% difference for all translations and started to decrease slightly above this number.

Loss The standard loss for semantic segmentation tasks is cross-entropy (CE), defined as the sum of the target entropy and the relative Kullback-Leibler divergence between the target and the prediction. Let n be the number of predicted elements, c the number of classes, p a softmax prediction of the network and y the ground truth. $p^i(k)$ denote the predicted probability that the i$^{\text{th}}$ element belongs to class k. $y^i(k) = 1$ if the true class of the i$^{\text{th}}$ elements is k and $y^i(k) = 0$ otherwise. The CE loss is computed as:

$$L_{CE}(p, y) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y^i(k) log(p^i(k)). \tag{4.3}$$

The principal limitation of CE is that it does not suitably handle imbalanced classes.

As stated previously, our case study holds highly imbalanced data, which prevents its adoption.

Many approaches have been proposed [257, 263] to cope with the class imbalance issue, such as the weighted cross-entropy [124] or focal loss [181]. We review below three losses for which we present the results in the following section.

The focal loss is built on the idea that rare classes are challenging to predict and thus result in low $p^i(k)$ values. It proposes to weight the cross-entropy inversely to the $p^i(k)$ value. A manually chosen α factor gives more or less importance to those low prediction values (when $\alpha = 0$, the focal loss is equivalent to CE).

$$L_{focal}(p, y) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y^i(k)(1 - p^i(k))^{\alpha} log(p^i(k)). \tag{4.4}$$

Additionally, region-based losses target to maximize the overlapping ratio between p and y. Among all these losses, we select the Dice loss [216]: it computes an approximation of the F1-score metric, which is vastly used in the remote sensing community. The Dice loss is computed as follows:

$$L_{DICE} = t - \sum_{k=1}^{c} \frac{2 \sum_{i=1}^{n} p^i(k) y^i(k)}{\sum_{i=1}^{n} p^i(k) + \sum_{i=1}^{n} y^i(k)}. \tag{4.5}$$

We combine the CE and the Dice loss, as suggested in [140, 161, 240, 328], to incorporate benefits from finer decision boundaries (Dice) and accurate data distribution (CE). This alleviates the problem of high variance of the Dice loss.

$$L_{CE+DICE} = L_{CE} + L_{DICE}. \tag{4.6}$$

Preprocessing Unlike image-to-land-cover classification, which takes continuous variables as input (pixel Digital Number), Land-cover map translation takes categorical variables (classes). Distances computed between two input values convey meaning on images, i.e. a 2 difference in radiometry is smaller than a 4, while it does not on raw land-cover, i.e. the 2 and 4 difference between a forest labeled "1" and water "3" or a forest and crops "5" can not be compared. The first pre-processing is to transform the input to ensure that distances between classes convey meaning. In this section, we assume no prior knowledge of interclass distances and thus consider that all pairs of classes should be equidistant. One-hot encoding is used to achieve this goal: each pixel is encoded as a vector with each dimension encoding for the presence (1) or the absence (0) of a given land-cover class. For instance in the previous examples $Forest = (1, 0, 0, 0, 0)$, $Water = (0, 0, 1, 0, 0)$ and $Crops = (0, 0, 0, 0, 1)$, thus the euclidean distance $D\ D(Forest, Water) = D(Forest, Crops) = D(Water, Crops) = 1$. Chapter 6 explores an another encoding paradigm for labels in which inter-class dis-

tances better reflect semantic differences. We underline that even though this step is not mandatory, it increases the network learning speed ($\approx$ 3 times).

Comparative assessment **UNetBili**, a method that processes the source map through a U-Net and resamples the predicted logits using bilinear resampling, is used as a comparative baseline to evaluate if the common nomenclature and resolution translation of **A-Unet** improve translation quality. **DeepLabV3** a DeepLabV3 architecture (see Section 2.3.1.2) with a ResNet50 backbone is used to evaluate if a deeper network achieving better results in most semantic segmentation tasks can improve translation results. Unlike **UNetBili**, the resampling problem is directly addressed by the Spatial Pyramidal Atrous module (see Figure 2.4).

4.2.2.3 Results

Qualitative analysis As the 26 possible France-wide translations can not be displayed, we selected a set of illustrative patches that enables discussing the strengths and weaknesses of the proposed solution. The translation results of **HardSem**, **HardStat** and **A-Unet** are presented in Figure 4.13.

The A-Unet spatial context-wise translation method obtains significantly different results than the baselines. As discussing all the qualitative improvements is impossible, we categorize those improvements (nomenclature or resolution based), illustrate them with some examples visible in Figure 4.13, and try to highlight the reasons for the observed improvement in order to enable a better understanding of the method strengths and weaknesses.

From a nomenclature point of view, the use of spatial context is beneficial for classes with:

- A characteristic shape, *e.g.* in the fourth row (OSO to CLC) the CLC class 124:*Airports* is partially predicted using the particular "crossing road" pattern observed in the OSO map.
- A characteristic pattern, *e.g.* mixed forest, dense urban (not illustrated in Figure but well visible in the France-wide result)
- An inside segment translation gradient, *e.g.* Forest is more likely to be denser in the middle than on the edges. Forests are translated as *Open Forest* on the edges and *Closed Forest* in the middle in row 2 and 3. This observation also concerns distinction such as CLC 111:*Continuous urban areas* and 112:*Discontinuous urban areas*.
- A spatial co-occurrence pattern, *e.g.* OCSGEc 221:*Herbaceous Formation* is translated in OSO 13:*Pastures* near forest and in 6:*Cereals* (the dominant crop type in France) far from it in row 6.

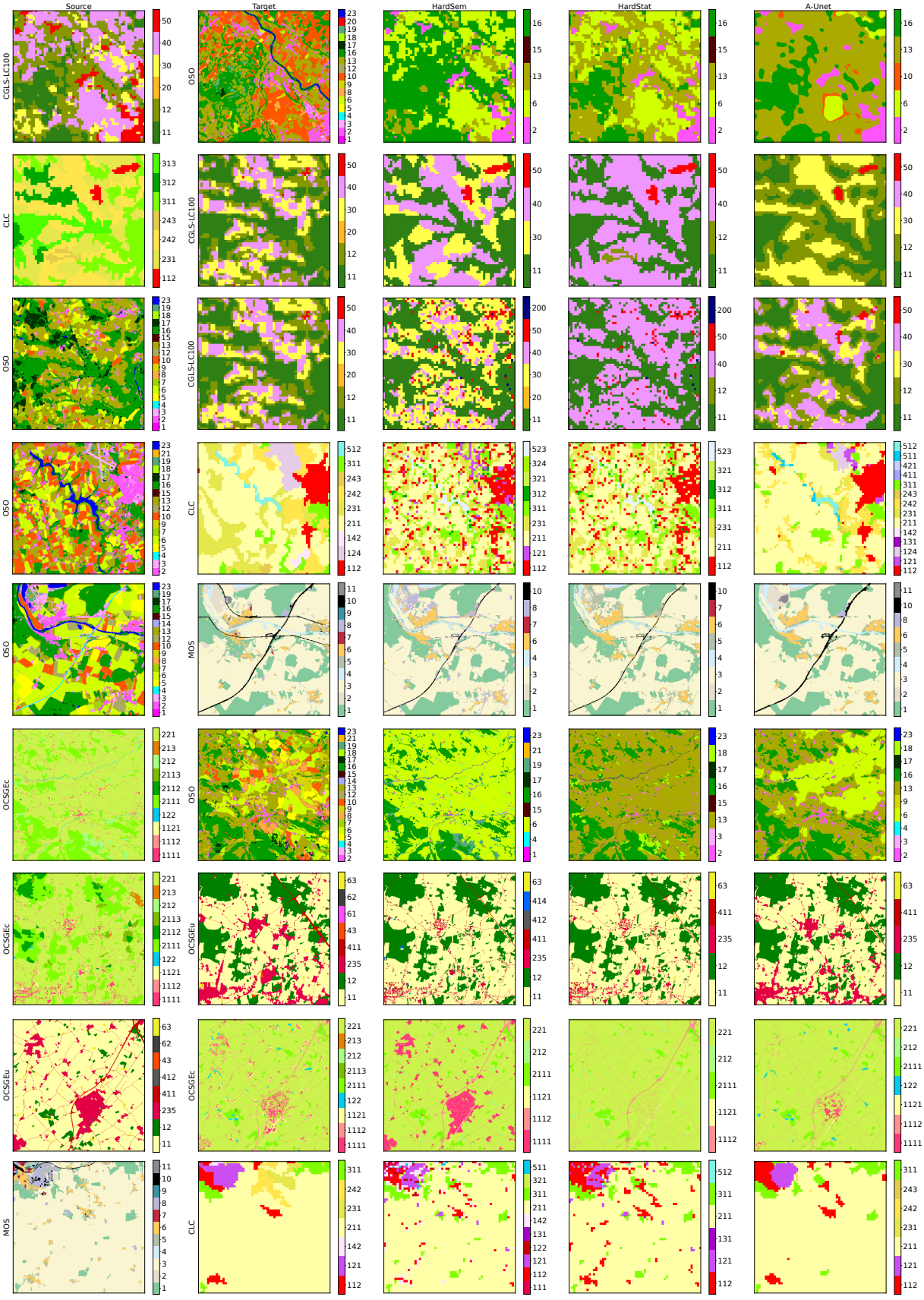

Figure 4.13: Translation results of the HardStat, HardSem and A-Unet method for various source/target couples

From a resolution point of view, spatial context contributes to:

- enable slight super resolution based on spatial patterns, *e.g.* in row 8, individual OCSGEc 1111:*Built-up areas* are partially retrieved from coarse OCSGEu 235:*Secondary or tertiary production and residential usage* based on 411:*Road networks* density.
- learn the CLC 25 adjacent pixels minimum mapping unit (MMU), *e.g.* A-Unet partially succeeds in learning MMU in rows 4 and 9.

The method exhibits the following qualitative limitations:

- The MMU is only partially learnt, *e.g.* single isolated pixels artefacts are still occasionally visible in rows 4 and 9.
- Linear shape structures are badly predicted, *e.g.* in the 5$^{\text{th}}$ row OSO 4:*Road surfaces* are only partially translated to MOS 10:*Transports*. In particular, even though baselines perform slightly better than the *A-Unet*. We link this behaviour to the fact that U-Net architectures are known to perform poorly on linear segment prediction.

.

Quantitative Improvement To better understand the potential of context-wise methods, Table 4.5 presents the method's quantitative results regarding the agreement between the translation and the target map. Those first results reveal an average +5% OA_{ag} and $mF1_{ag}$ improvement between the best baseline **SoftLearntGridPattern** and the A-Unet. Interestingly, this comes at the cost of reducing the average EPI by 3%. We link these paradoxical results to a qualitative observation that can be made in Figure 4.13, the A-Unet translation significantly alters the geometric shape compared to the baselines. For instance, the previously mentioned translation gradient of the forest to CGLS *Closed Forest* or *Open Forest* based on pixels distance to forest border results in a "buffer-like" translation agreeing highly to the target CGLS while losing precise edges delimitation.

Table 4.6 presents the quantitative results of the method between the translation and the manually built ground truth. A fine analysis requires the distinction of two groups of maps. **Group A** includes experiments where the source and target exhibit a France-wide spatial extent (CLC, OSO and CGLS). Conversely, **group B** includes those where the target covers a smaller extent (OCSGEc, OCSGEu, and MOS).

Group A OA_{ag} and OA_{gt} are computed on the same extent (France-wide) and should be identical ($\pm 1 to 2\%$) depending on the map. $OA_{ag} > OA_{gt}$ implies that the method is learning to replicate target map errors, which increases OA_{ag} and decreases OA_{gt}. Conversely, $OA_{gt} > OA_{ag}$ implies the correct translation of pixels that are erroneous in the target map. When the source or target is CGLS, the most erroneous map with 72% accuracy, the **A-Unet** OA_{ag} is on average superior to OA_{gt} from 4% underlying that the

network is prone to replicate target errors. In other cases, no significant differences are observed. Interestingly, **A-Unet** translations between Group A land-covers is in average 11% better in OA_{ag} than **SoftSem** ones while only 5% higher in OA_{gt}. As the **SoftSem** method is insensible to target noise, this underlines **A-Unet** lack of label-noise robustness.

Conversely, **Group B** OA_{ag} and OA_{gt} are not computed on the same extent (target extent vs France-wide) and thus not comparable. However, they enable studying the spatial generalization ability of the framework. As the OA_{gt} of baseline methods exhibit less than 1% difference with those obtained by the A-Unet, we conclude that the methods perform poorly on spatial generalization. Especially when translating one of the maps of group A into MOS, the obtained France-wide MOS is significantly better using a baselines instead of A-Unet. This can mainly be imputed to unseen spatial patterns. For instance, the original MOS spatial extent does not include sea areas corresponding to patches entirely covered by water; translating such patches can result in strange predictions such as *forest*).

The comparison between **A-Unet** and **UNetBili** reveals that performing jointly resolution and nomenclature translation performs slightly better (average +1% OA_{ag}, +2.5% $mF1_{ag}$).

The comparison with the agreement metrics obtained for the six translations using the manually defined shape feature (see Table 4.4) demonstrates that using a simple manually defined features procedure is sufficient to achieve results close from those obtained with the A-Unet when translating MOS to the coarse CLC or CGLS maps (1.5% OA_{ag} and 5.5% $mF1_{gt}$).

Interestingly, the A-Unet and its 3 million trainable parameters give almost identical results in most cases to the DeeplabV3 network (see Figure 2.4) on a ResNet50 backbone (40 million) parameters. It even performs significantly better when the source or target map is MOS, as the training set is especially tiny (around 250 patches) due to the quick overfitting of the Deeplab framework. Additionally, as achieving A-Unet performance does not benefit from increasing the number of feature maps, we choose to mainly focus on adding more information or changing the training characteristics (loss, goal) rather than increasing the architecture depth and complexity.

Source		P					C					O					G1				G2				M			Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	
OA_{ag}	SoftSem	52	42	56	70	**75**	65	49	67	77	79	62	59	69	76	82	56	41	34	87	57	40	31	75	80	76	59	62
	SoftLearntGridPattern	54	44	65	70	75	68	55	71	78	79	72	62	73	80	82	63	47	49	89	62	41	41	78	**84**	**82**	62	66
	UnetBili	**61**	53	67	76	75	**73**	57	71	79	78	75	66	**79**	**86**	84	66	52	**56**	**92**	67	48	**49**	**79**	**84**	**82**	61	70
	DeepLabV3	**62**	**55**	**68**	**78**	74	**73**	**58**	**72**	**80**	78	**77**	**68**	**79**	**86**	83	**68**	**54**	**56**	**92**	**68**	**49**	**49**	**79**	**84**	**82**	60	70
	A-Unet	**62**	**55**	**68**	**78**	**76**	**73**	**58**	**72**	**80**	**79**	**77**	**68**	**79**	**86**	**85**	**68**	**54**	**56**	**92**	**68**	**49**	**49**	**79**	**84**	**82**	**63**	**71**
$mF1_{ag}$	SoftSem	13	17	22	15	24	46	32	36	31	**42**	38	19	36	20	38	27	10	17	27	20	8	8	29	38	19	19	25
	SoftLearntGridPattern	13	18	19	16	24	47	32	33	30	**42**	47	24	34	20	42	37	15	20	27	28	12	10	27	43	25	18	27
	UnetBili	**26**	25	26	18	31	**56**	34	35	24	41	55	34	**45**	**27**	51	44	21	**29**	**44**	37	16	**22**	**40**	45	27	20	**35**
	DeepLabV3	**26**	26	**27**	**19**	29	53	**36**	35	**30**	40	**60**	**37**	**45**	**27**	50	48	**24**	**29**	**44**	38	19	**22**	**40**	45	26	18	34
	A-Unet	**26**	**28**	**27**	**19**	**32**	**56**	**36**	**35**	**30**	41	60	**37**	**44**	**27**	**53**	**48**	**24**	**29**	**44**	**40**	**20**	**22**	**40**	**46**	29	**21**	**35**
EPI	SoftSem	**28**	**5**	**6**	**5**	**17**	**30**	**6**	**6**	**7**	**20**	36	**36**	**28**	**32**	**58**	28	**32**	29	79	32	**36**	29	**77**	54	54	**38**	**31**
	SoftLearntGridPattern	27	**5**	3	**5**	16	**30**	**6**	5	**7**	**20**	**42**	35	25	31	57	**35**	31	**30**	**80**	**35**	32	**30**	**77**	**57**	**55**	**38**	**31**
	UnetBili	22	4	3	3	11	27	5	4	5	17	39	35	25	28	49	30	28	24	77	28	30	21	70	50	54	28	28
	DeepLabV3	22	4	3	4	9	27	6	5	6	15	39	**36**	25	28	46	31	29	24	77	29	30	21	70	48	53	17	28
	A-Unet	22	4	3	4	12	27	6	5	6	17	39	**36**	25	28	49	31	29	24	77	29	30	21	70	50	**55**	29	28

Table 4.5: Agreement between translation results and targeted maps, for the two best baselines, UNetBili A UNet followed by a resampling layer, A DeepLabV3 architecture and the A-Unet

.

Source		P					C					O					Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	
OA_{gt}	SoftSem	47	46	62	79	**81**	**72**	51	76	84	**85**	71	65	**86**	86	**92**	72
	SoftLearntGridPattern	52	45	68	80	**81**	68	57	77	85	**85**	73	68	**86**	89	**92**	**74**
	UnetBili	**57**	50	69	81	76	69	**58**	77	**86**	76	**74**	69	**86**	**91**	87	**74**
	DeepLabV3	**57**	**51**	**70**	**82**	73	69	**58**	**78**	**86**	74	**74**	**70**	**86**	**91**	82	73
	A-Unet	**57**	**51**	**70**	**82**	76	69	**58**	**78**	**86**	78	**74**	**70**	**86**	**91**	87	**74**
$mF1_{gt}$	SoftSem	12	18	29	22	**26**	**62**	**42**	**56**	**55**	**60**	47	22	51	30	**43**	**38**
	SoftLearntGridPattern	13	18	22	25	**26**	59	37	50	48	57	41	28	48	31	41	36
	UnetBili	**21**	22	30	**26**	22	51	36	44	42	38	52	**39**	**53**	**34**	37	37
	DeepLabV3	**21**	**24**	**31**	**26**	17	51	36	45	42	30	**53**	**39**	**53**	**34**	32	37
	A-Unet	**21**	**24**	**31**	**26**	22	51	36	45	42	38	**53**	**39**	**53**	**34**	39	37

Table 4.6: Comparison between translation results and ground-truth for the two best baselines, UNetBili (a UNet followed by a resampling layer), A DeepLabV3 architecture and the A-Unet. Unlike the agreement measure, which is computed on the target original spatial extent, this measurement is computed France-widely.

We compared the different loss functions on the OSO to CLC translation, which we considered the most interesting variety (44 labels in CLC) of predictable classes (OSO is highly resolved with 23 classes). Table 4.7 resumes the $MF1_{ag}$ and OA_{ag} of the different loss functions introduced earlier while Figure 4.14 presents the F1-score per class. Our observation is that CE obtains the highest OA_{ag} but the lowest $mF1_{ag}$. All our attempts to increase the $mF1_{ag}$ score worsens the OA_{ag} metrics; a compromise must be made between having a higher $OA_{ag}/mF1_{ag}$. A simple rule of thumb to compare the different losses is to use the CE as a reference for computing the ratio between the improvement of $mF1_{ag}$ and the decrease of OA_{ag}. For instance, the focal loss with α =2 increased the mF1 from 2% and decreased the accuracy of 1% resulting in a 2/1 ratio. With this strategy, we conclude that the combined CE and Dice loss offers the best compromise (5/2 ratio).

	CE	focal α =0.5	focal α =1	focal α =2	focal α =5	Dice	CE+Dice
OA_{ag}	68	68	67	67	65	65	66
$mF1_{ag}$	37	37	38	39	40	40	42

Table 4.7: OSO to CLC translation OA_{ag} and $mF1_{ag}$ for different loss functions. The focal loss is computed with different alpha values.

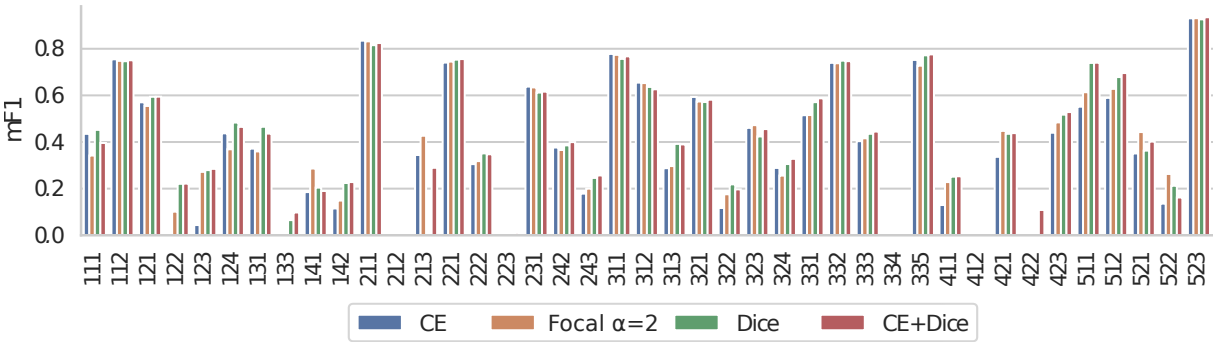

Figure 4.14: $mF1_{ag}$ computed between CLC translated from OSO and the original CLC map for various loss functions.

4.2.2.4 Conclusion

This section presented experiments on how to consider the spatial context for translation. We separated our analysis into two parts, the first one studying the potential of manually defined shape features to help the translation. Although significant improvements are observed compared to the baselines, we highlighted that using manually defined features instead of learnt features might be detrimental to the quality of translations. We secondly proposed a deep learning-based method that jointly analyses the shape of the object a pixel belongs to and the classes of neighbouring pixels. We obtained significant improvement in most cases. However, we underlined that the current method lacks spatial generalization compared to Semantic baselines and is sensitive to target noise. Those problems is addressed in the next sections.

4.3 Geographic context

Section 3.2 observed that land-cover classes are heterogeneously distributed over France for various anthropic and bio-physical reasons. The spatial heterogeneity is further increased by the uneven distribution of errors on the territory resulting from annotation by different human operators depending on the localisation or local fine-tuning of machine learning models.

We argue that information about wide-scale spatial heterogeneity, which we refer to as *geographic context*, is valuable for translation as information such as *"this pixel is in a mountainous area"* drastically influences the probable translation *e.g.* a forest is probably composed of coniferous stands. This spatial heterogeneity could theoretically be taken into account directly by the network if all of France were processed as a single input. However, the patch-based approach used to circumvent memory limitation prevents it. Answering how to incorporate geographic context is mandatory.

Section 4.3.1 focuses on identifying potential sources of geographic context suitable for translation. Section 4.3.2, proposes a methodology to incorporate one of these potential sources in the A-Unet architecture.

4.3.1 Determining a source of geographic context

To simplify our analysis, the experiments conducted in this section only focus on the CGLS to CLC translation. This choice is made based on the observation that this small to high number of classes is a France-wide challenging translation for which the impact of geographic context should be especially visible. Therefore it simplifies comparison between methods. The impact of geographical context on other translations is detailed in section 4.3.1.

4.3.1.1 Motivation

Section 2.3.2 points out that geographical context is difficult to apprehend as many elements influence class distribution (topography, climate, anthropogenic factors, pedology, error paterns ...). Additionally, abstract notions such as distance thresholds can help translation, *e.g.* inland wetlands are far from the seashore while maritime wetlands are not. Designing a geographical context-aware method using the vast amount of multimodal data sources needed to learn such notions is challenging.

State-of-the-art techniques rely on stratified training in which a separate classifier is trained for areas with different geographical context [139]. This alleviates the need to incorporate multiple data sources in training but assumes prior knowledge of homogenous geographical context areas. For instance, the OSO map is generated by training independent classifiers on eight different ecoclimatic areas (see Figure 3.3) based on climate variables by [154]. An implementation, termed **HardStatEcoCli**, is presented in the following section.

From a theoretical aspect, stratified training introduces the idea that each area is completely independent. For instance, if forests in an area are always composed of coniferous stands, then translating forest to broad-leaved or water are comparable errors. Indeed resulting models only use the information of their respective areas and thus do not have access to more general knowledge. This results in sharp edges between two different areas, *e.g.* all forests are translated into conifer in area 1 and broad-leaved in area 2. As geographic context can be a more progressive gradient, we propose directly feeding a unique model with using the area identifier as an input feature. An implementation of this method termed **LearntClassEcoCli** is presented in the following section.

The core limitation of the two previous approaches is that they often focus on a single source of geographical context, in the previous example, eco-climatic areas, neglecting all the other sources. This drastically constrains the potential of geographical context to the quality of the defined areas. In Section 2.3.2 we observe that land-cover may co-variate with its spatial coordinates (latitudes and longitudes). Close-by coordinates generally exhibit similar topography, climate and pedology. Conversely, those geographical variables are highly correlated to the coordinates. For instance, in France, the temperature tends to be higher for low latitude values while precipitation tends to be lower, the wind is more pronounced on the West side of the country due to the proximity to the seashore, and mountainous areas are only observed at specific latitude and longitude. One could directly use the spatial coordinates as a proxy to assess geographical context. However, unlike the eco-climatic areas, the link between the geographical context and the coordinates is not easy to analyse. We can not easily define a stratified approach in which different models are trained depending on the coordinates as we do not have direct knowledge of the links between coordinates and geographical context. Therefore using a machine-learning strategy to use geographical coordinates as features seems appropriate. A simple illustration is that once trained to learn the main characteristic of the land-cover map of a country, one can realise an educated guess on the nature of an element in an area only by knowing its

coordinates. For instance, if one is asked to guess what is the land-cover map at 48.8566° N, 2.3522° E, he would probably answer *buildings*, provided he previously learnt that some close-by coordinates belong to Paris. An implementation of this method termed **LearntClassLatLon** is presented in the following section. As the Stratified approach is the only common strategy in land-cover mapping literature, a comparison with the two other methods has to be conducted.

4.3.1.2 Experimental protocol

HardStat introduced in Section 4.1.2 assigns to each CGLS class the most frequently observed CLC correspondent on the training set, *i.e.* if CGLS *Closed forest* pixels are observed at the same place as CLC *Broad-leaved forest* pixels 80% of the time on the training set, then *"Broad-leaved forest* is the most probable translation for *Closed forest*. We underline that **HardStat** method gives identical results as a random trained to translate using only one feature (the class of the considered pixel). **HardStat** is used as a control to evaluate the gain or benefits of the other methods.

HardStatEcoCli is very similar to **HardStat**. However, instead of assigning to each CGLS class the most frequently observed CLC correspondent on the whole training set, it assigns to each CGLS class the most frequently observed CLC correspondent in the considered ecoclimatic zone. For instance, if *Closed forest* pixels correspond at 90% to *Broad-leaved forest* in an ecoclimatic area A and at 70% to *Coniferous forest* in an ecoclimatic area B, the translation of *Closed forest* is different depending if the considered pixels is in area A or B. As predefined ecoclimatic areas are required, we rely on the same ecoclimatic map as OSO [154]. This method is used as a comparison tool to assess if learning to analyse the geographical context from features is better than using those features to perform stratified learning.

LearntClassEcoCli proposes to use the directly ecoclimatic information as a feature used in a machine learning model. A random forest model predicts the CLC class of a pixel based on the CGLS class and the ecoclimatic zone it belongs to. This approach offers the possibility to avoid training multiple independent models as for **HardStatEcoCli**.

textbfLearntClassLatLon is based on the same idea as **LearntClassEcoCli** but processes the geographical coordinates directly instead of the ecoclimatic zone as a feature. A random forest is trained to predict the CLC class based on the CGLS class and the latitude and longitude. Three sub-methods are also proposed to study the influence of the coordinates alone. **LearntLatLon** only use the coordinates with no information of the CGLS class. **LearntClassLat**, only have access to CGLS class and the latitude. **LearntClassLon** only have access to the CGLS class and the longitude.

Implementation details: Class used as input features for **LearntClassEcoCli** and **LearntClassLatLon**, are one-hot encoded. Similarly the ecoclimatic areas used as input features of **LearntClassEcoCli** are also one-hot encoded. We underline that those methods are trained on the training patch of the MLULC dataset after applying an

exclusion buffer on the patch's edges to limit the risk that the use of coordinates breaks the independence between the train and test. An arbitrary 2.5km buffer is used based on the observation that the 25ha and 100m width CLC MMU does not guarantee independence below, *i.e.* each patch being 6×6 km wide only a 1*times*1km area in the centre of the patch is used for training. Additionally, instead of using the exact coordinates of the considered pixel, each pixel is annotated with the pixel coordinates in the centre of the patch it belongs to.

4.3.1.3 Results

	HardStat	HardStatEcoCli	LeanrtEcoCli	LearntLatLon	LearntClassLat	LearntClassLon	LearntClassLatLong
OA_{ag}	54	55	55	49	55	54	61
$mF1_{ag}$	13	16	17	16	18	17	28

Table 4.8: Agreement measurement computed between CLC and the translation result of various geographic context aware methods

Table 4.8 presents the results of each of those methods. First, we observe that geographical context-aware method obtains better results than **HardStat**. Note that the $mF1_{ag}$ remains unsurprisingly low as the translation from CGLS to CLC involves translating 12 classes into 44. Secondly, **LearntLatLon** perform globally worse than **HardStat** (-6% OA_{ag}) while its $mF1_{ag}$ score is higher (+3%). This underlines that a higher class diversity can be obtained using a pure coordinate analysis rather than the CGLS classes. More interestingly, the results demonstrate that the simultaneous use of coordinates and the CGLS class (**LearntClassLatLon**) can significantly improve the translation into CLC compared to the **HardStatEcocli** and **LearntClassEcoCli** baselines. Lastly, by comparing those results with those presented in the spatial context section, we observed that the obtained OA_{ag} and $mF1_{ag}$ for the **LearntLatLon** are identical between the **A-Unet** and this random forest with CGLS labels and coordinates. This is particularly interesting as it suggests that for this specific translation, the geographical context might have the same order of importance as the spatial context.

To better understand how the geographical context can improve the results, we provide the per-class F1-score histogram for the different methods in Figure 4.15. A first observation is that **HardStatEcocli** and **LearntClassEcoCli** give almost identical results except for the CLC *Bare rock* and *Water bodies*, for which the **LearntClassEcoCli** get slightly higher results. This underlines that learning to interpret geographical context, *i.e.*, how to analyse ecoclimatic areas, might slightly improve translation compared to a stratification into multiple independent models. The second observation is that the LearntLatLon, which, unlike other models, has no access to the CGLS labels, outputs a significantly more expansive number of classes than **HardStat**, **HardStatEcoCli** and **LearntClassEcoCli**. For instance, it predicts partially classes such as *Rice Fields*, which are difficult to translate using only CGLS labels or Climate based indicators. This is explained by the fact that

those classes are spatially constrained to a limited area, *i.e. Rice crops* are only observed in a particular region. Lastly, **LearntClassLatLon** widely outperforms any other methods.

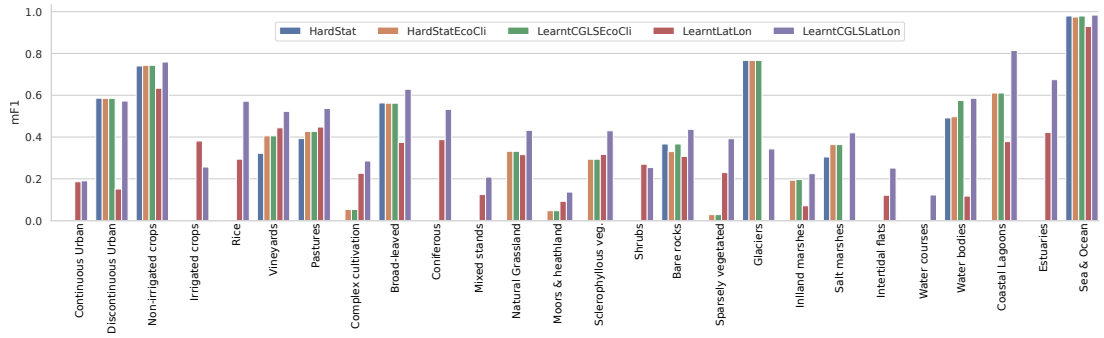

Figure 4.15: Per class F1-score of the translation into CLC for different geographical context methods. For visibility, only the 26/44 CLC class (non-zero f1-score) are displayed.

Figure 4.16 compares the OA_{ag} of the geographical context methods with the OA_{ag} of the **HardStat** method by summarising results on 30×30 km grid. Blue (reciprocally red) pixels denote the areas where the geographical context method obtains better (reciprocally worse) results than the **HardStat** method. As the aspect of the HardStatEcoCli and LearntClassEcoCli maps are identical, we only provide one for readability. The first observation is that adding the geographical context does not improve the translation homogeneously across France. Most of the improvements are observed in the Southern part of France, mainly in the mountainous areas and the Mediterranean coastline (which exhibits a specific climate). LearntClassLatLon improved the results in more areas. For instance, blue parts observed on the west coast are due to the successful translation from CGLS class *Closed Forest"* to CLC class *Coniferous forest* of a vast coniferous forest area, which is the only one of this size outside a mountainous area inside the French territory (otherwise it is translated into "Broad-leaved forest").

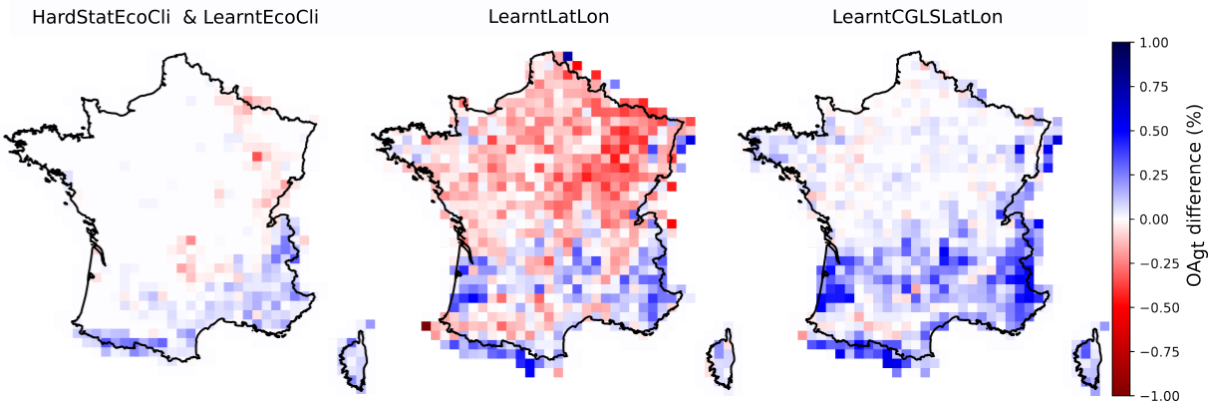

Figure 4.16: Spatial distribution of translation improvement using various techniques compared to the **HardStat** method. Blue areas denote the location where the considered method outperformed the **HardStat** method, and red areas give the opposite information.

4.3.1.4 Conclusion

We conclude that the best feature to assess geographical context is the geographical coordinates as it substantially improves the stratified ecoclimatic training and does not require incorporating multi-modal data. In particular, our experiments underline that both latitude and longitude are needed to get a refined analysis of the geographical context. The achieved translation improvement is not distributed homogeneously across the territory.

Moreover, in the case of the CGLS to CLC translation, the improvement of using coordinates is equivalent to the one fostered by using the spatial context with A-Unet. Therefore developing ways to incorporate geographical context into the A-Unet appears important.

4.3.2 Incorporating geographical coordinates into a convolutionnal neural network

4.3.2.1 Motivation

Section 2.3.2 underlined that little work had been conducted on incorporating geographical coordinates in machine learning frameworks, especially on convolutional neural networks.

The few works addressing this issue underlined the difficulty of using GPS coordinates directly as the input of a neural network [348]. However, they did not formulate hypotheses on the underlying reasons and simply observed that very fine location indicators such as GPS coordinates are difficult to analyse for the classifier [294].

In natural language processing, the position of words in a sentence is often used to improve the translation. These word positioning problems share with geographical coordinates the idea that training and testing values might differ, *i.e.* No sentence in the training data is 24 words long, but some sentences of the testing data are. To tackle this problem, a positional encoding strategy was proposed by [318] (See Section 2.3.2). It transforms the coordinates using a fixed set of size $2d$ composed of cosine and sine functions with different frequencies. Let x be the considered position in an input sentence, $\overrightarrow{p_x} \in \mathbb{R}^d$ be its corresponding encoding, and d be the encoding dimension. Then $f : \mathbb{N} \rightarrow \mathbb{R}^d$ is the function that produces the output vector $\overrightarrow{p_x}$ and is defined as shown in Equation 2.9. The attracting aspect of this transformation is that it ensures that all positions have a unique encoding while ensuring that this encoding value always remains in a constant range between -1 and 1. Moreover, close by values remains close by in the encoding; for instance, the cosine similarity between the encoding of position one and position 2 is higher than between position 1 and 9. This is interesting as it might help the network to find relations of proximity between objects which we previously underlined as important for land-cover map translation.

Section 2.3.2 pointed out that positional encoding is also used in computer vision to break the translation invariance of convolution by encoding pixel location *e.g.* to distinguish the

upper left pixel of an image from the right bottom one. However, no guarantee exists on the efficiency of positional encoding for geographical coordinates as the problem exhibits widely different characteristics. First, the proposed approach should not break pixel translation invariance as geographical context on 6k×6 km patches is homogeneous, *e.g.* a forest in the upper part of a patch should be treated the same way as one in the bottom. Consequently, all pixels of the same patch should have the same positional encoding. Instead, the approach aims to break the patch invariance to translation, *i.e.* the same patch in the North and South of France should be translated differently. Secondly unlike pixel coordinates which are the same for training and testing patches, *i.e.* the first pixel is a notion existing both for training and testing, geographical coordinates are different between training and testing. It requires a generalisation ability that is not required for pixel positional encoding. We present in the following sections the experiments conducted to evaluate the potential of positional encoding on geographical coordinates.

4.3.2.2 Methods

Figure 4.17 resumes the different experimented methods concisely. All the presented methods only have access to the coordinates of the patch's centre pixel instead of providing each pixel's geographical coordinates. This directly stems from the observation that geographic context is defined as wide-scale context, *i.e.* is the same on a 6×6 km^2 patch, and that using per pixel location would break the translation invariance as discussed previously.

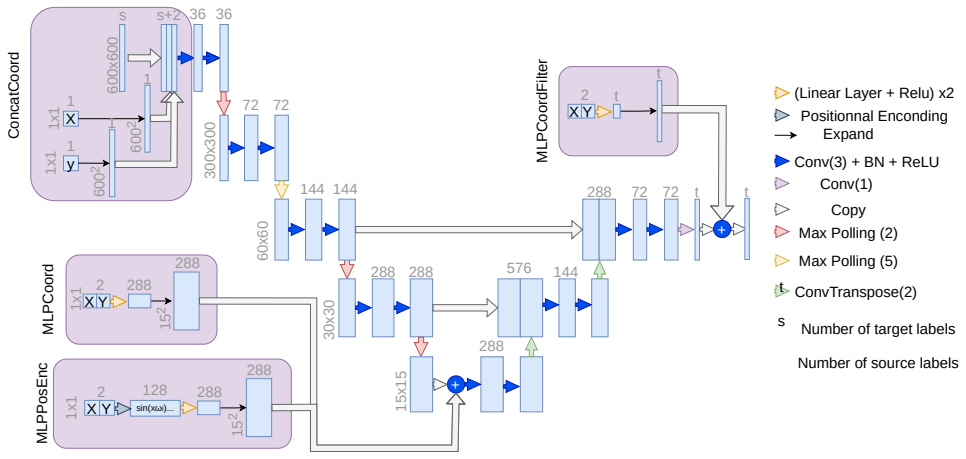

Figure 4.17: Presentation of four different methods (purple boxes) to add geographical coordinates to A-Unet. Each methods is used independently.

As a simple reference baseline, we propose to concatenate to the input map two new dimensions, one with the latitude and one with the longitude value. This simple baseline is referred to as the **ConcatCoord** method.

As a more evolved solution, we propose feeding the coordinates to a small multi-layer

perceptron to learn a coordinate embedding that can be added to the network. **MLPCoord** adds the output of the MLP to the bottleneck of A-Unet so that the positional information only influences a coarse scale using simple addition. **MLPCoordFilter** is identical to [49], one of the only approaches found in the literature using geographical coordinates. It adds the output of the MLP at the end of the network.

Lastly, we propose **MLPPosEnc** which use the same strategy as **MLPCoord** method, but using the positional encoding before the MLP discussed earlier. Unlike the traditional sequence-to-sequence architecture, the positional encoding must be in 2D to integrate both latitude and longitude. The authors of [235] proposed a strategy for image coordinates: rows and columns are independently encoded and then concatenated. We adopt the same strategy with the latitude and longitude coordinates. Let x and y be the longitude and latitude. $p_{x,y}$ is the corresponding positionally encoded matrix. d is the dimension of the encoded matrix, corresponding to the number of feature maps generated by the CNN layer where the positional encoding is added.

$$p_x = \begin{bmatrix} sin(x\omega_1) \\ cos(x\omega_1) \\ \vdots \\ sin(x\omega_{d/4}) \\ cos(x\omega_{d/4}) \end{bmatrix}_{d/2} \quad p_y = \begin{bmatrix} sin(y\omega_1) \\ cos(y\omega_1) \\ \vdots \\ sin(y\omega_{d/4}) \\ cos(y\omega_{d/4}) \end{bmatrix}_{d/2} \quad p_{x,y} = \begin{bmatrix} sin(x\omega_1) \\ cos(x\omega_1) \\ \vdots \\ sin(y\omega_{d/4}) \\ cos(y\omega_{d/4}) \end{bmatrix}_{d} \quad with\ \omega_i = \frac{1}{10000^{2i/d}}. \tag{4.7}$$

The resulting encoding $p_{x,y}$ is given to a one layer perceptron to preprocess the positional encoding. Adding more layers did not show significant improvement in our results.

4.3.2.3 Results

Table 4.9 presents the agreement measurements for the four different geographical coordinates encoding methods. We underline that for processing times concerns, we did not train each of the 26 models a dozen times to obtain a standard deviation estimation of each approach and thus can not conclude on the statistical significance of each value of the table. However from experience, we estimate the OA_{gt} and $mF1_{gt}$ variations between training to be between $\pm 0.3\%$ for OA_{gt} and $\pm 1\%$ for $mF1_{gt}$. We additionally provide in Figure 4.19 the per-class metrics of the OSO to CLC translations with error bars indicating the standard deviation over 10 independent training to give an insight into the training variability.

A first observation is that the **ConcatCoord** obtains slightly worse results than not using coordinates. This is directly linked to the fact that it enforces learning maps and coordinates features with the same first convolution layer despite being unrelated.

Interestingly, the **MLPCoord** approach obtains almost identical results as using the A-Unet without coordinates confirming that using geographical coordinates directly is

not efficient as explained in other studies [294]. Even though the **MLPCoordFilter** proposed by [49], slightly improve the results, the proposed **MLPPosEnc** achieve a more significant improvement (+1% OA_{ag} and +3% $mF1_{gt}$) encompassing wide per translation differences from -2% to +8% OA_{ag} and -3% to +12% $mF1_{ag}$. Explaining the reason for each improvement/non-improvement for each translation is unfeasible as it depends on the source and target nomenclature complementarity, but some general rules are observed.

In general, the biggest improvements are observed when translating OCSGE cover and use into CGLS, CLC and OSO. This can widely be imputed to the peculiar spatial extent of OCSGE (see Figure3.4), *i.e.* divided into two areas, one in the North and one in the South. The North and South parts belong to widely different ecoclimatic areas conditioning many translations. For instance, OCSGE use **Forestry** mainly concerns Broad-leaved trees in the North and Coniferous in the South, and OCSGE cover **Herbaceous vegetation** mainly includes pastures in the North and croplands in the South. Geographical context incorporation help to translate each part of the map independently. Even though this difference in geographic context also helps to translate CGLS, CLC and OSO into one another, their France wide-extent partially reduces the quantitative impact of this improvement. For instance, trees in the south part of OCSGE account for approximately 10 15% of CGLS, CLC and OSO trees (France wide maps) but for more than 50% of OCSGE. Thus the maximum improvement on tree translation by adding geographical context in the south part is 10 15% for France-wide maps while being close to 50% for OCSGE.

Conversely, using coordinates brings no improvement or worsens the results when the source map or target is the MOS map. We link this behaviour to the fact that the MOS is by far the one with the smallest spatial extent (around 2% of France's territory). The geographical context is almost the same in all areas, making the geographical coordinate encoding useless. Moreover, we observed some overfitting of the network due to the small number of coordinates which downgrades the results on the test set.

Source		CGLS (P)					CLC (C)					OSO (O)					OCSGE cover (G1)				OCSGE use (G2)				MOS(M)			Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	
OA	NoCoord	62	55	68	**78**	**76**	73	58	72	**80**	**79**	**77**	68	79	**86**	**85**	68	54	56	92	67	49	49	79	84	82	63	71
	ConcatCoord	62	54	68	**78**	**76**	72	58	72	79	**79**	**77**	67	79	**86**	**85**	68	54	55	92	66	48	49	79	83	82	63	70
	MLPCoord	63	55	68	**78**	**76**	73	58	72	**80**	**79**	**77**	68	79	**86**	**85**	69	55	57	92	67	51	50	79	85	82	63	71
	MLPCoordFilter	63	56	68	**78**	**76**	73	58	72	**80**	**79**	77	68	79	**86**	**85**	69	56	57	92	67	52	51	79	85	82	63	71
	MLPPosEnc	**65**	**57**	**69**	**78**	**76**	**75**	**59**	**73**	**80**	**79**	**77**	**69**	**80**	**86**	**85**	**71**	**58**	**58**	**93**	**69**	**54**	**53**	**79**	**86**	**84**	**64**	**72**
mF1	NoCoord	26	28	27	19	32	56	36	35	**30**	**41**	60	37	44	**27**	**53**	48	24	29	**44**	40	20	22	**40**	**46**	29	21	35
	ConcatCoord	26	28	27	19	32	56	34	35	**30**	**41**	60	37	44	**27**	**53**	47	25	29	**44**	41	20	22	**40**	**46**	28	21	35
	MLPCoord	27	28	27	19	32	56	36	36	**30**	**41**	60	38	44	**27**	**53**	48	27	29	**44**	43	22	23	**40**	45	29	21	35
	MLPCoordFilter	28	28	28	19	**33**	56	36	36	**30**	**41**	60	38	44	**27**	**53**	49	29	29	43	45	23	24	**40**	45	29	22	36
	MLPPosEnc	**30**	**29**	**30**	**20**	**33**	**58**	**37**	**38**	**30**	**41**	**61**	**40**	**45**	**27**	**53**	**52**	**34**	**31**	43	**52**	**29**	**25**	**40**	45	**30**	**23**	**38**

Table 4.9: Agreement between the translation results of the different geographical coordinate aware models and the target maps. The backbone network is A-Unet.

Table 4.10 presents the results of the different geographical coordinates encoding methods on the ground truth. Translation learnt on France-wide source to local extent target can not be used to produce France-wide target using geographical coordinates encoding methods as only local extent coordinates are seen during training. Therefore metrics computed on the France-wide ground truth can only be computed for the six France-wide sources to France-wide target translations. This, of course, represents a major limitation

to the method that is addressed in the next chapter. The observations of the ground truth results confirm the improvement observed in agreement metrics, ensuring that observed improvements are not due to a higher ability of the network to replicate target errors when using coordinates.

Source		CGLS (P)		CLC (C)		OSO (O)		Average
Target		C	O	P	O	P	C	
OA	NoCoord	57	51	69	58	74	70	63
	ConcatCoord	56	51	69	57	73	69	63
	MLPCoord	58	52	69	58	**74**	70	64
	MLPCoordFilter	58	53	69	58	74	70	64
	MLPPosEnc	**60**	**55**	**70**	**59**	**75**	**71**	65
mF1	NoCoord	21	24	51	36	53	39	37
	ConcatCoord	22	24	51	36	53	40	38
	MLPCoord	22	25	52	36	53	41	38
	MLPCoordFilter	24	25	52	36	53	42	39
	MLPPosEnc	**27**	**28**	**53**	**36**	**56**	**45**	41

Table 4.10: Comparison between the different geographical coordinates aware models translations and the ground truth. The backbone network is **A-Unet**. Only translations for which the source and target spatial extent are France-wide can be computed.

To evaluate the ability of the network to interpolate geographical context from a limited number of training patches, Figure 4.18 presents the evolution of the agreement metrics depending on the training size for **NoCoord** and **MLPPosEnc**. The validation and test patches are fixed and represent each 5% of patches. A first observation is that **NoCoord** and **MLPPosEnc** results stabilized at the same training size (between 70-80% of all patches). It underlines that using the coordinates does not imply increasing the training size compared to a no-coordinate approach. We also observe that even when only 5% of the patches are in the training set, the **MLPPosEnc** gives slightly better results than the **NoCoord** one. We conclude that using coordinates is never detrimental. Lastly, using 40% of patches to train **MLPPosEnc** gives the same results as using 90% with **NoCoord** underlying the vast potential of coordinates to reduce training size.

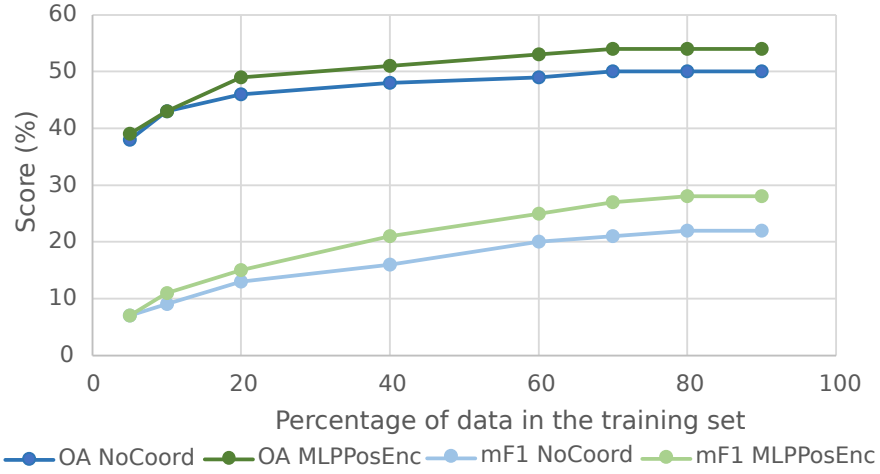

Figure 4.18: Agreement for OSCGEu to CLC translation depending on training set size.

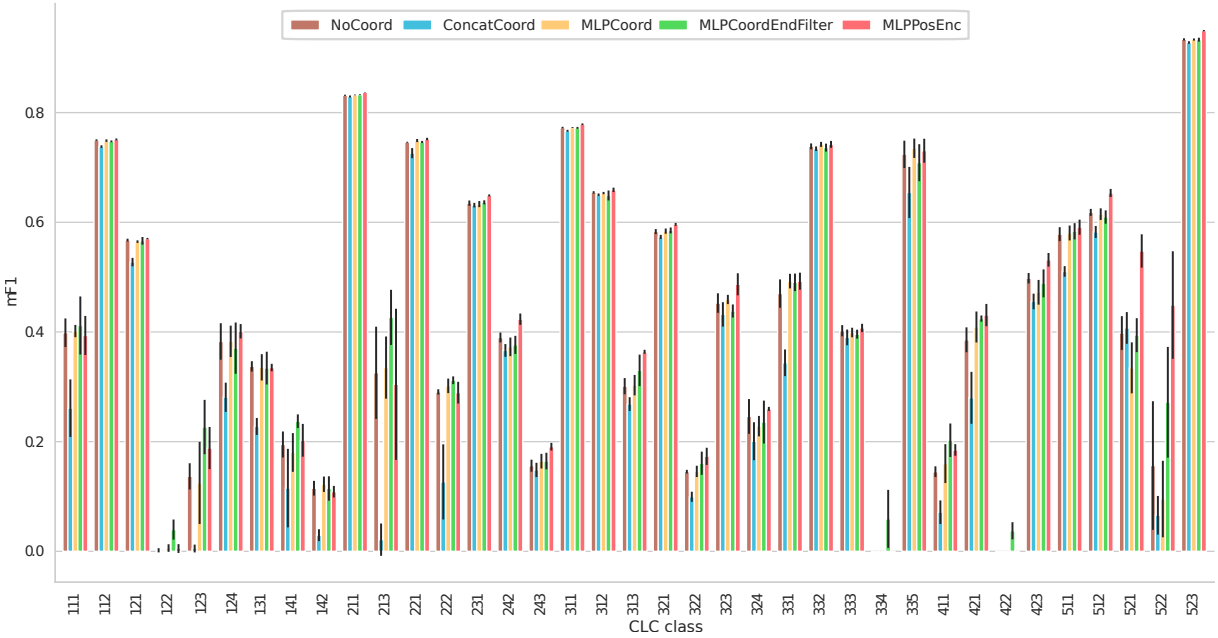

Figure 4.19: Per class agreement F1-score of the OSO to CLC translation using the different coordinate encoding methods. Errors bars are computed as the standard deviation over 10 indepent training. Only the 37/44 CLC classes with non zero F1-score are displayed for readability.

4.3.2.4 Conclusion

We first focused on determining potential data sources of contextual information. Instead of adding all the multi-modal data that could provide geographic context information, we propose to use the geographical coordinates directly to improve the translation. By comparing the result obtained by a random forest trained to use coordinates with one trained using Ecoclimatic areas, we obtained an intuition of the high potential of using a coordinate approach.

The second section investigates how to incorporate those coordinates to a convolutional neural network. Inspired by recent advances in natural language processing, we proposed to use positional encoding to ensure that the coordinate feature extraction learnt on the training set transfer well to the test set. The results demonstrate the effectiveness of the method in improving translation accuracy. However, this solution suffers one main limitation: it requires the training data to be more or less homogeneously dispatched over the studied area. This is not an issue as long as no spatial generalisation is needed but it becomes problematic if one wants to increase the spatial extent of a target map using a wide extent source. We address this issue in Section 5.1. Additionally, we point out that since no comparison between the MLPPosEnc method and a stratified training based on EcoClimatic areas was conducted during this PhD, no conclusion is made on the hypothetical superiority as results obtained using the random forest algorithm are not transferable to those obtained with a CNN. However, we underline that unlike the current

state of the stratified art training, the MLPPosEnc does not require pre-defined knowledge of the area and only needs to train a single model.

4.4 Temporal context

This section focuses on the impact of temporal context and how to integrate it. We distinguish two underlying kinds of temporal context. The semantic temporal context, discussed in Section 2.3.3.1, is the one directly implied by the source and target nomenclature definitions. For instance, CLC *Rice fields* definition states, "As part of regular cultivation cycle, rice fields are occasionally left fallow for 1-3 years. These parcels are considered to be rice fields, too." and the "Pastures" definition states "Lands that are permanently used (at least five years) for fodder production.". Conversely, the "Rice crops" definition for OSO only includes rice fields cultivated in the considered year. Thus the translation from OSO to CLC requires three years of OSO maps to perform the correct translation.

The second one, we refer to as the temporally constrained context describes the potential temporal gap between the source and target used for learning. For example, translation from OSGEc to CLC involves learning the translation from a 2014 map to a 2012 or 2018 map. This temporal gap constraint is not a problem for traditional semantic-based methods as they do not learn the translation on existing maps. Conversely, it could represent a considerable limitation if this temporal gap significantly affects the translation results. This section investigates those two temporal contexts separately. We exclude from this study the use of image time-series to preserve a pure land-cover translation approach.

The following experiments focus on the OSO to CLC translation, as (i) CLC is the maps with the highest number of classes with temporal context (ii) 3 consecutive years of OSO maps (2016, 2017, 2018) can be used for CLC 2018 translation. The CGLS-LC100 to CLC translation could have also been studied as four consecutive years of CGLS are available but would have been more challenging to analyse due to the lower diversity of predictable classes using CGLS and the high error rate of the original CGLS product. Eventually, since the OSO 2016 and 2017 nomenclature merges most of the agricultural classes into two superclasses, *Summer crops* and *Winter crops*, we perform similarly for the OSO 2018 products, *i.e.* all the OSO products included 17 classes.

4.4.1 Multitemporal source translation

4.4.1.1 Motivation

Section 2.3.3.1 underlines that state-of-the-art for incorporating multitemporal information into a classical land-cover map generation problem involves analysing a large time series of images. The literature shows the current superiority of correctly encoding the temporal context provided by satellite image time-series into a deep-based architecture (using

Transformers or Recurrent Neural Networks with adapted cells) w.r.t. manually defined features based on the variation of given spectral indices.

However, taking into account temporal context in the land-cover map translation case is widely different due to the difference in nature of the problem and data availability :

- Temporal ordering does not matter: nomenclature definition such as CLC "Rice Field": "[...] rice fields are occasionally left fallow for 1-3 years. [...]" does not include a notion of temporal ordering, *i.e.* it does not matter if the field was occupied by rice the last year, two years ago or three years ago but only that it was occupied by rice once during the three years.
- We do not have access to substantial temporal stacks of land-cover maps: most land-cover maps have been produced for less than ten different periods, many of them only once.

The limited and irregular temporal sampling through decades of LC maps prevents from designing a specific method for temporal information extraction. We assume assume that using a simple approach, such as concatenating multiple versions of the source maps, is sufficient to achieve improvement. We instead focus on understanding the reason behind the potential improvement. Especially as the fusion of multiple dates of a source requires frequently updated land-cover maps, the method discussed here is mainly intended to be used on automatically classified land-cover maps (ACLC). Section 2.3.4 pointed out that those ACLC exhibit particular spatial error patterns that might vary depending on the year due to changes in the classification algorithm or in the data used. When attempting a multitemporal analysis on such source, learning different features per year could improve translation results by better detecting source errors rather than exploiting temporal context. We argue that understanding the reason behind potential improvements is crucial for future studies on multi-temporal map denoising. Refer to Section 5.2.1 for a more elaborate way than concatenation to incorporate multiple sources of data information (such as images, maps, Digital terrain model).

4.4.1.2 Experimental protocol

We compare three methods with different temporal information extraction power:

Concat feeds the network with a temporally ordered concatenation of OSO 2016, 2017, and 2018. Since input is one-hot encoded, we underline that the first 17 channels describe OSO 2016, channels 17 to 34 describe OSO 2017, and the remaining 34 to 51 OSO 2018. In this setup, the network might learn different features for the same land-cover class depending on its year of observation.

Shuffle randomly exchanges the temporal order of the concatenation for each patch, *e.g.* in one patch, the first 17 pixels describe OSO 2018 while in the other patch, the first 17

channels describe OSO 2017. In this setup, the same features are learnt for all years for a given land-cover class. However, the learnt features can be oriented towards detecting some error patterns specific to a given year, *e.g.* focus on distinguishing OSO 2018 roads which are far less erroneous than other years and exhibits different spatial pattern.

Mean replaces the concatenation by the temporal average of one-hot encoding. For instance, in a two-class setup a pixel classified as class one during two years and as class two during one year is encoded as $e = [\frac{2}{3}, \frac{1}{3}]$. This encoding is independent from temporal ordering making it difficult for the network to learn year-specific features.

Theoretically in a scenario where a multi-temporal analysis only benefits from the temporal semantic context, the results of this three methods should be identical, *i.e.* temporal ordering should not matter. A higher score of the **Concat** compared to other methods would indicate that learning custom feature per year is beneficial for the translation, which could be linked to a higher ability to identify source errors.

To comfort the distinction between improvement fostered by semantic context from those fostered by a higher ability to compensate source errors, CLC classes with annual temporal context elements are manually identified (see definitions in Appendix D). This results in the selection of 10/44 CLC classes which should be the only to one to benefit from the multitemporal analysis under a no-noise compensation scenario. It mainly concerns, crop oriented classes such as*Non-Irrigated arable land* and *Rice fields* (crop rotation with fallow land period) or *Pastures* (used for fodder production at least for 5 years) and natural landscapes such as *Natural Grasslands* (no human influence for long period), *Moors and heathland* and *Sclerophyllous vegetation* (includes crops left fallow for 3 years and more), *Transitional woodland/shrub* (includes recently cut/reforested areas), *Burnt areas* (recently burnt) and *Glaciers and perpetual snow* (permanent over several years). Those classes are displayed in green in the per-class histograms of the results section.

Lastly, OSO 2016 to CLC 2018, OSO 2017 to CLC 2018 and OSO 2018 to CLC 2018 translation results are presented as control values to evaluate the benefit of a multi-temporal analysis over a mono-temporal one.

4.4.1.3 Results

	OSO2016	OSO2017	OSO2018	Mean	Shuffle	Concat
OA_{ag}	68	68	68	68	68	**69**
$mF1_{ag}$	37	39	37	39	40	**42**

Table 4.11: Average over ten independent training of the agreement metrics for the OSO to CLC translation in different temporal configurations

Table 4.11 presents the translation results from OSO to CLC using the multiple methods described above. A first observation is that the **Concat** multi-temporal analysis improves translation quality, especially in $mF1_{ag}$ (+5%). The fact that the **Mean** and **Shuffle**

methods perform worse than **Concat** underlines that temporal ordering does matter. As we pointed out that CLC class definitions don't include temporal ordering notions, we assume that the main benefit of ordered temporal sequences is to enable the network to learn year dependent features describing year specific noise patterns rather than exploiting the temporal semantic context. To obtain a rough estimate of the proportion of improvement due to better noise detection or temporal semantic context, we provide Figure 4.20 presenting per class results. CLC classes, including a multi-year temporal criterium in their definition, are displayed in green.

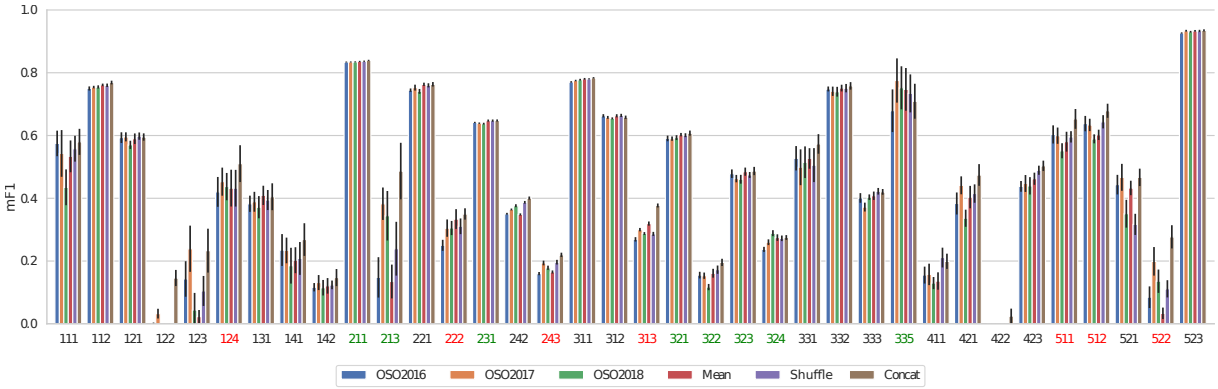

Figure 4.20: Agreement F1-score per class for the OSO to CLC translation in different temporal scenarios. CLC classes including multi-year temporal criteria in their definition, are in green. Red values indicates CLC classes with no such criteria but a high gap between one of the multitemporal method and the per-year methods. Only CLC classes with a non-zero F1 score are displayed for visibility. Errors bar are computed as the standard deviation of 10 independently trained models.

First, we observe that the three proposed method gives almost always better results than mono-temporal analysis. However, in most cases, the improvements are small, *e.g.* 112:Discontinuous urban fabric mF1 is, on average, 2% higher using **Concat** than only one date. Focusing on the "green" classes reveals that only three of the 8 classes benefit significantly from a multitemporal analysis: *211:Non permanently irrigated crops, 231:Pastures, 321:Natural Grasslands*. In the remaining 5 classes, the incorporation of multitemporal data did not increase the results. Additionally, looking at red classes reveals that the most significant improvements are observed on classes with no temporal constraints, such as 1*24:Airport* or *313:Mixed Forest*. We highlight that in those cases the Concat method usually performs better than the two other methods. We directly link those improvements to special error patterns varying across years of the OSO map, *e.g.* Airports roads in OSO are widely imprecise, and a change in data used for training fostered a difference in the road's spatial distribution in 2018 compared to other dates.

We comfort the constatation that learning custom per-dates features counteracts the noise in the source map by observing that only 4 out of the eight predicted "green" classes exhibit the same results for the three implemented methods: *231:Pastures, 311:Natural grassland, 313:Scelorypyllous vegetation, 314: Transitionnal Woodland/Shrubs*. In four

other cases, the **Concat** method performs better than the two other methods.

To sum up, we conclude that conducting multitemporal source land-cover translation is beneficial for the quality of translation. We underline that the main improvement is not due to a higher ability to analyse the temporal semantic context, as all three methods should perform equally (only ≈ 1-2% mF1 improvement). Instead, we link most of the improvement to a higher ability to identify source errors by enabling learning custom features per date (+5% mF1).

4.4.2 Temporal gap effects

4.4.2.1 Motivation

Ideally, source and target used to train our models should represent the same time period. Unfortunately, under operational uses, many land-cover translations can not be learnt on pairs of temporally matching sources and targets. This temporal gap inherently affects the training procedure by adding some label noise, *i.e.* each source pixel is learnt to be translated into what was or will be its corresponding target class in the past/future.

The amount of temporal gap induces noise correlated to the land cover change proportion observed between the two dates. This proportion depends, of course, on the temporal gap size and the spatial extent, *i.e.* some areas are prone to land-cover change. Additionally, land cover change proportion also highly correlated to the resolution, *i.e.* small changes are more likely, and nomenclature, *i.e.* the most probably changing classes over time, are crops-oriented classes.

In this section, we assume that land-cover map changes slightly between two dates as long as the period considered remains modest (within ten years), the considered spatial extent is wide enough (country scale), and the nomenclature is not focused on temporally changing classes. For instance, less than 2% changes are observed for CLC 2012/2018, MOS 2012/2017 and CGLS-LC100 2015/2018 over France. We hypothesise that under those assumptions learning land-cover map translation with a temporal gap between source and target should not bring detrimental changes to the translation quality as it induces a reasonable amount of noise (2%) compared to the noise inherently included in the land cover maps (*e.g.* CGLS product is only 72% accurate). This section focuses on testing this hypothesis by proposing three different operational setups with different temporal gaps to evaluate the robustness of the method to this temporal induced noise.

We underline that in the special case of the OSO, those assumptions are not met. Interestingly a stable 25% ±1% change is observed between all the different combinations of periods between 2016 and 2019, *i.e.* the 25% change between OSO 2016 and OSO 2017 is also observed between 2016 and 2019 or 2017 and 2018. We links this behavior to OSO crop nomenclature with an annual rotation between *summer crops* and *winter crops* explaining 15%. The remaining 10% is mainly observed on erroneous areas or edges

between different land-cover types. As those circumstances add a considerable amount of noise, we consider that a specific method robust to label noise must be designed and address this issue separately (see Section 4.5).

4.4.2.2 Experimental protocol

To evaluate the robustness of the method under a reasonable amount of changes, we present three scenarios corresponding to different time gaps between source and target land-cover maps and, subsequently, three specific operational use cases. Two reference years for CLC (2012 and 2018) and two for OSO (2016 and 2018) are selected. The experiments are conducted on CLC at level 2, which exhibits a 15 classes nomenclature, to simplify the analysis.

Scenario 1 corresponds to a scenario aiming for the automatic extension of CLC to a broader area, assuming that CLC has not yet been generated over a full area of interest. It could be particularly relevant for the forthcoming editions of CLC: one would only need to generate a high-quality sample version on specific areas, and our framework could fill the gaps. This methods trains and tests on OSO 2018 and CLC 2018. This scenario represents a no-temporal gap setup.

Scenario 2 corresponds to the updating operational setting. First, a translation model is trained on a pair of pre-existing OSO and CLC products. Then, the model is applied to the OSO product of the year for which the new CLC map is to be produced. We assume that the most recent OSO product is 2018 and that we want to produce CLC 2018. The translation model is trained using CLC 2012. We choose to pair it with OSO 2016 to minimize the disagreement in the training data caused by land-cover map changes. In this scenario, OSO 2018 is translated with the learnt model, and CLC 2018 is used as reference data for validation.

One limitation of **Scenario 2** is that if the learning algorithm can cope with discrepancies in the training data, it may be better to use the most recent OSO map in the training phase. This would allow taking into account the evolution of land-cover map, which does not affect the translation itself. For instance, one could assume that climate change makes wetland areas dryer. It would not introduce a change between CLC 2012 and CLC 2018 (dryer wetlands are still wetlands) but could introduce an evolution in OSO. Indeed, OSO does not have wetland classes, and wetlands in CLC correspond to water, sand, grasslands or moorlands in the OSO nomenclature. In this situation, some CLC wetlands could transition from water to sand, grassland or moorland. This would mean that the translation rule learned between OSO 2016 and CLC 2012 could not be applied to OSO 2018. In order to assess this situation, we propose **Scenario 3**, in which the translation model is trained by pairing CLC 2012 with OSO 2018. This model is then applied to OSO 2018. CLC 2018 is used for validation. The underlying assumption is that the model fails to learn some of the associations imposed by the training data that correspond to real changes from the point of view of the CLC nomenclature.

4.4.2.3 Results

As the scenarios gives almost identical results visually this section focuses on quantitative assessment. We provide a web interface to visualize results at a France-wide scale for scenario 2 at `https://oso-to-clc.herokuapp.com/`.

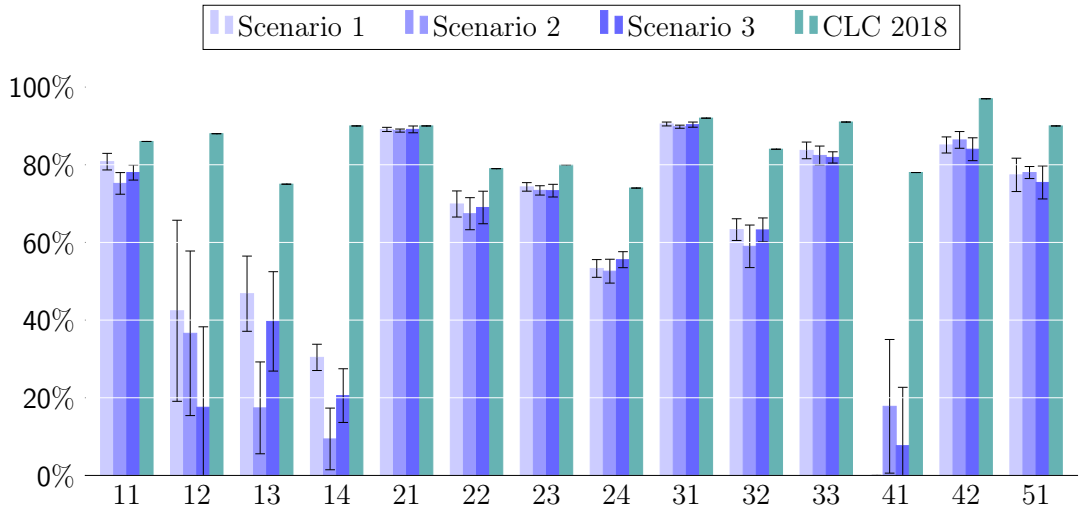

Figure 4.21: mF1 obtained by performing ten different initializations of the network for each scenario. Error bars indicate the standard deviation over the 10 trials.

Figure 4.21 presents the per-class results of the different scenarios computed on the ground truth. The original CLC 2018 level 2 map is also compared to our ground truth (achieves a 86% OA_{gt}, 85% $mF1_{gt}$). The first scenario, which translates OSO 2018 into CLC 2018 in both the train and test phase, obtains an 81% OA_{gt} and a 63% $mF1_{gt}$. The second scenario, training a model to translate OSO 2016 into CLC 2012 and then using it in the test phase to translate OSO 2018 into CLC 2018, achieves a 79% OA_{gt}, 60% $mF1_{gt}$, slightly under-performing the first scenario, especially on urban classes. This could partly be explained by the difference between the OSO 2016 and 2018 products in per class thematic accuracy. Additionally, it could also be imputed to the difference between the regularisation needed to correct land-cover map changes between OSO 2016 and CLC 2012 versus OSO 2018 and CLC 2012. The third scenario translates OSO 2018 into CLC 2012 with the underlying hypothesis that changing areas between 2012 and 2018 are only considered as additional noise achieves a performance qualitatively and quantitatively similar to the first scenario (81% OA_{gt}, 61% $mF1_{gt}$). This scenario obtains reasonably good results while not needing a common time stamp of the source and target for learning (scenario 1) and no older source map (scenario 2).

The analysis of the three scenarios reveals that, under the mentioned constraint of learning the translation with a limited temporal gap between the source and the target and on a wide spatial extent, the temporal gap does not deteriorate the results. Interestingly this statement holds for all classes.

4.5 Cartographic Context

4.5.1 Motivation

The image to the landcover classification paradigm often faces reasonable amount of noise in the input image while the label noise in the target used for training is often unknown. Conversely, translation between cartographic products involves working with significant label-noise in both source and target (5 to 30% errors) for which the noise distribution is most of the time known in the form of a confusion matrix provided by the land-cover maps producer.

Precisely estimating the impact of source and target label noise on the translation training is unfeasible at large scale as it would require noise-free data. However, we still can estimate the proportion of errors replicated by the translation as the gap between OA_{ag} and OA_{gt} for translation between France-wide maps (-2% to +7% taking into account a $\pm 2\%$ uncertainty on OA_{gt}). We argue that developing a method that could reduce this overfitting to target errors could increase the results by approximately up to 7%.

This section focuses on how to tackle target label noise which we believe more impactful than source label noise. Source errors are already partly compensated by learnt translation as i) they define translation on real data, *i.e.* they also learn to translate erroneous source pixels into the expected target classes ii) the approach relies on spatial context analysis, *e.g.* roads misclassified as urban areas can be identified based on their linear shape and correctly translated. Moreover source errors can also partially be compensated by the use of multi-temporal inputs (previous section), or multi-modal data (next chapter).

4.5.2 Idea

As underlined in the literature review, learning with noisy target labels has been a core subject in the machine learning community for decades. However, limited work has been conducted on the specific case of semantic segmentation (per-pixel classification) and land-cover mapping field. Using the classification of [284] mentioned in Section 2.3.4, we underline that current methods used in the land-cover field are mostly based on building architecture performing sample selection [5, 179, 338] (detect noisy elements and corrects or remove them from training) or regularization [132]. We believe this is detrimental as some interesting specificities of land-cover mapping and semantic segmentation could help build a custom loss adjustment-based solution to improve the quality of the translation.

First, semantic segmentation exhibits a considerable difference from other classification problems. When performing a per-image classification, the network can only access a very constrained sample of classes per iteration. For instance, the current state of the art on the Image-Net dataset used a 128 batch size on a dataset with 120 classes. This implies that, at best, each class has a unique sample at each iteration. This is not the case in

semantic segmentation. For instance, in our case, with a batch size of 30 and a 600×600 patch size, we might have access to multiple thousands (even millions) of examples for almost all classes (except the very spatially constraint one) as each pixel corresponds to a single classification.

Secondly, when classifying standard computer vision datasets such as image-net, no information is provided on the per class noise as it was never assessed. Conversely, one specificity when working with land-cover maps is that most of them are provided with a confusion matrix to assess the quality of the product. We refer to the confusion matrix as a noise transition matrix for coherence with works on noisy label correction [237]. Interestingly we did not find previous works that integrate those noise transition matrices in the loss function in the land-cover field, while this strategy was proposed several years ago in computer vision. Those strategies apply the idea that the loss function can be corrected using the noise transition matrix as per-instance information, *i.e.* if one element is T_i in the target map but is 80% of the time truly T_i and 20% T_j, then the network should predict this element as 80% T_i and T_j 20%.

Those strategies appear widely sub-optimal as we have access to numerous examples of each class. Indeed the knowledge that T_i is 80% of the time truly T_i and 20% T_j does not truly reflect per-element information. In reality T_i elements should either be T_i at 100% in 80% of cases and T_j at 100% in 20% of cases. Instead of performing a per element loss correction, the loss correction should use the distribution of target labels. For instance, if there are 10^5 pixels of class T_i in the target map, then the network should predict 8×10^4 T_i and 2×10^4 T_j each with a 100% confidence. To our knowledge, this strategy, which involves knowing the noise transition matrix and having access to large enough class distribution at each iteration, has never been explored.

4.5.3 Experimental Protocol

As we evaluate the quality of the translation on our ground truth the source map usable in this experiment only includes CGLS, OSO and CLC. Experiments are conducted on the OSO to CGLS translation as translating into CGLS is the most complicated from a noise point of view (28% noise) and is not ill-defined due to a higher target spatial resolution (as the CLC to CGLS translation).

Implementing a distribution-based loss correction error The proposed solution is inspired by [202] consist in matching distribution for label super-resolution. We process in in three steps.

First, we need to evaluate the class distribution from the network output in a differentiable way. We first apply a softmax layer to obtain per-class confidence between 0 and 1, with all the per-class confidence summing to one for each pixel. Let $p_j^i(X = k)$ be the predicted confidence value for the i^{th} pixel to be of class k while being annotated as class j in the target, then $O_j(k)$ defined in Equation 4.8 approximates the number of pixels labeled as j

in the target and k by the network:

$$O_j(k) = \sum_{i=0}^{n_j} p_j^i(X = k) \quad \text{with } n_j \text{ the total number of pixels of class } j \text{ in the target.} \tag{4.8}$$

As the number of pixels labeled j might vary from one iteration to the other, we normalize $O_j(k)$ by the n number of pixels labeled j, which gives the probability that a pixel labeled j in the target is labeled k by the network (Equation 4.9):

$$P_j(k) = \frac{O_j(k)}{n}. \tag{4.9}$$

Secondly, we need to have access to the probability that an element labeled j in the target is, in truth k. This is given by the noise transition matrix available with most land-cover maps. We denote this value $N_j(k)$. In our case, we can either evaluate this noise transition matrix directly using the manually built ground truth or the official noise transition matrix provided with the dataset. Using the same ground truth for obtaining $N_j(k)$ than the one used for evaluating the results might overestimate the proposed method's potential by informing the network of the ground truth distribution used at test time. We always use the official confusion matrix unless explicitly mentioned otherwise.

Thirdly, we evaluate the difference between the previously defined discrete probability functions P_j and N_j. We compute Kullback-Leibler divergence between the predicted class distribution P_j and the expected one N_j for all pixels denoted j in the target as defined in Equation 4.10:

$$KL_j(P_j||N_j) = \sum_{k=1}^{c} N_j(k) \log\left(\frac{N_j(k)}{P_j(k)}\right) \quad \text{with c the total number of classes.} \tag{4.10}$$

This loss could theoretically be computed for each target class, outputting a divergence for each class distribution and averaged. However, this method would give the same weight to all class distributions, which would be widely detrimental to the OA_{gt} by giving equal weight to rare and frequent classes. We propose to proceed to a weighted by class proportion mean of those Kullback-Leibler divergence measurements giving the Noise Divergence loss defined as:

$$L_{ND} = \sum_{j=1}^{c} \frac{KL_j(P_j||N_j)n_j}{\sum_{l=1}^{c} n_l}. \tag{4.11}$$

Our first experiment consist in using the Noise Divergence loss to train a network and compare the obtained OA_{gt} and $mF1_{gt}$ with the one obtained with the Cross-Entropy.

Studying the importance of the noise quality matrix As this method relies on a noise matrix obtained from the land-cover map provider, the quality of this transition

matrix is uncertain. More specifically, as most confusion matrices are computed on size-limited ground truth, per-class confusion might be widely inaccurate.

Moreover, the confusion matrix provided by the land-cover map is not necessarily computed to the same spatial extent as the one on which the translation is performed. For instance, the CGLS official confusion matrix for level 2 is only available at global scale. As error patterns can vary depending on the regions, it might be significantly different from one computed on France only. Thus it appears crucial to test the method's robustness to the quality of the noise confusion matrix.

We designed two experiments. The first compares the OA_{gt} and $mF1_{gt}$ obtained when the L_{ND} use the official noise transition matrix or the ground truth noise transition matrix. L_{ND} using the ground truth should provide an upper bound of possible results as it partially breaks the independence between training and testing (should only be used as a comparison tool). The results obtained by the L_{ND} using the official CGLS noise transition matrix, which appears significantly different from the one obtained on ground truth (Figure 4.22), provide a first insight into the method robustness to errors in the noise matrix.

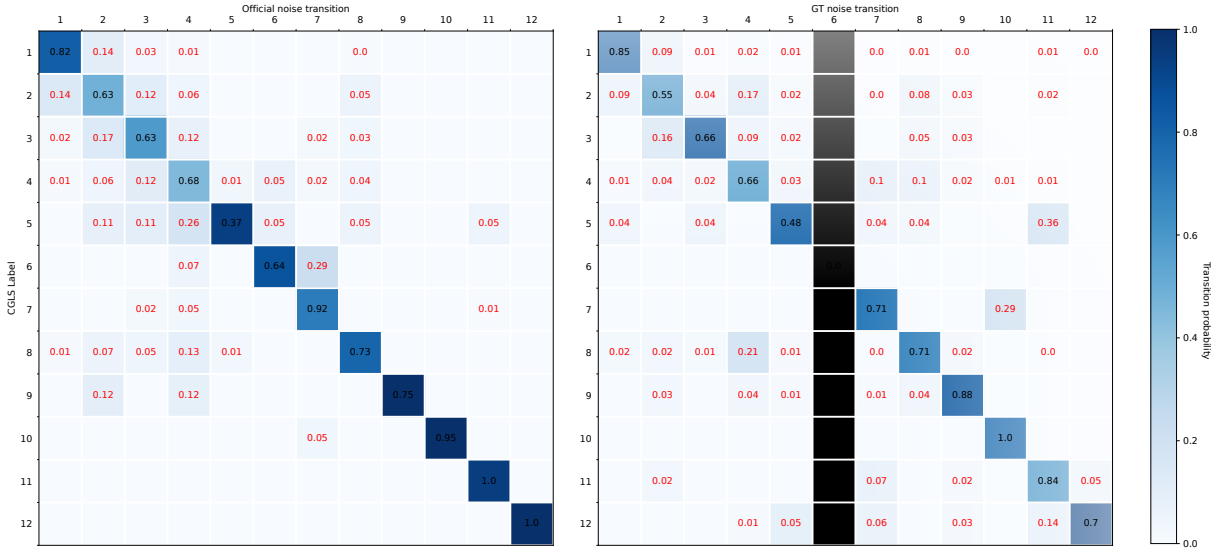

Figure 4.22: Probability for each CGLS class (in row) to be confused with one another (in column) according to the official confusion computed worldwide (left) or on France-wide ground truth (right). The Black line indicates that there is no "Moss and Lichens" on the France-wide extent.

A second experiment introduces various noise levels in the official noise transition matrix. Noise in land-cover maps is directional, *i.e.* all errors are not equiprobable. To preserve the erroneous class distribution, we redistribute the additional noise in the same proportion as the one observed in the original matrix, *i.e.* if we increase the CGLS official noise transition matrix by a 5% ratio, CGLS 1:"Closed Forest" will have $N'_1(1) = 0.95N_1(1) = 0.95 * 0.82$. Let c_{CGLS} denotes the number of CGLS classes (12), the remaining transitions are computed

according to Equation 4.12:

$$N_1'(k) = N_1(k) + (N_1(1) - N_1'(1)) \frac{N_1(k)}{\sum_{i=2}^{c_{CGLS}} N_1(i)}. \tag{4.12}$$

Comparison with other methods We propose comparing four loss-based strategies to improve results when training on noisy labels.

The Mean Absolute Error (**MAE**) has been mathematically demonstrated to be a noise-robust loss function [99]. As its implementation is easy and not computationally intensive, most current loss-based noise correction compares to it, which makes it an interesting baseline to compare to other works.

$$L_{MAE} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} |y^i(k) - (p^i(k))|. \tag{4.13}$$

The Symmetric cross-entropy [326] (**SCE**) has also been proposed to tackle label noise. Unlike cross-entropy, reverse cross-entropy is a noise-robust loss function. The authors propose a linear combination of cross-entropy (good convergence) and reverse cross-entropy (noise robust) to achieve better results than cross-entropy alone. Using the same notation as those used for Equation 4.3 we obtain Equation 4.14.

$$L_{SCE} = \alpha L_{CE} + \beta L_{RCE} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} \alpha y^i(k) log(p^i(k)) + \beta p^i(k) log(y^i(k)). \tag{4.14}$$

We obtain the best results for this technique using $\alpha = 0.1$ and $\beta = 1$. Solely those results are presented.

The simplest loss correction approach to take into account the per-pixel classification uncertainty is to consider a uniform class noise. This strategy refered as the Label Smoothing regularisation [242] (**LSR**), uses the cross-entropy loss (Equation 4.3) but replaces the one-hot encoded label by a soft encoding version *i.e.* instead of having $y^i(k)$ if the i$^{\text{th}}$ element true class is k and 0 otherwise, $y^i(k)$ follow Equation 4.15 with $smooth < 0.5$.

$$p^i(k) = \left\{ \begin{array}{ll} 1 - smooth, & \text{if i}^{\text{th}} \text{ true class is k} \\ \frac{smooth}{c-1}, & \text{else} \end{array} \right\}. \tag{4.15}$$

The last experimented strategy relies on the noise transition matrix to compensate for errors at a pixel level. Known as the Forward Correction [237] (**FC**), this strategy multiplies the network prediction with the ground truth noise matrix during the training phase to obtain a corrected prediction (see Equation 4.16).

$$L_{FC} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y^i(k) log(p^i(k) N_{j_i}(k)) \quad \text{with } j_i \text{ target class of i.} \tag{4.16}$$

4.5.4 Results

Table 4.12 presents the effects of the different loss both in terms of agreement and on the GT. In terms of agreement with the target, the cross-entropy is, as expected, the best-performing function. Interestingly the noise correction loss tends to give comparable results with the CE regarding the result/target agreement except for the noise transition matrix aware losses (FC and ND).Our solution is by far the worst in terms of resemblance to the target with -8% in OA_{ag} and -6% in OA_{mF1}. Conversely, compared to the ground truth, the best method for OA_{gt} is the SCE and the best in $mF1_{gt}$ is our ND loss. Even though the results obtained are stable across multiple independent trainings ($mF1_{gt}$ standard deviation is around 2% for each method), we underline that those results are not statistically significant due to the limited ground truth size (2300 points for OA_{gt} and 2700 $mF1_{gt}$). The results presented here should be taken cautiously, as only global tendencies can be observed. Comparing the results of ND_{off} and ND_{gt} is also interesting as despite the vast difference between the official and the ground truth confusion matrix, the results are not significantly different. This demonstrates partial robustness to an imperfect confusion matrix as the observed drop down in $mF1_{gt}$ is only 2%.

	CE	MAE	SCE	LSR	FC	ND_{off}	ND_{gt}
OA_{ag}	**77**	76	**77**	**77**	73	69	68
OA_{gt}	72	72	**73**	72	69	72	72
$mF1_{ag}$	**61**	38	60	60	39	55	48
$mF1_{gt}$	53	36	55	56	39	**60**	62

Table 4.12: Comparison of various noise correction methods.

Figure 4.23 compares the per class F1-score of the different method. We observed a global tendency to achieve higher results using the **ND loss** than any other baselines in seven (out of eleven) of the reported classes. The largest improvement is observed on the 60: "Bare / Sparse vegetation", reaching nearly a 40% F1-score while being close to 0% for all other methods. However, we can not conclude further due to sample size limitation.

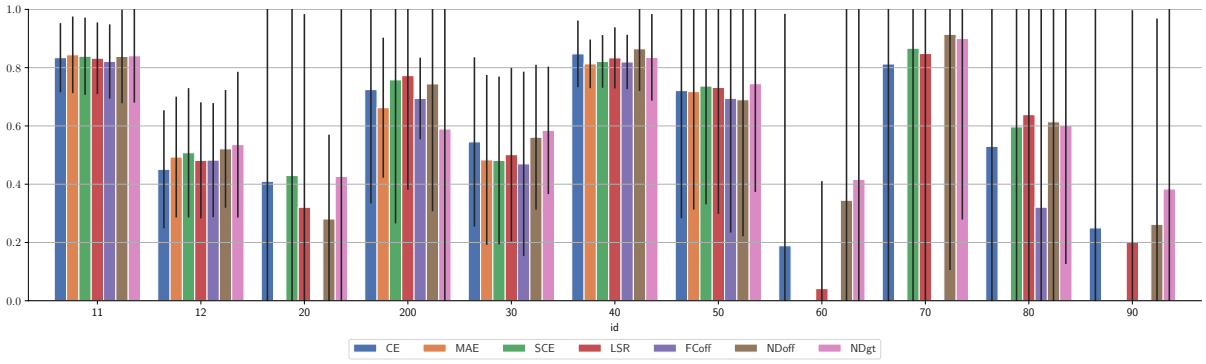

Figure 4.23: Per-class $F1_{gt}$ for the OSO to CGLS translation. Error bars are estimated using the Equation F.14 presented in Appendix F.

Figure 4.24 presents the evolution of the $mF1_{gt}$ depending on the noise level inserted inside the official confusion matrix. Increasing the noise matrix with a 15% noise ratio, the ND_{off} loss remains better than the traditional CE loss. This demonstrates the usability of our method even with imperfect approximations of the noise matrix.

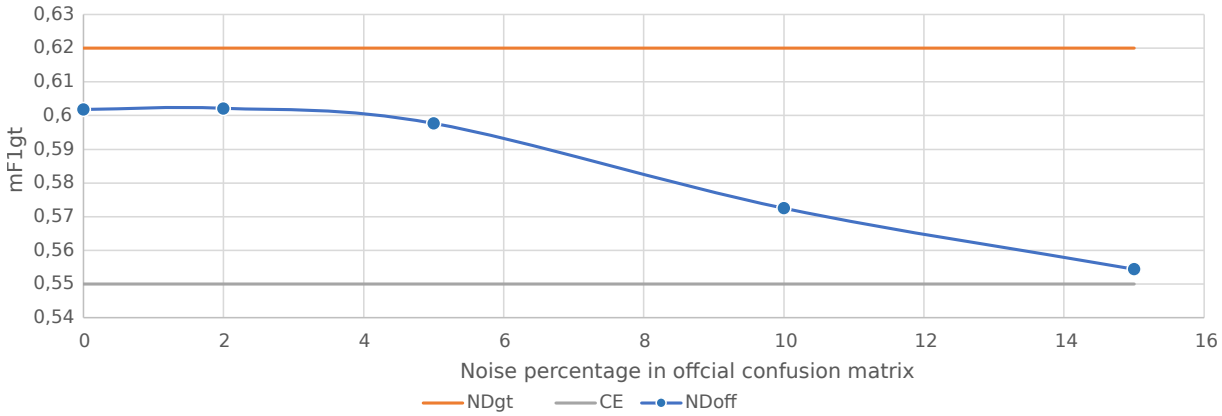

Figure 4.24: $mF1_{gt}$ obtained by training the network with the ND_{off} with an increasing directional noise ratio. We provide the mF1 score for the CE and ND_{gt} as respective expected lower and upper bounds of the method.

4.5.5 Conclusion

We introduced a Noise Divergence loss built on the idea that land-cover translation is a specific problem in which we have access at each iteration to a vast amount of labels for all classes and an accurate transition matrix. We propose to compute the divergence between the per-target class distribution of network prediction and the expected distribution (the noise transition matrix). The results demonstrate a global tendency of the method to outperform other loss correction methods in terms of $mF1_{gt}$. However, due to the small ground truth size, we can not conclude on statistical significance. We underline that a common strategy would be to work on simulated label noise. Albeit results obtained from a noise label simulation are not transferable to real noise scenarios as noise simulation techniques are unable to reproduce accurately real land-cover noise, which is directional, often systemic, and location dependent (*i.e.* edges of object), class dependent and not homogeneously distributed geographically. We also observed that our method is quite resilient to errors in the noise translation matrix used for learning, making it usable in real operational cases. Lastly we point out that more experimentation should be conducted to determine the amount of noise required to make such loss correction approach better than cross entropy. Some preliminary experiment conducted on the translation from CGLS to CLC suggest that the method is not suitable for small amount of noise as under the 10% noise of CLC cross entropy seems to give better results than our noise corrected loss.

4.6 Conclusion

The translation of land-cover maps is traditionally performed at the scale of the map nomenclature by associating each source class and target class. Starting from the observation that two elements of the same source class might have highly different translations according to their spatial, geographical and temporal context, we proposed to replace the nomenclature-levelled translation with methods performing translation at the scale of the pixel. Pixel-scale translation should allow two source pixels of the same class to have a different translation in the target nomenclature. To assess the quality of the proposed methods, we first proposed six baselines inspired by the current literature on map translation. The first two **HardSem** and **HardStat** are the most commonly found in the literature and associate a single target class with each source class. The other four methods are based on the idea that when the target map has a lower resolution than the source, it is possible to carry out a combined translation of several source pixels, slightly improving the translation quality. The key takeaway of this section is that statistical translation generally achieves better results than semantics. Indeed, when a source class can be translated in several ways, the semantic translation does not guarantee that the translation carried out is the most probable. This effect is all the stronger as the translation is carried out in an area of limited spatial extent. Thus, the apparent semantic translation of OSO's *Broad-leaved forest* class in CLC nomenclature is *Broad-leaved forest*. However, if the translation is performed on the Paris area, the statistically most probable translation is *Green urban areas*. This observation is crucial because current methods such as LCCS are based on the idea that the translation from one class to another must rely on proximity between semantic attributes of source and target classes rather than an on a probability of association. It leads to a considerable deterioration in the quality of map translations. The second lesson of these experiments was to emphasise that statistical methods favouring the global resemblance to the target map tend to provide a lower diversity of classes than semantic methods. This underlines the importance of considering solutions to preserve this diversity as much as possible.

Secondly, we explored the potential of spatial context for translation. We first hypothesise that knowing the shape of the object to which the pixel to be translated belongs increases the quality of the results of a translation significantly. To confirm this hypothesis, we started by estimating the shape of the objects using shape indicators commonly used in the GIS community but rarely used on land-cover maps. As the nomenclature does not include in the definitions information on the expected object shape (*e.g.* it does not indicate that a river is a thin elongated shape) we proposed to learn directly on the dataset how to use these indicators to perform a translation. Results demonstrated that the shape information allows a significant increase in the diversity of the predicted classes when the target resolution is equivalent to or lower than the source. Starting from the observation that these shape indicators are probably not perfectly tailored to the translation, we introduce a method capable of learning by itself to exploit the shape of objects and include

the influence of neighbouring pixel classes. This method, called A-Unet, is based on adapting a convolution network to enable it to translate the resolution and nomenclature simultaneously. The analysis of the results highlights several important points. First, the joint use of the shape and the neighbourhood significantly improves the quality of the translation both visually and qualitatively: +5% OA_{ag} and +8% $mF1_{ag}$ on average compared to the best baselines. However, we point out several limitations of the method. It is generally less good at preserving translation geometry than baselines and does not generalise well spatially. Furthermore, we observe that the map with the smallest spatial extent (MOS) does not benefit significantly from using context. We explain this mainly by the small size of the training set used. Observing that a change of architecture does not improve the results, we postulate that efforts should focus on adapting the training procedure.

The following section focuses on the geographical context. We start with the hypothesis that using geographic coordinates is a simple way to learn this context based on the observation that two close coordinates share the same geographical context. Experiments show that this method is more likely to improve the translation than considering a predefined ecoclimatic zone. Based on this observation, we propose incorporating the geographical coordinates into the network. Making the same observation as other works, we show that integrating the coordinates directly into a network does not work. We propose an approach based on positional encoding, a strategy that has proven itself in NLP. The results demonstrate the method's effectiveness, even when the size of the training set does not allow fine learning of the geographical context. In particular, we demonstrate that the proposed approach is always superior to coordinate-free translation.

The fourth section focuses on the temporal context of translation from two different angles. We first question the relevance of using multiple dates from a source map to improve the translation quality. We start from the observation that multiple classes include a notion of temporality, *i.e.* it is class A if class A has at least been observed once in the last three years. We hypothesise that the temporal order of the source maps does not matter because the definitions never precisely stipulate a date but always a period. We show that the incorporation of this temporality allows a significant increase in the quality of the translation, mainly for the target classes, including a notion of temporality, but also, in a lesser way, for others. We explain this by the fact that combining several source maps allows the method to distinguish errors in the source data better improving the translation quality. Secondly, we emphasise the existence of a temporal context specific to the "learnt" translation paradigm: the temporal spacing between the source and target data used to learn. We hypothesise that the temporal spacing between the source and target maps adds only little noise to the learning compared to the noise generated by the errors in the original maps, as long as the learning is carried out on a large area and that the time spacing remains reasonable. We propose three distinct learning scenarios corresponding to different operational situations. Results show that the temporal spacing only very slightly deteriorates the results. In particular, there is no significant difference between training a

network to translate two maps with the same date (scenario 1) and two maps with a time gap (scenario 3). In both cases, the resulting map gives a translation corresponding to the date of the source map. This critical finding guarantees the operational usability of learnt translation methods.

Lastly, we focused on tackling the cartographic context of translation, which involves working on, by nature, noisy data. We first underlined that creating a source noise resilient translation method is a complicated task as it involves evaluating instance-specific noise probabilities. We argued that the detrimental impact of source noise on translation was limited as errors tend to exchange two semantically close classes that tend to have the same translation. Moreover, as our spatial and temporal context-aware methods can already partially evaluate those source noise characteristics, we believe that efforts must concentrate on tackling the target noise effect. We presented a method for tackling target noise based on the idea that the land-cover translation exhibits specific characteristics: a large amount of annotation at each iteration and an available noise transition matrix. Thus we propose to compute the divergence between the expected per class distribution (noise transition matrix) and the observed one. Despite being unable to assess statistical significance due to ground truth size, our method seems to outperform commonly used loss-based correction strategies in terms of output class diversity ($+7\%$ $mF1_{gt}$ compared to cross-entropy). Furthermore, the methods appear to be resilient to a significant noise level in the noise transition matrix used for training, making it usable in real operational scenarios. We believe additional experiments should be conducted on this loss function using a broader ground truth and more datasets to comfort the results.

As the experiments were carried out on a wide variety of translations (up to 26 different), we are confident in the ability to generalise these results to other land-cover maps than those in the data set. By focusing only on these 26 translations, we can identify some limitations of the proposed methods. In particular, the use of context to translate CLC to any other map generally offers only slight improvement, which is explained by the fact that the difference in resolution is considerable with the other maps (up to a factor of 2500) due to CLC's MMU. Therefore, it seems essential to propose data insertion methods that are more spatially resolved to improve the translation quality. This would also improve the preservation of target geometry which, as evidenced by the value of EPI, is poorly preserved by the methods presented. This would also significantly increase the values of OA_{gt} and $mF1_{gt}$, which remain, with the notable exception of the OSO to CGLS translation, lower than that of the original products. Lastly, the methods suffer from a significant lack of generalisation ability which should be compensated for applications such as extending a land-cover map on a broader spatial extent using a source map with a broad spatial extent.

CHAPTER 5

Multi-land-cover, Multi-modal translation

5.1 Multi-land-cover map translation

5.1.1 Motivation

In the previous chapter, we introduced a Convolutional Neural Network based encoding-decoding strategy to translate land-cover maps using spatial context. The core idea lies in the possibility for each map pixel to be translated differently depending on its close surrounding pixels. However, this supervised method is designed to perform mono-land-cover map translation, *i.e.* a single source is translated in a single target. It requires the two maps to overlap, at least partially, spatially. When impossible, a pivotal map that spatially overlaps the two others might be used, but this is likely to drastically lessen the translation performance by requiring two translations instead of a single one. Multiple translations require separate training phases. With few training samples, the performances are likely to be limited.

Recently, deep learning methods have achieved state-of-the-art results in natural language processing and, more precisely, in language translation [59, 307]. State-of-the-art methods have shown the superiority of multi-lingual trained models against their mono-lingual counterparts [59], especially for languages with a small number of translation examples. Multi-language training seems to benefit from the obtained multi-language common representation space [245]. Finding shared representations is also frequently addressed by the remote sensing community for combining multi-modal data from various sensors, with varying resolutions and information into a compact and discriminative embedding [11, 127, 128, 223]. Surprisingly, this question remains unaddressed for the land-cover map translation task at a pixel or object level. Therefore, this section tries to answer the following question: can we find a shared space for multi-land-cover map translation that would be beneficial for their individual generation?

Even though a current trend in computer vision is to address the projection into common representation space with vision transformers, we adopted a Convolutional Neural Network-based solution that requires far less training and is thus suitable even for land-cover with

small spatial extent (Figure 5.1). Independent U-Nets are trained to project each land-cover map into a shared representation space. This, tailored for translation representation space, enables both translation and the self-reconstruction of the input land-cover maps to ensure that no information is lost during the mapping to the shared space.

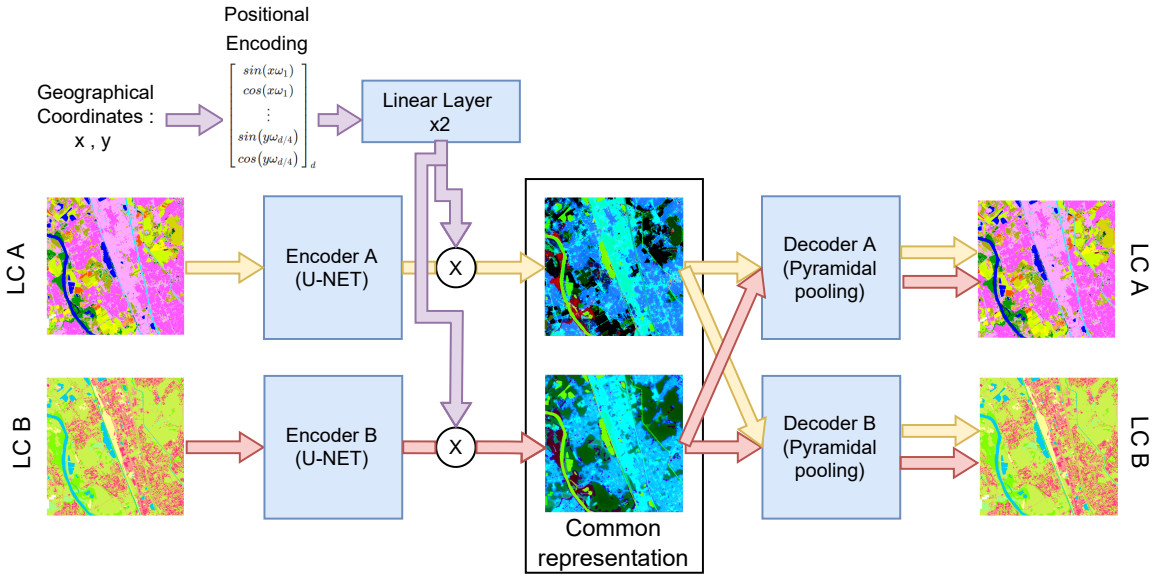

Figure 5.1: Overall multi-land-cover map translation architecture. Our network (blue boxes) is trained to perform both self-reconstruction and translation. There is no restriction in the number of maps that can be embedded into the common representation. For convenience, we only represent two maps (A and B). Red and orange arrows represent the possible paths for maps A and B, *e.g.* A can be either translated into B or self reconstructed into A. At inference, only one source map is required.

The key contributions of this section are summarised as follows:

- We propose a method to perform multiple translations at the same time using a single translation model.
- We ensure that the method projects all land-cover maps into a shared representation space in which two different maps of the same extent are represented closely. This ensures that the shared representation space encodes elements with comparable characteristics in terms of semantics and context independently from the land-cover map used, increasing spatial generalisation for maps with limited spatial extents.
- We conduct a comparative evaluation of the approach with the mono-land-cover map translation presented in the previous chapter.

5.1.2 Method

5.1.2.1 Training protocol

We aim to find a simultaneous transformation of the nomenclature and spatial resolution of the six maps (see Section 3.2). Inspired by the literature review conducted in Section 2.4, we enforce the translation to use a intermediate common representation space for all maps. This representation in a common space is referred as an "*embedding*". This leads to reach two consecutive objectives: 1) project each map into a shared embedding space; 2) decode this embedding into each of the six maps.

Inspired by recent works on multi-modal data representation [40, 134, 152, 337, 352], we propose to train separate encoders and decoders for each map, and subsequently use cross-reconstruction to enforce common representations of land-cover maps representing the same spatial extent (see Figure 5.2). We train our network to both reconstruct a given land-cover map with one decoder and to translate into the desired target land-cover map with another decoder. This dual objective enforces the embedding to be rich enough to preserve all source map information (reconstruction) while encoding it suitably for translation.

Even though cross-reconstruction encourages the learnt embedding to be comparable for all land-cover map, it does not guarantee it. Therefore, many of the previously cited works also included a constraint on embedding pairs of corresponding data (*e.g.*, using adversarial training or a loss term for embedding comparison). We adopt the latter strategy which avoids the convergence complexity of adversarial training. A Mean Square Error between embeddings covering the same spatial extent is thus computed.

Instead of computing the loss for all available maps covering one spatial extent, the network is trained by computing the loss for only one pair of maps at each optimisation step. This pair-wise optimisation is used as a workaround for GPU memory limitations. Indeed, land-cover map translation requires large image patches (600×600 pixels× the number of classes) to account for the MMU of some of the maps. Also, simultaneously training multiple networks is memory consuming. This pair-wise optimisation enables the use of a larger batch size and achieves a better result than optimising all different maps simultaneously on smaller batches. This iterative pair-wise approach is also the one generally used in multi-lingual model training [59].

To sum up, at each iteration, two patches, A and B belonging to different maps but representing the same spatial extent are encoded by distinct U-Net producing Encoded A and Encoded B. A first loss term L_{emb} (detailed in Section 5.1.2.3) is computed between the two embeddings to estimate their resemblance, then each embeddings is processed by each decoder resulting in two translations (Encoded A into the decoder of B, and Encoded B into the decoder of A) and two self reconstructions (Encoded A into the decoder of A, and Encoded B into the decoder of B). Two loss terms evaluating the quality of the two translations L_{tra} and the two reconstructions L_{rec} are then computed.

5.1.2.2 Network

The design of our Multiple Land-Cover Translation Network (MLCT-Net) is made according to the following observations:

1. The encoder must have a sufficient receptive field to encode each object using its surroundings. Thus, the architecture is constrained by the MMU of each map. Since CLC has a 250,000 m^2 MMU, the theoretical receptive field should at least have a 250,000 m^2 width. An embedding with a ground resolution of 10 m leads to at least a 250 pixels wide receptive field. Achieving this size of receptive field using only convolution is unfeasible as it would require very large networks and would be inefficient due to gradient vanishing problem. Pooling strategies appears well suited to increase the receptieve field as land-cover often offers wide homogeneous regions limiting the information loss.

2. The decoder should remain as simple as possible to ensure that the learnt embedding remains as identical as possible for all land-cover maps. Decoders with high capacity may lead to a latent space with small information content [47].

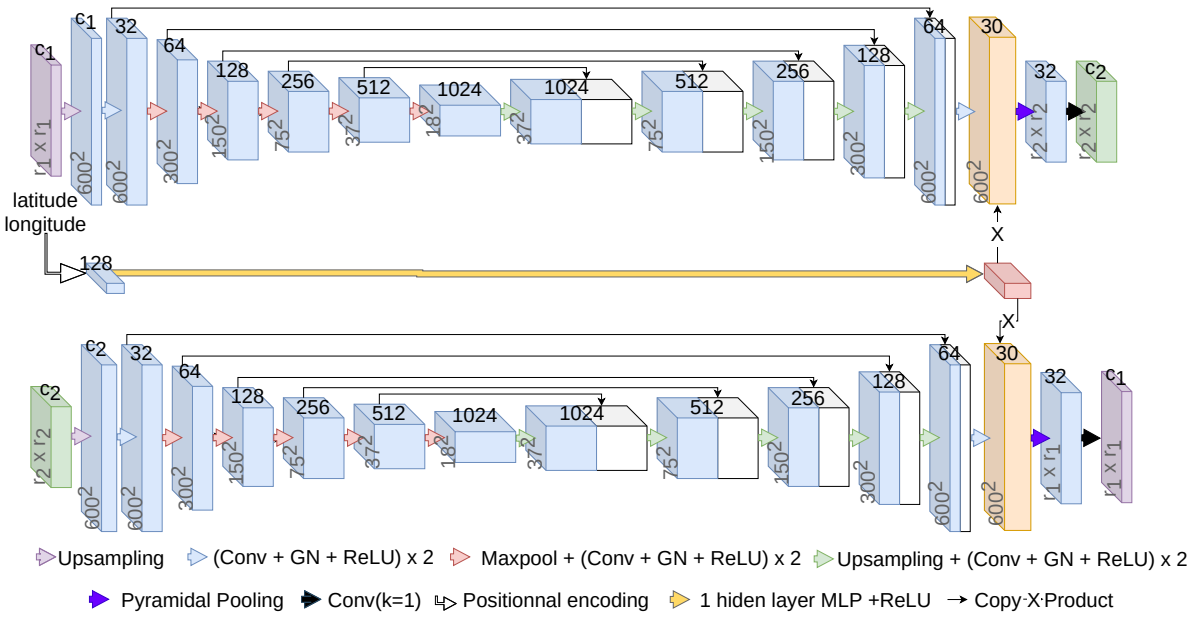

Figure 5.2: The proposed cross-encoder architecture. In purple and green, two land-cover maps with respectively c_1 and c_2 classes and $r_1 \times r_1$ and $r_2 \times r_2$ pixels. We represent in orange the common embedding space.

We develop the architecture illustrated in Figure 5.2. It is mainly composed, for each map, of a (1) a nearest neighbour resampling to the highest spatial resolution (10 m), (2) a U-NET [257] encoder, (3) and a pyramidal spatial pooling [44] followed by a 1-pixel wide kernel convolution layer as a decoder. This architecture meets each of the above criteria.

The resampling enables using the same architecture for each map. However, this strategy is only feasible if the gap between the lowest and the highest resolutions remains reasonable compared to the GPU memory. A low resolution enforces patches to cover a wide area to get a grasp of the spatial context. This results in very large patches for the maps with higher resolutions. The U-Net addresses the receptive field size by down-sampling the input multiple times, which is more memory efficient than increasing the network depth. Our encoder architecture exhibits only two differences with the original U-Net architecture. The first one is the use of Group Normalisation [332] instead of Batch Normalisation. This enables a stable normalisation, even on a small batch size. The second one is the use of 5 down-sampling blocks, instead of 4 in the original paper, to widen the receptive field.

5.1.2.3 Losses

At each optimiser step, the loss is computed for one pair of maps using Equation 5.1:

$$L = L_{rec} + L_{tra} + L_{emb}. \tag{5.1}$$

Let $E_A(A)$ denote the the result of encoding A with its dedicated encoder and $D_A(E_A(A))$ denote the processing of A by its dedicated encoder-decoder (self reconstruction). $D_A(E_B(B))$ denotes the encoding of B is processed through the decoder A (translates B into A). We denote $E_A^i(A)$ the i$^{\text{th}}$ dimension out of w dimensions of the encoded version of A.

$L_{rec} = L_{CE}(D_A(E_A(A)), A) + L_{CE}(D_B(E_B(B)), B)$ is the reconstruction loss enforcing the embedding to preserve map-specific information: computed as the sum of cross-entropies (L_{CE} see Equation 4.3) between the two self-reconstructed and their respective sources.

$L_{tra} = L_{CE}(D_A(E_B(B)), A) + L_{CE}(D_B(E_A(A)), B)$ evaluates translation quality: computed as the sum of the two cross-entropies of the two translated maps and their respective targets.

$L_{emb} = \frac{1}{n}\sum_{i=1}^{w}(E_A^i(A) - E_B^i(B))^2$ is the Mean Square Error (MSE) loss between the embedding of the two source maps which enforces the representation to be shared between maps. The global loss is theoretically minimal when the three following assumptions are simultaneously met: 1) the self-reconstruction of each map is perfect; 2) the translation is also perfect; 3) embeddings on the same areas are identical.

The Cross entropy combined with Dice loss explored in the previous chapter can be used here instead of traditional cross entropy to improve mF1. However it complexifies the loss significantly by adding four more terms making the optimisation more challenging. As the optimisation take more times, using such loss would have reduced the number of experiment presented in this section with this combined loss all the experiments presented in this section are carried using solely the cross entropy.

5.1.2.4 Geographical context

The same geographical coordinate encoding strategy as in the previous chapter is used. However, we add a softmax transformation after the MLP and replace the addition by a multiplication between the geographical context and the embedding of each map (Figure 5.2). Softmax followed by a multiplication aims to maintain the generalisation ability to areas unseen during training.

Since each translation does not necessarily need the same geographical context information, one could learn one geographical context per pairwise translation. However, it would be impossible to generalise the translation to an area of the target map unseen during training. For example, learning a specific geographical context for the OSO-to-MOS translation is only possible on the Paris region. To preserve the common representation space, we train a unique MLP on the set of coordinates of the patches. This unique geographical context representation slightly worsens the translation quality compared to learning a per translation representation. However, it is the only way to maintain the ability to translate maps on areas not included in their original spatial extent.

5.1.2.5 Comparisons

Since, to the best of our knowledge, no other multi-land-cover map translation method has been published, we compare our approach to the methods introduced in Chapter 4: the baselines and the 26 A-UNet trained independently. We refer to those multiple A-UNet as the mono-land-cover map translation method. Results with the multi-map approach are expected to be better than with the two first non-contextual translation methods. They should be at least on par with the mono-map translation method, and better on land-cover maps with few training patches, as observed in natural language processing.

5.1.3 Results

5.1.3.1 Qualitative assessment

Figure 5.3 presents the 12 translation results obtained on a patch belonging to the Paris region in Figure 3.4. Each row corresponds to the translation of one source map into the four others available for this area. Unsurprisingly, translations from coarse to high resolution maps results in almost similar performance than associating one unique target class to all the pixels of given source. External data (such as satellite imagery) could participate in increasing the translation performances. The second observation is that our network may face some difficulties in learning the MMU of CLC (25 pixels) as shown by the small three pixel wide urban areas (in red) in Figure 5.3 (first column). Commonly, network training leads to replicate in the predictions the bias observed in the original data. The most striking example is OSO *road* class, which has a 45% recall in the original data. It is often confused with *Industrial and commercial units (ICU)*. When learning to translate

a road from a given land-cover map source map to the OSO map, the corresponding class has a high probability of being an OSO *ICU*, as illustrated with the MOS-OSO translation case in Figure 5.3 (3^{th} row, 2^{nd} column). This also increases the difficulty in quantitatively assessing the quality of the results using the target data as reference.

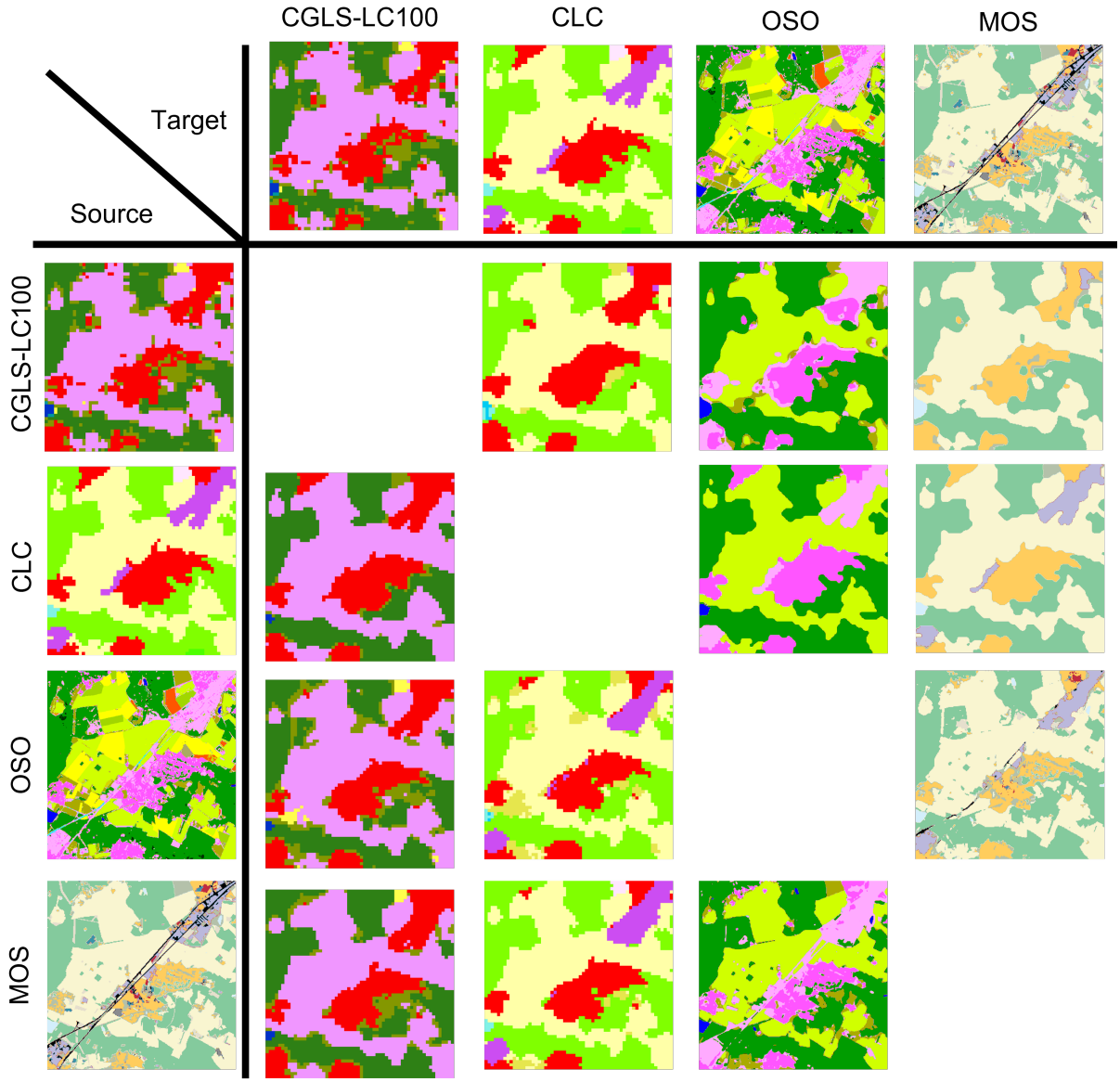

Figure 5.3: Translation scenarios for all land-cover map available on a 6×6 km^2 patch belonging to the Paris region.

Figure 5.5 presents a set of patches selected for their representativeness of the behaviour of the multi-LC. The first observation is that the spatial context influences the translation mainly on object edges, especially when the source exhibits a low resolution, *e.g.* in the first row, the border of a CLC *Discontinuous urban* area is translated into an OSO *pasture* area. The second observation is that when the gap between spatial resolutions remains

limited, the translation provides a successful context-dependent translation (*i.e.*, the same class is translated differently depending on its neighbourhood), as shown for example in Figure 5.5 (second row): OSO *sparse urban* and *Industrial and Commercial Units* are satisfactorily translated into either MOS *Individual housing*, *Collective housing* or *Activity areas*, depending of each source class density or on the third row, where MOS *Forest* is translated into CGLS-LC100 *Open forest* or *Closed forest*, depending on the elongated shape of the object. Thirdly, despite context, some cases remain difficult without external data. This could for example be used in the fifth row where an OCS-GEc *Water* area must be translated into the land-use counterpart. Most of the time, such areas are classified as *No-use* in OSC-GEu. However this water lake is used for farming which the network fails to predict. This difficult case illustrates well the main limitation of this method.

Figure 5.6 presents the qualitative comparison of Statistic, Semantic, Mono-LC, and MLCT-Net on the same spatial extents. A first observation is that mono-land-cover map and MLCT-Net gives visually almost identical results. The two methods outperform the semantic and statistic baselines when source classes have multiple probable translations. For instance, for OCS-GEu (G2) to CGLS translation, "*Agriculture areas*" are translated solely into "*croplands*" by the semantic method while being translated quite accurately both into *cropland* and *pastures* by the context-aware methods. The same observation holds for urban areas in the OCS-GEu to OCS-GEc translation (and OCS-GEc to OCS-GEu). A second observation is that pure semantic based translation outperforms other methods on erroneous classes in the original target data. For MOS to OSO translation, roads (black on the MOS map) are always translated into industrial and commercial units except by the semantic baseline. This behaviour is learnt from the original OSO map, which often presents this confusion. Conversely, the learnt methods outperform the semantic baseline when the source map is erroneous. In the reverse translation case, the erroneous industrial and commercial units (truth: *roads*) are correctly translated into *roads* in the MOS maps by all methods except the semantic one.

To assess if land-cover maps are all embedded in a shared representation space, Figure 5.4 presents the embedding of a 6×6 km^2 patch of the test set for five different maps. As displaying 30 dimension embedding is unfeasible, a dimension reduction technique is used to reduce to 3 dimensions for RGB visualisation. Principal Component Analysis (PCA) is used as, unlike methods such as T-SNE, it enables to re-use the same reduction model (trained using a subset of the training set) for various patches of test set, ensuring that the representation remains stable for all test patches. The first observation is that all embeddings look similar, which was expected through the double constraint of cross-reconstruction and the MSE computation between embeddings. The second is that edges have a particular behaviour in the embedding. This is particularly visible on coarse resolution maps (such as CLC) with a gradient on each object near the edges. We link this behavior to the higher uncertainty of the translation near object boundaries. The third observation is that the learnt embedding for coarse resolution maps has a blurrier aspect than high resolution ones. This is, for example, clear on the CGLS-LC100 embedding,

especially on **Built up** areas. We relate this behaviour to the relative broad semantic content of such class in a low resolution map compared to a higher resolution one (*i.e.*, a *Built up* area might simultaneously include trees, dense or sparse urban and roads). Another observation is that close values in the embedding space for two classes often reflect close semantic values. For instance, all artificial surfaces appear in light blue, all forest types in light to strong red, all crops and pastures in dark blue. This closeness might be beneficial for tasks such as zero shot learning since semantically close elements are represented closely in the embedding space. For instance, the model is never trained to translate CGLS *Ocean* in one of the MOS class as there is no ocean in the Paris area. However translation is easy to perform as the CGLS *Ocean* representation in the embedding space is closer to MOS *Water* than any other MOS class.The last observation is that when one class of a land-cover map establishes a complex semantic relationship with another land-cover map, it is often visible in the embedding. For example, the OSC-GE cover class *Herbaceous vegetation* mixes cultivated areas and natural grasslands while all other land-cover maps make a clear distinction between those two vegetation types. This leads to distinct embeddings.

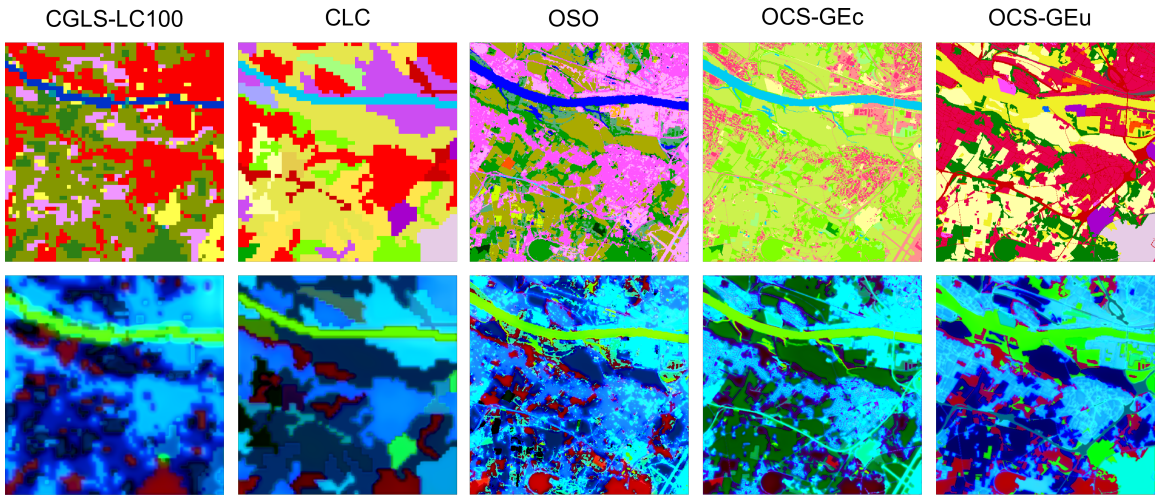

Figure 5.4: Shared embeddings (*below*) for five land-cover maps of interest (*top*). Colors result from a dimension reduction from the original 30-dimension embedding to 3 dimensions (RGB) using Principal Component Analysis.

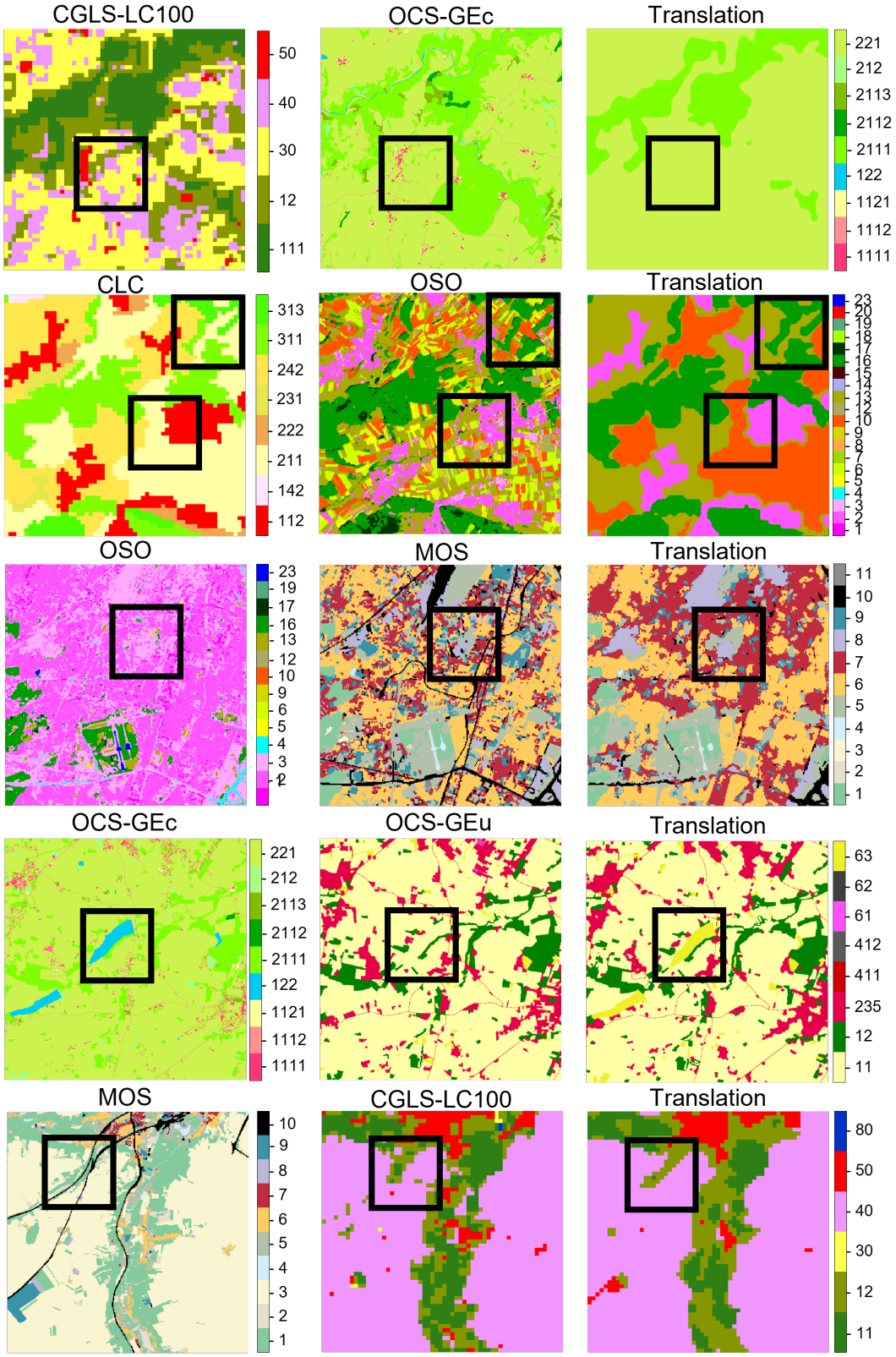

Figure 5.5: Benefits and limitations of multi-land-cover map translation. Each square highlights an area with meaningful spatial context (see text for more details).

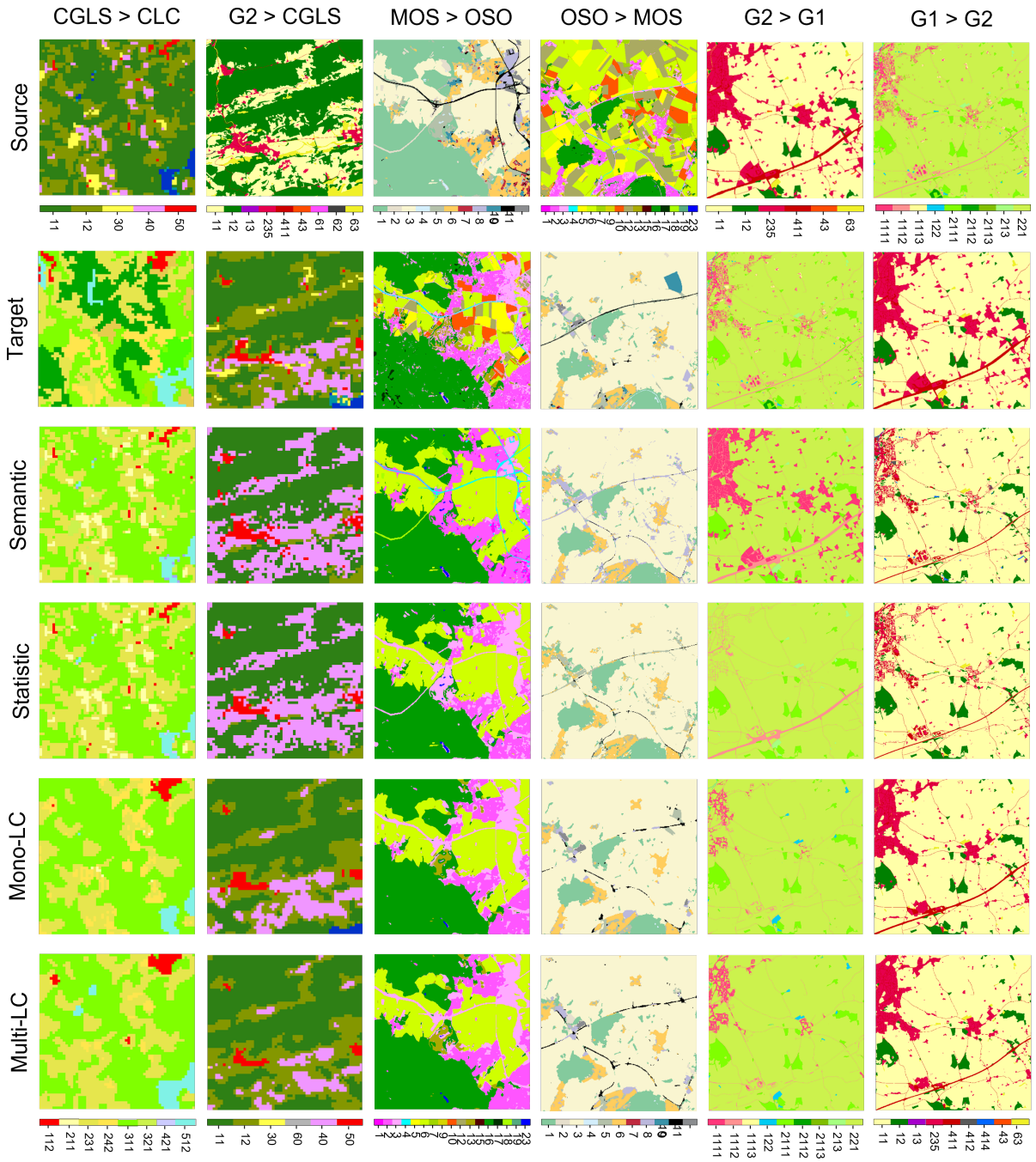

Figure 5.6: Visual comparison between the output of MLCT-Net and existing baselines.

5.1.3.2 Quantitative assessment

All conceivable translation scenarios are tested. Table 5.1 reports the agreement metrics. We remind the reader the agreement can only be computed on the intersection of the source and the target extents. Context-aware translation methods have higher agreements than their semantic and statistical counterparts. The improvement between contextual

and non-contextual methods ranges from 1% to 17%. The smallest differences are usually observed when the source map has a coarser spatial resolution than the target. It is impossible to obtain high scores on a spatial super-resolution task without adding fine geometric and spatial information (*e.g.*, high resolution images on the same extent). In practice, a good rule of thumb is to estimate that the MMU of the target maps is always of the same magnitude than the source one (*i.e.*, translation a 25 ha MMU land-cover map results in a more or less 25 ha MMU). Conversely, significantly better results are observed when a high resolution map is translated into a coarser one.

Source		CGLS (P)					CLC (C)					OSO (O)					OCS-GEc (G1)				OCS-GEu (G2)				MOS (M)			Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	
OA	SoftSem	52	42	56	70	**75**	65	49	67	77	79	62	59	69	76	82	56	41	34	87	57	40	31	75	80	76	59	62
	SoftLearntGridPattern	54	44	65	70	75	68	55	71	78	79	72	62	73	80	82	63	47	49	89	62	41	41	78	84	82	62	66
	mono-LC	**65**	**57**	**69**	**78**	**76**	**75**	**59**	**73**	80	**79**	**77**	**69**	**80**	**86**	85	**71**	58	**58**	**93**	69	54	53	79	**86**	**84**	**64**	72
	multi-LC	**65**	**57**	**69**	**78**	**76**	74	**59**	**73**	**81**	**79**	76	68	**80**	**86**	**86**	**71**	**60**	**58**	**93**	**70**	**57**	**54**	**81**	**86**	**84**	**64**	**73**
mF1	SoftSem	13	17	22	15	24	46	32	36	**31**	**42**	38	19	36	20	38	27	10	17	27	20	8	8	29	38	19	19	25
	SoftLearntGridPattern	13	18	19	16	24	47	32	33	30	**42**	47	24	34	20	42	37	15	20	27	28	12	10	27	43	25	18	27
	mono-LC	30	29	**30**	**20**	33	**58**	**37**	**38**	30	41	**61**	40	**45**	**27**	53	52	34	31	43	**52**	29	25	40	45	30	23	38
	multi-LC	**35**	**30**	**30**	**20**	**34**	57	**37**	37	29	41	57	**41**	44	**27**	**54**	**53**	**39**	**34**	**44**	50	**34**	**27**	**46**	**48**	**33**	**24**	**39**
EPI	SoftSem	**28**	**5**	**6**	**5**	**17**	**30**	**6**	**6**	**7**	**20**	36	36	**28**	**32**	**58**	28	**32**	29	79	32	**36**	29	**77**	54	54	**38**	**31**
	SoftLearntGridPattern	27	**5**	3	**5**	16	**30**	**6**	5	**7**	**20**	**42**	35	25	31	57	**35**	31	**30**	**80**	**35**	32	**30**	**77**	**57**	**55**	**38**	**31**
	mono-LC	22	4	3	4	12	27	**6**	5	6	17	39	36	25	28	49	31	29	24	77	29	30	21	70	50	**55**	29	28
	multi-LC	23	4	3	4	13	28	**6**	4	6	18	41	**38**	26	51	31	31	31	25	79	30	31	23	72	52	**55**	29	29

Table 5.1: Agreement between translations and target maps. P: CGLS-LC100, C: CLC, O: OSO, G1: OCS-GEc, G2: OCS-GEu, M: MOS. Best values are in bold.

mF1 differences revealts that mono and multi-land-cover map translations successfully use spatial context to outperform the simpler counterparts in terms of number of predicted classes. Wr provide the observed per-class f1-score in Figure 5.7. Since displaying all the 26 possible configurations would be counterproductive, we add the confusion matrices of all maps for each target, resulting in one confusion matrix per target map. We then compute the per-class f1-score, *e.g.* CLC per-class f1-score is computed on the fused confusion matrix of OSO-to-CLC, MOS-to-CLC, PROBA-to-CLC, OCS-GEc-to-CLC and OCS-GEu-to-CLC. In Figure 5.7, a high f1-score is reached when the translation from all sources to the considered target is successful. The well predicted classes are identical for all methods. The translation into CLC is the one for which context-wise methods are the most beneficial, as it significantly increases the number of partially predictable classes, compared to the semantic and statistical baselines. The insertion of context mainly helps specific classes, especially those defined by a spatial pattern such as CLC *Heterogeneous crops* (mix between arable and permanent crops), and on spatially correlated classes. Forests in mountainous areas mainly include coniferous stands. Thus, a forest in this area is more likely to be translated as *Coniferous* than *Broad-leaved*.

Our approach has a similar agreement to the mono-land-cover map scenario, exhibiting close scores in most cases. However, it tends to slightly under-perform on the OSO-any other configuration. This is mainly due to the fact that our MLCT-Net tends to have more difficulties in learning the MMU than the mono-land-cover map counterpart. This observation is comforted by noticing that the mean area of errors in the multi-land-cover map model is significantly smaller. This can partly be explained by the difficulty in learning the concept of MMU in a shared representation space, due to the risk of also applying

the same MMU when translating finer resolution land-cover map. One could argue that learning the MMU only requires estimating the area occupied by classes and filtering non adequate small areas. However, this would overlook that estimating areas is not a trivial task for a network fed with image patches due to the lack of information on edges (ideally, this would require processing the whole data at once, which is unfeasible). Furthermore, undetected areas in the target data act like a generalisation procedure neglecting some of the information. While this last statements affect both multi and mono-land-cover map models, the difficulty in learning the MMU naturally increases as the number of generalisation rules (and errors) increases, explaining the poorer MMU learning of the multi-land-cover map model compared to its mono-land-cover map counterpart. Since OSO is the highest resolution map used in this study, translation, from OSO are the most prone to MMU errors explaining the observed slight under-performance compared to the mono-land-cover map model.

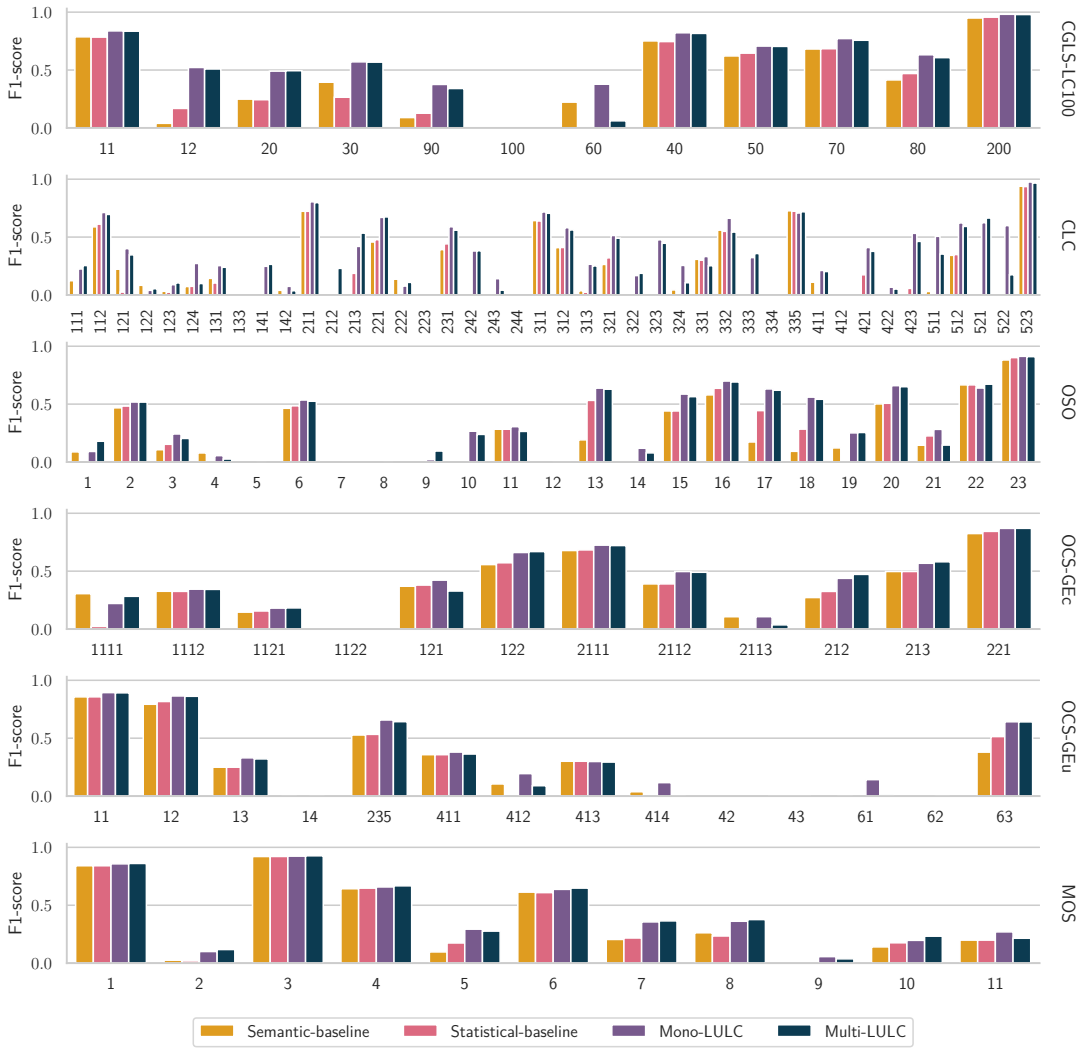

Figure 5.7: Per-class f1 agreement computed on the sum of the translation confusion matrices of all the sources to one target.

5.1.3.3 Impact of the number of input land-cover maps

We propose to analyse the influence of the number of maps fed into MLCT-Net on the quality of the translation. Figure 5.8 displays the accuracy measured on ground truth depending on the number of target maps used for learning. Each histogram represents the stacking of translation results from all three France-wide maps (CGLS, CLC, OSO) to the considered target. For instance, the first histogram presents the average translation results of CLC to CGLS and OSO to CGLS-LC100 for different models trained to perform mono-LC (1 map) or multi-LC translation using (2 to 6 maps). The three histograms OSC-GE cover, OSCGE use, and MOS are directly dependent on the network's spatial generalisation ability as they result from the average of translation performed to a broader extent than their original one. For instance, the MOS histogram is obtained by averaging France-wide translated MOS using CGLS, CLC or OSO as input, while the actual MOS only covers the Paris region. Error bars are computed as the mean of uncertainties estimated using Equation 5.2.

$$u(t) = \frac{1}{n} \sum_{s=1}^{m} z \sqrt{\frac{OA_s(1 - OA_s)}{n}}, \tag{5.2}$$

where u(t) is the uncertainty for a target map t, s is the considered source map, OA_s is the estimated accuracy of the translation from source s to map t, $z = 1.96$ for 95% confidence, and n is the ground truth sample size.

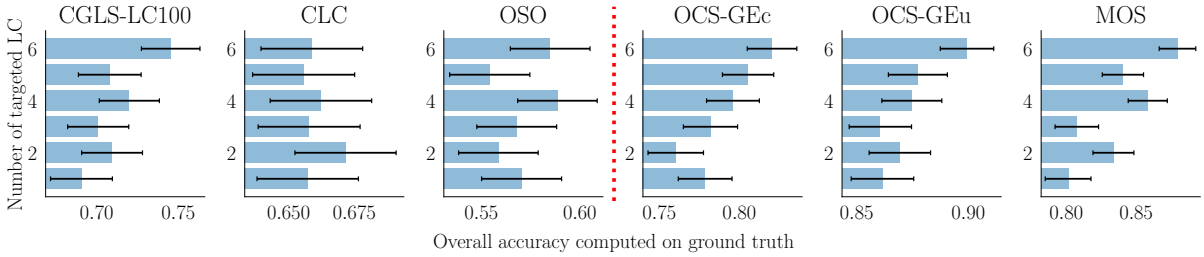

Figure 5.8: Mean accuracy per target land-cover map for different models trained with one (mono-land-cover map) up to six maps. The red-dotted line separates land-cover map available France wide (left) from those with smaller spatial extent (right).

Our first observation is that although the model trained on six maps tends to perform better in the majority of cases, the performance variations observed on CGLS-LC100, CLC, and OSO remain insignificant given the size of the ground truth sample. This statement prevents us from concluding on a real advantage of using a multi-land-cover map model for these three maps. This observation is further supported by the fact that there is no stable trend of a performance increase when going from 2 to 6 maps. Thus additional experiments with more maps are required for further analysis. On the other hand, a more significant trend is observed on the MOS, OCS-GEc and OCS-GEu maps, which all initially covered

only a fraction of the territory. The progressive performance growth with the increasing number of maps comforts the previous analysis of greater robustness to generalisation to new landscapes of multi-LC models compared to the mono-land-cover map model.

5.1.3.4 Land-cover map extension

The generalisation ability of a deep neural network is a key feature when studying the representativeness of the shared space and subsequently the "universality" of such learnt representation. A universal representation should be able to represent all land cover types independently from the spatial extent enabling to use a model trained on small spatial extent maps to broader extents. This would allow to generate only high quality land-cover maps on a restricted area without spending too much time.

We propose to evaluate the ability to retrieve the target MOS, OCS-GEc and OCS-GEu over France from the sources OSO, CLC and CGLS land-cover maps, while the 3 target land-cover maps originally covers less than 20% of the country. Each source map is translated into one of the targets at a France-wide scale. The translation may face unseen classes during training in both source and target maps (*e.g.*, there is no *glacier* on the original MOS spatial extent) resulting in wrong translations. Therefore, for each pair of source/target maps, the semantic baseline is used to translate source classes unseen during training. Unseen target classes during training are ignored. OCS-GEc *Snowfields and glaciers* and *Other non-woody formations* are considered unseen due to numerous errors in the OCS-GE data. In this setup, mono-land-cover map models cannot be trained with the geographical coordinates sub-module since they are trained solely on the original target spatial extent. To be able to assess if differences between the mono and multi-land-cover map models are due to the use of the geographical coordinates sub-module, we provide the multi-land-cover map results with and without it.

Source		P					C					O					Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	
OA_{gt}	SoftSem	47	46	62	79	81	**72**	51	76	84	85	71	65	86	86	92	72
	SoftGridPattern	52	45	68	80	81	68	57	77	85	85	73	68	86	89	92	74
	mono-LC no c	57	51	70	82	76	69	58	78	**86**	78	74	70	86	91	87	74
	mono-LC	**60**	**55**	X	X	X	70	**59**	X	X	X	75	**71**	X	X	X	X
	multi-LC no c	57	52	71	**83**	82	71	**59**	78	**86**	**86**	**78**	70	**87**	91	**93**	76
	multi-LC	**60**	53	**74**	**83**	**83**	71	**59**	**79**	**86**	**86**	**78**	70	**87**	**92**	**93**	**77**
$mF1_{gt}$	SoftSem	12	18	29	22	26	**62**	**42**	**56**	**55**	**60**	47	22	51	30	43	38
	SoftGridPattern	13	18	22	25	26	59	37	50	48	57	37	20	48	31	42	36
	mono-LC no c	21	24	31	26	22	51	36	45	42	38	53	39	53	34	39	37
	mono-LC	**27**	**28**	X	X	X	53	36	X	X	X	56	**45**	X	X	X	X
	multi-LC no c	21	18	33	**27**	29	57	34	51	41	55	58	34	**57**	32	48	40
	multi-LC	**27**	22	**37**	**27**	**30**	56	33	52	43	46	**59**	**43**	**57**	**33**	50	**41**

Table 5.2: Translation results for the 3 full France maps computed on the ground truth. "*no-c*" corresponds to ablation cases where the geographical coordinate sub-module is removed. X denotes impossible translations with the geographical context encoding.

Table 5.2 presents the results computed on the ground truth. We observe that the multi-land-cover map model outperforms the mono model, especially in terms of $mF1_{gt}$. Satisfactory MOS map translation heavily relies on the coordinate sub-module (0.39

for mono-land-cover map with no coordinates, 0.48 for multi-land-cover map with no coordinates, 0.5 for multi-land-cover map with coordinates). This can be explained by the fact that the geographical context is most useful when translating unseen objects during training (such as sea). The maps with limited spatial extent, exhibiting a lower diversity of classes and objects, benefit the most from the geographical context. On the contrary, the coordinate sub-module seems less useful on the two OCS-GE maps, which perform almost the same with and without it due to training on larger and more diverse areas.

5.1.3.5 Geographical context encoding

Visualising the learnt geographical encoding is crucial to better understand its effect on translation accuracy. Figure 5.9 compares the encoding using the six maps and the mono-land-cover map approach trained on the OSO-to-CLC translation. It is obtained by applying a PCA to the output of the MLP. First, the representation obtained for the multi-land-cover map approach does not seems correlate with the number of maps or the nature of maps covering each area (see Figure 3.4). Secondly, it seems that this encoding correlates well with major French geographical landscapes such as Alpes and Pyrenees mountains, the Paris basin, and the Mediterranean seashore. These results underline that the learnt geographical encoding through a multi-land-cover map approach learns to discriminate translation based on specific characteristics, by representing comparable geographic context such as high mountains in a comparable way even when they are spatially far-away. Conversely the results with the geographical encoding obtained by the mono-land-cover map model seems to be proner to learn a land-cover specific representation to compensate for local source and target errors.

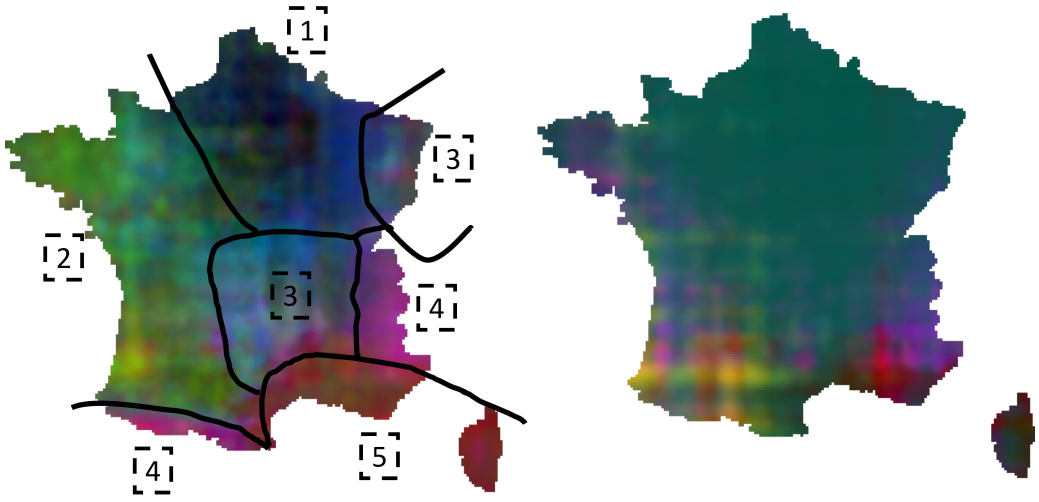

Figure 5.9: PCA representation of the learnt geographical context embedding for our multi-land-cover map model (left) and the mono-land-cover map OSO to CLC model (right). One may easily delineate the main French landscapes, namely 1) Paris basin, 2) Atlantic seacoast, 3) Medium mountains, 4) High mountains, 5) Mediterranean seashore.

5.1.4 Conclusion

We comprehensively investigated the potential of country-wide multi-land-cover map translation with our novel MLCT-Net model. In order to obtain higher-quality translation than models trained on specific translations or non-spatial-context-aware translation methods, we inspired ourselves by recent work on multi-task and multi-modal deep learning models. Namely, we designed a multi-encoder decoder network incorporating a three-term loss: 1) a translation loss to evaluate the quality of the land-cover map translation, 2) a self-reconstruction loss to ensure that the embedding preserves each map information, 3) a maximum distance loss on the embedding to ensure that similar features of different maps are encoded the same way to ensure high-quality results even on unseen spatial extents. Each encoder learns to project a specific map into a representation space shared between all land-cover map. Conversely, each decoder translates this shared representation space into one target land-cover map. Our key contribution is such a universal country-wide representation space which demonstrates an increase in translation generalisation.

We comprehensively evaluated our method by comparing the obtained translations to the original land-cover map and a manually annotated ground truth. Our method outperforms the standard semantic and statistical methods that only focus on exploring per-class associations instead of defining context-aware ones. The average improvement is about 9.5% in overall agreement between source and translation compared to the semantic baseline (6.2% for the statistical baseline). In contrast with the mono-land-cover map method, although we do not learn specific translation parameters, the multi-land-cover map method is only 0.4% worse in terms of overall agreement. Furthermore, statistics computed on ground truth reveal that the multi-map model outperforms the mono-map when computing the translation of maps on a spatial extent that they do not initially cover. These results demonstrate that learning a universal representation for multiple land-cover maps improves the robustness of the translation.

5.2 Multi-modal land-cover translation

The previous contextual translation methods considerably improve the translation compared to traditional methods. However, they fail to correctly predict many classes, as evidenced by the low average $mF1_{gt}$ (39%). We attribute this problem to the lack of information necessary for the translation in the source map. We define the notion of the *semantic gap* as the abstract measurement of the gap between the source semantic information available (the classes and context) and the information necessary to predict all target classes. For instance, translating a source map with a single agricultural class (*e.g.* CGLS) to a 12 agricultural classes target map (*e.g.* OSO) suffers from a wide semantic gap that a contextual analysis cannot fill. Similarly, we define the notion of the *geometric gap* as the abstract measurement of the gap between the local resolution of an available source

object and its target resolution. For instance, the translation of a CGLS *Forest* whose geometric precision is about 100 meters to an OSO *Forest* whose geometric resolution is between 10 and 20 meters suffers from a wide geometric gap. The low EPI values particularly underline the importance of this geometric gap.

Overcoming these two gaps requires fusing additional sources of information with the source map. A large variety of sources can theoretically be used to reduce the two gaps. Most current literature focuses on fusing several maps with various nomenclature and resolution. On the other hand, few studies have focused on the fusion of other data types, such as remote-sensing images. These works generally simply use the source map as masking data to improve a classification entirely carried out on the images, *e.g.* the image pixel can only be classified as wetland if it is classified as a wetland in this reference map [33]. They neglect the context information by relying only on nomenclature.

This section proposes to replace this "pure image classification filtered with a reference map" with a joint image and map fusion classification. Based on the experiments conducted in the previous parts, we limit the study to developing a method compatible with the MLCT-Net architecture that has shown good translation robustness even on land-cover map with small spatial extent. The nature of additional data can widely change depending on operational constraints such as the target nomenclature and resolution and the data availability. Therefore a versatile architecture is developed based on giving selectively more weight to the data the most well-suited for a given prediction.

We propose a comparison of different multi-modal sources of information to evaluate the versatility of the proposed method. Unable to test all potential data sources, we focus on three types of data: optical image, radar image, and Digital Elevation Models (DEM). Those data being the most commonly used to produce land cover maps seem well-tailored to bring semantic information. However, their capacity to fill the semantic gap is still to be discussed as it depends directly on the amount of information that is not already analysable using a single source map. The selected data have a resolution close to the highest resolved map to fill the geometric gap. We deliberately exclude the analysis of multi-temporal remote-sensing data from the study to avoid an additional layer of complexity.

To sum up, the key contributions of this section include:

- An improved translation tool based on a versatile image and map fusion framework built on the MLCT-NET architecture.
- An evaluation of the ability to fill the semantic and resolution gap of three mainstream data sources.

The first section presents various possible image/map fusion methods based on the MLCT-net architecture and the state-of-the-art. The second section introduces the three additional data incorporated into our study and discusses their potential to reduce the semantic and geometric gaps. The third section presents the experimental protocol designed to compare

the different fusion approaches and evaluate the potential of the different additional data. Lastly, we present the results of the experiments.

5.2.1 Data fusion architecture for translation

This section presents various MLCT-Net-based architectures to fuse a source map with remote-sensed data. First, from the literature review conducted in Section 2.4, we identify the stage of insertion in the architecture as a critical architecture component that may vary depending on the data and task.

Secondly, based on the observation that the relative importance of the considered additional data and source map should vary depending on the considered translation, we propose to investigate the attention mechanism mentioned in Section 2.4.

Lastly, inspired by the MLCT-net representation space shared between all maps, we propose to learn a representation space shared between the additional data and the maps. This could enable training a single network per additional data independently from the considered translation *e.g.* an image is encoded through a single encoder, fused with a given source map representation, and decoded into any land-cover map without requiring one image encoder per translation. We term *Specificity*, the additional data model dependence on the translation. A not-specific model is observed when a single model encodes the additional data independently from the considered translation. A target-specific model learns a different model depending on the target translation.

5.2.1.1 Fusion stage

Traditionally, three fusion strategies are identified: early/ mid/ late-stage [320]. There are no best strategies, and the best methods depend on the data, architecture and training task [134, 244]. As Land-cover translation has rarely been studied no previous work can give insight on the strategy to adopt: we propose to compare those three strategies.

The **Early** fusion approach (see Figure 5.10) is constant across all literature and always consists of a simple concatenation of the LC with the image before feeding it to the encoder. This approach assumes that features from the LC and the images are similar. In our case, we expect this method to perform poorly as CNN on images intensely focuses on texture, while low-texture land-cover maps are rich in terms of geometric information.

The **Mid** fusion (see Figure 5.12) consist in processing the map and additional data in separate encoders to fuse progressively shared features. Plethoric solutions have been proposed. Those presented in this section are based on solutions known to perform reasonably for multiple tasks. The **Mid-1** is based on the Fuse-net architecture [113], easily adaptable to work with the U-NET encoder architecture by simply adding a max-pooling layer to the image encoder. It processes the image and map representations in parallel in two identical architectures, fusing the image representation with a simple addition at each

encoding block of the U-Net. The **Mid-2** architecture improves the Fuse-net architecture by replacing the addition with a simple channel-wise attention module based on the Squeeze-and-Excite (SE) block (described separately in Section 5.2.1.2). Coincidentally, this makes the methods comparable to the Multi-Modal Transfer Module architecture (MMTM) proposed by [156]. Lastly, we introduce **Mid-3**, which fuses higher-level features in the decoder part of the U-Net using the same SE block as **Mid-2**.

Late fusion methods (see Figure 5.11) consist of training simultaneously one network per modality and either concatenating or summing their results just before the last classification layer. We use the same architecture for learning the representations, as commonly performed when the two modalities are identical. Like **Mid-2** and **Mid-3** architectures, we propose to fuse the different representations using SE.

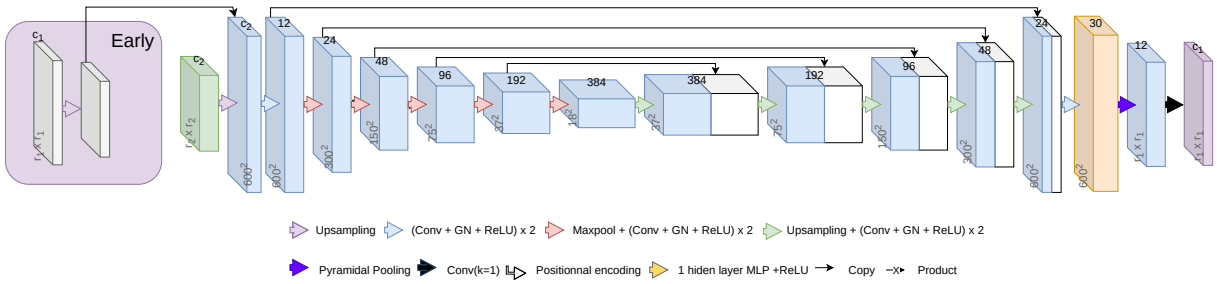

Figure 5.10: Early fusion strategy (purple box). For convenience, only one branch of the MLCT-Net is presented.

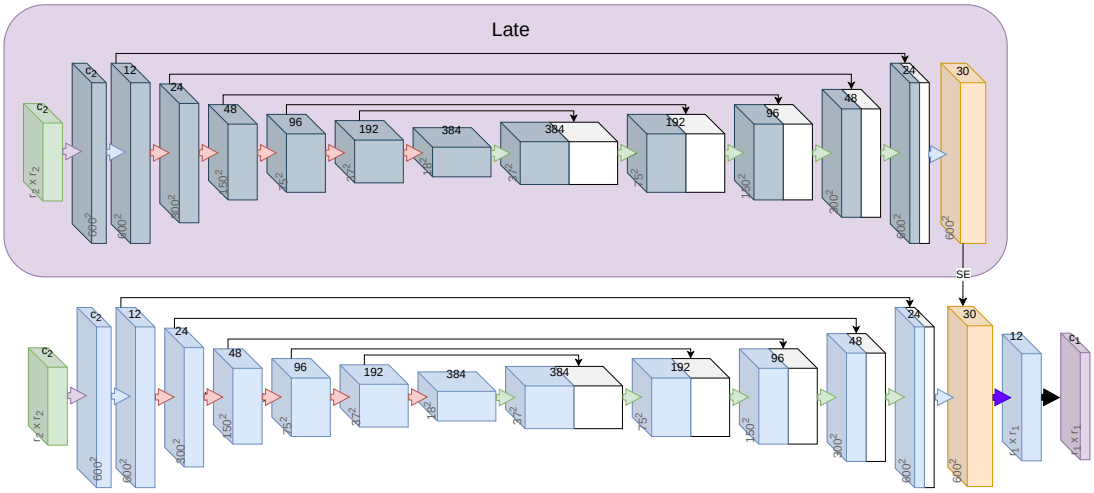

Figure 5.11: Late fusion strategy (purple box). For convenience, only one branch of the MLCT-Net is presented.

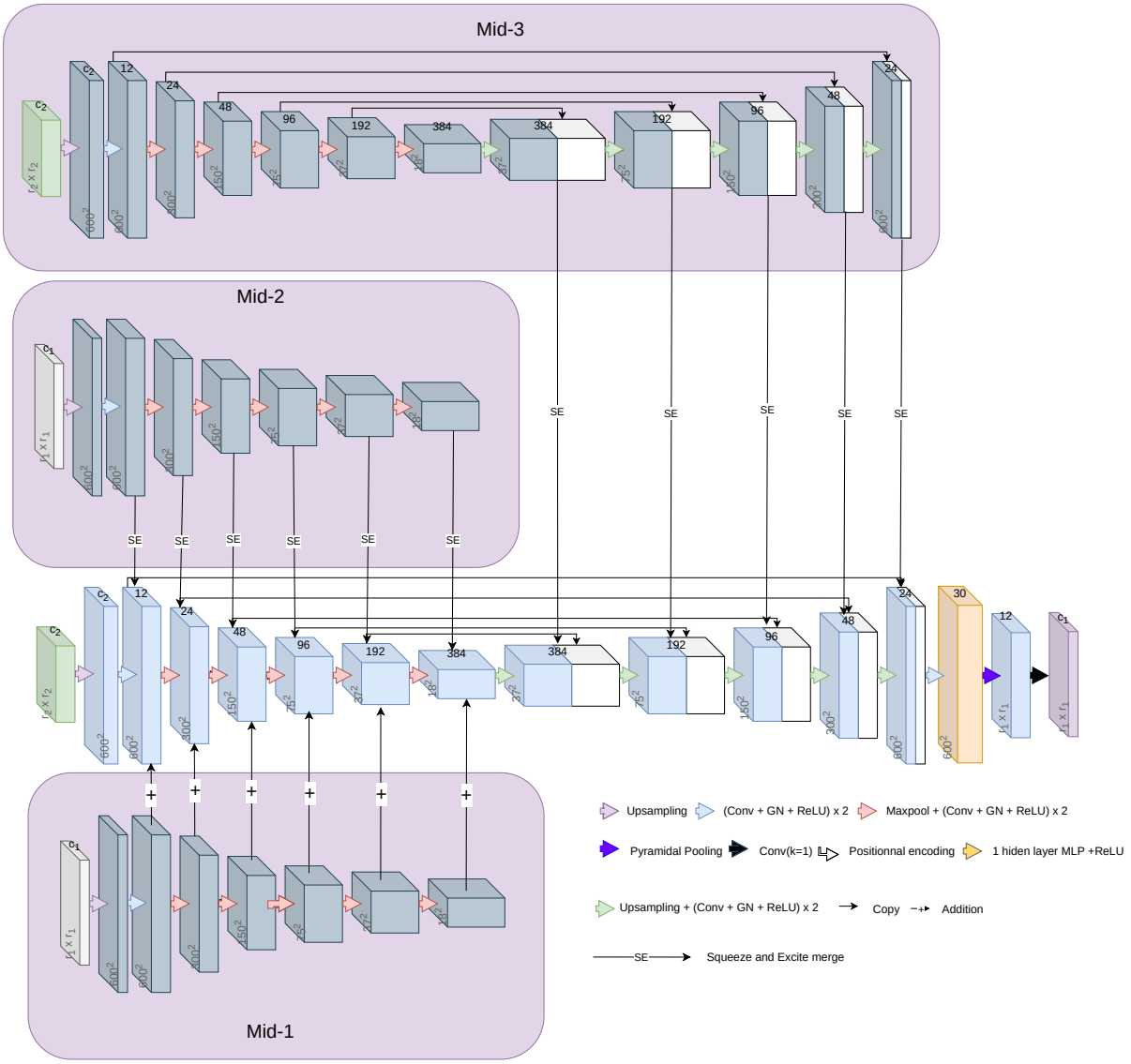

Figure 5.12: Three different intermediate fusion strategies. For convenience, only one branch of the MLCT-Net is presented.

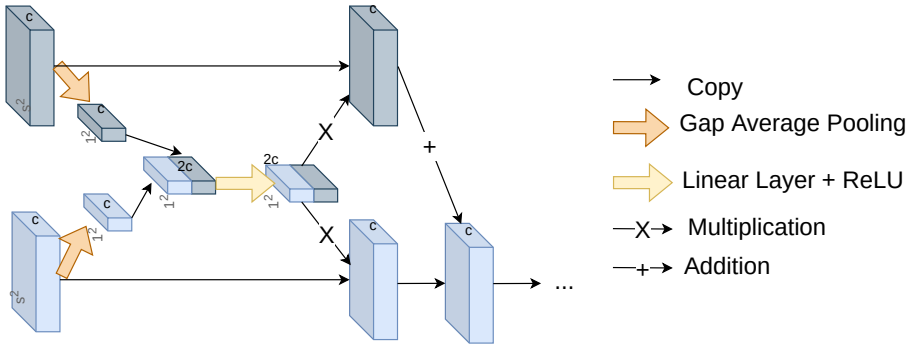

Figure 5.13: Squeeze and excite block [131].

5.2.1.2 Channel attention

A current trend in computer vision is to incorporate attention mechanisms into the networks, *i.e.* methods to give more importance to some features than others. We distinguish two sorts of mechanisms based on the nature of the features considered: (i) spatial attention, when features are considered spatially (ii) channel-wise attention, when features are considered per-pixel. Spatial attention, for instance, implemented in Vision transformers [70], leverages the spatial sequence of pixels to give more importance to some parts of the feature maps. In our case, we deliberately ignore those methods as they require vast amounts of training data to achieve satisfying results. Moreover, as they are memory-consuming, training simultaneously with the six land-cover maps is unfeasible. On the other hand, Channel-wise attention appears very interesting, given the properties of the MLCT-Net. Indeed, different channels of the MLCT-net common representation space encode for different land-cover types (see Figure 5.4). When fusing image and map features at the common embedding stage (**Late** fusion), it could be beneficial to give more influence to some image or map representation channels depending on their reliability on the land cover encoded by the considered channel. The experimented channel attention mechanism relies on Squeeze-and-Excite introduced in [131]. It consists of a three-step procedure presented in Figure 5.13.: i) a global average pooling computes separately the per-channel mean for the image and map representation ii) the two per-channel means are concatenated and processed by a linear layer outputting separate per channel weights for the image and map representation iii) the image and map representation channels are multiplied separately by the previous weights and then added together.

5.2.1.3 Image representation specificity

MLCT-Net aims to perform multiples translations using a single model. Its core idea is to translate one source map into multiple target ones and thus can be seen as a multi-task network. A core question when fusing multi-modal data into a multi-task model relies on the dependency between each task and the multi-modal data. This requires answering the question "should the multi-modal fusion model be dependent on the task ?". For instance, when fusing an image with a source land-cover map, should the model used to encode the image be the same if the source land-cover is A or B? Should it be the same if the target land cover is A or B ? To our knowledge, no previous works have been conducted to evaluate this multi-modal multi-task fusion in the remote sensing field. However, some works in the medical imagery field suggest that training a single multi-modal model to perform multiple tasks can outperform training a separate multi-modal model per task [357].

The learnt representation can be distinguished into four levels of specificity. We illustrate this idea with the following example: we target translating LC A, B and C into one another and use an optical image as additional data.

Level-1 learns a unique image representation, *i.e.* a unique image representations is

combined with the representations of A, B, and C. We refer to this method as the **Unique** method as a unique universal image representation is learnt.

Level-2 learns one image representation per source land-cover map, *i.e.* the image representation combined with the source map A representation is different from the one combined with the B or C representation. We refer to this method as the **SourceSpe** method, as one specific image model is learnt for each source map.

Level-3 learns one image representation per target land-cover map, *i.e.* the image representation combined with the A representation is not the same if the goal is to obtain B or C. The image representation fused with A and with B is identical if the target for A and B is C. We refer to this method as the **TargetSpe** method, as one specific image model is learnt for each target map.

Level-4 learns one image representation per source/target pairs. We do not present level four results implying training $n \times (n-1)$ image models.

5.2.2 Remote-sensing data sources

Our experiments incorporate three different sources of remote-sensing data, presented in Section 3.4, to alleviate the semantic and geometric gaps. Data are standardised using their respective mean and standard deviation.

5.2.2.1 Optical imagery: Sentinel-2

The first source is optical images. As land-cover maps are mainly produced by analysing optical images, they should be the best fitted to increase land-cover translation ability. They can both bring semantic and geometric information. For instance, such images could help retrieve the previously mentioned OSO agricultural classes. However, we underline that this data is insufficient to achieve perfect translation, as (i) we only process mono-temporal images, (ii) multiple land-cover maps such as CLC or OCSGE cover and use are not solely based on image analysis but also stem from fusing with independent databases. The core question is to identify to which extent a single image provides complementary information to source map. All experiments are conducted on Sentinel-2 cloudless synthesis.

5.2.2.2 Radar imagery: Sentinel-1

The second source is Synthetic-Aperture Radar imagery which offers some interesting properties. In particular, the high wavelength used in radar acquisition conveys a different radiometric information that could partially fill the semantic gap in classes related to water, urban areas, or glaciers. From an operational point of view, SAR images are more readily available for a task involving a mapping with temporal constraints or under tropical

conditions as they are insensible to weather conditions. All experiments are conducted on the Sentinel-1 mosaic averaged temporally to reduce speckle noise.

5.2.2.3 Digial Elevation Model: ALOS World 3D

The last source is a Digital Elevation Model (DEM). DEM provides interesting semantic information for very specific classes correlated to topographic specificities (*e.g.* conifers and snow are often observed at high altitudes). It could also help fill the geometric information for classes exhibiting a geometry correlated to topography. For instance, tree density might change abruptly on the ridge of mountains (due to both luminosity and humidity variations on both sides), resulting in a change in land-cover classification. All experiments are conducted on ALOS World 3D and the related features computed (exposure, TPI, roughness).

5.2.3 Experimental protocol

5.2.3.1 Finding the best architecture

Many architectures can be obtained by combining the Fusion stage and Specificity characteristics presented earlier. We test all possible Fusion stage - Specificity combinations to determine the best configuration. For instance, we test the **Late** fusion either with a **Unique**, **SourceSpe** or **TargetSpe** specificity. We underline that the **Early** fusion methods are always Source Specific as the image and source map are concatenated from the start and thus can not be analysed separately. This results in 13 network configurations: 1 for **Early** and 3 for **Mid-1**, **Mid-2**, **Mid-3**, **Late**.

As we aim to obtain a versatile architecture usable with other data than the three experimented, we consider best architecture the one that obtains the best results on average across the three additional data. Subsequently, we test each architecture with each additional data separately, *i.e.* the 13 configurations are trained using optical images, SAR and DEM, resulting in 39 models.

5.2.3.2 Assessing the potential of multi-modal translation

Multi-modal translations should obtain a higher quality than using only the map or classifying the remote-sensed data without any source map information. We compute the classification of the additional data without any source map using the same U-Net architecture as the one used for fusion: we train to classify the S1 and S2 data into each of the target maps (the DEM-only classification is not presented as it gives poor results). We underline that a separate model is trained for each of the specific spatial extents of the dataset. For instance, three models are trained to classify S2 into CLC, one France-wide (corresponding to OSO to CLC or CGLS to CLC translation), one on the OCSGE spatial

extent, and one on the MOS spatial extent.

5.2.4 Results

5.2.4.1 Finding the best architecture

Table 5.3 present a comparative study of the 13 configurations in terms of OA_{gt} using Sentinel-2 data. The that the same conclusion can be drawn using the other additional data (see Appendix I). A first observation is that the specificity does not affect the results for a given fusion stage. This underlines that training a unique model, independent from the considered translation, is not detrimental compared to training multiple independent ones. This is interesting for operational cases, as training a single model is more time efficient. It indicates that a single model can learn to map remote-sensed images into a universal representation space.

Moreover, we observe that the **Late** fusion always performs better than the other configurations. This statement holds for the three data. We hypothesise that this could be attributed to the necessity to considerably transform the map and image features to make them complementary, preventing fusion at earlier stages. In the following section we only consider the **Late** - **Unique** method as it obtains the best results in the least computationally expensive way.

Source		CGLS (P)					CLC (C)					OSO (O)					OCSGEc (G1)				OCSGEu (G2)				MOS (M)			Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	
Early	SourceSpe	68	70	78	85	85	77	70	79	86	85	76	67	79	86	86	72	61	66	92	72	60	66	84	87	85	70	77
Mid-1	Unique	68	70	78	85	85	77	70	79	86	85	76	67	80	87	86	73	60	66	92	72	60	66	84	**87**	**85**	71	77
	SourceSpe	67	69	78	85	85	76	69	78	85	85	75	66	80	87	86	72	59	66	92	72	59	65	84	**87**	84	71	76
	TargetSpe	67	69	78	86	85	77	70	79	86	85	76	67	80	86	86	72	60	67	92	73	60	66	84	**87**	**85**	71	77
Mid-2	Unique	67	69	77	85	85	77	70	79	86	85	76	67	80	86	86	73	60	67	92	72	58	65	82	**87**	**85**	71	76
	SourceSpe	67	70	78	85	85	77	69	79	86	84	77	67	80	86	86	72	60	68	92	72	59	65	84	86	**85**	71	77
	TargetSpe	67	69	78	85	85	77	69	80	86	84	76	67	80	87	86	73	60	68	92	72	58	65	83	86	**85**	71	76
Mid-3	Unique	68	71	79	86	86	78	70	79	86	86	77	68	81	87	86	74	61	68	**93**	73	59	66	85	86	**85**	71	77
	SourceSpe	67	71	79	86	86	77	70	80	86	85	77	67	81	86	86	72	61	69	92	73	59	66	85	86	**85**	71	77
	TargetSpe	68	71	79	86	86	77	70	79	86	84	77	67	80	86	86	73	61	68	92	73	59	66	84	**87**	**85**	71	77
Late	Unique	**71**	**73**	**81**	**87**	**86**	**79**	**73**	**81**	**87**	**86**	**78**	**70**	**82**	**88**	**87**	**75**	**64**	**70**	**93**	**75**	**64**	**70**	**87**	**87**	**85**	**72**	**79**
	SourceSpe	**71**	**73**	**81**	**87**	**86**	**79**	**73**	**81**	**87**	**86**	77	69	**82**	**88**	**87**	75	**63**	**70**	**93**	**75**	63	69	86	**87**	**85**	**72**	**79**
	TargetSpe	**71**	**73**	**81**	**87**	**86**	78	**73**	**81**	**87**	**86**	77	**70**	**82**	**88**	**87**	**75**	**64**	**70**	**93**	74	63	69	86	**87**	**85**	**72**	**79**

Table 5.3: Comparison of various Fusion strategies using a Sentinel-2 cloudless synthesis, the 6 land cover maps and the MLCT-net in terms of OA_{ag}.

5.2.4.2 Comparing the remote-sensed data sources

Figure 5.14 compares the results obtained for the three different remote-sensed images using the **Late** - **Unique**. We first observe that for the the coarse-to-fine resolution case, additional data sources considerably increases the qualitative aspect of the result compared to the pure map translation approach. This is, for example, well illustrated by the first and last row of Figure 5.14 (respectively, CLC to MOS and CGLS to OCSGEu translation). When the source resolution is equivalent to or finer than the target one, more data only marginally increases the semantic accuracy, *e.g.* the S2 mosaics improves the MOS to OSO translation by helping to retrieve agricultural classes. We observed that

fusion gives significantly different results than using each separately. For instance, in the CGLS to OCSGEu translation, the fusion-based approach is the only one able to retrieve *411: "Road Networks"*.

Those results are comforted by the analysis of the OA_{ag} and $mF1_{ag}$ (Table 5.4). In particular, the observed average +2% OA_{ag} and $mF1_{ag}$ increase between the map+S2 and S2 only strategy encompass widely heterogeneous per-translation results. When the source map is CGLS or CLC (the coarsest maps), the difference between using only S2 or fusing S2 with one of those two maps is almost nonexistent and slightly detrimental (-0.6% $mF1_{ag}$) due to the difficulty of optimising the MLCT-Net: coarse information is often of low interest when predicting a fine-resolved map, as one source object might encompass multiple target land-cover types. Conversely, when the source map is identically or more resolved than the target, the fusion performs significantly better than using the image-alone strategy (+4.6% $mF1_{ag}$). We conclude that land-cover to additional data fusion is only relevant when the source resolution is comparable or higher than the target.

Another observation is that fusing S1 still performs reasonably well and is often preferable to the S2-only approach. Therefore it appears possible to perform our land-cover fusion strategy even in areas where optical imagery suffers from cloud occlusion.

The OA_{ag} and $mF1_{ag}$ of the map-only translation approach are always lower or equivalent to the S2-only strategy. Consequently, land-cover translation could appear inefficient for operational use as using a single image often outperforms the land-cover translation approach. However, we argue that this higher OA_{ag} and $mF1_{ag}$ is mostly due to a higher capability of the network to replicate target errors when using an image as source data instead of the map. This behaviour is demonstrated in Table 5.5 using the ground truth. We observe that the map-only approach often exhibits higher metrics than the S2 only, especially when the fine-resolved OSO map is translated into one of the others. Furthermore this two overall metrics encompass significant per-class differences, illustrated in Figure 5.15. The map-only translation approach significantly outperforms the S2-only approach for classes with various semantic definitions such as *CGLS 70: "Snow and Ice"*, *CLC 212: "Permanently irrigated land"*, *OCSGEc 1112: "Undeveloped areas"*. Secondly, we observe that fusion strategies seem to keep "the best of the two worlds" when either the map or the additional data obtain results used alone while the other performs well as well illustrated by the OCSGEc 1112 class example.

Additionally, Table 5.5 shows that the fusion between the image and map is most of the time beneficial compared to using each data separately exception made of the translation into the MOS map. As mentioned previously, those measurements on the France-wide ground truth test in the case of the translation into MOS, the ability of the method to spatially generalise from an original MOS extent. We conclude that the proposed fusing method does not guarantee efficient spatial generalisation ability, *i.e.* the universal image+map representation space is non-homogenous across the territory preventing spatial generalisation.

Lastly, we observe that the S1 and DEM data are rarely more beneficial than the S2 data except on very specific classes such as "Snow and Ice" equivalent classes for CGLS, CLC and OSO.

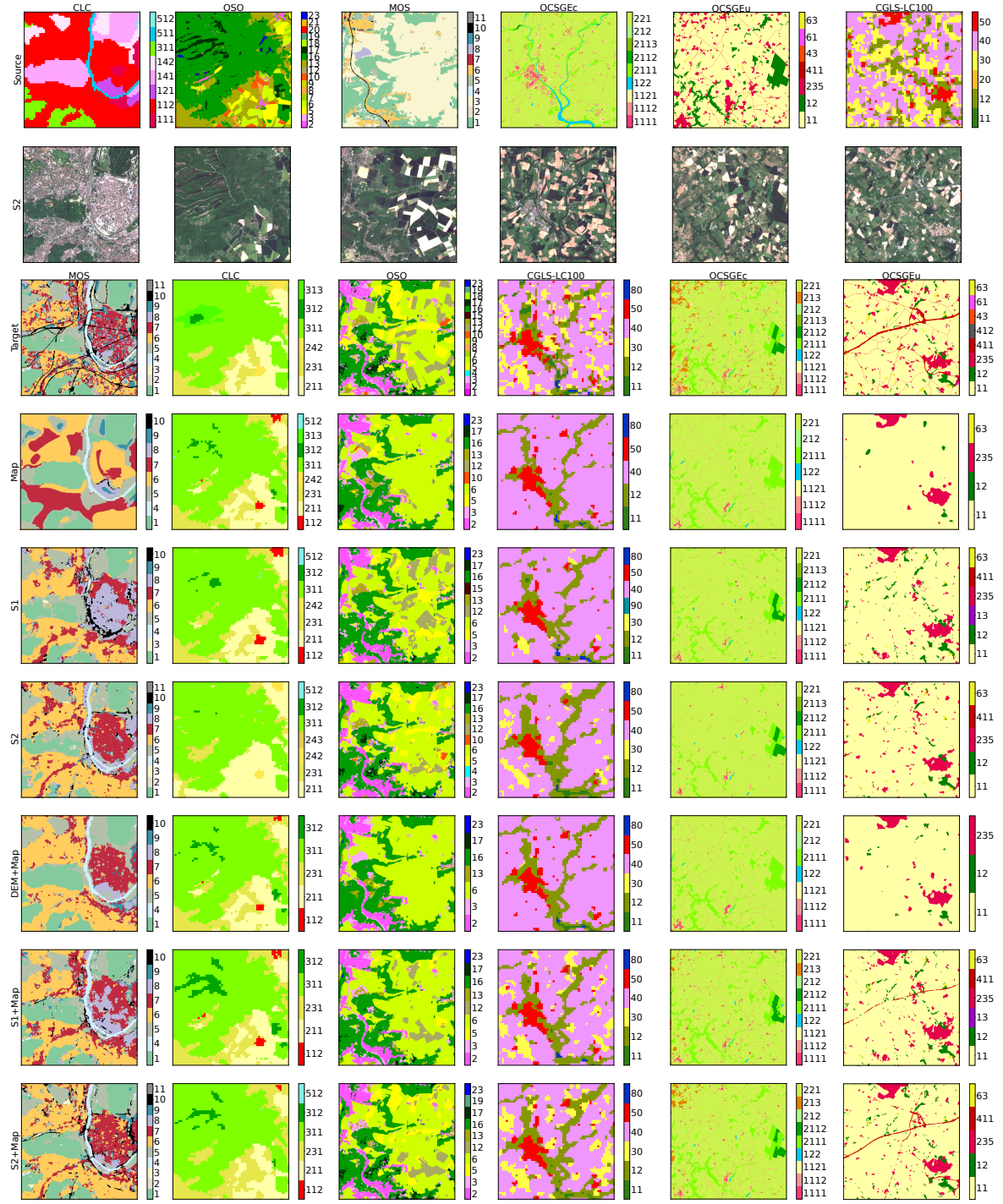

Figure 5.14: Comparison of the translation results using various remote-sensed images and the Late-Unique fusion strategy. Map, s1 and s2 results are obtained using respectively only the source map, only the Sentinel-1 data, and only the Sentinel-2 data. map+s2 correspond to results obtained by fusing the image and map representation using the Late-Unique fusion strategy.

Source		CGLS (P)					CLC (C)					OSO (O)					OCSGEc (G1)				OCSGEu (G2)				MOS (M)			Total
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	average
OA	Map only	65	57	69	78	76	75	59	73	80	79	77	69	80	86	85	71	58	58	93	69	54	53	79	86	84	64	72
	S1 only	69	67	77	84	83	77	67	77	84	83	77	69	77	84	83	71	57	59	85	71	57	59	77	86	82	65	74
	S2 only	70	**74**	**81**	**87**	**86**	77	**74**	**81**	**87**	**86**	77	**70**	81	87	86	74	60	68	87	74	60	68	80	86	83	70	77
	Map + DEM	64	59	72	80	80	73	60	74	82	81	76	69	80	87	86	72	61	60	93	71	58	56	82	86	84	64	73
	Map + Aspect	66	57	70	79	77	76	60	73	81	80	78	68	80	86	86	73	61	59	93	72	57	55	81	**87**	84	64	73
	Map + DEM +Aspect	67	60	72	80	81	75	60	74	81	82	79	69	80	86	86	74	61	60	93	73	59	55	81	**87**	84	63	74
	Map + S1	70	68	79	86	85	78	68	79	86	85	77	**70**	**82**	**88**	**87**	73	63	65	93	73	62	63	85	**87**	**85**	69	77
	Map + S2	**71**	73	80	**87**	**86**	**79**	73	80	**87**	**86**	**78**	**70**	**82**	**88**	**87**	**75**	**64**	**70**	**94**	**75**	**63**	**69**	**86**	**87**	**85**	**73**	**79**
mF1	Map only	30	29	30	20	33	58	37	38	30	41	61	40	45	27	53	52	34	31	43	52	29	25	40	45	30	23	38
	S1 only	39	44	39	26	43	55	44	39	26	43	55	38	39	26	43	50	28	32	27	50	28	32	38	46	27	29	36
	S2 only	43	**53**	**48**	**36**	**58**	59	**53**	**48**	**36**	**58**	59	43	48	36	58	54	36	**45**	37	54	36	**45**	48	50	40	35	47
	Map + DEM	35	35	34	20	40	58	37	39	29	45	58	41	47	28	55	54	38	35	44	52	37	30	49	50	39	24	41
	Map + Aspect	35	30	32	20	35	61	37	37	24	43	63	40	44	23	53	55	39	33	35	53	35	27	47	49	37	23	39
	Map + DEM +Aspect	35	35	33	21	41	62	38	40	28	46	62	42	48	28	54	56	38	36	44	54	37	29	48	50	40	25	41
	Map + S1	41	43	42	26	52	60	46	43	31	54	60	43	47	27	57	55	41	39	41	54	40	35	53	**51**	**44**	32	45
	Map + S2	**44**	50	47	30	57	**65**	52	**48**	**36**	57	**64**	**47**	**50**	**38**	**60**	**62**	**44**	**45**	**49**	**61**	**44**	44	**57**	49	**44**	**36**	**49**

Table 5.4: OA_{ag} and $mF1_{ag}$ comparison for various additional data fusing. Grey columns denotes translation for which the source map is higher resolved than the target one

Source		CGLS (P)					CLC (C)					OSO (O)					Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	
OA_{gt}	Map	60	53	74	83	83	**71**	59	79	86	**86**	**78**	70	87	**92**	**93**	**77**
	s2	63	70	82	89	81	69	70	82	89	81	69	63	82	89	81	**77**
	Map+s2	**70**	**72**	**84**	**91**	67	**71**	**72**	**85**	**91**	66	72	**72**	**88**	**92**	69	**77**
$mF1_{gt}$	Map	27	22	37	27	**30**	56	33	**52**	**43**	**46**	**59**	43	**57**	33	**50**	**41**
	S2 only	37	40	46	31	39	52	40	46	31	39	52	37	46	31	39	40
	Map+s2	**44**	**41**	**47**	**37**	**30**	**58**	**48**	50	42	29	**59**	**47**	55	**43**	30	**44**

Table 5.5: OA_{gt} and $mF1_{gt}$ comparison for the map only, s2 only, map+s2.

5.2.5 Conclusion

We presented a strategy to fuse remote-sensed images and map representation to both improve results and study the potential of re-using existing land-cover maps as input of the traditional image-to-land-cover classification procedure. As we consider that both the image and map should be fused using a context-aware method to increase the amount of information available, we proposed to adapt the MLCT-Net. The first section introduced how to fuse additional data inside our MLCT-net architecture. From our experiment we observe that the fusion should be performed at the land-cover representation space level. Since the land-cover representation discriminates different land-cover using different dimensions, the land-cover and image representation dimensions should be weighted differently depending on their content. Moreover, learning a single representation space for the additional data, independent from the source and target map, appears sufficient. Secondly, we focused on the impact of different sources on the quality of land cover translation. Experiments demonstrate that fusing additional data mainly the quality of translation for all source/target map couples. However, we also pointed out that the difference between classifying the remote-sensed data alone and fusing dit not significantly improve the results when the source is significantly coarser than the target. Therefore we argue that fusion between image and representation should mainly be conducted using a source map with an equivalent or finer resolution than the target. Note that those

Figure 5.15: Highest observed per class $F1_ag$ for each source to target translation.

conclusions are drawn from experiments on a limited number of maps and that other factors could have played a role. Especially one could argue that the fact that results using the remotely sensed image and low-resolved map fusion are not better than just using the image could be a mere coincidence. Instead, one could argue that the core reason lies in the complementary between the nomenclature of the source and target maps, *e.g.* CLC to OSO translation does not give bad results because CLC has a coarse resolution but because CLC nomenclature brings no more information to obtain OSO than the image. However, we found it unlikely as the $\approx 50\%$ mF1 obtained by classifying the S2 image into CLC or CGLS indicates that many classes are not obtainable from the image data alone. Lastly, we observed that the S2 imagery always achieved the best results amongst the three additional data experimented on, both used alone or merged. However, in multiple cases, the difference with S1 images remains modest, which underlines the possibility of performing this on areas where S2 images suffer from significant cloud occlusion.

5.3 Conclusion

This chapter focused on improving the translation methods to make them usable in multiple operational cases. Two research axis are developed: i) building a multi-translation framework to enable predicting multiple land-cover maps translation from a single source while ensuring good spatial generalisation properties and ii) fusing additional data to increase the quality of the translation.

The A-UNet trained to perform a single translation suffered both from a lack of spatial generalisation and was sensible to the noise in the source and, more importantly, in the target map. To alleviate those issues, we proposed to comprehensively investigate multi-land-cover map translation with our novel MLCT-Net model driven by the observation that models trained to perform multi-lingual translation outperformed models trained to perform each language separately. In order to increase the translation ability of our model on maps with small spatial extent by reducing both over-fitting and increasing the spatial generalizability, a multi-encoder decoder network was designed, incorporating a three-term loss. A universal country-wide representation space of the six land-cover map was obtained and demonstrated good translation results compared to training separate A-UNet. In particular, the multi-land-cover map model obtained a higher OA_{gt} and $mF1_{gt}$ ($+2$ and $+3\%$ respectively) than the mono-land cover baselines while achieving almost the same scores in terms of OA_{ag} and $mF1_{ag}$ demonstrate better noise robustness. Moreover, the multi-LC training enables significantly increased cartographic generalisation ability (on average $+7\%$ OA_{gt} and $+9\%$ $mF1_{gt}$ when translating CLC, CGLS or OSO into MOS).

Secondly, we investigated the potential of fusing additional data sources to increase the quality of the translation, especially in the case where the source map is coarser than the target. Land-cover maps are rarely used as raw data to predict other land-cover maps and are not jointly analysed with images. We proposed to adapt our multi-land-cover map

model to perform this fusion. In particular, learning a single universal image representation is sufficient to achieve the best translation results, *i.e.* all land cover can be inferred from a single image representation. Moreover, the late strategies appear to be the best suited in the specific case of land-cover translation, with a shared representation space for all land-covers. We thoroughly examined the effects of three different kinds of additional data (optical, SAR, DEM) and observed that fusing additional data with a map representation significantly improves the results, provided that the source map is not too coarsely resolved compared to the target. Indeed, a coarsely resolved source map conveys little information on the expected output class and does not bring extra information.

To sum up, this section presented significant improvement to the learnt spatial and geographical context translation method by increasing the quality of the translation, the spatial generalisation ability, and the noise robustness. However, the current multi-landcover strategy appears widely inefficient as (i) multiple source classes with close or identical content are learnt separately, *e.g.* the "coniferous forest" of the CLC, OSO and OCSGEc map are all mapped using a separate encoder, (ii) multiple target classes with close or identical content are learnt separately, *e.g.* the network learns separately the concept that "herbaceous near water" is likely wetlands for the OSO to CLC translation and the OCSGEu to CLC translation. This fosters multiple operational constraints: a source map can only be translated into one of the target maps used during training and only if the source map was also included in the training. For instance, the MLCT-Net we trained on the six land cover is able to translate one of those maps into one of the other five but can not use as a source or target any other maps even if it exhibits close semantic and resolution characteristics. We address this problem in the following chapter.

CHAPTER 6

Building a semantically continuous land-cover representation

6.1 Introduction

Previous sections considered the translation as the transformation between a discrete representation space with n dimensions formed by the c_S source classes to another with c_T target classes. In those sections, the CGLS *Closed Forest* and OCSGE *Broad-leaved trees* translations are learnt distinctly using two different one-hot encodings (see Equation 6.1). As each nomenclature is considered an independent representation space, adding a new source or target nomenclature requires retraining the algorithm from scratch.

$$\begin{aligned} CGLS_{ClosedForest} &= [1,0,0,0,0,0,0,0,0,0,0,0,0] \\ OCSGEcover_{BroadLeavedTrees} &= [0,0,0,0,0,0,0,1,0,0,0,0,0,0] \end{aligned} \tag{6.1}$$

We observe that a person taught that CGLS *Forest* is translated either into OCSGE use *Forestry* or *Without use*, depending on the context, is able, without additional information, to suggest that MOS *Forest* can be translated similarly. This ability to generalise a translation model to new source and target nomenclatures without specific training is termed *zero-shot translation*.

Contrary to our methods, the person is aware of the semantic links between the different classes. She/He relies on a global understanding of class definitions to estimate a resemblance between classes.She/He assumes that a unknown class must be approximately translated to classes that resemble it the most amongst the ones he learned. We assimilate this understanding to a capacity to represent the translation not as the transformation from a discrete representation space to another but as a change in the interpretation of a single representation space including all possible land-cover labels.

Finding a way to train the model on such a common nomenclature representation space could avoid retraining the network with each new source or target nomenclature. This could be particularly useful when no preexisting training samples of the new nomenclature is available.

The approaches currently representing multiple nomenclatures in a unified space have been presented in Section 2.1.1.1 under the term *standardisation* (LCCS, EAGLE). However, those semantic spaces are unsuitable for our paradigm, as their representation spaces are discretely defined. For instance, EAGLE offers more than 570 dimensions (expandable), describing distinct attributes with seven discrete values according to how mandatory the attribute is/is not. The notion of distance between two classes is not clearly defined. For instance, the distance between classes for which the second attribute is either valued x (The attribute is not relevant for the class) or 0 (the attribute is not in the definition) or 2 (The attribute is mandatory) require defining subjective rules. Moreover, it requires weighting each attribute/dimension arbitrarily.

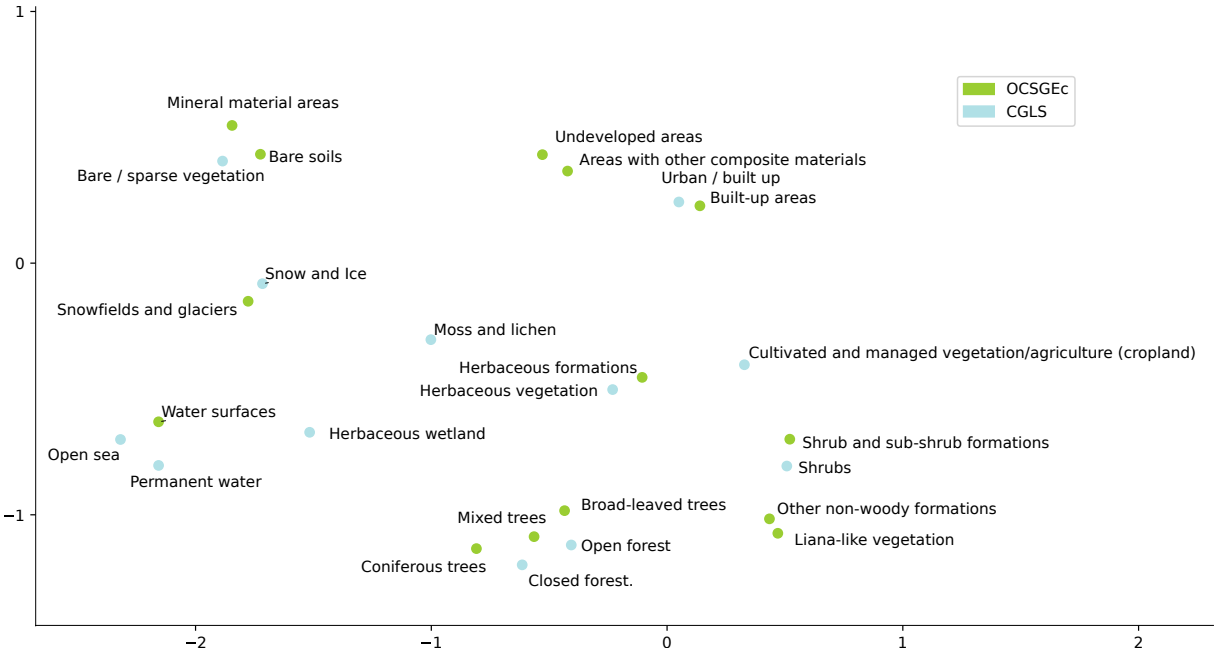

Figure 6.1: Manually defined 2D semantic representation space for the OCSGE cover and CGLS maps. The closer two classes definitions are, the closer their semantic encoding.

We propose a method projecting a class into a continuous nomenclature representation space based on its definition. Like the person's mental representation space, it should bring two semantically close classes near in such a space. The metric space in which semantically similar classes are close to each other is called *semantic space* in the literature. Instead of one-hot encoding CGLS *Closed Forest* and OCSGE cover *Broad-leaved trees*, we encode them in a single space in which they are closer to one another than classes such as OCSGE cover *Bare soil*. We illustrate this idea by proposing a 2D manually defined semantic representation space with the classes of these maps in Figure 6.1. The resulting encoding of the two previous classes is then given in Equation 6.2.

$$
\begin{aligned}
CGLS_{ClosedForest} &= \begin{pmatrix} -0.6 \\ -1.2 \end{pmatrix} \\
OCSGEcover_{Broad-leaved} &= \begin{pmatrix} -0.45 \\ -1 \end{pmatrix}
\end{aligned}
\tag{6.2}
$$

This semantic encoding can replace the standard one-hot encoding as input and output data of the networks, as shown in Figure 6.2. To date, the few examples of the application of semantic spaces in land-cover mapping are mainly in zero-shot image classification or segmentation. A model is trained to project pixels into a semantic space obtained by applying a pre-trained language model [246]. Unlike image classification, in translation, the semantic representation space can be applied to both source data (a map) and target data (another map). Encoding the source maps provided to **MLCT-Net** during training could allow spatial context-wise translation of maps never seen during training at the time of inference (*e.g.* US NLCD $\rightarrow$ CLC) based on the idea that the translation of the unseen classes fallows approximately the same contextual rules than the classes seen during training with an encoding close to the unseen class. We refer to this scenario as *zero-shot source translation*. Encoding the target maps, could allow each source map object to be projected differently into the semantic space depending on its spatial context. For instance, a pixel of OCSGE *Herbaceous formations* could be projected closer to *Herbaceous wetlands* or *Herbaceous vegetation* depending on its distance from a water point. Instead of outputting a fixed nomenclature with n classes, we could obtain a continuous representation that can be derived into a nomenclature never seen during training (CLC $\rightarrow$ NCLD). We refer to this scenario as a *target zero-shot* solution.

Training models or comparing methods producing semantic representation space requires assessing the quality of the nomenclatures encoding on a ground truth. As no public dataset with multiple nomenclatures has been released yet, we created a small land-cover definition dataset (LCDD). Ten nomenclatures are selected for their semantic diversity: the 6 of the MLULC dataset, the one of ESRI land-cover map[1], and three different nomenclature of the MODIS Land Cover Type MCD12Q1 [290] using different standardisation approach. We introduce the LCDD 169 class definitions in Appendix D. The following section discusses the expected properties desired for a semantic space and related evaluation criteria. We underline that the small size of the dataset prevents training complex models but can be used to train very simple models.

[1] `https://livingatlas.arcgis.com/landcover/`

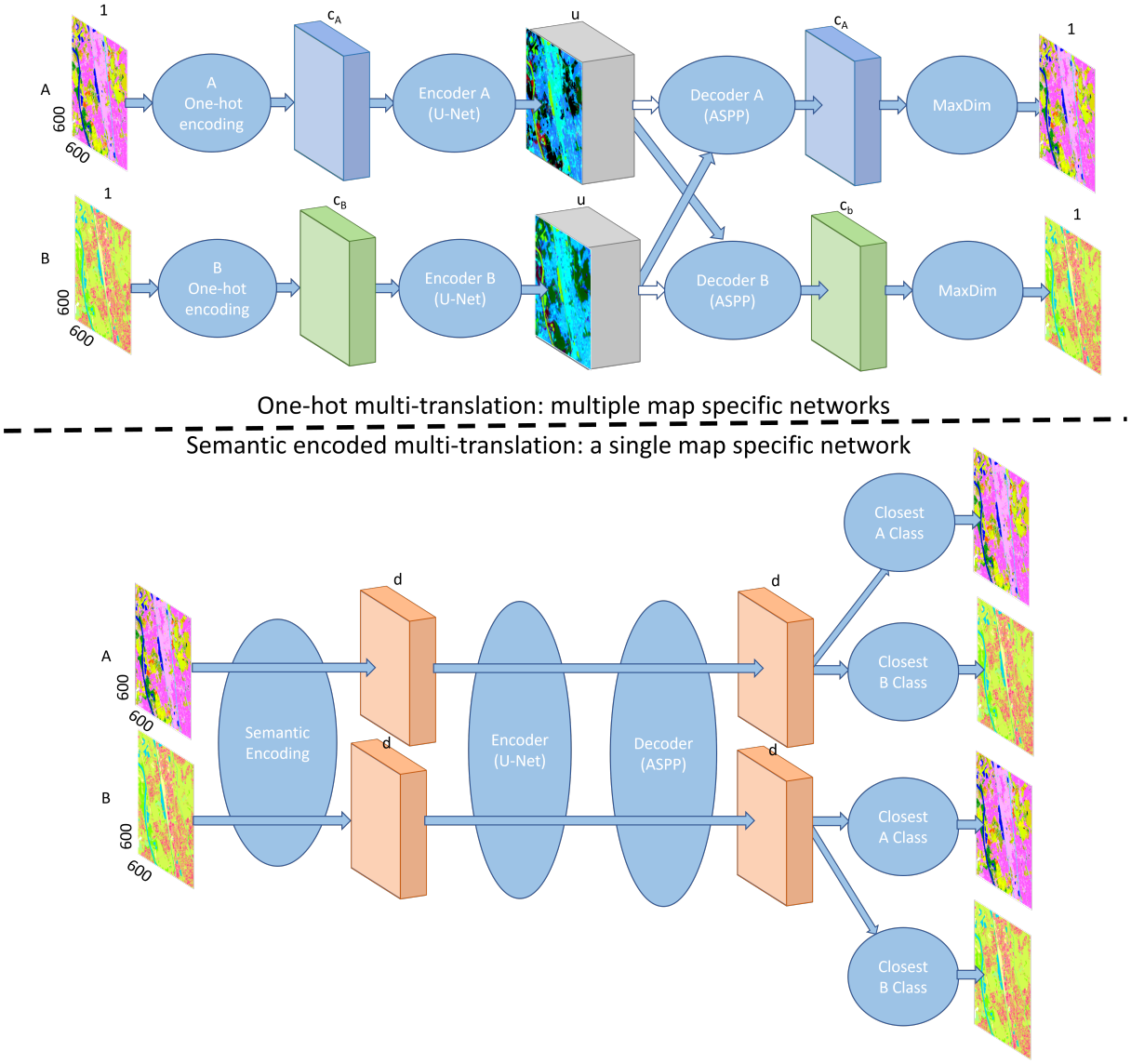

Figure 6.2: Comparison between a translation with one-hot encoding using a **MLCT-Net** with a u-dimension shared representation space (top) and a d-dimension semantic encoding (bottom) for land-cover map A (c_A classes) and B (c_B classes). A similar colour represents a comparable representation space, *i.e.* same number of dimensions encoding comparable features. Semantic encoding requires a single network since both the input and output share the same representation space. **MaxDim**: classifies the final map as the class with the highest logit value. **ASPP**: atrous spatial pyramidal pooling.

Our zero-translation framework based on semantic representation space can be decomposed into three main steps:

1. Class definitions are encoded using a language model in a continuous semantic representation space. Two approaches have been proposed in the literature so far: a bag-of-words [58] trainable using a few land-cover classes definitions such as those

of the LCDD and a Word2Vec [246]) model trained on large generic text corpora not land-cover specific. We compare those two approaches and propose a third one based on a transformer architecture trained generic text corpus.

2. The three explored techniques result in very high dimensional representation spaces (300 - 1 000 dimensions) that are unusable to encode land-cover data fed to a network for memory concerns. Moreover, models trained on generic text corpora encode for semantic notions irrelevant from a land-cover point of view *e.g.* the word water is close to surf. Different supervised or unsupervised dimension reduction techniques are explored to reduce the dimensions. In particular, we introduce a new supervised dimension reduction technique based on a multi-layer perception trained using the small LCDD dataset.

3. Land-cover maps used to train a translation CNN are first semantically encoded using the dimension-reduced SRS. At inference, an unseen map is encoded using the same semantic space and fed to the network.

.

The first section defines a set of desired properties for a land-cover semantic representation space and related evaluation criteria. We point out that this first section is part of our original contributions, as the only two previously existing works [58, 246] did not define them explicitly. The second section introduces and compares the different natural language models and dimension reduction techniques based on the evaluation criteria presented in the first section. Finally, we propose the first example of applications of these semantic spaces to the zero-shot translation case.

6.2 What is an ideal semantic representation space ?

A **Semantic Representation Space (SRS)** aims to reflect the proximity between different concepts in a space with a finite number of dimensions [195]. In our land-cover case, we aim to obtain a continuous metric space endowed with a distance metric (Euclidean distance, for example) in which semantically close classes are also close in the sense of the metric. This section first defines the notion of land-cover SRS in a broad way independently from our desired application (zero shot translation). In particular, we discuss, how and what to encode in class definitions and the notion of proximity in the SRS and its link semantic similarity. We then define additional characteristics that the space should exhibit to be usable in the specific case of zero-shot translation.

For simplification, we use the notations of Section 2.1 (see Figure 6.3). S_i and T_j are sets holding all the source pixels classified i (out of c_S classes) and all the target pixels labelled j (out of c_T classes). s_i, t_j denote the descriptors of the source class i and target class j in a considered standardisation system. $s_i(v)$ denotes the v$^{\text{th}}$ descriptor of the source

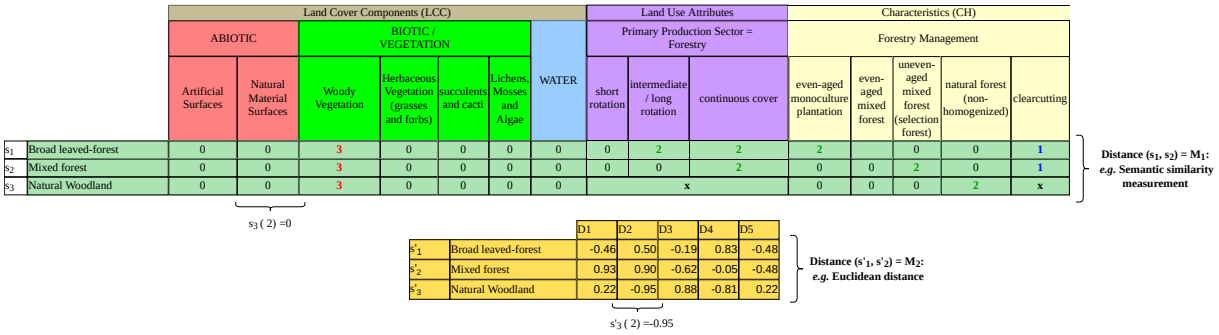

Figure 6.3: Illustration of encoding of classes using a standardisation system (EAGLE) and a continuous semantic representation space. Standardisation systems use discrete variables (x: "should not be included", 0: "can be included but not in definition ..."), while SRS use abstract conitnuous variables (here 5). For clarity only 15 discrete variables are displayed for EAGLE while the matrix used for CLC includes more than 350 variables.

class i in the standardisation system. s'_i, t'_j denote their descriptors in the SRS, with $s'_i(u)$ being the u$^{\text{th}}$ dimension out of q dimensions of the SRS. We underline that $s_i(v)$ might be a discrete encode variable, depending on the standardisation system, while $s'_i(u)$ is a continuous variable from the SRS.

6.2.1 Encoding class definitions

Our semantic space should be able to encode for all information present in the class defintion. The space must include semantic properties of the objects, including their cover and use. Additionally, it should integrate spatial resolution information (commonly provided in the form of a threshold). Lastly, it should differentiate between inclusion and exclusion statements, *e.g.* if *Crop* definition excludes pastures, *Crop* should not be close to *Pastures* even though the term pasture appears in its definition.

In the particular case of hierarchical nomenclature, we consider each hierarchical level as a distinct nomenclature, accumulating the semantic content of the previous levels. We do not aim to create multiple sub-spaces encoding the notion of class hierarchy separately, as the hierarchical division of nomenclatures is often widely different from one map to another.

As underlined before, the SRS is used to encode land-cover maps before feeding them two Convolution neural networks such as those presented in the previous sections. As such, the definitions should be encoded with continuous variables that enables straightforward distance computation or interpolation. We insist that standardisation approaches such as LCCS or EAGLE result in high-dimension discrete spaces encoding with discrete variables. In the remainder, we assume a continuous space, thereby neglecting the standardisation approaches.

6.2.2 Computing proximity between classes

Class SRS encoding must reflect the semantic proximity between class definitions. Analysing the quality of a semantic space requires: (i) defining a measure M_1 of proximity between class definitions and a measure M_2 of proximity between classes in SRS, and (ii) defining the correlation link expected between M1 and M2.

6.2.2.1 M1: Proximity between classes

Computing the proximity between class definitions can be achieved using three different measurement methods.

The **Binary Overlap Measure (BOM)** is based on a manual distinction between classes sharing elements in common with others. This binary metric is implicitly used by nomenclature translation system relying on expert knowledge [1]. As the binary behaviour does not allow a satisfactory comparison of the degree of resemblance, a few more evolve techniques such as [55, 56] assess it using a three level granularity (expected, uncertain, or unexpected). As there is little doubt about the existence or not of an element in common between to two classes, BOM is rarely false.

$$BOM(s_i, t_j) = \begin{cases} 1 & \text{if } S_i \cap T_J \neq \emptyset, \\ 0 & \text{else.} \end{cases} \tag{6.3}$$

Section 2.1.1.2 introduced the notion of **Semantic Similarity Measure (SSM)** (Equation 2.8). Often based on a proportion of attributes shared between two classes, it is adopted by most semantic approaches. It measures the resemblance with a higher granularity than BOM. However, it is often very biased because it depends enormously on the relative weight given to each attribute, arbitrarily fixed in the method.

The **Statistical Co-Occurrence Measure (SCOM)** computes the proportion of pixels of each target class corresponding to a given source. This measurement is normalised by the proportion of each target class. It offers a high granularity while avoiding arbitrary weights. Source/target map errors might reduce the correlation between this observed statistic relation and the expected semantic one. Conversely to SSM and BOM, SCOM is only based on measurement conducted on real map samples and ignores class definitions.

6.2.2.2 M2: Measuring proximity in the representation space

We restrict this manuscript scope to SRS in Euclidean geometry. We underline that approaches encoding in non-Euclidean geometries, such as hyperbolic spaces, have shown encouraging results, especially for encoding hierarchical data in low-dimensional spaces [162, 227]. They require a very high numerical precision for distance computation (64 or even 128-bit): we consider them unsuitable and deliberately ignore them.

In Euclidean space, the Euclidean distance is commonly adopted for distances (Equation 6.4). Many works have pointed out that this measure is unsuitable for high-dimensional spaces, which may be problematic for some of our semantic spaces.

$$L_2(s'_i, t'_j) = \sqrt{\sum_{u=1}^{q} (s'_i(u) - t'_j(u))^2}. \tag{6.4}$$

The Manhattan distance is often used as a replacement for higher-dimensional spaces (see Equation 6.5). As the Euclidean distance, the Manhattan one is not scale-invariant. This is not an issue as long as the amplitude of variation of each dimension correlates to the semantic gap.

$$L_1(s'_i, t'_j) = \sum_{u=1}^{q} |s'_i(u) - t'_j(u)|. \tag{6.5}$$

In other cases, a commonly solution in high dimensional spaces is the cosine distance (Equation 6.6). Instead of focusing on the difference s'_i and t'_j, the cosine distance computes the cosine of the angle between both. $CosineDistance = 0$ when vectors are collinear, 1 when they are orthogonal, and 2 when they are opposite.

$$CosineDistance(s'_i, t'_j) = 1 - \frac{\sum_{u=1}^{q} s'_i(u) t'_j(u)}{\sqrt{\sum_{u=1}^{q} (s'_i(u))^2} \sqrt{\sum_{u=1}^{q} (t'_j(u))^2}}. \tag{6.6}$$

6.2.2.3 Relating M1 and M2 scores

Let $M_1(s_i, t_j)$ be the measure of the proximity between s_i and t_j definitions. We denote $M_2(s'_i, s'_j)$ the proximity of their respective encodings. C is the constraint the semantic space should minimize. We propose the distinction between two kinds of semantic spaces.

Distance-conservative semantic spaces (DCSS) aim to ensure the distance between two encodings is identical to one of the continuous proximity measurements (SSM, SCOM, Equation 6.7). Training a natural language model or a dimension reduction model to obtain a DCSS is a difficult task as it requires a training dataset associating to each couple of source target class an expected distance value. No such dataset currently exist for significant number of maps and would be difficult to define as current SSM and SCOM measures does not ensure the preservation of the triangular inequality: some inter-class distance might be mutually exclusive.

$$C_{DNCSS}(s'_i, t'_j) = |M_1(s_i, t_j) - M_2(s'_i, t'_j)|. \tag{6.7}$$

Neighbour-conservative semantic spaces (NCSS) aim to ensure that classes with close definitions are encoded more closely in the semantic space than classes with far definitions. NCSS only constraints the neighbourhood without constraining the distance: the distance between *Water* and *Cereals* can be 0.2 or 3000 as long as it is higher than the distance with *Maize*. This makes models targeting NCSS easy to optimize even for NCSS with low number of dimensions. We distinguish two subcategories of NCSS, Absolute-Neighbour conservative semantic spaces (ANCSS) and Relative-Neighbour conservative semantic spaces (RNCSS). For ANCSS, target classes partially corresponding to a given source class ($BOM = 1$) are closer to this class in the semantic space than any other target class (Equation 6.8). They distinguish close/far classes in a binary manner. For instance, *Cereals*, *Maize*, and *Tubers* should be closer to the source class *Agriculture* than any other, but there is no constraint on how far each class should be or if *Maize* is closer than *Cereals*. Conversely, RNCSS (see Equation 6.9) ensures that each target class encoding respects the correct proximity ordering to a source class by defining each neighbour relatively. For instance, *Maize* should be closer to *Agriculture* than *Cereals* which should be closer to *Agriculture* than *Water*.

$$C_{ANCSS}(s'_i, t'_j) = \begin{cases} M_2(s'_i, t'_j) & \text{if } BOM(s_i, t_j) = 0. \\ 0 & \text{else.} \end{cases} \tag{6.8}$$

$$C_{RNCSS}(s'_i, t'_j, t'_k) = \begin{cases} |M_2(s'_i, t'_j) - M_2(s'_i, t'_k)| & \text{if } M_2(s'_i, t'_j) > M_2(s'_i, t'_k) \ \& \ M_1(s_i, t_j) < M_1(s_i, t_k). \\ |M_2(s'_i, t'_j) - M_2(s'_i, t'_k)| & \text{if } M_2(s'_i, t'_j) < M_2(s'_i, t'_k) \ \& \ M_1(s_i, t_j) > M_1(s_i, t_k). \\ 0 & \textit{else.} \end{cases} \tag{6.9}$$

For land-cover translation, an ANCSS is sufficient as translating an unknown source nomenclature can be achieved by simply translating the unknown class similarly to the few closest classes in the SRS. As training, a model to produce an ANCSS is significantly simpler (does not require knowing the precise semantic distances, nor the relative ones), this manuscript only considers ANCSS.

6.2.2.4 Evaluation of the SRS

We evaluate the SRS on its ability to ensure that classes determined as a neighbour in the LCDD dataset (BOM=1) are also neighbours in the SRS. We first describe how we determined neighbour classes in the LCDD dataset and then introduce two quantitative metrics assessing the neighbour preservation quality.

Determining neighbours classes in the LCDD dataset Neighbourhood inside the LCDD dataset (BOM) is determined by hand independently for each source/target couple of nomenclature. Following notations of Section 2.1, we consider that S_i has a single neighbour T_j when they include the same objects (Equation 2.3) or when all the objects

of S_j are included in T_j (Equation 2.4). When T_j is included in S_i (Equation 2.5), S_i has multiple neighbours, all classes included in S_i. Without overlap (Equation 2.6), the classes are considered as non-neighbours. As mentioned in Section 2.1, determining by hand the pairs of classes establishing these relationships is relatively simple and leads most of the time in consensus results between several analysts [339].

The distinction of neighbour classes is more difficult to apprehend when a source class partially overlaps with several target classes [57] (see Equation 2.7). Generally, we consider here that all the target classes establishing a partial overlap are neighbours. However, we place two notable exceptions. When the partial overlap is due to a difference in the spatial resolution of the analysis, S_i may have the constraint of necessarily including a certain number of target classes (T_j, T_k...) without the converse being true. For instance, road and airport are partially overlapping as some road pixels do not belong to airports, and some airport pixels do not belong to roads. However, an airport necessarily contains roads (the runway), while the reverse statement is invalid. In this case, we consider an asymmetric link in which S_i (airport) is close to T_j (road) without the converse being true. The second exception is when the partial overlap is due to an apparent lack of correspondence between a source class and the set of target classes. The CLC *Wetland* class has no equivalent in the OSO, MOS, or OCSGE nomenclatures. Then, an asymmetric neighbour link is also used (*Wetland* is considered a neighbour with *Water* or *Herbaceous areas* in the OCSGE, while the converse is invalid).

We point out that determining a neighbourhood relationship between a source class and the classes of several nomenclatures simultaneously is not considered. This would imply some difficult determination such as the *Forest* of map A is closer to the *Forest* of map B than the *Forest* of map C. Consequently MOS *Water* necessarily has at least five neighbours among the MLULC nomenclature (one per land-cover map). Using the above protocol, we manually define the neighbors of the 169 classes of the LCDD independently for each target nomenclature (Appendix B).

Evaluation of SRS neighborhood preservation We derive two distinct neighbour conservation quality metrics (Figure 6.4). First, we define the **Closest Neighbour** metric (**CN**) as the proportion of source class encoding having for closest target encoding one of the expected neighbour (BOM=1). We simply average Equation 6.10 for all source and target couples. In practice, this metric directly reflects the proportion of classes for which the closest neighbour in the SRS is one of the possible neighbours.

$$\begin{aligned} CN(S,T) &= \frac{1}{c_S c_T} \sum_{i=1}^{c_S} \sum_{j=1}^{c_T} f(S_i, T_j) \\ \text{with} \quad f(S_i, T_j) &= \begin{cases} 1 & \text{if } argmin_{t'} M_2(s'_i, t') = t'_j \text{ and } BOM(s_i, t_j) = 1 \\ 0 & \text{else} \end{cases} \end{aligned} \tag{6.10}$$

Secondly, the **Neighbor preservation** (**NP**) metric evaluates the overall preservation of

all neighbours. The proportion of target encoding labeled as not neighbouring the encoded source class but closer to it than at least one of the expected neighbours. For instance, if the representation of *Cereals* has for nearest neighbour *Corn*, then *Water* and *Rapeseed*, whereas *Corn* must be the neighbour of *Corn* and *Rapeseed*, this measurement equals 2/3. In practice, a source class with x neighbours in the target nomenclature achieve a 1 NP when the x closest element of the source class in the SRS are only those neighbours. A measure close to 0 indicates that multiple non-expected neighbouring classes are closer than one of the expected neighbours.

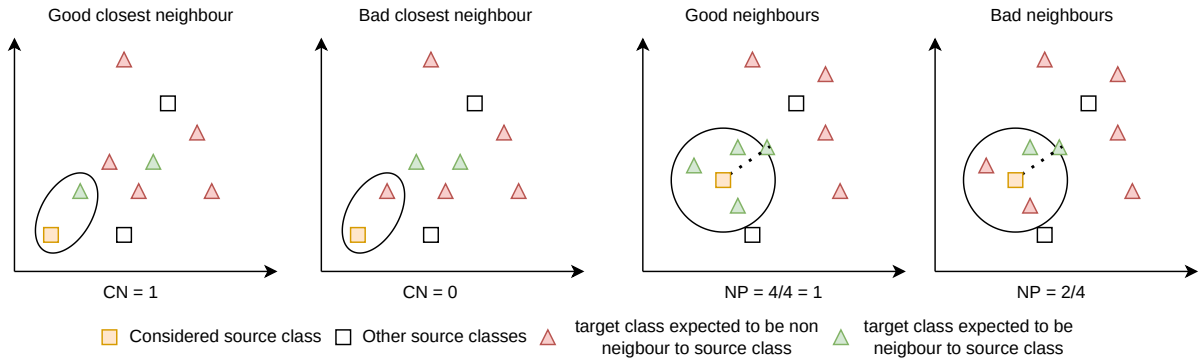

Figure 6.4: Illustration of the Closest Neighbor (CN) and Neighbor preservation metrics (NP) considering a single source class example. The metrics over a nomenclature are computed by the averaging the results over all source classes.

6.2.3 Constraints specific to the zero-shot use-case

This section identifies additional criteria when the SRS is aimed specifically to be used to encode map feed to a CNN to achieve zero-shot translation.

First, a model trained to project n nomenclatures in an SRS should be able to encode a new nomenclature unseen during training without changing the encoding of the n first nomenclatures. Machine learning-based models that project class definition into the SRS should thus be trained inductively and not transductively.

A second immediate consequence is that it must guarantee that the encoding of two different nomenclatures is not spatially disjoint. As nomenclatures have very different description schemes, the model is likely to encode different nomenclatures at different locations in the SRS space. This representation shift could prevent the network from interpolating the translation of the class of an unseen nomenclature based on the position of the seen one, as it could be encoded in a completely different way. We propose two simple unsupervised shift measurements, **Inertia** and **CH**, that aim to quantify the average shift between two nomenclature representations.

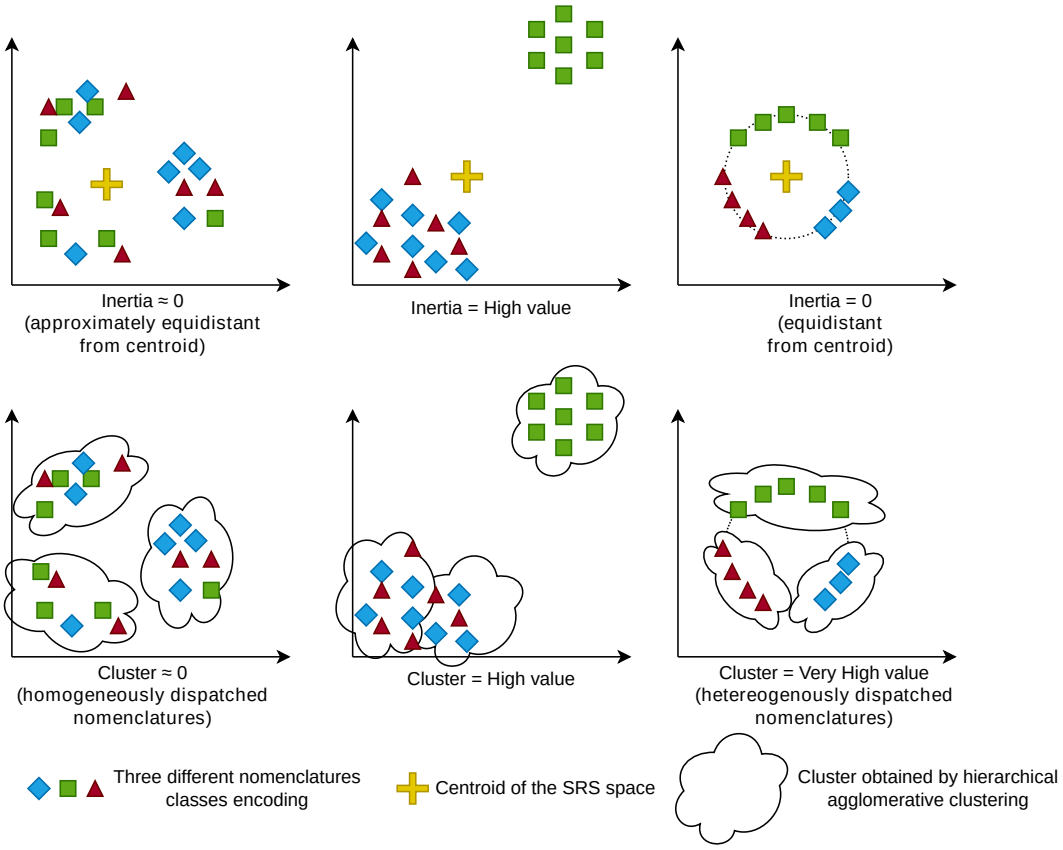

Figure 6.5: Illustration of the Inertia and Cluster homogeneity (CH) metrics (three nomenclatures). *Inertia* compares the average variation of distance of the nomenclatures to the SRS. *CH* evaluates the homogeneity of the nomenclature distribution into the SRS space. The desired SRS should look like the one in the first column.

Under a no-shift between nomenclature scenario, all nomenclatures should approximately have the same centroid coordinates as they are on average at the same location in the SRS (as illustrated in the first row, the first column of Figure 6.5). The inter-nomenclature **Inertia** measurement gives an order of magnitude of the dispersion around the empirical centre of SRS. It compares the average distance between the centroid of the SRS (centroid of all classes) and the classes of a given nomenclature with the average distance between the centroid and the classes of all other nomenclatures. We normalise this value in order to get a distance percentage of variation. For instance, a nomenclature A with Inertia of 0.1 indicates that the classes of A are, in average, 10% further from the centroid of all classes than the classes of other nomenclatures. Let V denote the centroid of all the classes independently from their nomenclature, and N denote the set of all the v nomenclatures available. We denote N_h the h^{th} nomenclature, N_{h_i} the i^{th} of the h^{th} nomenclature, and c_h the number of classes of N_h. Inertia is given by Equation 6.11. This metric is directly inspired by the notion of intra-class and inter-class inertia used by the clustering algorithm, which aims either to minimise intra-class inertia or to maximise inter-class inertia (achieving one objective achieves the other). Our inter-nomenclature inertia measure is equivalent to

the notion of inter-class inertia, but should be minimised rather than maximised.

$$Inertia(N_h) = \frac{\frac{1}{c_h}\sum_{i=1}^{c_h} M_2(N_{h_i}, V)}{\frac{1}{v}\sum_{\substack{j=1 \\ j\neq h}}^{v} \frac{1}{c_h}\sum_{i=1}^{c_h} M_2(N_{j_i}, V)}. \tag{6.11}$$

The second, based on the hierarchical clustering algorithm [222], aims to verify that the classes of each nomenclature are approximately homogeneously dispatched in the SRS. The hierarchical clustering algorithm is tasked to find the same number of clusters as the number of nomenclatures. In a homogeneous space, each cluster should include approximately the same proportion of each nomenclature as the proportion observed globally *e.g.*. If OSO represents 10% of the classes, each cluster should include approximately 10% of OSO classes. This shift measurement evaluates the distance to this homogeneity objective.
Let $P(N_h) = \frac{c_h}{\sum_{j=1}^{v} c_j}$ represent the proportion of classes belonging to the h nomenclature.
Let AG_i denote the i$^{\text{th}}$ cluster, $AG_i(N_h)$ denote the number of classes belonging to h nomenclatures in the i$^{\text{th}}$ cluster. The cluster homogeneity value **CH** (Equation 6.12) is only used to compare different SRS: there is no theoretical obligation that each cluster exhibits precisely the same proportion of each nomenclature. However, it can be used to compare different SRS with comparable CN or NP measurement: a lower cluster value indicates a more homogeneous space in which each nomenclatures is homogeneously dispatched in the space .

$$CH(N_h) = \sum_{i=1}^{v} M_2\left(\frac{AG_i(N_h)}{\sum_{j=1}^{v} AG_i(N_h)}, P(N_h)\right). \tag{6.12}$$

The **Neighbor Stability** (NS) evaluates the potential to interpolate encoding between two neighbouring classes while preserving neighbourhood links (Figure 6.6). Let s'_i and t'_j be two neighbouring classes. Any encoding between s'_i and t'_j should be closer to s'_i, t'_j or one of their neighbours than any non-neighbour classes. For instance, an encoding between *Cropland* and *Wheat* should be closer to *Corn* than *Water*. To verify this characteristic, we regularly sampled 20 points on a segment going from the source to the target class. We estimate the proportion of the segment closer to one of those neighbours by computing the proportion of those 20 points closer to one of the potential neighbours than a non-neighbouring element. A 100% NS implies that approximately 100% of the segment between two neighbouring elements respects the neighbour's preservation. In those conditions, one could imagine using simple arithmetic combinations between classes while ensuring that the results are still be semantically meaningful *e.g.* $\frac{Water+Crops}{2}$ should be closer to Rice than Forest when NS is high.

Lastly, we point out that an SRS need to have a small number of dimensions (less than

100) to avoid excessive memory consumption in perspective to encode land-cover map feed to CNN. To give an order of magnitude of the problem, a 600×600 pixel patch encoded in a 100-dimensional space using a float 32 precision represent a 1.15 GB memory usage alone. The 32GB GPU used in this thesis would only be able to load 26 patches at a time, decomposed in 13 input and 13 output, without even considering the memory consumption of the network computation itself. From a theoretical point of view, small dimensional SRS are sufficient to encode fairly the notion of proximity between classes, as evidenced by the encoding in only two dimensions proposed for two nomenclatures in Figure 6.1.

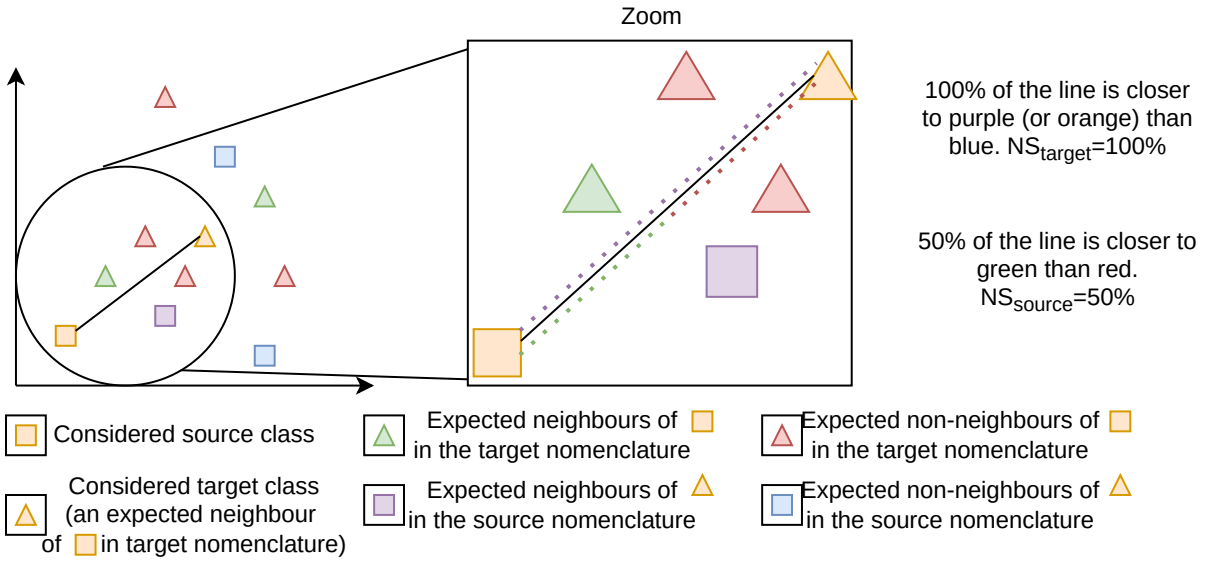

Figure 6.6: Neighbour Stability (NS) metric illustration between one source and one target class. The considered target class is one of the expected neighbors of the source. NS estimates the percentage of elements on a segment from source to target that are closer to one of the expected neighbours of source ($BOM(s_i, t_j) = 1$) in the target nomenclature NS_{source} than any other target class. Reciprocally NS_{target} estimates the elements that are closer to one of the expected neighbor of the target class in the source nomenclature ($(BOM(t_i, s_j) = 1)$ than any other source class.

6.3 How to built a SRS for land-cover translation ?

The main difficulty faced when aiming to obtain a model projecting land-cover classes in a continuous semantic representation space is the lack of wide land-cover definitions dataset preventing to train task-specific complex language model. To alleviate this issue we propose a two step procedure. First we project the land-cover definition into an SRS by using either a simple language model that can be trained on our small LCDD dataset or more complex models trained using generic text corpora unrelated to land-cover. We evaluate those models based on their capacity to respect the properties mentioned in the previous section. The different obtained SRS exhibits the two same limitations: (i) some

of the abstract variables used to encode classes takes into account concepts irrelevant from a land-cover point of view (ii) the SRS are to high dimensional to be usable to encode maps fed to a convolution neural network (>300 dimensions). Therefore, the second step compares different dimension reduction techniques aiming to recombine the numerous encoding variables into a smaller number of well designed one. In particular we explore a supervised dimension reduction technique trained to optimise the neighbourhood constraints defined in the previous section using the LCDD dataset. In the remainder, we denote HDSRS, the high-dimensional SRS obtained as the output of one of the three previous methods and LDSRS, the low-dimensional SRS we aim to obtain.

6.3.1 Using language models to obtain high-dimensional SRS

As only a few works have been conducted in obtaining SRS from land-cover definitions, this section evaluates the current main semantic embedding models. We first present a Bag-of-Word model proposed for land-cover translation by [58] that we train directly on the LCDD dataset. Then we present a Word2Vec model proposed by [246] for zero-shot land-cover semantic segmentation from images trained on large text corpora. Lastly, we propose replacing those two models (the only one found in the literature for land-cover class embedding) with a transformer-based architecture trained on text corpora that we believe more adapted to work the complex sentences of class definitions.

6.3.1.1 Bag-of-words

Overall idea The bag-of-word approach [112] (BOW) proposed by [58] is the only model applied to the land-cover translation task in the literature. This model determines all the words used in the land-cover definitions to build a land-cover term dictionary. Each definition is then encoded as the count of the words inside it. For instance, in a setup where the only available land cover definitions are (5) and (6), the dictionary is {"Forest","natural","tree","vegetation","with","higher","than", "5m","Orchards","Cultivated"}.

"Forest : natural tree vegetation with tree higher than 5m" (5)

"Orchards: cultivated tree " (6)

Using each word of the full dictionary as one of the encoding dimensions of the BoW with a value equal to the number of word occurrences in the definition, we obtain Figure 6.7.

	Forest	natural	tree	vegetation	with	higher	than	5m	Orchards	Cultivated
Forest	1	1	2	1	1	1	1	1	0	0
Orchards	0	0	1	0	0	0	0	0	1	1

Figure 6.7: Bag-of-word encoding of two schematic classes Forest (5) and Orchards (6)

In practice, this encoding is very sensible to the class definition length thus [58] normalise the encoding values by the number of words in the class definition *e.g.* Orchard encoding becomes $Orchards = [0, 0, \frac{1}{3}, 0, 0, 0, 0, 0, \frac{1}{3}, \frac{1}{3}]$. In order to emphasise the importance of class-specific words, they adopted a common total frequency times inverse document frequency weighting scheme. Each term is weighted such that a term that frequently appears in one definition but not in the other exhibits a high weight. Let $n_{s_i}(t)$ be the number of times the word t appears in the definition of class i, L_{s_i} be the number of words in class i definition, $h(t)$ be the number of classes including t in their definitions and D be the total number of classes. The resulting weight is obtained using Equation 6.13.

$$s'_i(t) = \frac{n_{s_i}(t)}{L_{s_i}} ln \frac{D}{h(t)} \tag{6.13}$$

Implementation details We follow [58]. Each class description is first converted into a list of terms that are used to build the dictionary. In most cases, a term represents a single word but it also includes small phrases for distances/surfaces informations (*e.g.* "5 m"), and locations (*e.g.* "Alpine grasslands"). Unlike [58], we add a few more constraints due to the small number of definitions and their limited size: each word is considered case insensitive, and plural and singular words are considered as a single key. This results in a 987-term dictionary inducing a 987-dimension HDSRS.

Strengths and weaknesses of the method The core advantage is that the model is directly built on existing land-cover definitions. The HDSRS only considers concepts related to land cover and ignores unrelated concepts. However, this approach suffers multiple limitations:

- the method is entirely dependent on the exhaustivity of the dictionary. As we considered that the HDSRS should be obtained inductively to enable zero shot land-cover translation, a nomenclature not used to build the HDSRS can include terms not encodable using the original dictionary. This results in a partial loss of class information.

- the method ignores grammar and order. This prevents learning concepts such as inclusion and exclusion. For instance, a statement such as "this class includes A and excludes B" is encoded as the reverse statement.

- the distance is based on the weighted co-occurrence of words between the definitions. It does not consider word meaning or semantic importance. Classes "Forest: high trees" and "Conifers: needle-leaved stands" have no common words but describe the same object. Consequently, the notion of threshold is also not encoded. For instance, a Forest defined as "with at least 80% coverage" is not closer from one with a 70% threshold than one with a 10%.

6.3.1.2 Word2Vec

Overall idea Instead of producing an encoding per class based on word frequencies like the BOW, the Word2Vec approach [215], aims to provide an encoding per word taking into account semantic meaning. The core idea is that a word semantic meaning can be inferred by analysing the words frequently used next to it. For instance, *Coniferous* is often located close to the word *tree* in a sentence, semantically linking those two terms. Word2Vec have been explored to perform many zero-shot classification tasks, especially semantic segmentation [32]. To our knowledge, the only attempt in the land-cover field deals with remote-sensed image zero-shot semantic segmentation [246]. This work did not focus on translation and encoded solely the class name and not the class definition, making it hardly generalisable to translation: many classes can have the same name (thus the same encoding) but different content.

The word embedding procedure consists in building a dictionary of all possible words leading to a one-hot encoded version of each word (the same 10-word dictionary as above results in $tree = [0, 0, 1, 0, 0, 0, 0, 0, 0, 0]$). A small 1-hidden layer perceptron processes sentences with a sliding window trying to predict the n^{th} word given the x previous and x later words. For instance, in the previous example, using $x = 2$, one iteration of the algorithm on "Forest : natural tree vegetation with trees higher than 5m" could consist in training the perceptron to predict *vegetation* given the 2 previous (*natural*, *tree*) and 2 later words (*with*, *tree*) (Figure 6.8). The first layer of the perceptron processes each of the four one-hot encodings iteratively in a d-dimensional space (using d neurons). The four d-dimensional vectors are averaged and fed to the second linear-layer, predicting the middle word.

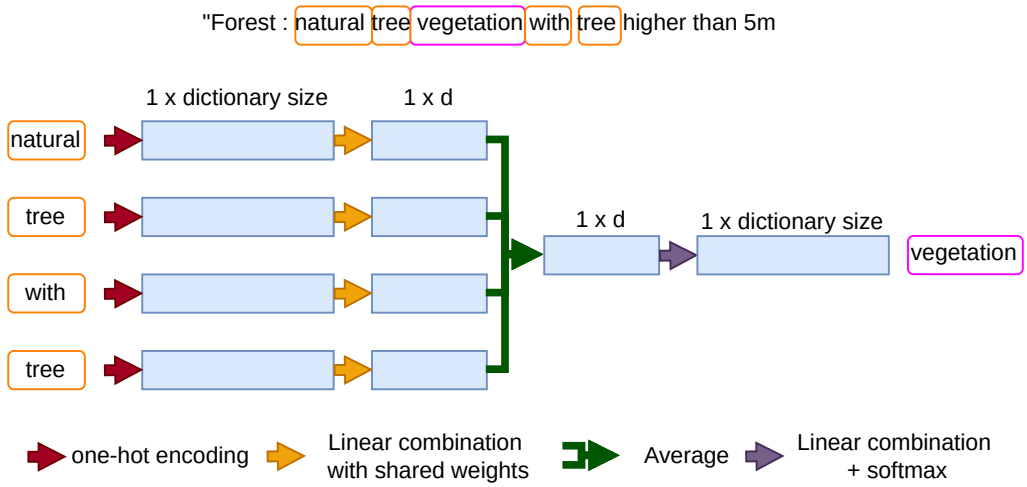

Figure 6.8: Illustration of the Word2Vec algorithm using a window of size 2 . Once trained, the word embedding is the output of the first linear layer (orange arrows).

Once trained on several sentences, the word embedding is then considered to be the d-dimensional space obtained after the first layer. We obtain the class definition encoding

as the average of all its words encoding.

Implementation details We used the Word2Vec implementation available at[2]. The model is pre-trained on the Google News datasets, including more than 300 billion words. As for the BoW method, we observe that some word combinations are essential. However, instead of defining manually a set of rules of which words should be gathered, the Word2Vec developers identify terms that are frequently similarly ordered based on n-gram models [214]. The pre-trained Word2Vec model uses a 10-dimensional window and encoded in a 300-dimensional HDSRS.

Strengths and weaknesses This method partially alleviates the BoW limitations by replacing definition encoding based on terms frequency with an average of estimated word semantic meaning. The computed distance in the semantic space conveys more semantic meaning than the BoW approach. However, Word2Vec ability to reflect semantic meaning widely depends on the window size, the number of words used for training and d. The small size of the LCDD dataset prevents from training the Word2Vec method: a pre-trained Word2Vec is thus used. Some of the resulting encoding dimensions might be irrelevant for land cover as the pretraining is conducted on a generic dataset. For instance the closest term embedding of *Coniferous* is *In your garden* and the second closest of Ocean is *Neptune*. Additionally, this method still disregards word order as the class encoded is realised by averaging all the words encoding of the definition independently from their order.

6.3.1.3 Transformer based encoding

Overall idea Alleviating the limitations of the previous methods requires: (i) developing a method to address words outside the dictionary encountered at inference, (ii) taking into account word ordering, (iii) proposing a context-wise encoding (*e.g.* the terms *water* in the definition of *rice*, *wetland*, *water course*, and *sea* have close yet not identical meaning). In 2018, BERT was introduced in [66] to tackles all those issues simultaneously by combining several independent works handling each of those problems. Since various models have been proposed to improve the results but are based on the same overall idea. For clarity, we introduce concepts based on the original Bert implementation. We then discuss the difference with the recent model we used.

First they rely on tokenisation, *i.e.* sub-words units. The most straightforward tokenisation algorithm is to work directly on a per-character basis [50]. The dictionary includes each character of the training set individually. Per-character dictionaries are often complete even when using a small training dataset, *e.g.* considering only letters (26 characters). They do not suffer from out-of-dictionary issues at inference. However, each element of the dictionary conveys a very high diversity of semantic meaning depending on its context, making per-character tokenisation hardly usable. For instance, the letter t in the word

[2] `https://radimrehurek.com/gensim/models/word2vec.html`

tree or *water* has a completely different meaning. Various strategies proposing an trade-off between word and character level tokenisation have been proposed based on sub-word decomposition, $water = wa + ter$ for instance. The BERT and subsequent methods are mostly based on the sub-word tokenisation algorithm proposed by [331].

Secondly, to consider word ordering and achieve context-wide encoding, the BERT paper adopts the transformer architecture proposed by [318] that incorporates the positional encoding module. The key contribution of the BERT is to propose a way to consider the word context in a non-directional manner. They replace the traditional natural language processing objective of predicting the next word given the sequence of the previous ones with the task of predicting randomly masked words on the whole sentence. We underline that those architectures jointly produce an encoding per-word context-wisely (the same word is encoded differently depending on the sentence) as well as a sentence encoding. Hereafter, we consider that a class definition encoding is the average of all the word encodings.

The transformer architecture often comes with a hundred million parameters (approximately 300 million for BERT's original implementation) making them untrainable on the tiny LCDD dataset due to overfitting concerns. Thus we rely on an available pre-trained model. We underline that comparing the available models is difficult as most of them are assessed on different tasks and datasets.

Implementation details As comparing all the available pre-trained models is unfeasible, we rely on a comparison of different top-performing models on 20 different datasets conducted by [251] and regularly updated[3]. We choose the MINILM [325] model, whose training uses two networks, a teacher (a regular BERT), and a student (a small BERT), enforcing the student to have the same attention parameters as the teacher. This distillation-based principle [122] enables a small student network to achieve comparable results to the big one. We rely on the student network. The implementation used is the official implementation provided by the hugging face library [4] trained on a 1Billion word dataset [5]. This results in a 387-dimension HDSRS.

Strengths and weaknesses Even though this strategy alleviates most of the limitations encountered by the previous methods, the unavailability of a large land-cover-oriented dataset prevents from training on pure land-cover-oriented tasks: the distances computed in the semantic space still consider some irrelevant concepts.

[3] `https://www.sbert.net/docs/pretrained_models.html`

[4] `https://huggingface.co/microsoft/MiniLM-L12-H384-uncased?text=I+like+you.+I+love+you`

[5] `https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2`

6.3.1.4 Comparison on the LCDD dataset

Quantitative assessment Unlike the two pretrained models (Word2Vec and MiniLM), the BoW is built using the LCDD dataset. To simulate the operational setup of a zero-shot land-cover translation, we iteratively use nine nomenclatures of the LCDD to train the model and use the last remaining one (termed *unseen nomenclature*) at inference. As some words of the unseen nomenclature can be out of the BoW dictionary, each model is different.

We compute the metrics mentioned in Section 6.2 for each model considering the unseen nomenclature as the source or the target. For instance, a BoW model trained using all nomenclatures except CLC, is evaluated using the metrics considering CLC either as the source or target nomenclature (CN_{source}, CN_{target}). We resume the results of those metrics using different M_2 measures in Table 6.1. Results from Inertia and CH values describe different notions depending on the M_2 measures. An inertia of 0.12 computed using Euclidean distance implies that in average the unseen nomenclature exhibits a 12% difference in terms of distance to the HDSRS centroid compared to seen nomenclatures. A 0.12 inertia computed using cosine distance implies a 12% angular variation. Thus, the best Inertia and CH values are compared independently per metric in this case. Conversely CN, NP and NS which are percentage based on M_2 metrics are comparable even when using different metrics thus we jointly assessed their best value for all metrics.

First, as the M_2 measure maximising the CN, NP and NS is almost exclusively the Cosine Distance, we consider it the most relevant metric. We underline that this was expected as the three spaces are high dimensional (from 300 to 987 dimensions) and some have been explicitly optimised for this metric (Word2Vec). Thus all the following observations are conducted by considering the results obtained with cosine distances.

A second observation is that the transformer-based encoding better preserves the closest neighbour (CN), the nearest neighbours (NP), and the overall shape (NS) at the cost of a worst inertia and CH value than the BoW approach. Therefore the transformer produces a better HDSRS from a pure proximity preserving point of view but introduces a slight semantic shift between land-cover maps. As the difference in CN and NN is very high (10 to 15% higher), we still consider Transformers as the most suitable solution.

		BOW			PreWord2Vec			MiniLM		
		euclidean	manhattan	cosine	euclidean	manhattan	cosine	euclidean	manhattan	cosine
CN	source	39	37	60	58	59	64	60	59	**76**
	target	32	37	58	58	59	64	60	59	**76**
NP	source	46	47	63	62	61	67	62	62	**76**
	target	43	47	61	62	61	67	62	62	**76**
NS	source	83	91	**96**	88	88	92	90	90	95
	target	59	64	68	65	65	67	66	66	**70**
Inertia		**0.13**	**0.05**	**0.07**	0.16	0.16	0.15	0.14	0.14	0.18
CH		0.4	**0.13**	**0.12**	0.31	0.38	0.278	**0.21**	0.47	0.17

Table 6.1: Quantitative of the three HDSRS obtained by three different semantic embedding methods

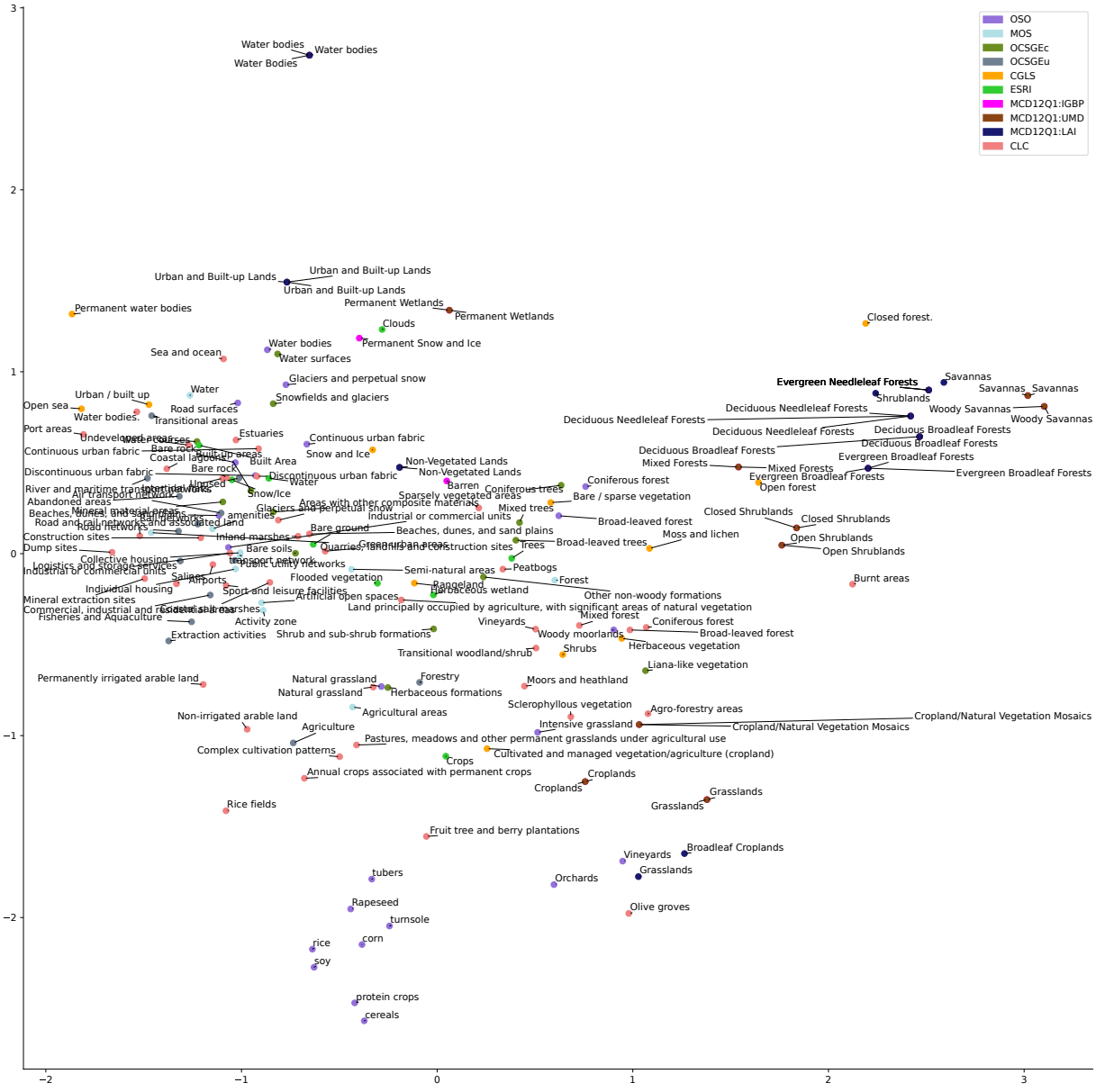

Figure 6.9: PCA representation of the HDSRS obtained with a trained on generic text corpora Transformer.

Qualitative assessment Two different dimension reduction techniques are used to evaluate qualitatively those HDSRS. PCA is used to evaluate the global aspect of the HDSRS as it preserves approximately all pairwise distances. Conversely, T-SNE preserves the topology of the neighbourhood.. Only the results for the transformer-based method are illustrated in the main text (see Appendix J for other configurations).

Globally the HDSRS appears well structured (see Figure 6.9) with clusters of similar land-cover nature: croplands are mainly concentrated in the bottom part of the figure, forest on the right side, unvegetated areas in the middle and urban areas on the left side. However, many local inconsistencies can be detected, *e.g.* MCD12Q1 *Urban and Built-up Lands* is surrounded by water areas. Additionally a slight semantic shift between the

various maps is observed. For instance, MOS and the two OCSGE are only represented on the left side, MCD12Q1 on the right. Some clusters are also observed, such as the OSO crops, far from the rest of arable crops and next to *fruit tree and berry plantation* of CLC. This shift between nomenclatures directly stems from nomenclature-specific ways to describe comparable land-cover. For instance, the CLC nomenclature details many inclusion and exclusion criteria, while those notions do not appear in CGLS.

Figure 6.10: T-SNE representation of the HDSRS obtained with a trained on generic text corpora Transformer.

Figure 6.10 illustrates that despite having coherent local neighbours association from a generic point of view, observed association are not necessarily those expected from a land-cover one. For instance, *Natural grassland* and *Pasture* are nearest neighbors in the

HDSRS while most semantic based standardisation system would have consider them far apart as one describes a natural area and the other croplands.

The same observation can be drawn for the Word2Vec encodings provided (Appendix J). The BoW encodings distinguishes itself from the others by a pronounced outlier like visual aspect underlying that a few classes are encoded very differently than the others. The BoW on average better Inertia and CH thus hides that a few rare classes are encoded poorly. We link this behaviour to the use of word proportion instead of semantic meaning: some classes exhibits unique words resulting in a specific encoding.

6.3.2 SRS optimisation through dimension reduction

6.3.2.1 Motivation

The initial results of Step 1 are already satisfactory albeit neighborhood conservation is not guaranteed. The main reason is that the encoding task is not land-cover specific. It could be interesting to only extract the relevant features from the 384 dimensions of the SRS. In particular since such a high dimension leads to technical constraints (32GB memory GPU). We propose to investigate how to reduce the number of dimensions while keeping only land-cover relevant elements.

Numerous techniques have been proposed so far [316]: feature selection (some dimensions are removed), matrix projection (linear combination of dimensions), manifold learning (non-linear combination), or auto-encoders (the data is reprojected in fewer dimensions with the constrain of ensuring that the original space can be retrieved). Most are self-supervised exhibiting principally three different objective functions: preserving the distance observed between elements in the HDSRS in the LDSRS (*e.g.* Isomap [300], Locally Linear Embedding [259]), the nearest neighbours (Neighborhood Preserving Embedding [116]), the information (auto-encoder reconstructs the original HDSRS from the LDSRS). All those objectives try to keep the information of the HDSRS while we point out that many of them are irrelevant from a land-cover point of view. Therefore the LDSRS obtained by those techniques is unlikely to exhbits better metrics than the HDSRS it stems from.

Instead, we propose to investigate supervised dimension reduction techniques using the LCDD dataset to keep only relevant information. Unlike self-supervised techniques which target that closeby elements in the HDSRS remain closeby in the LDSRS, we directly target that closeby elements in LDSRS are closeby in terms of definition (BOM). Additionally, the proposed method must ensure that no shift is observed between nomenclatures to ensure good zero-shot properties. As the dimensionality reduction techniques should be usable under the zero-shot translation setup, the dimension reduction model should be trained inductively. Consequently, transductive solutions as T-SNE will not be discussed.

6.3.2.2 Method

Overall idea Instead of preserving HDSRS neighbours like self supervised methods, we propose supervising the dimensional reduction by enforcing the preservation of neighbours in terms of class definition. Thus we aim to create an Absolute neighbor-preserving space. This constraint is implemented by training a dimension reduction model with the same information on the neighbourhood (BOM) required to compute the CN metric. For instance, the dimension reduction model is trained to represent closely in the LDSRS OSO *Broad-leaved* and MOS *Forest* as we labelled them as a neighbour (BOM=1). We detail the chosen dimension reduction model and how to apply neighbourhood constraints on the LDSRS below. Figure 6.11 illustrates the overall framework.

Dimension reduction model Th model design is based on two principal observations. First, the model is only trained using a small dataset (less than 169 class definitions, as one nomenclature is kept for testing) on a high dimensional space (at least 300). Therefore, the model should exhibit a small number of learnable parameters to limit overfitting. Secondly, the model should be able to apply a non-linear transformation in order to obtain better neighbour preservation in the LDSRS than those observed in the HDSRS. We rely on a small backbone based on a simple 1-hidden layer perceptron. This MLP takes for input the 384 dimensions and reduces them to x-dimension with $x < 100$. The best x value is experimentally determined and presented in the next section.

We adopt an adversarial training setup, in which a second MLP is tasked to predict the land-cover nomenclature from its encoding in the LDSRS. This setup drastically reduces the risk of nomenclature shift in the resulting LDSRS. The two MLP include dropout layers with high dropout rates (0.3) to regularize the loss and limit overfitting.

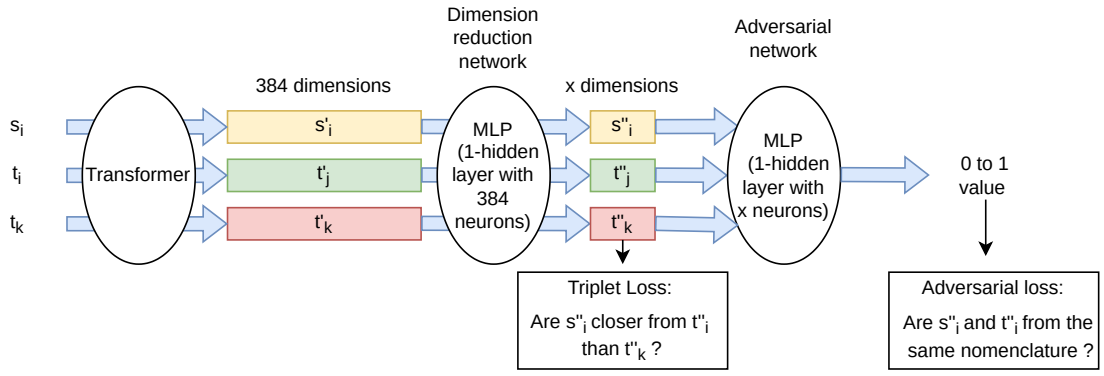

Figure 6.11: Overall dimension reduction architecture framework in the case a source class s_i expected to be closer from t_j than t_k. During optimisation the transformer parameters are fixed (only the two MLP are trained)

Loss function The model should learn the notion of neighbour in a binary fashion. Let s''_i denote the representation of source class i in the LDSRS. The loss uses three elements: the source class representation s''_i, a neighbour target t''_j, and a non-neighbour one t''_k.

We aim to obtain $M_2(s_i'', t_j'') < M_2(s_i'', t_k'')$. Instead of constraining the distance value, this constraint focuses on pairwise relative value, which is less restrictive. This weakly constrained optimisation enables obtaining low dimension SRS pretty easily. We implement this distance constraint using the Triplet Loss proposed by [270] given in Equation 6.14. ϵ represents a margin value arbitrarily fixed. We underline that in its original form, the M_2 metric used is the Euclidean distance, but any other distance metric could also be used.

$$TripletLoss_{M_2}(s_i'', t_j'', t_k'') = max(0, M_2(s_i'', t_j'') - M_2(s_i'', t_k'') + \epsilon). \tag{6.14}$$

Optimising both Euclidean and Cosine distance is possible without deteriorating the quality of the results, we combine two triplet loss functions to enable using both metrics in Equation 6.15:

$$TripletLoss(s_i'', t_j'', t_k'') = TripletLoss_{Euclidean}(s_i'', t_j'', t_k'') + TripletLoss_{Cosine}(s_i'', t_j'', t_k'') \tag{6.15}$$

The arbitrary margin value is set to 1 for Euclidean and 0.5 for cosine. The results are insensible to the margin for Euclidean TripletLoss but must be between 0 and 1 for the cosine as a cosine between 0 and one denotes positively correlated variables.

The adversarial training ensures that the classes of different nomenclatures are mapped into the same part of the space. The dimension reduction MLP is optimised to make the Adversarial MLP predicts that the source and target encodings belong to the same nomenclature ($Adversarial_{mistake}$). In contrast, the adversarial MLP is optimised to predict the contrary ($Adversarial_{true}$). Let $A(s_i'', t_j'', t_k'')$ denote the output of the adversarial network. We will consider that the Adversarial network should output 1 when considering two classes belonging to the same nomenclature and 0 otherwise. $Adversarial_{mistake}$ and $Adversarial_{true}$ are given in Equation 6.16 and 6.17.

$$Adversarial_{mistake}(s_i'', t_j'') = L_{CE}(A(s_i'', t_j''), 1). \tag{6.16}$$

$$Adversarial_{true}(s_i'', t_j'', t_k'') = L_{CE}(A(s_i'', t_j''), 0) + L_{CE}(A(s_i'', t_k''), 1). \tag{6.17}$$

To sum up, the dimension reduction MLP is optimised using the full loss (Equation 6.18), while the adversarial network is only optimised using the $Adversarial_{true}$ loss.

$$DimRedLoss(s_i'', t_j'', t_k'') = TripletLoss(s_i'', t_j'', t_k'') + Adversarial_{mistake}(s_i'', t_j''). \tag{6.18}$$

Training procedure The model is trained iteratively using nine of the ten nomenclatures and we evaluate the quality metric on the last nomenclature. The training is carried out at each iteration on approximately 150 definitions embedded into the 384 dimensions. Therefore the model is inherently overfitting despite its small size and the dropout regularisation.

Comparison with other dimension reduction techniques We compare our method to the two most commonly used unsupervised dimension techniques: (i) Principal component analysis (PCA) which linearly combines the HDSRS dimensions in a few new uncorrelated dimensions and (ii) Isomap, a non-linear combination of dimension that aims to preserve local neighbourhoods based on geodesic distance measurement.

We also compare to a supervised dimension reduction technique called Latent Dirichlet allocation (LDA). LDA projects the input data into a linear subspace made of the directions maximising the separation between labeled groups. As we do not have access to groups but to neighbours, we define a group as a set of classes of different nomenclatures that are all pairwise neighbours. For instance, OSO Coniferous, MOS forest and CGLS Open Forest are all pairwise neighbours, which can be considered as one group. In order to optimise the results, we reduce the number of classes belonging to multiple groups as much as possible. In particular, we will consider that two classes of the same nomenclature can belong to the same group provided they are neighbours of the same classes in the other nomenclature *e.g.* OSO *Broad-leaved* and *Coniferous* belong to the same group.

6.3.2.3 Results

We first evaluate the number of dimensions required to achieve the best metrics for each of the methods (Figure 6.12). PCA quickly reaches a plateau with few dimensions (around 10), with metrics very close to those obtained for the HDSRS space. The ISOMAP method performs significantly worse than the other methods. PCA performs well but does not provide a LDSRS with better characteristics than the HDSRS.

LDA method achieves better metrics than those of the HDSRS making it an adequate solution. The method is very efficient in small dimension space (10 to 20) and worsens when more dimensions are used, especially Inertia and CH metrics. Under 40 dimensions, the CH metric is better than the MLP one, indicating that the classes of different nomenclatures are slightly more homogeneously distributed.

The MLP dimension reduction consistently outperforms other solutions on all other metrics. Interestingly this statement holds even when comparing the 5-dimensional space obtained by MLP with less reduced space obtained by other techniques. Even though the best results with the MLP method are observed for 60 dimensions, the difference with the 5-dimensional space is minimal $\pm$ 2% for all metrics. In particular, we observe an improvement of 10% in CN, and NP compared to the metrics computed directly in the HDSRS space.

We underline that the average 85% CN is significantly lower from the metrics observed on the training nomenclatures (97% CN with five dimensions, 99% with 100). This is easily understandable as the tiny size of our dataset inherently leads to overfitting. Thus, results could be significantly higher with a more extensive dataset.

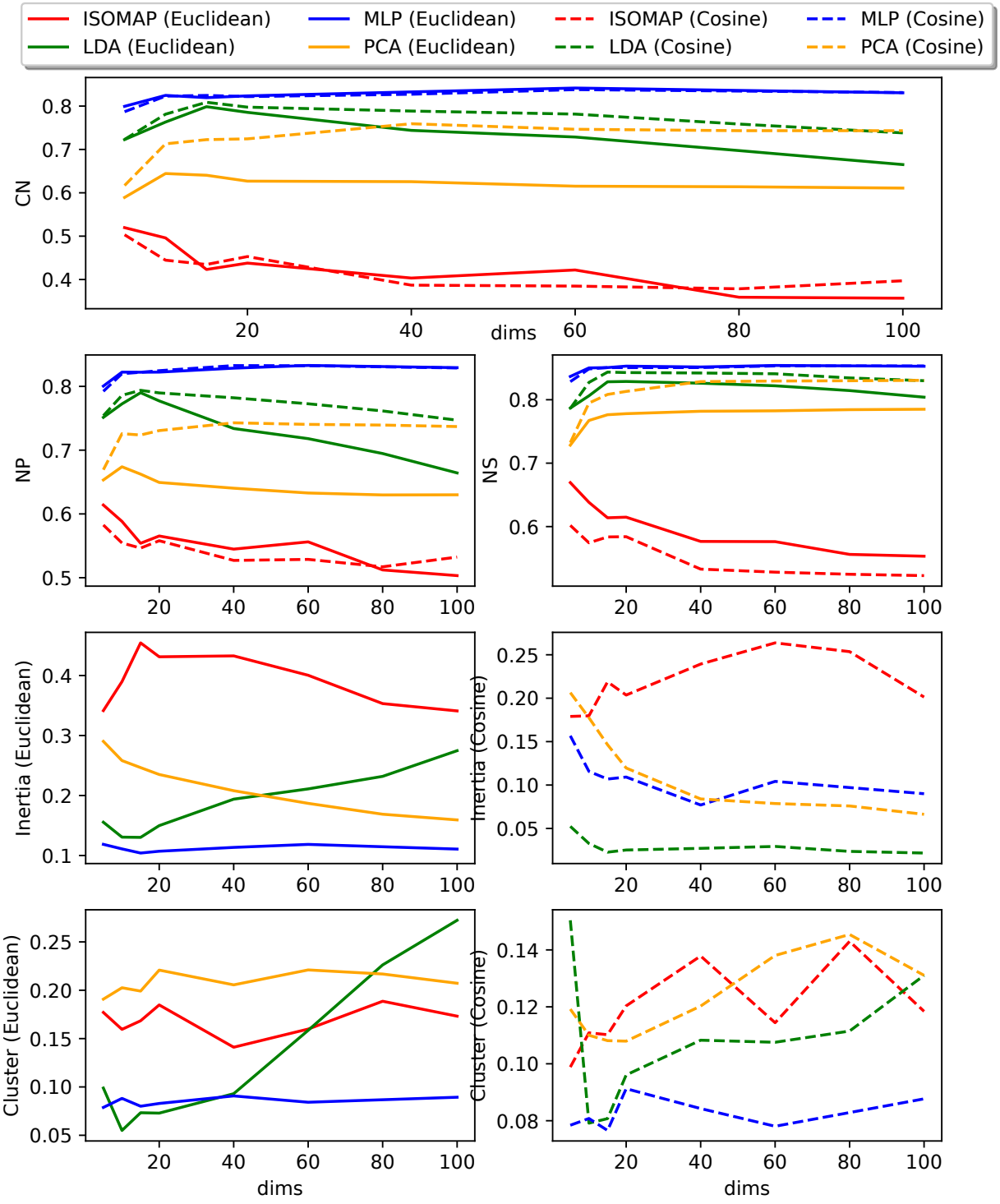

Figure 6.12: Dimension reduction technique metrics under zero-shot constraint depending on the number of output dimensions. For Inertia and CH, lower values denote the best metrics. For clarity metrics from source and target point of view are averaged (*e.g.* $CN = (CN_{source} + CN_{target})/2$)

6.4 Applications to land-cover translation

We highlighted two applications of SRS: zero-shot source and zero-shot target translations. At the time of writing, zero-shot target is still under experiments and will be included later in Appendix. This Section focuses solely on the zero-shot source, *i.e.* encoding the input land-cover maps into the SRS and training a network to translate them into various targets. At inference time, unseen source maps are encoded and provided to the network to translate them into one of the target map seen during training. It avoids training new translation models when an unseen source map is provided. Additionally, as the unseen map is never used for training, it alleviates the constraint of translating only source maps with partial spatial overlap. The following section first presents the architecture and training procedure used for our zero-shot source configuration. Results and comparison with other methods are discussed in the second part.

6.4.1 Using SRS for zero-shot source translation

6.4.1.1 Architecture

The overall architecture, termed **OneEncoder**(Figure 6.13), encode each map into a 15-dimension SRS. For instance, all pixels of OSO forest are replaced by the semantic encoding of forest obtained from the Transformer+MLP. Secondly, All maps are resampled using a bilinear resampling algorithm to the maximum resolution (10x10m). We underline that using a continuous resampling algorithm instead of a discrete one (nearest neighbour) is directly made possible by semantic encoding. Lastly, the maps are fed to a single U-Net encoder and then decoded by map-specific decoders. The use of a single-encoder instead of multiple ones like for the **MLCT-Net** (Section 5.1) is made possible by the use of semantically encoded inputs instead of one-hot encodings.

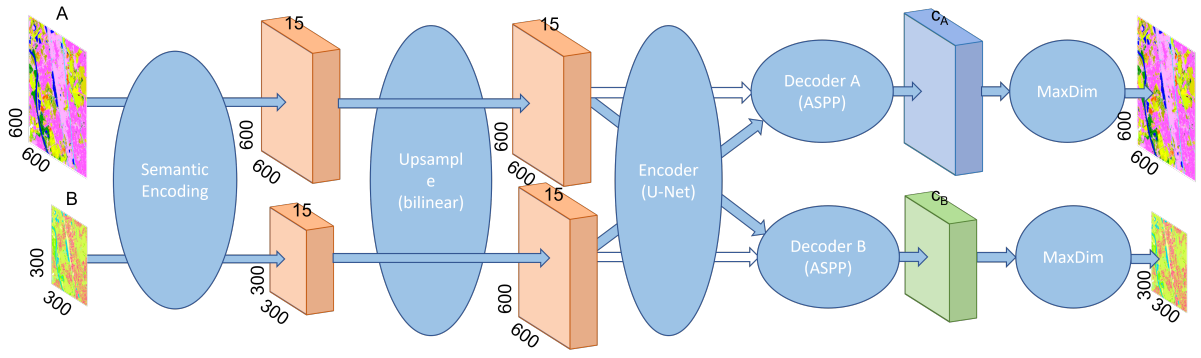

Figure 6.13: OneEncoder architecture using a 15 dimensional semantic representation space (in orange). A unique encoder is used for all maps while decoders are target specific. Only 2 maps are represented for readability but the architecture can take any number of maps. White arrows denote the self-reconstruction goal.

6.4.1.2 Losses

Similarly to the **MLCT-Net**, we choose to perform q dual objective training: translation and self-reconstruction. Self-reconstruction is important as it enforces the encoder to avoid losing classes that have no direct equivalent in the other training nomenclatures (such as CLC Wetlands). The total loss is obtained by summing the translation and self reconstruction loss computed using Cross-Entropy. Let S be the source map and T be one of the target maps. Let $E(S)$ denote the encoded version of S and $D_T(E(S))$ denote this encoded-decoded into the T nomenclature, the loss is given in Equation 6.19).

$$L(S,T) = L_{tran} + L_{self-rec} = L_{CE}(D_T(E(S)), T) + L_{CE}(D_S(E(S)), S). \tag{6.19}$$

6.4.1.3 Training procedure

In an ideal setup, the SRS dimension reduction MLP and the network would be trained with hundreds of different maps, enabling the encoder to be trained with thousands of different encoding values. However, the dimension reduction MLP is trained only on 9 nomenclatures (all the LCDD nomenclatures except the one used at inference.) Similarly, the network is only trained on 5 out of the 6 land-cover maps of the MLULC dataset. Consequently, the network is likely to quickly overfit the training data, restraining the ability to analyse unseen during training semantic encodings.

A noise-based data augmentation technique is used to artificially compensate for the small number of maps. Slight jitter is added around each embedding of the HDSRS to regularize the learning of the MLP. We assume that for a given source nomenclature, all the elements closer to class s_i than all the other classes of the source nomenclature belong to this class. Instead of encoding a map with c classes with c distinct semantic encodings, the c classes can be encoded with an infinity of encodings that respect the previous distance constraint. In practice, each source patch is semantically encoded with this random distance-constrained noise using a single noisy-encoded value for each pixels of a given class. Conversely, the same class is encoded differently in two different patches.

This noise is strong from a translation point of view, as it might change the closer class in the target nomenclature. This leads to insert a additional noise constraint. Since the SRS is built using a triplet loss, all elements below half the margin distance from the source encoding are closer to one of the original neighbours than to the non-neighbour elements. By combining with the previous noise definition, we constraint the noise to both ensure that the noisy-encoding is closer to the original encoding than any other source class encoding and at a maximum distance from it of half the margin.

We will denote **OneEncoder** the OneEncoder method used at inference on the same nomenclature used for training (not zero-shot like in previous chapters). We denote **ZeroShot** the OneEncoder used at inference on a different nomenclature than those used for training (zero-shot source).

6.4.1.4 Comparisons

SemEnc computes the translation directly from the semantically encoded source map using the Transformer followed by the MLP. Each source class is translated into the closest target class in the SRS without considering the spatial context. It is analogous to the **HardSem** method (Section 4.1) as the translation is performed at the nomenclature level (all pixel of source class i are translated the same way and based on semantic assumptions, here on proximity in the SRS space).

HardSem is the only baseline that does not require the computation of a model based on pairs of spatially overlapping patches and is thus used as the current reference in terms of zero-shot translation.

HardStat aims to associate each source class to each most frequently co-occurring targets. This method is not zero-shot, as the co-occurrence is directly determined from pairs of overlapping maps. It provides an interesting comparison to evaluate if the zero-shot source context-wise model only uses the knowledge acquired during training on target class probability of occurrence. For instance, a model trained to perform CLC/OSO translations will learn that amongst the 11 agricultural classes of the OSO map *Cereals* and *Pastures* are the most frequent. At inference, when translating MOS to OSO, the model could preferably translate all MOS *Crops* pixels to *Cereals* based on this prior knowledge.

The **MLCT-Net** provides an intuition on the upper bound of obtainable results as trained on all nomenclatures. At inference, with unseen source map, **ZeroShot** results should be at best equivalent but more probably lower than the **MLCT-Net**. Conversely, when processing seen source nomenclatures, **OneEncoder** and **MLCT-Net** results should be equivalent. Obtaining on-par results on seen during training maps would indicate that using a single encoder is sufficient to encode any map despite a resolution shift.

6.4.2 Results

6.4.2.1 Qualitative analysis

Figure 6.14 and Figure 6.15 present the results of the zero-shot translation. **SemEnc** results illustrate well the quality and limitation of the SRS space as they are obtained by associating each source class to the closest target class. First, **SemEnc** translations are often different from those of **HardSem**. Most of the time, differences are observed for source classes that establish multiple semantically equivalent translations. For instance, CGLS *Cropland* are respectively translated into OSO *Cereals* or *Tubers* by **HardSem** and **SemEnc** (second column of Figure 6.14). From a purely semantic point of view, any of those translations is equally valid (as the translation into *Soy* or *Corn*). In practice, one unique translation must be chosen when performing non context-wise translations. We underlined in Section 4.1 that we made the arbitrary choice to choose the most frequently observed class in order to maximize the *HardSem* statistic when multiple equally valid

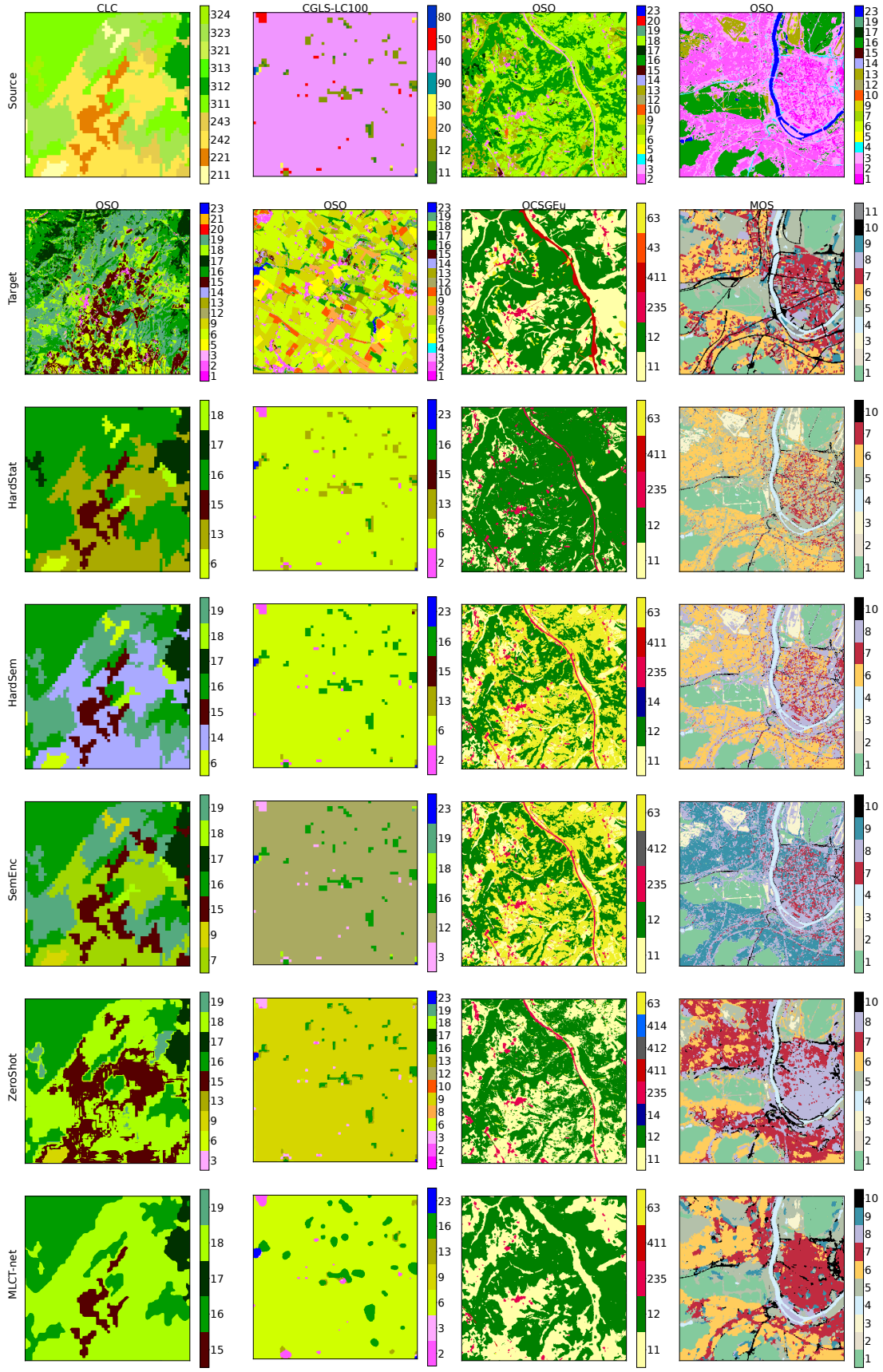

Figure 6.14: Comparison between translation using the baselines and the **ZeroShot** model.

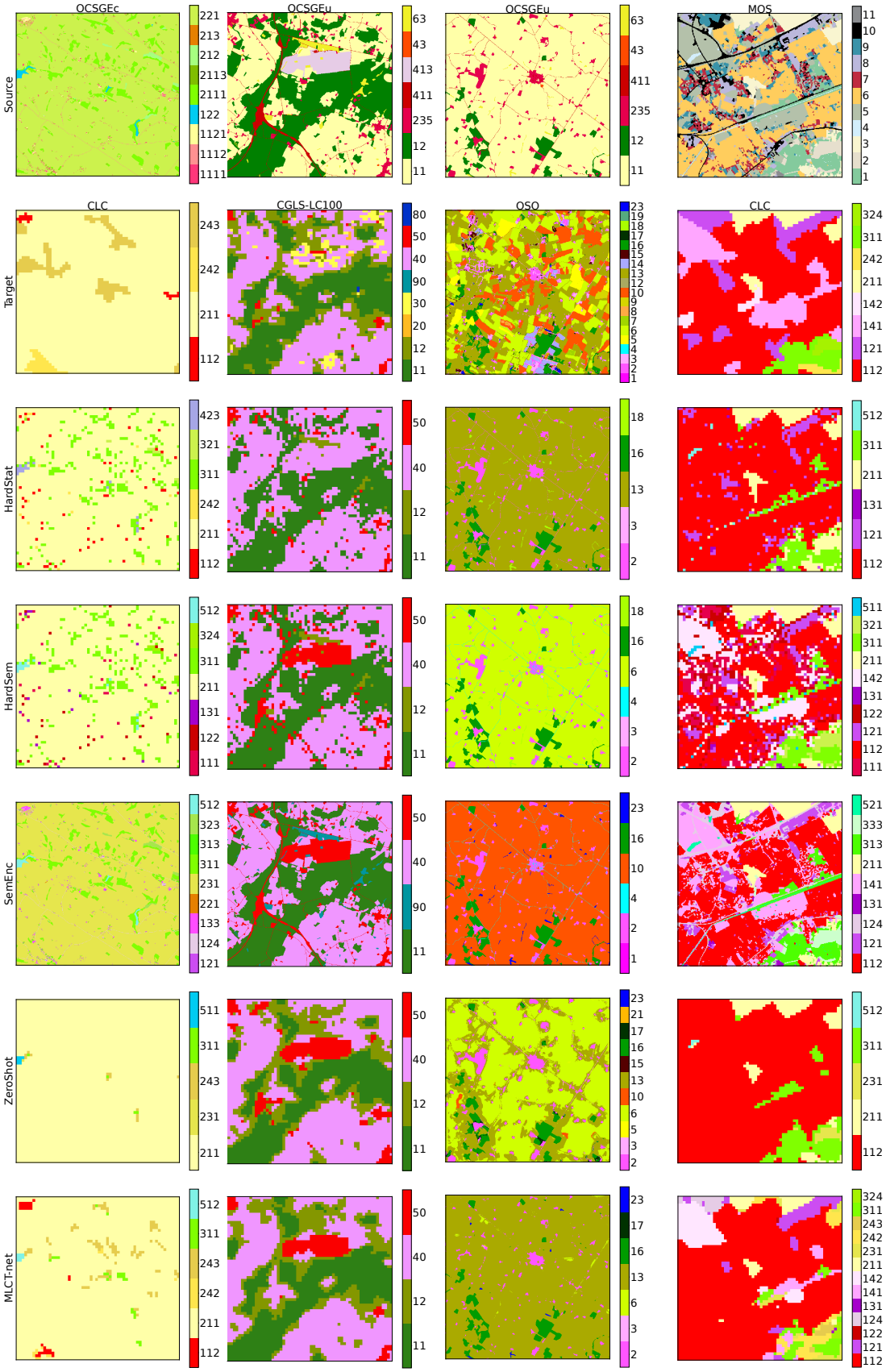

Figure 6.15: Comparison between translation using the baselines and the **ZeroShot** model.

semantic translations are possible. Conversely, no such choice is made for the **SemEnc** method: it outputs a semantically consistent class while not necessarily being the most frequent one. Consequently, from a purely quantitative point of view, the **HardSem** method will perform better than the simple **SemEnc** translation.

ZeroShot results are highly differs from the other methods. They efficiently retrieve some spatial context information (*e.g.* third column of Figure 6.14, an OSO patch localized in a mountainous area with *Broad-leaved* and *Natural Grasslands* is translated into OCSGE use). From a purely semantic point of view, *Natural Grasslands* should be translated as *No-use* in the OCSGE use and is translated as such by the **HardSem** and **SemEnc** method. As most OSO *Natural Grasslands* are spatially co-occurring with OCSGE use *Forestry areas*, the **HardStat** method translates them into it. The **MLCT-Net** and **ZeroShot** methods translates them into *Forestry areas* when close to forested areas or when being narrow and linear shaped. They also translate accurately the parts corresponding to grassland in valley areas into *Agriculture*. The comparison between **MLCT-Net** and **ZeroShot** results reveals that the **ZeroShot** framework tends to output less buffered results with a finer geometric aspect. This is particularly visible on linear structures such as the MOS *roads* (last column of Figure 6.14). This context-retrieval ability is observed on numerous other examples such as the distinction between CGLS open and Closed forest (second column of Figure 6.15), or the convincing translation of OSO *Pastures* area between forest patches, near cities and alongside the roads (third column of Figure 6.15).

However, we observe that **ZeroShot** performs poorly when the target is CLC (first and the last column of Figure 6.15). Unlike the other methods that do not apply (**HardSem**, **HardStat**) or underestimate CLC minimum mapping unit (**MLCT-Net**), **ZeroShot** tends to eliminate too much information. We link this behaviour to the fact that learning such coarse MMU rules using a single encoder trained with various resolutions is a difficult task. Conversely, when CLC (first column of Figure 6.14) is translated into a fine resolved map (OSO), we observe a strange linear filamentous-like pattern in the **ZeroShot** results for Vineyards. Once again, we impute this behaviour to the significant resolution gap, making it difficult for a single encoder to perform accurate shape retrieval.

6.4.2.2 Quantitative analysis

Table 6.2 compares the results of the **MLCT-Net** (without geographic coordinates encoding) and the **OneEncoder** method. **OneEncoder** results appear almost identical to the **MLCT-Net** method ($\pm$ 1% OA_{avg} and $mF1_{avg}$). When the two networks are trained and tested using the six LULC of the MLULC dataset, replacing the six encoders with a single one is only very slightly deteriorating the quality of the results while dividing the number of parameters by 6 (the lightweight decoders are negligible). This underlines that comparable features can be used to translate all maps despite the resolution gap between the land covers. This observation seems in adequation with traditional image classification convolution neural networks that are often trained using multiple zoom levels

of comparable objects on a single architecture.

Source		P					C					O					G1				G2				M		
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O
OA_{ag}	OneEncoder	**62**	54	67	77	**76**	**73**	**58**	71	79	78	**77**	67	78	85	**85**	67	**54**	**56**	91	67	47	48	**79**	**84**	81	62
	MLCT-Net	**62**	**55**	**68**	**78**	**76**	**73**	**58**	**72**	**80**	**79**	**77**	**68**	**79**	**86**	**85**	**68**	**54**	**56**	**92**	**68**	**49**	**49**	**79**	**84**	**82**	**63**
$mF1_{ag}$	OneEncoder	25	26	**27**	18	31	54	**36**	34	29	40	59	35	43	26	**53**	46	23	27	43	39	19	21	38	**46**	27	**21**
	MLCT-net	**26**	**28**	**27**	**19**	**32**	**56**	**36**	**35**	**30**	**41**	**60**	**37**	**44**	**27**	**53**	**48**	**24**	**29**	**44**	**40**	**20**	**22**	**40**	**46**	29	**21**

Table 6.2: Translation using multiple encoders (**MLCT-Net**) or a single universal one (**OneEncoder**)

Table 6.3 compares the results of **ZeroShot** with other models. **ZeroShot** obtains, on average better results than **HardSem** method (+6% OA_{ag},+3% $mF1_{ag}$), demonstrating its ability to use context. It is also obtains better than **HardStat** (+3% OA_{ag},+3% $mF1_{ag}$), demonstrating that this higher ability is not solely based on a higher capacity to translate into the global statically most frequent class but on real spatial context analysis. Interestingly, when studying individual source-to-target translation, we observe that the zero-shot translation from CLC to other nomenclatures achieves significantly worse results than using **HardStat** or **HardSem** baselines. We link this behaviour to the fact that the 44 classes of CLC account for almost a third of all the classes of the ten nomenclatures of the LCDD dataset. Thus, under zero-shot circumstances, the SRS built by training on the nine other nomenclatures have considerable difficulties to accurately represent CLC classes. Circumventing this issue would require significantly increasing the LCDD dataset size. Conversely, other translations, such as the **ZeroShot** from OCSGE use to CGLS, often outperforms the baselines (+12%OA_{ag}). **ZeroShot** performs significantly worse than **MLCT-Net** trained on the unseen map, underlying that room for improvement still exists.

Source		P					C					O					G1				G2				M			Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	
OA_{ag}	HardSem	52	42	56	70	**75**	65	49	67	77	**79**	57	54	69	76	80	53	39	34	87	54	37	31	75	77	72	59	61
	HardStat	54	44	**65**	70	**75**	68	**55**	**71**	**78**	**79**	61	54	**73**	80	**81**	55	41	49	**89**	54	37	41	**78**	80	77	**62**	64
	ZeroShot	**57**	**49**	62	**72**	72	**69**	54	70	**78**	77	**71**	**60**	72	**81**	**81**	**65**	**45**	**51**	83	**66**	**45**	**47**	75	**85**	**81**	**62**	**67**
	MLCT-Net	62	55	68	78	76	73	58	72	80	79	77	68	79	86	85	68	54	56	92	68	49	49	79	84	82	63	71
$mF1_{ag}$	HardSem	**13**	17	22	15	**24**	46	**32**	**36**	**31**	**42**	36	17	36	20	38	24	9	17	**27**	19	8	8	29	35	**17**	**19**	24
	HardStat	**13**	18	19	16	**24**	**47**	**32**	33	30	**42**	33	16	34	20	38	26	10	20	**27**	19	8	10	27	31	15	18	24
	ZeroShot	11	**23**	**24**	**18**	**24**	38	31	34	20	36	**39**	**21**	**38**	**24**	**39**	**43**	**13**	**22**	20	**35**	**15**	**20**	**31**	**41**	14	**19**	**27**
	MLCT-net	26	28	27	19	32	56	36	35	30	41	60	37	44	27	53	48	24	29	44	40	20	22	40	46	29	21	35

Table 6.3: Translation under zero-shot constraint. **MLCT-Net** results are displayed as upper bound performance limit. Bold values denote best performing method among HardSem, HardStat and ZeroShot.

6.5 Discussion

This section investigated the semantic encoding of land cover classes to increase the potential of learnt land-cover translation models. In particular, we build a semantic representation space encoding the class definitions in the form of a continuous variable ensuring the proximity between semantically close classes and proposed a list of criteria and metrics to evaluate it.

A trained on generic text corpora transformer model is used to represent each class definition in the form of 384 continuous variables. We show that this encoding method analysing the whole definition and not sensitive to words outside the dictionary, represents the land cover classes significantly better than the current methods used in literature (BOW, Word2Vec).

Nonetheless, the obtained encoding is insufficient for map translation as class proximity is imperfectly applied and exhibits too many dimensions to be used in one of the convolutional neural network architectures explored in the previous Chapters. We investigate a supervised dimension reduction method to select the semantically meaningful information from a land cover map translation point of view. We optimise the low-dimensional space based on relative proximity between neighbouring and non-neighbouring classes rather than bare distances. The low-dimension encoding obtained adequately represents classes belonging to nomenclatures never seen during training, demonstrating the viability of the process.

We then propose the application of this low-dimensional semantic representation space for the zero-shot source translation problem. A model is trained to translate several semantically encoded maps and tested on unseen ones. We demonstrate that such a model obtains better results than the baselines. However, the difference with the methods trained on all nomenclatures reveals that improvements are still possible.

The proposed SRS embedding model could be further improved. In particular, it is necessary to significantly increase the size of the LCDD definition dataset on which the dimension reduction method is based. Its limited size prevents correctly encoding specific concepts of classes seldom present in the nomenclatures used, such as wetlands. In addition, it is also necessary to increase the size of the data set used to train the MLULC translation model (here, five nomenclatures are used during training and one for testing). Indeed, a translation model trained on numerous maps will better generalise to unseen during training.

From an application point of view, efficiency on the zero-shot target problem is still to be demonstrated. The translation model is directly trained to represent the source land-cover map in the semantic representation space. Thus, the contextual translation map obtained would no longer be a simple representation in the form of a fixed number of classes but a contextual representation in the semantic space. A pixel of a herbaceous area near water would be placed in the semantic space between the two classes allowing to classify the pixel as herbaceous, water or wetland.

Lastly we point out that SRS could be used for other applications than zero-shot translation. In particular, we believe that such continuous representation is useful for land-cover fusion or change-resolution as it enables deriving new classes through simple arithmetic operations *e.g.* $\mathit{water} + \mathit{grassland} = \mathit{wetland}$.

CHAPTER 7

Conclusion

This last chapter presents our concluding remarks. In particular, we summarise this work's key takeaway and offer some perspectives.

7.1 Long story short

Context Land-cover products exhibit a fixed nomenclature and resolution restraining their potential adoption and re-use to situations where the nomenclature and resolution of the considered product are adapted. Additionally, as each product exhibits widely different specifications, they are hardly interoperable, *i.e.* one can not easily replace a land-cover product with another one. Numerous standardisation approaches have been conducted to alleviate those issues, increasing their re-usability. However, none of those frameworks has been universally adopted by the remote-sensing community. Therefore, multiple attempts to harmonise existing nomenclatures have been proposed based on the semantic analysis of class definitions. Even though acknowledging multiple possible translations between a source class and all the classes of the target nomenclature (*e.g. Forest* can be translated either into *Broad-leaved* or *Coniferous*), those methods only proceed to the translation into the semantically closest as they do not define in which context a translation is more relevant than another (*e.g. Coniferous* is the best translation in mountainous areas). Moreover, the resolution translation is addressed independently from the nomenclature translation while those notions are intertwined.

Solutions We proposed to replace the actual nomenclature-level land-cover translation (NLLCT) with pixel-level land-cover translation. We investigated context-wise solutions that enable reclassifying and changing the resolution of existing land-cover maps. Specifically, we first focused on which information should be considered for land-cover translation. We proposed a convolution neural network (**A-UNet**) to account for spatial context and incorporate wide-scale geographical information using learnt geographical coordinate features. Even though those methods significantly increase the number and quality of predictable classes compared to NLLCT, they lacked spatial generalisation ability for maps with a small spatial extent. Hence, inspired by the outstanding results of language

models trained to translate multiple languages, we introduced **MLCT-Net**, a first multi-land-cover translation framework, significantly improving the spatial generalisation ability. Since the translation between a source and target nomenclatures might be ill-defined due to a lack of either semantic information in the source map or a higher spatial target resolution, we present a versatile solution able to merge context-wisely various highly resolved image sources with map representation. We observed that image and map fusion considerably improve translation quality. Lastly, we claim that training models to translate each land cover as the transformation from an independent representation to another is a real operational burden and requires retraining the translation model for each map. We underlined that, theoretically, one could infer the translation of an unseen map based on knowledge acquired on previous maps. We investigated the potential of semantic representation spaces and introduced metrics to qualify them. We compared different solutions and proposed a way to adapt them for land-cover translation. We present an application for zero-shot source translation, better performing than standard NLLCT methods. Table 7.1 offers a comprehensive view of the principal results evaluated on the six land covers of the MLULC dataset. For simple one-source to one-target translation, our **A-Unet** improved the state-of-the-art by 10 points both in OA_{ag} and $mF1_{ag}$ without taking into account geographical coordinates and by respectively 11 points and 13 points with them. The multi-land-cover translation training of **MLCT-net** increases the generalisation ability of 3 points OA_{gt} and 5 points in $mF1_{gt}$ compared with **A-Unet**. The versatile image fusion architecture significantly improves the quality of the result when the source is finer resolved than the target (+8% $mF1_{ag}$ compared with **MLCT-NeT**, +5% compared with the image only approach). However, it is irrelevant when the source map is very coarse compared to the target, as the pure image classification performs identically (+1% compared to the image-only approach). Lastly, the zero-shot model offers a compromise between the constraint of fully supervised training and the results obtained by NLLCT methods. We believe that the results presented in this manuscript are generalisable to other land-cover maps as our experiments were conducted at a wide scale with very different land-cover maps.

Translation difficulty Beyond the global results, the high heterogeneity of per translation results (from 0 more than 25% $mF1_{ag}$ improvement) illustrates that performance mainly depends on the source and target complementarity. As this manuscript deliberately focuses on challenging translation cases with poorly compatible nomenclature and resolution, we gain a broad understanding of the potential of land-cover translation and its limitations. In particular, we pointed out several factors weakening translation performance, such as classes defined by specific temporal patterns or errors in the source and target maps used to train the model. Even though enabling improvement, the straightforward multi-temporal source translation framework and the divergence-based loss for noise correction still require far more research to allow translation at its full potential. Especially investigating how to consider the errors in source and target simultaneously could tremendously improve the translation. That said, in more straightforward and more

realistic translation setups, in which the source and target characteristics are relatively comparable, translation can quickly produce products of quality comparable to image classification-based ones. For instance, when translating OSO into the ESRI land-cover map (23 classes 10m to 9 classes 10m, both obtained from sentinel imagery), the OA and mF1 go up to 90% and 70% respectively without using additional data while needing less than 12 hours to train. This is interesting for operational purposes as it could be used to increase the update rate of some land-cover maps, such as CLC products released only once each six years.

	OA_{ag}			$mF1_{ag}$			OA_{gt}			$mF1_{gt}$			Time
	Average	DownRes	UpRes	Average	DownRes	UpRes	Average	SameExt	SpaGe	Average	SameExt	SpaGe	
HardSem*	61	58	63	25	23	26	72	58	81	38	33	40	<1 hours
HardStat*	62	61	63	25	24	25	72	58	82	35	30	39	<3 hours
SoftLearntFreq*	X	66	X	X	27	X	X	X	X	X	X	X	<3 hours
A-UNet	71	71	71	35	40	33	74	63	82	37	37	37	≈ 12 hours
A-UNet + coord	72	**73**	**72**	38	44	34	X	**65**	X	X	**41**	X	≈ 12 hours
MLCT-Net	**73**	**73**	**72**	**39**	**45**	**35**	**77**	**65**	**85**	**41**	40	**42**	≈ 24 hours
S2 only	77	75	79	47	48	46	**77**	67	**84**	40	43	39	≈ 12 hours
MLCT-Net +S2	**79**	**76**	**80**	**49**	**53**	**47**	**77**	**72**	81	**44**	**50**	**40**	≈ 24 hours
ZeroShot	67	67	67	30	31	30	74	61	82	37	35	38	≈ 12 hours

Table 7.1: Main results on the MLULC dataset. Methods with a * are baseline methods existing prior to our work. X denotes metrics not computable for a given method. $_{ag}$ metrics are computed by comparing the translation with the target while $_{gt}$ metrics are computed on ground truth. Average metrics are computed accross all the 26 experimented translations for $_{ag}$ and 15 for $_{gt}$. DownRes (reciprocally UpRes) metrics corresponds to the average metric value taken only on translations where the source is finer (coarser) than the target. SameExt (reciprocally SpaGe) corresponds to the average metric value taken only on translation where the source and target maps have the same spatial extent (reciprocally when the target has a smaller spatial extent), *i.e.* $OA_{gt}SpaGe$ indicates the spatial generalisation ability of the method. As each SameExt/SpaGe/UpRes/DownRes are average results of different translation. Intercomparison between those values is irrelevant.

7.2 Insight on potential improvement for operational use

The quality of the translated products obtained using NLLCT approaches was generally too low to enable translation in real operational setups. Thus, most works either required to fuse multiple source land-cover translations or assumed knowledge of target segmentation to retrieve products with decent quality. We believe the various solutions proposed in this manuscript could enable new use cases introduced in Figure 7.1. As we needed more time to test every one of them, we outlined below different perspectives and improvements that could be applied to our method to achieve each of them.

Change Detection and Updating One has access to a source and a target map covering the same spatial extent with a temporal gap and wants to ensure that they have the same nomenclature and resolution to make them comparable. For instance, one wants

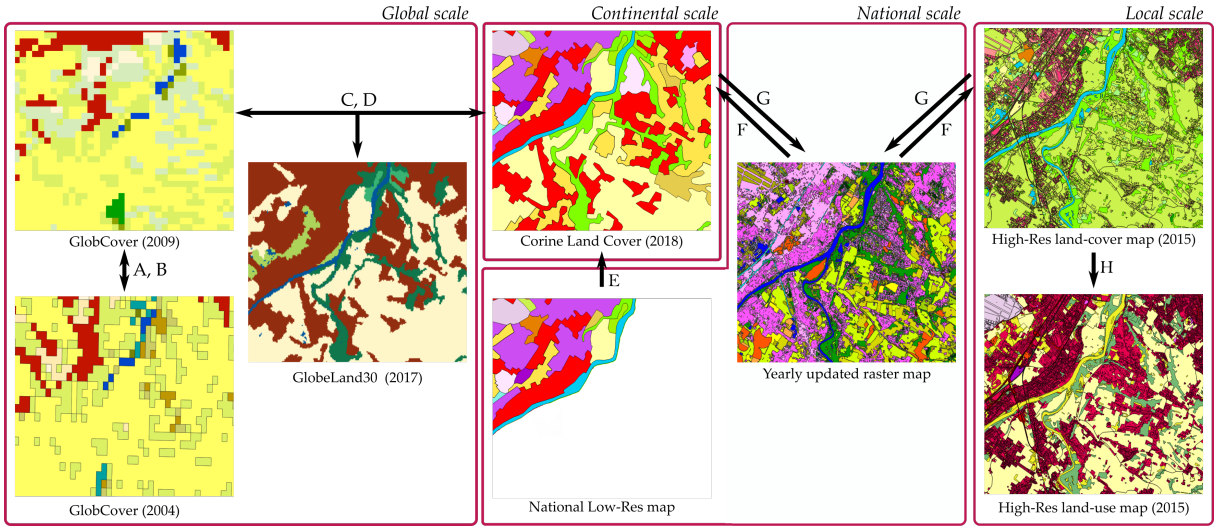

Figure 7.1: Main possibilities for land-cover map translation. (A) change detection, (B) updating, (C) validation, (D) harmonisation, (E) completion, (F) spatial simplification, (G) spatial improvement, (H) semantic modification.

to study land-cover change between a 2000 and a 2020 map produced using different nomenclature and resolution. We addressed this case in the example of OSO 2018 to CLC 2012 and obtained accuracy up to 82% at level 2 and 69% at level 3. As the OSO map is produced each year, one could thus create a translated version of CLC each year instead of each six years and reduce the year-long process needed to develop the CLC map. We believe the temporal gap is a reasonable amount of noise (the change between the two dates) compared to the noise already in the source and target maps. This is often the case for vast spatially overlapping maps with reasonable temporal gaps like those experimented on in this manuscript. However, more research should be conducted on other situations, especially on informing the network of the temporal gap-induced noise characteristics. In particular, we believe that an active learning-based solution, such as the one proposed by [179], could be used to learn to differentiate temporal noise from other map errors automatically. This could pave the way to proceed to translation on maps with significant temporal gaps (20 years or more) while increasing noise robustness.

Validation One could translate high-quality reference land-cover maps into the exact nomenclature and resolution as a target map one wishes to validate. The comparison between the target and this high-quality translation reference could be used to evaluate the quality of the target product. However, as we pointed out, that translation has highly heterogenous per-class results, so providing per-pixel translation uncertainty with the translated map would be necessary. At the time of writing, raw classifier confidence scores are available. However, we believe in the importance of using approaches such as the one proposed by [108] to calibrate this confidence score into a real probability of errors.

Harmonisation One aims to translate multiple source maps into a single target nomenclature and resolution. This is mainly used to compare different maps, but it could also

be used the other way around to produce several maps from a single one. For instance, multiple worldwide land-cover products have been obtained in the last decade based on Landsat and Sentinel imagery with comparable resolutions and nomenclatures. Therefore, a translation could be used to bring all those products accurately from a single source map instead of independently reclassifying those image time series. However, as the zero-shot target procedure still needs to be fully operational, the translation method will not alleviate the need for existing training samples of each target class.

Completion One aims to extend the spatial extent of a target map. We train a model to translate source maps with the desired spatial extent into a target on the target map extent. At inference, the model is used to translate the whole source map and obtain a broad extent target. We compared our results with those obtained by training to classify a single image into the target and showed that translation gives better results than image classification when the target has a minimal spatial extent. On a bigger extent, we showed that the fusion of image and map translation performs significantly better than single-image classification. We attribute this to the fact that translation suffers less from domain adaptation problems, *e.g.* a "forest" class remains "forest" in the South or North of France while the radiometry and features vary deeply. Thus we believe in the high potential of translation methods for completion. However, we pointed out that translation baselines are more robust in many cases than our machine-learnt solutions. This pledge for more research on improving our method's robustness to spatial generalisation. In particular in the potential of data-augmentation techniques as they are commonly used to tackle generalisation in the image classification field [299]. This would require developing new land-cover-specific data augmentation methods by proposing, for instance, a realistic label-noise augmentation that could be applied to the source land-cover maps.

Spatial simplification/improvement One aims to change the resolution of source land-cover maps. Our experiment showed that the method works well for downsampling maps with some minor problems learning very coarse minimum mapping units. Conversely, we did not obtain satisfactory results in upsampling maps. The upsampling without any additional data is ill-defined and thus naturally gives poor results. Our attempts to improve the translation results using highly resolved image data did not provide better results than the image data alone, making the map translation approach obsolete. We point out that no satisfactory solutions have been proposed apart from [202] only targeting a 4-class nomenclature. Inspired by recent advances in image super resolution [34, 350], we believe that generative adversarial networks provide an exciting solution to explore.

Semantic modifications One aims to change the nomenclature of a map. This setup is used in all our experimentations and gives satisfactory results. As for now, the translation methods can translate any source nomenclature (even unseen during training) into one of the target nomenclature seen during training. However, we need to provide tools to combine different target nomenclatures efficiently. For instance, one wants to obtain OSO nomenclature with the wetland of CLC and the Open forest of CGLS. As

the method currently outputs the OSO, CLC and CGLS translations, one could define postprocessing rules to merge the different classes. This would face cases where a pixel can be assigned multiple classes. A second, more advanced solution could be to rely on per-class prototypes [282] as they could be combined to derive maps with a unique set of classes while ensuring a unique class for each pixel. A third solution relies on targeting the Semantic representation space directly as the output of the translation framework, as it provides an off-the-shelf method to pick the desired output classes. We point out that quantitatively assessing those solutions' results is difficult as no ground truth data of such recombined nomenclature is available.

Bibliography

[1] Maria Adamo et al. "Expert knowledge for translating land cover/use maps to general habitat categories (ghc)". In: *Landscape Ecology* 29.6 (Apr. 2014), pp. 1045–1067. DOI: 10.1007/s10980-014-0028-9.

[2] Japan Aerospace Exploration Agency. *Alos world 3d - 30m*. 2016. DOI: 10.5069/G94M92HB.

[3] Alem-meta Assefa Agidew and K. N. Singh. "The implications of land use and land cover changes for rural household food insecurity in the northeastern highlands of ethiopia: the case of the teleyayen sub-watershed". In: *Agriculture & Food Security* 6.1 (Oct. 2017). DOI: 10.1186/s40066-017-0134-4.

[4] O. AHLQVIST. "Using uncertain conceptual spaces to translate between land cover categories". In: *International Journal of Geographical Information Science* 19.7 (Aug. 2005), pp. 831–857. DOI: 10.1080/13658810500106729.

[5] Ahmet Kerem Aksoy et al. "A consensual collaborative learning method for remote sensing image classification under noisy multi-labels". In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sept. 2021. DOI: 10.1109/icip42928.2021.9506236.

[6] James R. Anderson et al. *A land use and land cover classification system for use with remote sensor data*. Research rep. US Geological Survey, 1976. DOI: 10.3133/pp964.

[7] André Araujo, Wade Norris, and Jack Sim. "Computing receptive fields of convolutional neural networks". In: *Distill* 4.11 (Nov. 2019). DOI: 10.23915/distill.00021.

[8] Stephan Arnold et al. "The eagle concept - a vision of a future european land monitoring framework". In: *33th EARSeL Symposium : "Towards Horizon 2020"*. Ed. by Biscione M. Lasaponara R. Masini N. 2013.

[9] Stephan Arnold et al. "The eagle concept: a paradigm shift in land monitoring". In: *Land Use and Land Cover Semantics*. CRC Press, July 2015, pp. 107–144. DOI: 10.1201/b18746-7.

[10] Peter M. Atkinson. "Super-resolution target mapping from softclassified remotely sensed imagery". In: *Proc. of the 6th International Conference on Geocomputation, Brisbane, University of Queensland*. 2001.

[11] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. "Beyond RGB: very high resolution urban remote sensing with multimodal deep networks". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 140 (June 2018), pp. 20–32. DOI: 10.1016/j.isprsjprs.2017.11.011.

[12] Reza Azad et al. "On the texture bias for few-shot cnn segmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 2674–2683.

[13] Hasi Bagan et al. "Land cover classification from modis evi times-series data using som neural network". In: *International Journal of Remote Sensing* 26.22 (2005), pp. 4999–5012.

[14] R. Bamler. "Principles of synthetic aperture radar". In: *Surveys in Geophysics* 21.2/3 (2000), pp. 147–157. DOI: 10.1023/a:1006790026612.

[15] Moshe Bar and Shimon Ullman. "Spatial context in recognition". In: *Perception* 25.3 (Mar. 1996), pp. 343–352. DOI: 10.1068/p250343.

[16] Michael J. Barnsley and Stuart L. Barr. "Distinguishing urban land-use categories in fine spatial resolution land-cover data using a graph-based, structural pattern recognition system". In: *Computers, Environment and Urban Systems* 21.3-4 (May 1997), pp. 209–225. DOI: 10.1016/s0198-9715(97)10001-1.

[17] E. Bartholomé and A. S. Belward. "GLC2000: a new approach to global land cover mapping from earth observation data". In: *International Journal of Remote Sensing* 26.9 (May 2005), pp. 1959–1977. DOI: 10.1080/01431160412331291297.

[18] Melih Basaraner and Sinan Cetinkaya. "Performance of shape indices and classification schemes for characterising perceptual shape complexity of building footprints in GIS". In: *International Journal of Geographical Information Science* 31.10 (July 2017), pp. 1952–1977. DOI: 10.1080/13658816.2017.1346257.

[19] Luc Baudoux, Jordi Inglada, and Clément Mallet. "Toward a yearly country-scale CORINE land-cover map without using images: a map translation approach". In: *Remote Sensing* 13.6 (Mar. 2021), p. 1060. DOI: 10.3390/rs13061060.

[20] Benjamin Bechtel, Matthias Demuzere, and Iain D. Stewart. "A weighted accuracy measure for land cover mapping: comment on johnson et al. Local climate zone (LCZ) map accuracy assessments should account for land cover physical characteristics that affect the local thermal environment. Remote sens. 2019, 11, 2420". In: 12.11 (May 2020), p. 1769. DOI: 10.3390/rs12111769.

[21] Paola Benedetti et al. "M^{3}fusion: a deep learning architecture for multiscale multimodal multitemporal satellite data fusion". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.12 (Dec. 2018), pp. 4939–4949. DOI: 10.1109/jstars.2018.2876357.

[22] J. R. Bergado, C. Persello, and A. Stein. "LAND USE CLASSIFICATION USING DEEP MULTITASK NETWORKS". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B3-2020 (Aug. 2020), pp. 17–21. DOI: 10.5194/isprs-archives-xliii-b3-2020-17-2020.

[23] Mario Bertero, Patrizia Boccacci, and Christine De MoI. *Introduction to inverse problems in imaging*. CRC Press, Oct. 2021. DOI: 10.1201/9781003032755.

[24] Elisabetta Binaghi et al. “A fuzzy set-based accuracy assessment of soft classification”. In: *Pattern Recognition Letters* 20.9 (Sept. 1999), pp. 935–948. DOI: 10.1016/s0167-8655(99)00061-6.

[25] Joan Blaeu. *Atlas maior of 1665. Gallia - france*. 1665.

[26] Stephen H. Boles et al. “Land cover characterization of temperate east asia using multi-temporal vegetation sensor data”. In: *Remote Sensing of Environment* 90.4 (2004), pp. 477–489.

[27] Paul Bolstad. *Gis fundamentals : a first text on geographic information systems : 5th ed.* Eider, 2016. 784 pp. ISBN: 978-1-50669-587-7.

[28] Isabel del Bosque González et al. “Creación de un sistema de información geográfico de ocupación del suelo en españa. Proyecto siose”. In: *XI Congreso Nacional de Teledeteccion*. AGE y CCAA. Universidad de La Laguna, Puerto de la Cruz, Tenerife, 2005, pp. 255–262.

[29] Alexandre Boucher and Phaedon C. Kyriakidis. “Super-resolution land cover mapping with indicator geostatistics”. In: *Remote Sensing of Environment* 104.3 (Oct. 2006), pp. 264–282. DOI: 10.1016/j.rse.2006.04.020.

[30] Leo Breiman et al. *Classification and regression trees*. Routledge, Oct. 2017. DOI: 10.1201/9781315139470.

[31] M. Brown, H. G. Lewis, and S. R. Gunn. “Linear spectral mixture models and support vector machines for remote sensing”. In: *IEEE Transactions on Geoscience and Remote Sensing* 38.5 (2000), pp. 2346–2360. DOI: 10.1109/36.868891.

[32] Maxime Bucher et al. “Zero-shot semantic segmentation”. In: *NEURIPS*. arXiv, 2019. DOI: 10.48550/ARXIV.1906.00817.

[33] Marcel Buchhorn et al. “Copernicus global land cover layers—collection 2”. In: *Remote Sensing* 12.6 (Mar. 2020), p. 1044. ISSN: 2072-4292. DOI: 10.3390/rs12061044.

[34] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. “To learn image super-resolution, use a gan to learn how to do image degradation first”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[35] G. F. Byrne, P. F. Crapper, and K. K. Mayo. “Monitoring land-cover change by principal component analysis of multitemporal landsat data”. In: *Remote Sensing of Environment* 10.3 (Nov. 1980), pp. 175–184. DOI: 10.1016/0034-4257(80)90021-8.

[36] Pierre Cantelaube and Marie Carles. “Le registre parcellaire graphique: des données géographiques pour décrire la couverture du sol agricole”. In: *Cahier des Techniques de l’INRA* Méthodes et techniques GPS et SIG pour la conduite de dispositifs expérimentaux (2014), pp. 58–64.

[37] Hugo Carrão, Paulo Gonçalves, and Mário Caetano. “Contribution of multispectral and multitemporal information from modis images to land cover classification”. In: *Remote Sensing of Environment* 112.3 (2008), pp. 986–997.

[38] Marcela Carvalho et al. “Multitask learning of height and semantics from aerial images”. In: *IEEE Geoscience and Remote Sensing Letters* 17.8 (Aug. 2020), pp. 1391–1395. DOI: 10.1109/lgrs.2019.2947783.

[39] Alessandro Cecconi. “Integration of cartographic generalization and multi-scale databases for enhanced web mapping”. en. PhD thesis. ETH Zurich, 2003. DOI: 10.3929/ETHZ-A-004553772.

[40] Punarjay Chakravarty, Praveen Narayanan, and Tom Roussel. “GEN-SLAM: generative modeling for monocular simultaneous localization and mapping”. In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, May 2019. DOI: 10.1109/icra.2019.8793530.

[41] Yue Chang et al. “Review of land use and land cover change research progress”. In: *IOP Conference Series: Earth and Environmental Science* 113 (Feb. 2018), p. 012087. DOI: 10.1088/1755-1315/113/1/012087.

[42] Robin L. Chazdon et al. “When is a forest a forest? Forest concepts and definitions in the era of forest and landscape restoration”. In: *Ambio* 45.5 (Mar. 2016), pp. 538–550. DOI: 10.1007/s13280-016-0772-y.

[43] Jun Chen et al. “Analysis and applications of GlobeLand30: a review”. In: *ISPRS International Journal of Geo-Information* 6.8 (July 2017), p. 230. ISSN: 2220-9964. DOI: 10.3390/ijgi6080230.

[44] Liang-Chieh Chen et al. “DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (Apr. 2018), pp. 834–848. DOI: 10.1109/tpami.2017.2699184.

[45] Liang-Chieh Chen et al. “Rethinking atrous convolution for semantic image segmentation”. In: *CVPR*. Vol. abs/1706.05587. 2017. arXiv: 1706.05587.

[46] Pengfei Chen et al. “Beyond class-conditional assumption: a primary attempt to combat instance-dependent label noise”. In: *Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*. arXiv, 2021. DOI: 10.48550/ARXIV.2012.05458.

[47] Xi Chen et al. “Variational lossy autoencoder”. In: 2017.

[48] Jianpeng Cheng, Li Dong, and Mirella Lapata. “Long short-term memory-networks for machine reading”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016. DOI: 10.18653/v1/d16-1053.

[49] Grace Chu et al. “Geo-aware networks for fine-grained recognition”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, Oct. 2019. DOI: 10.1109/iccvw.2019.00033.

[50] Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. "A character-level decoder without explicit segmentation for neural machine translation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016. DOI: 10.18653/v1/p16-1160.

[51] Matthew L. Clark, T. Mitchell Aide, and George Riner. "Land change for all municipalities in latin america and the caribbean assessed from 250-m modis imagery (2001–2010)". In: *Remote Sensing of Environment* 126 (2012), pp. 84–103.

[52] William G. Cochran. *Sampling techniques*. John Wiley & Sons, 1977.

[53] A. J. Comber. "Land use or land cover?" In: *Journal of Land Use Science* 3.4 (Nov. 2008), pp. 199–201. DOI: 10.1080/17474230802465140.

[54] A. J. Comber, P. F. Fisher, and R. A. Wadsworth. "Land cover: to standardise or not to standardise? Comment on "evolving standards in land cover characterization" by herold et al." In: *Journal of Land Use Science* 2.4 (Jan. 2008), pp. 283–287. DOI: 10.1080/17474230701786000.

[55] Alexis Comber, Peter Fisher, and Richard Wadsworth. "Assessment of a semantic statistical approach to detecting land cover change using inconsistent data sets". In: *Photogrammetric Engineering & Remote Sensing* 70.8 (Aug. 2004), pp. 931–938. DOI: 10.14358/pers.70.8.931.

[56] Alexis Comber, Peter Fisher, and Richard Wadsworth. "Comparing statistical and semantic approaches for identifying change from land cover datasets". In: *Journal of Environmental Management* 77.1 (Oct. 2005), pp. 47–55. DOI: 10.1016/j.jenvman.2005.02.009.

[57] Alexis Comber, Peter Fisher, and Richard Wadsworth. "Comparing the consistency of expert land cover knowledge". In: *International Journal of Applied Earth Observation and Geoinformation* 7.3 (Nov. 2005), pp. 189–201. DOI: 10.1016/j.jag.2005.02.001.

[58] Alexis Comber, Peter Fisher, and Richard Wadsworth. "Text mining analysis of land cover semantic overlap". In: *Land Use and Land Cover Semantics*. CRC Press, July 2015, pp. 191–210. DOI: 10.1201/b18746-10.

[59] Alexis Conneau et al. "Unsupervised cross-lingual representation learning at scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.747.

[60] J. W. Coulston et al. *The spatial scan statistic: a new method for spatial aggregation of categorical raster maps*. 2014. DOI: 10.48550/ARXIV.1408.0164.

[61] David J. Cunningham et al. "Geocover lc - a moderate resolution global landcover database". In: *ESRI User Conference*. 2002.

[62] John C. Curlander and Robert N. McDonough. *Synthetic aperture radar*. Vol. 11. Wiley, New York, 1991.

[63] Rodrigo Caye Daudt et al. "Multitask learning for large-scale semantic change detection". In: *Computer Vision and Image Understanding* 187 (Oct. 2019), p. 102783. DOI: 10.1016/j.cviu.2019.07.003.

[64] Pierre Defourny et al. "Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the sen2-agri automated system in various cropping systems around the world". In: *Remote Sensing of Environment* 221 (Feb. 2019), pp. 551–568. DOI: 10.1016/j.rse.2018.11.007.

[65] Dongpo Deng. "Measurement of semantic similarity for land use and land cover classification systems". In: *SPIE Proceedings*. Ed. by Deren Li, Jianya Gong, and Huayi Wu. SPIE, Dec. 2008. DOI: 10.1117/12.815965.

[66] Jacob Devlin et al. "Bert: pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/n19-1423.

[67] Antonio Di Gregorio. *Land cover classification system: classification concepts and user manual: lccs*. Vol. 2. Rome: Food and Agriculture Organization of the United Nations, 2005. Chap. 2, p. 3. ISBN: 92-5-105327-8.

[68] Antonio Di Gregorio and Louisa J. M. Jansen. "A new concept for a land cover classification system". In: *The Land* 2.1 (1998), pp. 55–65.

[69] Xiaohu Dong et al. "Remote sensing object detection based on receptive field expansion block". In: *IEEE Geoscience and Remote Sensing Letters* 19 (2022), pp. 1–5. DOI: 10.1109/lgrs.2021.3110584.

[70] Alexey Dosovitskiy et al. "An image is worth 16x16 words: transformers for image recognition at scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[71] Iryna Dronova. "Object-based image analysis in wetland research: a review". In: *Remote Sensing* 7.5 (May 2015), pp. 6380–6413. DOI: 10.3390/rs70506380.

[72] M. Drusch et al. "Sentinel-2: ESA's optical high-resolution mission for GMES operational services". In: *Remote Sensing of Environment* 120 (May 2012), pp. 25–36. DOI: 10.1016/j.rse.2011.11.026.

[73] C. D. Elifrits et al. *Mapping land cover from satellite images: a basic, low cost approach*. Tech. rep. NASA, Jan. 1, 1978.

[74] Abolfazl Farahani et al. "A brief review of domain adaptation". In: *Advances in Data Science and Information Engineering*. Springer International Publishing, 2021, pp. 877–894. DOI: 10.1007/978-3-030-71704-9_65.

[75] Hannes Feilhauer et al. “Let your maps be fuzzy!—class probabilities and floristic gradients as alternatives to crisp mapping for remote sensing of vegetation”. In: *Remote Sensing in Ecology and Conservation* 7.2 (Nov. 2020). Ed. by Kate He and Mat Disney, pp. 292–305. DOI: 10.1002/rse2.188.

[76] C.-C. Feng and D. M. Flewelling. “Assessment of semantic similarity between land use/land cover classification systems”. In: *Computers, Environment and Urban Systems* 28.3 (2004), pp. 229–246.

[77] Yu Feng, Frank Thiemann, and Monika Sester. “Learning cartographic building generalization with deep convolutional neural networks”. In: *ISPRS International Journal of Geo-Information* 8.6 (May 2019), p. 258. DOI: 10.3390/ijgi8060258.

[78] Jan Feranec et al. “Analysis and expert assessment of the semantic similarity between land cover classes”. In: *Progress in Physical Geography: Earth and Environment* 38.3 (Apr. 2014), pp. 301–327. DOI: 10.1177/0309133314532001.

[79] Jan Feranec et al. *European landscape dynamics*. Taylor & Francis Ltd., Aug. 19, 2016. 367 pp. ISBN: 1482244683. URL: https://www.ebook.de/de/product/26629117/european_landscape_dynamics.html.

[80] Peter Fisher, Alexis Comber, and R. Wadsworth. “Land use and land cover: contradiction or complement”. In: *Re-Presenting GIS*. Ed. by David Unwin Peter Fisher. WILEY, Oct. 2006. ISBN: 0470848472. URL: https://www.ebook.de/de/product/3606471/re_presenting_gis.html.

[81] Katherine Fitzpatrich-Lins. “Accuracy and consistency comparisons of land use and land cover maps made from high-altitude photographs and landsat multispectral imagery”. In: *Journal of Research of the U. S. Geological Survey* 6.1-2 (1978), pp. 23–40.

[82] T. Foerster, J. Stoter, and B. Köbben. “Towards a formal classification of generalization operators.” In: *23rd International Cartographic Conference*. International Cartographic Association. Moscow, Russia, 2007.

[83] G. M. FOODY and C. O. X. D. P. “Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions”. In: *International Journal of Remote Sensing* 15.3 (Feb. 1994), pp. 619–631. DOI: 10.1080/01431169408954100.

[84] Giles M. Foody. “Estimation of sub-pixel land cover composition in the presence of untrained classes”. In: *Computers & Geosciences* 26.4 (May 2000), pp. 469–478. DOI: 10.1016/s0098-3004(99)00125-9.

[85] Giles M. Foody. “Status of land cover classification accuracy assessment”. In: *Remote Sensing of Environment* 80.1 (Apr. 2002), pp. 185–201. DOI: 10.1016/s0034-4257(01)00295-4.

[86] M. R. B. FORSHAW et al. “Spatial resolution of remotely sensed imagery a review paper”. In: *International Journal of Remote Sensing* 4.3 (Jan. 1983), pp. 497–520. DOI: 10.1080/01431168308948568.

[87] Benoit Frenay and Michel Verleysen. “Classification in the presence of label noise: a survey”. In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5 (May 2014), pp. 845–869. DOI: 10.1109/tnnls.2013.2292894.

[88] Mark A. Friedl et al. “Global land cover mapping from modis: algorithms and early results”. In: *Remote sensing of Environment* 83.1-2 (2002), pp. 287–302.

[89] Mark A. Friedl et al. “Modis collection 5 global land cover: algorithm refinements and characterization of new datasets”. In: *Remote sensing of Environment* 114.1 (2010), pp. 168–182.

[90] Zhenyong Fu et al. “Zero-shot object recognition by semantic manifold distance”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. DOI: 10.1109/cvpr.2015.7298879.

[91] S. Furby et al. “Continental scale land cover change monitoring in australia using landsat imagery”. In: *Earth Conference—Studying modeling and sense making of planet Earth*. 2008.

[92] Robert H. Gardner et al. “A new approach for rescaling land cover data”. In: *Landscape Ecology* 23.5 (Mar. 2008), pp. 513–526. DOI: 10.1007/s10980-008-9213-z.

[93] Vivien Sainte Fare Garnot and Loic Landrieu. “Leveraging class hierarchies with metric-guided prototype learning”. In: *BMVC*. 2021.

[94] Vivien Sainte Fare Garnot et al. “Satellite image time series classification with pixel-set encoders and temporal self-attention”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[95] Steffen Gebhardt et al. “Mad-mex: automatic wall-to-wall land cover monitoring for the mexican redd-mrv program using all landsat data”. In: *Remote Sensing* 6.5 (2014), pp. 3923–3943.

[96] Robert Geirhos et al. “Imagenet-trained cnns are biased towards texture; Increasing shape bias improves accuracy and robustness”. In: *ICLR*. 2019.

[97] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. “Variable selection using random forests”. In: *Pattern Recognition Letters* 31.14 (Oct. 2010), pp. 2225–2236. DOI: 10.1016/j.patrec.2010.03.014.

[98] Pedram Ghamisi et al. “Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art”. In: *IEEE Geoscience and Remote Sensing Magazine* 7.1 (Mar. 2019), pp. 6–39. DOI: 10.1109/mgrs.2018.2890023.

[99] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. "Robust loss functions under label noise for deep neural networks". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 1919–1925.

[100] Paulo Gonçalves et al. "Land cover classification with support vector machine applied to modis imagery". In: *Global developments in environmental earth observation from space* (2006), pp. 517–525.

[101] Ian Goodfellow, Joshua Bengio, and Aaron Courville. *Deep learning*. Ed. by MIT Press. MIT Press Ltd, Nov. 18, 2016. 800 pp. ISBN: 0262035618. URL: https://www.ebook.de/de/product/26337726/ian_goodfellow_joshua_bengio_aaron_courville_deep_learning.html.

[102] Joseph W. Goodman. "Some fundamental properties of speckle". In: *Journal of the Optical Society of America (1917-1983)* 66.11 (Nov. 1976), pp. 1145–1150. URL: http://adsabs.harvard.edu/abs/1976JOSA...66.1145G.

[103] Sucharita Gopal, Curtis E. Woodcock, and Alan H. Strahler. "Fuzzy neural network classification of global land cover from a 1 avhrr data set". In: *Remote sensing of Environment* 67.2 (1999), pp. 230–243.

[104] Kass Green, Dick Kempka, and Lisa Lackey. "Using remote sensing to detect and monitor land-cover and land-use change". In: *Photogrammetric Engineering and Remote Sensing* 60 (1994), pp. 331–337.

[105] Antonio Gregorio. *Land cover classification system : lccs : classification concepts and user manual*. Rome: Food and Agriculture Organization of the United Nations, 2000. ISBN: 92-5-104216-0.

[106] George Grekousis, Giorgos Mountrakis, and Marinos Kavouras. "An overview of 21 global and 43 regional land-cover mapping products". In: *International Journal of Remote Sensing* 36.21 (Oct. 2015), pp. 5309–5335. DOI: 10.1080/01431161.2015.1093195.

[107] Rong Gui et al. "A generalized zero-shot learning framework for PolSAR land cover classification". In: *Remote Sensing* 10.8 (Aug. 2018), p. 1307. DOI: 10.3390/rs10081307.

[108] Chuan Guo et al. "On calibration of modern neural networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1321–1330.

[109] Olivier Hagolle, David Morin, and Mohamed Kadiri. "Detailed processing model for the weighted average synthesis processor (wasp) for sentinel-2". en. In: (2018). DOI: 10.5281/ZENODO.1401360.

[110] Zayd Mahmoud Hamdi, Melanie Brandmeier, and Christoph Straub. "Forest damage assessment using deep learning on high resolution remote sensing data". In: *Remote Sensing* 11.17 (Aug. 2019), p. 1976. DOI: 10.3390/rs11171976.

[111] M. C. Hansen et al. “Global land cover classification at 1 km spatial resolution using a classification tree approach”. In: *International Journal of Remote Sensing* 21.6-7 (Jan. 2000), pp. 1331–1364. DOI: 10.1080/014311600210209.

[112] Zellig S. Harris. “Distributional structure”. In: *WORD* 10.2-3 (Aug. 1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

[113] Caner Hazirbas et al. “FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture”. In: *Computer Vision – ACCV 2016*. Springer International Publishing, 2017, pp. 213–228. DOI: 10.1007/978-3-319-54181-5_14.

[114] Kaiming He et al. “Deep residual learning for image recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. DOI: 10.1109/cvpr.2016.90.

[115] Kaiming He et al. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *ECCV 2014*. Springer International Publishing, 2014, pp. 346–361. DOI: 10.1007/978-3-319-10578-9_23.

[116] Xiaofei He et al. “Neighborhood preserving embedding”. In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. IEEE, 2005. DOI: 10.1109/iccv.2005.167.

[117] Katherine Hermann, Ting Chen, and Simon Kornblith. “The origins and prevalence of texture bias in convolutional neural networks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 19000–19015. URL: https://proceedings.neurips.cc/paper/2020/file/db5f9f42a7157abe65bb145000b5871a-Paper.pdf.

[118] Txomin Hermosilla et al. “Land cover classification in an era of big and open data: optimizing localized implementation and training data selection to improve mapping outcomes”. In: *Remote Sensing of Environment* 268 (Jan. 2022), p. 112780. DOI: 10.1016/j.rse.2021.112780.

[119] M. Herold et al. “Evolving standards in land cover characterization”. In: *Journal of Land Use Science* 1.2-4 (Dec. 2006), pp. 157–168. DOI: 10.1080/17474230601079316.

[120] M. Herold et al. “Some challenges in global land cover mapping: an assessment of agreement and accuracy in existing 1 km datasets”. In: *Remote Sensing of Environment* 112.5 (May 2008), pp. 2538–2556. DOI: 10.1016/j.rse.2007.11.013.

[121] Yves Heymann. *Corine land cover : technical guide*. Luxembourg: European Commission, Directorate-General, Environment, Nuclear Safety and Civil Protection, 1994. ISBN: 9282625788.

[122] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *NIPS*. arXiv, 2015. DOI: 10.48550/ARXIV.1503.02531.

[123] Stephen C. Hirtle. "Representational structures for cognitive space: trees, ordered trees and semi-lattices". In: *Spatial Information Theory A Theoretical Basis for GIS*. Ed. by Andrew U. Frank and Werner Kuhn. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 327–340. ISBN: 978-3-540-45519-6. DOI: 10.1007/3-540-60392-1_21.

[124] Y. Ho and S. Wookey. "The real-world-weight cross-entropy loss function: modeling the costs of mislabeling". In: *IEEE Access* 8 (2020), pp. 4806–4813.

[125] Collin Homer et al. "Conterminous united states land cover change patterns 2001–2016 from the 2016 national land cover database". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (Apr. 2020), pp. 184–199. DOI: 10.1016/j.isprsjprs.2020.02.019.

[126] Collin Homer et al. "Development of a 2001 national land-cover database for the united states". In: *Photogrammetric Engineering & Remote Sensing* 70.7 (2004), pp. 829–840.

[127] Danfeng Hong et al. "More diverse means better: multimodal deep learning meets remote-sensing imagery classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.5 (May 2021), pp. 4340–4354. DOI: 10.1109/tgrs.2020.3016820.

[128] Danfeng Hong et al. "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 178 (Aug. 2021), pp. 68–80. DOI: 10.1016/j.isprsjprs.2021.05.011.

[129] Andrew J. Hoskins et al. "Downscaling land-use data to provide global 30' estimates of five land-use classes". In: *Ecology and Evolution* 6.9 (Mar. 2016), pp. 3040–3055. DOI: 10.1002/ece3.2104.

[130] M. Hossain and Dongmei Chen. "Segmentation for object-based image analysis (obia): a review of algorithms and challenges from remote sensing perspective". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), pp. 115–134. DOI: 10.1016/J.ISPRSJPRS.2019.02.009. URL: https://www.semanticscholar.org/paper/23c240bf9d7ce46dcf4e96835337e27fcf342cdf.

[131] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[132] Yuansheng Hua et al. "Learning multi-label aerial image classification under label noise: a regularization approach using word embeddings". In: *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Sept. 2020. DOI: 10.1109/igarss39084.2020.9324069.

[133] Bo Huang, Bei Zhao, and Yimeng Song. "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery". In: *Remote Sensing of Environment* 214 (Sept. 2018), pp. 73–86. DOI: 10.1016/j.rse.2018.04.050.

[134] Wen-Chin Huang et al. "Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 4.4 (Aug. 2020), pp. 468–479. ISSN: 2471-285X. DOI: 10.1109/tetci.2020.2977678.

[135] George C. Hurtt et al. "Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6". In: *Geoscientific Model Development* 13.11 (Nov. 2020), pp. 5425–5464. DOI: 10.5194/gmd-13-5425-2020.

[136] Dino Ienco et al. "Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 158 (2019), pp. 11–22.

[137] Dino Ienco et al. "Land cover classification via multitemporal spatial data by deep recurrent neural networks". In: *IEEE Geoscience and Remote Sensing Letters* 14.10 (Oct. 2017), pp. 1685–1689. DOI: 10.1109/lgrs.2017.2728698.

[138] S. Inamdar et al. "Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 46.4 (Apr. 2008), pp. 1243–1252. DOI: 10.1109/tgrs.2007.912445.

[139] Jordi Inglada et al. "Operational high resolution land cover map production at the country scale using satellite image time series". In: *Remote Sensing* 9.1 (Jan. 2017), p. 95. DOI: 10.3390/rs9010095.

[140] Fabian Isensee et al. "nnU-net: self-adapting framework for u-net-based medical image segmentation". In: *Informatik aktuell*. Springer Fachmedien Wiesbaden, 2019, pp. 22–22. DOI: 10.1007/978-3-658-25326-4_7.

[141] Furkan Isikdogan, Alan C. Bovik, and Paola Passalacqua. "Surface water mapping by deep learning". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.11 (Nov. 2017), pp. 4909–4918. DOI: 10.1109/jstars.2017.2735443.

[142] Koki Iwao et al. "Creation of new global land cover map with map integration". In: 03.02 (2011), pp. 160–165. DOI: 10.4236/jgis.2011.32013.

[143] OIli Jaakkola. "Automatic generalization of land cover data". In: *16th International Cartographic Conference*. Internation Cartographic Association, 1995. ISBN: 84-393-3583-0.

[144] Louisa J. M. Jansen and Antonio Di Gregorio. "Parametric land cover and land-use classifications as tools for environmental change detection". In: *Agriculture, Ecosystems & Environment* 91.1-3 (Sept. 2002), pp. 89–100. DOI: 10.1016/s0167-8809(01)00243-2.

[145] Louisa J. M. Jansen, Geoff Groom, and Giancarlo Carrai. "Land-cover harmonisation and semantic similarity: some methodological issues". In: *Journal of Land Use Science* 3.2-3 (Oct. 2008), pp. 131–160. DOI: 10.1080/17474230802332076.

[146] John R. Jensen. "SPECTRAL AND TEXTURAL FEATURES TO CLASSIFY ELUSIVE LAND COVER AT THE URBAN FRINGE". In: *The Professional Geographer* 31.4 (Nov. 1979), pp. 400–409. DOI: 10.1111/j.0033-0124.1979.00400.x.

[147] Martin Rudbeck Jepsen and Gregor Levin. "Semantically based reclassification of Danish land-use and land-cover information". In: *International Journal of Geographical Information Science* 27.12 (Dec. 2013), pp. 2375–2390. DOI: 10.1080/13658816.2013.803555.

[148] Kun Jia et al. "Land cover classification of finer resolution remote sensing data integrating temporal features from time series coarser resolution data". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 93 (2014), pp. 49–55.

[149] Jia et al. "Super-Resolution Land Cover Mapping Based on the Convolutional Neural Network". In: *Remote Sensing* 11.15 (Aug. 2019), p. 1815. DOI: 10.3390/rs11151815.

[150] Weimin Jiang et al. "Simplifying Regional Tuning of MODIS Algorithms for Monitoring Chlorophyll-a in Coastal Waters". In: *Frontiers in Marine Science* 4 (May 2017). DOI: 10.3389/fmars.2017.00151.

[151] L. Jiao and Y. Liu. "ANALYZING THE SHAPE CHARACTERISTICS OF LAND USE CLASSES IN REMOTE SENSING IMAGERY". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* I-7 (July 2012), pp. 135–140. DOI: 10.5194/isprsannals-I-7-135-2012.

[152] Dae Ung Jo et al. "Associative Variational Auto-Encoder with Distributed Latent Spaces and Associators". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 11197–11204. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i07.6778.

[153] J. Michael Johnson and Keith C. Clarke. "An area preserving method for improved categorical raster resampling". In: *Cartography and Geographic Information Science* 48.4 (Apr. 2021), pp. 292–304. DOI: 10.1080/15230406.2021.1892531.

[154] Daniel Joly et al. "Les types de climats en France, une construction spatiale". In: *Cybergeo* (June 2010). DOI: 10.4000/cybergeo.23155.

[155] M. I. Jordan and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (July 2015), pp. 255–260. DOI: 10.1126/science.aaa8415.

[156] Hamid Reza Vaezi Joze et al. "MMTM: Multimodal Transfer Module for CNN Fusion". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. DOI: 10.1109/cvpr42600.2020.01330.

[157] Martin Jung et al. "Exploiting synergies of global land cover products for carbon cycle modeling". In: 101.4 (Apr. 2006), pp. 534–553. DOI: 10.1016/j.rse.2006.01.020.

[158] T KASETKASEM, M ARORA, and P VARSHNEY. "Super-resolution land cover mapping using a Markov random field based approach". In: *Remote Sensing of Environment* 96.3-4 (June 2005), pp. 302–314. DOI: 10.1016/j.rse.2005.02.006.

[159] Marinos Kavouras and Margarita Kokla. "A method for the formalization and integration of geographical categorizations". In: *International Journal of Geographical Information Science* 16.5 (July 2002), pp. 439–453. ISSN: 1365-8816. DOI: 10.1080/13658810210129120.

[160] Bahador Khaleghi et al. "Multisensor data fusion: A review of the state-of-the-art". In: *Information Fusion* 14.1 (Jan. 2013), pp. 28–44. DOI: 10.1016/j.inffus.2011.08.001.

[161] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. "Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers". In: *Medical Image Analysis* 51 (Jan. 2019), pp. 21–45. DOI: 10.1016/j.media.2018.10.004.

[162] Valentin Khrulkov et al. "Hyperbolic Image Embeddings". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. DOI: 10.1109/cvpr42600.2020.00645.

[163] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. arXiv, 2015. DOI: 10.48550/ARXIV.1412.6980.

[164] J.F. Knight and R.S. Lunetta. "An experimental assessment of minimum mapping unit size". In: *IEEE Transactions on Geoscience and Remote Sensing* 41.9 (Sept. 2003), pp. 2132–2134. DOI: 10.1109/tgrs.2003.816587.

[165] Vasiliki Kosmidou et al. "Harmonization of the Land Cover Classification System (LCCS) with the General Habitat Categories (GHC) classification system". In: *Ecological Indicators* 36 (Jan. 2014), pp. 290–300. DOI: 10.1016/j.ecolind.2013.07.025.

[166] S. B. Kotsiantis. "Supervised Machine Learning: A Review of Classification Techniques". In: *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. NLD: IOS Press, 2007, pp. 3–24. ISBN: 9781586037802.

[167] Wouter M. Kouw and Marco Loog. "A Review of Domain Adaptation without Target Labels". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.3 (Mar. 2021), pp. 766–785. DOI: 10.1109/tpami.2019.2945942.

[168] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90. DOI: 10.1145/3065386.

[169] Nataliia Kussul et al. "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data". In: *IEEE Geoscience and Remote Sensing Letters* 14.5 (May 2017), pp. 778–782. DOI: 10.1109/lgrs.2017.2681128.

[170] Dana Lahat, Tulay Adali, and Christian Jutten. "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects". In: *Proceedings of the IEEE* 103.9 (Sept. 2015), pp. 1449–1477. DOI: 10.1109/jproc.2015.2460697.

[171] Eric F. Lambin, Helmut J. Geist, and Erika Lepers. "Dynamics of Land-Use and Land-Cover Change in Tropical Regions". In: *Annual Review of Environment and Resources* 28.1 (Nov. 2003), pp. 205–241. DOI: 10.1146/annurev.energy.28.050302.105459.

[172] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. "Learning to detect unseen object classes by between-class attribute transfer". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009. DOI: 10.1109/cvpr.2009.5206594.

[173] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: *biometrics* (1977), pp. 159–174.

[174] Rasim Latifovic, Darren Pouliot, and Ian Olthof. "Circa 2010 Land Cover of Canada: Local Optimization Methodology and Product Development". In: *Remote Sensing* 9.11 (Oct. 2017), p. 1098. DOI: 10.3390/rs9111098.

[175] Rasim Latifovic et al. *North American Land-Change Monitoring System*. In: *Remote Sensing of Land Use and Land Cover*. Ed. by Chandra P. Giri. CRC Press, 2012. DOI: 10.1201/b11964.

[176] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (May 2015), pp. 436–444. DOI: 10.1038/nature14539.

[177] Tien Ming Lee and Walter Jetz. "Future battlegrounds for conservation under global change". In: *Proceedings of the Royal Society B: Biological Sciences* 275.1640 (Feb. 2008), pp. 1261–1270. DOI: 10.1098/rspb.2007.1732.

[178] Guangbin Lei et al. “The roles of criteria, data and classification methods in designing land cover classification systems: evidence from existing land cover data sets”. In: *International Journal of Remote Sensing* 41.14 (Apr. 2020), pp. 5062–5082. DOI: 10.1080/01431161.2020.1724349.

[179] Jiayi Li, Xin Huang, and Xiaoyu Chang. “A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 163 (May 2020), pp. 1–17. DOI: 10.1016/j.isprsjprs.2020.02.022.

[180] Zhan Li et al. “Land cover harmonization using Latent Dirichlet Allocation”. In: *International Journal of Geographical Information Science* 35.2 (July 2020), pp. 348–374. DOI: 10.1080/13658816.2020.1796131.

[181] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. DOI: 10.1109/iccv.2017.324.

[182] Feng Ling and Giles M. Foody. “Super-resolution land cover mapping by deep learning”. In: *Remote Sensing Letters* 10.6 (Mar. 2019), pp. 598–606. DOI: 10.1080/2150704x.2019.1587196.

[183] Feng Ling et al. “Interpolation-based super-resolution land cover mapping”. In: *Remote Sensing Letters* 4.7 (July 2013), pp. 629–638. DOI: 10.1080/2150704x.2013.781284.

[184] Feng Ling et al. “Super-resolution land-cover mapping using multiple sub-pixel shifted remotely sensed images”. In: *International Journal of Remote Sensing* 31.19 (Oct. 2010), pp. 5023–5040. DOI: 10.1080/01431160903252350.

[185] J. Liu et al. “Seasonal variation of land cover classification accuracy of Landsat 8 images in Burkina Faso”. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-7/W3 (Apr. 2015), pp. 455–460. DOI: 10.5194/isprsarchives-xl-7-w3-455-2015.

[186] Qinghui Liu et al. “Dense Dilated Convolutions’ Merging Network for Land Cover Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.9 (Sept. 2020), pp. 6309–6320. DOI: 10.1109/tgrs.2020.2976658.

[187] Rosanne Liu et al. “An intriguing failing of convolutional neural networks and the CoordConv solution”. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Samy Bengio et al. 2018, pp. 9628–9639. URL: https://proceedings.neurips.cc/paper/2018/hash/60106888f8977b71e1f15db7bc9a88d1-Abstract.html.

[188] Hualou Long and Yi Qu. “Land use transitions and land management: A mutual feedback perspective”. In: *Land Use Policy* 74 (May 2018), pp. 111–120. DOI: 10.1016/j.landusepol.2017.03.021.

[189] Yang Long et al. “On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 4205–4230. DOI: 10.1109/jstars.2021.3070368.

[190] T. R. Loveland and A. S. Belward. “The IGBP-DIS global 1km land cover data set, DISCover: First results”. In: *International Journal of Remote Sensing* 18.15 (Oct. 1997), pp. 3289–3295. DOI: 10.1080/014311697217099.

[191] T. R. Loveland et al. “Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data”. In: *International Journal of Remote Sensing* 21.6-7 (Jan. 2000), pp. 1303–1330. DOI: 10.1080/014311600210191.

[192] Thomas R. Loveland. “History of Land-Cover Mapping”. In: *Remote Sensing of Land Use and Land Cover: Principles and Applications*. CRC PR INC, May 2012, pp. 13–22. ISBN: 1420070746. URL: https://www.ebook.de/de/product/6921907/remote_sensing_of_land_use_and_land_cover_principles_and_applications.html.

[193] Thomas R. Loveland and Ruth S. DeFries. “Observing and monitoring land use and land cover change”. In: *Ecosystems and Land Use Change*. American Geophysical Union, 2004, pp. 231–246. DOI: 10.1029/153gm18.

[194] George Lukes. *A review of computer-assisted photo interpretation research at USAETL (U.S. Army Engineer Topographic Laboratories)*. Research rep. USAETL, Jan. 16, 1987.

[195] Kevin Lund and Curt Burgess. “Producing high-dimensional semantic spaces from lexical co-occurrence”. In: *Behavior Research Methods, Instruments, & Computers* 28.2 (June 1996), pp. 203–208. DOI: 10.3758/bf03204766.

[196] Wenjie Luo et al. “Understanding the Effective Receptive Field in Deep Convolutional Neural Networks”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 4905–4913. ISBN: 9781510838819.

[197] Wenjie Luo et al. “Understanding the Effective Receptive Field in Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/c8067ad1937f728f51288b3eb986afaa-Paper.pdf.

[198] Riccardo de Lutio et al. “Guided Super-Resolution As Pixel-to-Pixel Transformation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[199] Wanli Ma, Oktay Karakuş, and Paul L. Rosin. “AMM-FuseNet: Attention-Based Multi-Modal Image Fusion Network for Land Cover Mapping”. In: *Remote Sensing* 14.18 (Sept. 2022), p. 4458. DOI: 10.3390/rs14184458.

[200] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. "Presence-only geographical priors for fine-grained image classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9596–9606.

[201] Gengchen Mai et al. "A review of location encoding for GeoAI: methods and applications". In: *International Journal of Geographical Information Science* 36.4 (Jan. 2022), pp. 639–673. DOI: 10.1080/13658816.2021.2004602.

[202] Kolya Malkin et al. "Label super-resolution networks". In: *ICLR*. 2019.

[203] Ioannis Manakos and Matthias Braun. *Land Use and Land Cover Mapping in Europe*. Springer Netherlands, 2014. DOI: 10.1007/978-94-007-7969-3.

[204] Ioannis Manakos et al. "Comparison of Global and Continental Land Cover Products for Selected Study Areas in South Central and Eastern European Region". In: *Remote Sensing* 10.12 (Dec. 2018), p. 1967. DOI: 10.3390/rs10121967.

[205] N. Manwani and P. S. Sastry. "Noise Tolerance Under Risk Minimization". In: *IEEE Transactions on Cybernetics* 43.3 (June 2013), pp. 1146–1151. DOI: 10.1109/tsmcb.2012.2223460.

[206] Diego Marcos et al. "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 145 (2018), pp. 96–107.

[207] F. J. Marschner. *Major land uses in the United States*. Ed. by Agricultural Research Service U.S. Dept. of Agriculture. map. 1950. URL: https://www.loc.gov/item/2006627627/ (visited on 06/20/2022).

[208] J.-F. Mas. "Monitoring land-cover changes: A comparison of change detection techniques". In: *International Journal of Remote Sensing* 20.1 (Jan. 1999), pp. 139–152. DOI: 10.1080/014311699213659.

[209] Tbd Max Born Emil Wolf. *Principles of Optics*. Cambridge University Press, Dec. 19, 2019. 994 pp. ISBN: 1108477437. URL: https://www.ebook.de/de/product/38080828/max_born_emil_wolf_tbd_principles_of_optics.html.

[210] Philippe Mayaux et al. "A new land-cover map of Africa for the year 2000". In: *Journal of biogeography* 31.6 (2004), pp. 861–877.

[211] Robert B McMaster. *Generalization in Digital Cartography Resource Publications in Geography*. English. Ed. by K S Shea. Association of America Geographers, 1992.

[212] Thomas Mensink, Efstratios Gavves, and Cees G.M. Snoek. "COSTA: Co-Occurrence Statistics for Zero-Shot Classification". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014. DOI: 10.1109/cvpr.2014.313.

[213] K. C. Mertens et al. "Using genetic algorithms in sub-pixel mapping". In: *International Journal of Remote Sensing* 24.21 (Jan. 2003), pp. 4241–4247. DOI: 10.1080/0143116031000159507 3.

[214] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119.

[215] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *ICLR* (2013).

[216] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *3DV*. IEEE, Oct. 2016. DOI: 10.1109/3dv.2016.79.

[217] Tom M. Mitchell. *Machine learning, International Edition*. McGraw-Hill Series in Computer Science. McGraw-Hill, 1997. ISBN: 978-0-07-042807-2. URL: https://www.worldcat.org/oclc/61321007.

[218] Adrien Moiret-Guigand et al. *CLC2018 / CLCC1218 VALIDATION REPORT*. Tech. rep. GMES Initial Operations / Copernicus Land monitoring services, Jan. 2021. URL: https://land.copernicus.eu/user-corner/technical-library/clc-2018-and-clc-change-2012-2018-validation-report/at_download/file.

[219] Alberto Moreira et al. “A tutorial on synthetic aperture radar”. In: *IEEE Geoscience and Remote Sensing Magazine* 1.1 (Mar. 2013), pp. 6–43. DOI: 10.1109/mgrs.2013.2248301.

[220] H. Al-Mubaid and H. A. Nguyen. “Measuring semantic similarity between biomedical concepts within multiple ontologies”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 39.4 (July 2009), pp. 389–398. ISSN: 1094-6977. DOI: 10.1109/tsmcc.2009.2020689.

[221] Hannes Müller et al. “Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape”. In: *Remote Sensing of Environment* 156 (2015), pp. 490–499.

[222] Daniel Müllner. “Modern hierarchical, agglomerative clustering algorithms”. 2011. DOI: 10.48550/ARXIV.1109.2378.

[223] M. Dalla Mura et al. “Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing”. In: *Proceedings of the IEEE* 103.9 (Sept. 2015), pp. 1585–1601. DOI: 10.1109/jproc.2015.2462751.

[224] Nagarajan Natarajan et al. “Learning with Noisy Labels”. In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.

[225] K. Neumann et al. “Comparative assessment of CORINE2000 and GLC2000: Spatial analysis of land cover data for Europe”. In: *International Journal of Applied Earth Observation and Geoinformation* 9.4 (Dec. 2007), pp. 425–437. DOI: 10.1016/j.jag.2007.02.004.

[226] Behnam Neyshabur et al. “Exploring Generalization in Deep Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/10ce03a1ed01077e3e289f3e53c72813-Paper.pdf.

[227] Maximillian Nickel and Douwe Kiela. “Poincaré Embeddings for Learning Hierarchical Representations”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.

[228] Allan A. Nielsen and Morton J. Canty. “Kernel principal component and maximum autocorrelation factor analyses for change detection”. In: *SPIE Proceedings*. Ed. by Lorenzo Bruzzone, Claudia Notarnicola, and Francesco Posa. SPIE, Sept. 2009. DOI: 10.1117/12.829645.

[229] *Ocs ge version 1.1*. Tech. rep. French mapping institute (IGN), 2016.

[230] Pontus Olofsson et al. “Good practices for estimating area and assessing accuracy of land change”. In: *Remote Sensing of Environment* 148 (May 2014), pp. 42–57. DOI: 10.1016/j.rse.2014.02.015.

[231] Esam Othman et al. “Domain Adaptation Network for Cross-Scene Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 55.8 (Aug. 2017), pp. 4441–4456. DOI: 10.1109/tgrs.2017.2692281.

[232] Fabio Pacifici. “Foreword to the Special Issue on Optical Multiangular Data Exploitation and Outcome of the 2011 GRSS Data Fusion Contest”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5.1 (Feb. 2012), pp. 3–7. DOI: 10.1109/jstars.2012.2186733.

[233] Lina Papšienė and Kęstutis Papšys. “CHANGES AFFECTING GENERALIZATION OF LAND COVER FEATURES IN A SMALLER SCALE”. In: *Geodesy and Cartography* 38.3 (Oct. 2012), pp. 98–105. DOI: 10.3846/20296991.2012.728045.

[234] Claudia Paris, Lorenzo Bruzzone, and Diego Fernandez-Prieto. “A Novel Approach to the Unsupervised Update of Land-Cover Maps by Classification of Time Series of Multispectral Images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 57.7 (July 2019), pp. 4259–4277. DOI: 10.1109/tgrs.2018.2890404.

[235] Niki Parmar et al. “Image Transformer”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 4055–4064. URL: https://proceedings.mlr.press/v80/parmar18a.html.

[236] Mohammad Pashaei et al. “Review and Evaluation of Deep Learning Architectures for Efficient Land Cover Mapping with UAS Hyper-Spatial Imagery: A Case Study Over a Wetland”. In: *Remote Sensing* 12.6 (2020), p. 959.

[237] Giorgio Patrini et al. “Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. DOI: 10.1109/cvpr.2017.240.

[238] Derek R. Peddle, Philippe M. Teillet, and Michael A. Wulder. “Radiometric Image Processing”. In: *Remote Sensing of Forest Environments*. Springer US, 2003, pp. 181–208. DOI: 10.1007/978-1-4615-0306-4_7.

[239] Charlotte Pelletier, Geoffrey Webb, and François Petitjean. “Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series”. In: *Remote Sensing* 11.5 (Mar. 2019), p. 523. DOI: 10.3390/rs11050523.

[240] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. “End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++”. In: *Remote Sensing* 11.11 (2019).

[241] Maria Joao Pereira and Amilcar Soares. “Mapping spatial distribution of land cover classification errors”. In: *2011 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, July 2011. DOI: 10.1109/igarss.2011.6048959.

[242] Gabriel Pereyra et al. “Regularizing neural networks by penalizing confident output distributions”. In: *ICLR* (2017).

[243] A. Perez-Hoyos, A. Udias, and F. Rembold. “Integrating multiple land cover maps through a multi-criteria analysis to improve agricultural monitoring in Africa”. In: *International Journal of Applied Earth Observation and Geoinformation* 88 (June 2020), p. 102064. DOI: 10.1016/j.jag.2020.102064.

[244] Juan-Manuel Pérez-Rúa et al. “MFAS: Multimodal Fusion Architecture Search”. In: *CVPR*. 2019, pp. 6959–6968.

[245] Telmo Pires, Eva Schlinger, and Dan Garrette. “How Multilingual is Multilingual BERT?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1493.

[246] Biswajeet Pradhan et al. “Unseen Land Cover Classification from High-Resolution Orthophotos Using Integration of Zero-Shot Learning and Convolutional Neural Networks”. In: *Remote Sensing* 12.10 (May 2020), p. 1676. DOI: 10.3390/rs12101676.

[247] Markus Probeck et al. “CLC+ Backbone: Set the Scene in Copernicus for the Coming Decade”. In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, July 2021. DOI: 10.1109/igarss47720.2021.9553252.

[248] Chunping Qiu et al. “Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 154 (Aug. 2019), pp. 151–162. DOI: 10.1016/j.isprsjprs.2019.05.004.

[249] Julien Radoux et al. “Automated training sample extraction for global land cover mapping”. In: *Remote Sensing* 6.5 (2014), pp. 3965–3987.

[250] S. Rajesh et al. “Land Cover/Land Use Mapping of LISS IV Imagery Using Object-Based Convolutional Neural Network with Deep Features”. In: *Journal of the Indian Society of Remote Sensing* 48.1 (Nov. 2019), pp. 145–154. DOI: 10.1007/s12524-019-01064-9.

[251] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/d19-1410.

[252] M. Andrea Rodriguez and Max J. Egenhofer. “Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure”. In: *International Journal of Geographical Information Science* 18.3 (Apr. 2004), pp. 229–256. DOI: 10.1080/13658810310001629592.

[253] M. Andrea Rodriguez, Max J. Egenhofer, and Robert D. Rugg. “Assessing Semantic Similarities among Geospatial Feature Class Definitions”. In: *Interoperating Geographic Information Systems*. Springer Berlin Heidelberg, 1999, pp. 189–202. DOI: 10.1007/10703121_16.

[254] M.A. Rodriguez and M.J. Egenhofer. “Determining semantic similarity among entity classes from different ontologies”. In: *IEEE Transactions on Knowledge and Data Engineering* 15.2 (Mar. 2003), pp. 442–456. DOI: 10.1109/tkde.2003.1185844.

[255] John Rogan and DongMei Chen. “Remote sensing technology for mapping and monitoring land-cover and land-use change”. In: *Progress in Planning* 61.4 (May 2004), pp. 301–325. DOI: 10.1016/s0305-9006(03)00066-7.

[256] Marcus Rohrbach, Michael Stark, and Bernt Schiele. “Evaluating knowledge transfer and zero-shot learning in a large-scale setting”. In: *CVPR 2011*. IEEE, June 2011. DOI: 10.1109/cvpr.2011.5995627.

[257] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[258] F. Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6 (1958), pp. 386–408. DOI: 10.1037/h0042519.

[259] Sam T. Roweis and Lawrence K. Saul. “Nonlinear Dimensionality Reduction by Locally Linear Embedding”. In: *Science* 290.5500 (Dec. 2000), pp. 2323–2326. DOI: 10.1126/science.290.5500.2323.

[260] D.P. Roy. “The impact of misregistration upon composited wide field of view satellite data and implications for change detection”. In: *IEEE Transactions on Geoscience and Remote Sensing* 38.4 (July 2000), pp. 2017–2032. DOI: 10.1109/36.851783.

[261] Anne Ruas. “Map Generalization”. In: *Encyclopedia of GIS*. Ed. by Shashi Shekhar and Hui Xiong. Springer US, 2008, p. 631. ISBN: 978-0-387-35973-1. DOI: 10.1007/978-0-387-35973-1.

[262] F. Russo. “New Method for Performance Evaluation of Grayscale Image Denoising Filters”. In: *IEEE Signal Processing Letters* 17.5 (2010), pp. 417–420. DOI: 10.1109/LSP.2010.2042516.

[263] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. “Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks”. In: *Machine Learning in Medical Imaging*. Springer International Publishing, 2017, pp. 379–387. DOI: 10.1007/978-3-319-67389-9_44.

[264] J. R. Santillan and M. Makinano-Santillan. “VERTICAL ACCURACY ASSESSMENT OF 30-M RESOLUTION ALOS, ASTER, AND SRTM GLOBAL DEMS OVER NORTHEASTERN MINDANAO, PHILIPPINES”. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B4 (June 2016), pp. 149–156. DOI: 10.5194/isprs-archives-xli-b4-149-2016.

[265] F. Sattar et al. “Image enhancement based on a nonlinear multiscale method”. In: *IEEE Transactions on Image Processing* 6.6 (June 1997), pp. 888–895. ISSN: 1057-7149. DOI: 10.1109/83.585239.

[266] C. O. Sauer. “Mapping the Utilization of the Land”. In: *Geographical Review* 8.1 (July 1919), p. 47. DOI: 10.2307/207319.

[267] S. Saura. “Effects of minimum mapping unit on land cover data spatial configuration and composition”. In: *International Journal of Remote Sensing* 23.22 (Jan. 2002), pp. 4853–4880. DOI: 10.1080/01431160110114493.

[268] Konrad Schindler. “An Overview and Comparison of Smooth Labeling Methods for Land-Cover Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 50.11 (Nov. 2012), pp. 4534–4545. DOI: 10.1109/tgrs.2012.2192741.

[269] M. Schmitt, L. H. Hughes, and X. X. Zhu. “THE SEN1-2 DATASET FOR DEEP LEARNING IN SAR-OPTICAL DATA FUSION”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-1 (Sept. 2018), pp. 141–146. DOI: 10.5194/isprs-annals-iv-1-141-2018.

[270] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. DOI: 10.1109/cvpr.2015.7298682.

[271] Angela Schwering. "Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey". In: *Transactions in GIS* 12.1 (Feb. 2008), pp. 5–29. DOI: 10.1111/j.1467-9671.2008.01084.x.

[272] Linda See et al. "Building a hybrid land cover map with crowdsourcing and geographically weighted regression". In: 103 (May 2015), pp. 48–56. DOI: 10.1016/j.isprsjprs.2014.06.016.

[273] Oliver Sefrin, Felix M. Riese, and Sina Keller. "Deep Learning for Land Cover Change Detection". In: *Remote Sensing* 13.1 (Dec. 2020), p. 78. DOI: 10.3390/rs13010078.

[274] Jeong Chang Seong. "Modelling the accuracy of image data reprojection". In: *International Journal of Remote Sensing* 24.11 (Jan. 2003), pp. 2309–2321. DOI: 10.1080/01431160210154038.

[275] M. Sester, Y. Feng, and F. Thiemann. "BUILDING GENERALIZATION USING DEEP LEARNING". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-4 (Sept. 2018), pp. 565–572. DOI: 10.5194/isprs-archives-xlii-4-565-2018.

[276] Monika Sester. "Cartographic generalization". In: *Journal of Spatial Information Science* 21 (Dec. 2020). DOI: 10.5311/josis.2020.21.716.

[277] Atharva Sharma et al. "A patch-based convolutional neural network for remote sensing image classification". In: *Neural Networks* 95 (Nov. 2017), pp. 19–28. DOI: 10.1016/j.neunet.2017.07.017.

[278] K. Stuart Shea and Robert B. Mcmaster. "Cartographic Generalization in a Digital Environment: When and How to Generalize". In: *Proceedings Auto-Carto 9*. 1989, p. 5667.

[279] Yaxin Shi et al. "Label Embedding with Partial Heterogeneous Contexts". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 4926–4933. DOI: 10.1609/aaai.v33i01.33014926.

[280] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6.1 (July 2019). DOI: 10.1186/s40537-019-0197-0.

[281] Julius Sim and Chris C Wright. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements". In: *Physical Therapy* 85.3 (Mar. 2005), pp. 257–268. DOI: 10.1093/ptj/85.3.257.

[282] Jake Snell, Kevin Swersky, and Richard S. Zemel. "Prototypical Networks for Few-shot Learning". In: *NIPS*. 2017. DOI: 10.48550/ARXIV.1703.05175.

[283] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. "SELFIE: Refurbishing Unclean Samples for Robust Deep Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 5907–5915. URL: https://proceedings.mlr.press/v97/song19b.html.

[284] Hwanjun Song et al. "Learning From Noisy Labels With Deep Neural Networks: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* (2022), pp. 1–19. DOI: 10.1109/tnnls.2022.3152527.

[285] Xiao-Peng Song, Chengquan Huang, and John R. Townshend. "Improving global land cover characterization through data fusion". In: *Geo-spatial Information Science* 20.2 (Apr. 2017), pp. 141–150. DOI: 10.1080/10095020.2017.1323522.

[286] Daniel R Steinwand. "A new approach to categorical resampling". In: *American Congress on Surveying and Mapping Spring Conference*. 2003.

[287] Sebastien Strebelle. "Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics". In: *Mathematical Geology* 34.1 (2002), pp. 1–21. DOI: 10.1023/a:1014009426274.

[288] J. Stuckens, P.R. Coppin, and M.E. Bauer. "Integrating Contextual Information with per-Pixel Classification for Improved Land Cover Classification". In: *Remote Sensing of Environment* 71.3 (Mar. 2000), pp. 282–296. DOI: 10.1016/s0034-4257(99)00083-8.

[289] Martin Sudmanns et al. "Assessing global Sentinel-2 coverage dynamics and data availability for operational Earth observation (EO) applications using the EO-Compass". In: *International Journal of Digital Earth* 13.7 (Feb. 2019), pp. 768–784. DOI: 10.1080/17538947.2019.1572799.

[290] Damien Sulla-Menashe and Mark A Friedl. "User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product". In: *USGS: Reston, VA, USA* 1 (2018), p. 18.

[291] Peijun Sun, Russell G. Congalton, and Yaozhong Pan. "Improving the Upscaling of Land Cover Maps by Fusing Uncertainty and Spatial Structure Information". In: *Photogrammetric Engineering & Remote Sensing* 84.2 (2018), pp. 87–100. DOI: 10.14358/PERS.84.2.87. URL: https://www.semanticscholar.org/paper/76832db356f5a36bc650081187eac9cfa01ef59a.

[292] Zoltan Szantoi et al. "Addressing the need for improved land cover map products for policy support". In: *Environmental Science & Policy* 112 (Oct. 2020), pp. 28–35. DOI: 10.1016/j.envsci.2020.04.005.

[293] Shiteng Tan et al. "Upscaling approach to land cover based on priority and semantic proximity rules". In: *Remote Sensing for Natural Resources* 28.1, 50 (2016), p. 50. DOI: 10.6046/gtzyyg.2016.01.08. URL: https://www.gtzyyg.com/EN/abstract/article_1932.shtml.

[294] Kevin Tang et al. “Improving Image Classification with Location Context”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Dec. 2015. DOI: 10.1109/iccv.2015.121.

[295] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[296] Agnieszka Tarko et al. “Producing consistent visually interpreted land cover reference data: learning from feedback”. In: *International Journal of Digital Earth* 14.1 (Feb. 2020), pp. 52–70. DOI: 10.1080/17538947.2020.1729878.

[297] Ryutaro TATEISHI, Cheng-Gang WEN, and L. Kithsiri PERERA. “Global Four-minute Land Cover Data Set.” In: *Journal of the Japan society of photogrammetry and remote sensing* 36.4 (1997), pp. 62–74. DOI: 10.4287/jsprs.36.4_62.

[298] A. J. Tatem et al. “Increasing the spatial resolution of agricultural land cover maps using a Hopfield neural network”. In: *International Journal of Geographical Information Science* 17.7 (Oct. 2003), pp. 647–672. DOI: 10.1080/1365881031000135519.

[299] Luke Taylor and Geoff Nitschke. “Improving Deep Learning with Generic Data Augmentation”. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Nov. 2018. DOI: 10.1109/ssci.2018.8628742.

[300] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. “A Global Geometric Framework for Nonlinear Dimensionality Reduction”. In: *Science* 290.5500 (Dec. 2000), pp. 2319–2323. DOI: 10.1126/science.290.5500.2319.

[301] Frank Thiemann and Monika Sester. “An Automatic Approach for Generalization of Land-Cover Data from Topographic Data”. In: *Geotechnologies and the Environment*. Springer International Publishing, Oct. 2017, pp. 193–207. DOI: 10.1007/978-3-319-52522-8_10.

[302] Truong Van Thinh et al. “How Does Land Use/Land Cover Map’s Accuracy Depend on Number of Classification Classes?” In: *SOLA* 15.0 (2019), pp. 28–31. DOI: 10.2151/sola.2019-006.

[303] Mark Thompson. “A standard land-cover classification scheme for remote-sensing applications in South Africa”. In: *South African Journal of Science* 92.1 (1996), pp. 34–42. DOI: 10.10520/AJA00382353_7698. eprint: https://journals.co.za/doi/pdf/10.10520/AJA00382353_7698. URL: https://journals.co.za/doi/abs/10.10520/AJA00382353_7698.

[304] M. W. Thornton, P. M. Atkinson, and D. A. Holland. “Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping”. In: *International Journal of Remote Sensing* 27.3 (Feb. 2006), pp. 473–491. DOI: 10.1080/01431160500207088.

[305] Valeria Tomaselli et al. “Translating land cover/land use classifications to habitat taxonomies for landscape monitoring: a Mediterranean assessment”. In: *Landscape Ecology* 28.5 (Mar. 2013), pp. 905–930. DOI: 10.1007/s10980-013-9863-3.

[306] Guillaume Touya, Xiang Zhang, and Imran Lokhat. "Is deep learning the new agent for map generalization?" In: *International Journal of Cartography* 5.2-3 (May 2019), pp. 142–157. DOI: 10.1080/23729333.2019.1613071.

[307] Chau Tran et al. "Facebook AI WMT21 News Translation Task Submission". In: *CoRR* abs/2108.03265 (2021). arXiv: 2108.03265. URL: https://arxiv.org/abs/2108.03265.

[308] A. Trlica et al. "Albedo, Land Cover, and Daytime Surface Temperature Variation Across an Urbanized Landscape". In: *Earth's Future* 5.11 (Nov. 2017), pp. 1084–1101. DOI: 10.1002/2017ef000569.

[309] Nandin-Erdene Tsendbazar, Sytze de Bruin, and Martin Herold. "Integrating global land cover datasets for deriving user-specific maps". In: 10.3 (Aug. 2016), pp. 219–237. DOI: 10.1080/17538947.2016.1217942.

[310] Nandin-Erdene Tsendbazar et al. *Copernicus Global Land Service: Land Cover 100m: version 3 Globe 2015-2019: Validation Report*. en. Tech. rep. 2020. DOI: 10.5281/ZENODO.3938974.

[311] Mao-Ning Tuanmu and Walter Jetz. "A global 1-km consensus land-cover product for biodiversity and ecosystem modelling". In: 23.9 (Apr. 2014), pp. 1031–1045. DOI: 10.1111/geb.12182.

[312] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. "Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances". In: *IEEE Geoscience and Remote Sensing Magazine* 4.2 (June 2016), pp. 41–57. DOI: 10.1109/mgrs.2016.2548504.

[313] Amos Tversky. "Features of similarity." In: 84.4 (1977), pp. 327–352. DOI: 10.1037/0033-295x.84.4.327.

[314] Arnas Uselis, Mantas Lukoševičius, and Lukas Stasytis. "Localized Convolutional Neural Networks for Geospatial Wind Forecasting". In: *Energies* 13 (July 2020), p. 3440. DOI: 10.3390/en13133440.

[315] Ava Vali, Sara Comai, and Matteo Matteucci. "Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review". In: *Remote Sensing* 12.15 (Aug. 2020), p. 2495. DOI: 10.3390/rs12152495.

[316] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. "Dimensionality reduction: a comparative review". In: *Journal of Machine Learning Research* 10 (2009), pp. 66–71.

[317] Christelle Vancutsem et al. "Harmonizing and Combining Existing Land Cover/Land Use Datasets for Cropland Area Monitoring at the African Continental Scale". In: *Remote Sensing* 5.1 (Dec. 2012), pp. 19–41. ISSN: 2072-4292. DOI: 10.3390/rs5010019.

[318] Ashish Vaswani et al. “Attention Is All You Need”. In: *NIPS*. 2017. arXiv: 1706.03762 [cs.CL].

[319] J Verhoeye. “Land cover mapping at sub-pixel scales using linear optimization techniques”. In: *Remote Sensing of Environment* 79.1 (Jan. 2002), pp. 96–104. DOI: 10.1016/s0034-4257(01)00242-5.

[320] Valentin Vielzeuf et al. “Multilevel Sensor Fusion With Deep Learning”. In: *IEEE Sensors Letters* 3.1 (Jan. 2019), pp. 1–4. DOI: 10.1109/lsens.2018.2878908.

[321] Guillermo Villa et al. “Land Cover Classifications: An Obsolete Paradigm”. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B4. Beijing* (July 2008).

[322] James E Vogelmann, T Sohl, and Stephen M Howard. “Regional characterization of land cover using multiple sources of data”. In: *Photogrammetric Engineering and remote sensing* 64.1 (1998), pp. 45–57.

[323] James E Vogelmann et al. “Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources”. In: *Photogrammetric Engineering and Remote Sensing* 67.6 (2001).

[324] Wei Wang et al. “A Survey of Zero-Shot Learning”. In: *ACM Transactions on Intelligent Systems and Technology* 10.2 (Mar. 2019), pp. 1–37. DOI: 10.1145/3293318.

[325] Wenhui Wang et al. “MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS’20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[326] Yisen Wang et al. “Symmetric Cross Entropy for Robust Learning With Noisy Labels”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019. DOI: 10.1109/iccv.2019.00041.

[327] C. G. Wen and R. Tateishi. “30-second degree grid land cover classification of Asia”. In: *International Journal of Remote Sensing* 22.18 (Jan. 2001), pp. 3845–3854. DOI: 10.1080/01431160010014783.

[328] Ken C. L. Wong et al. “3D Segmentation with Exponential Logarithmic Loss for Highly Unbalanced Object Sizes”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, 2018, pp. 612–619. DOI: 10.1007/978-3-030-00931-1_70.

[329] P. D. Wu et al. “AGGREGATION IN LAND-COVER DATA GENERALIZATION CONSIDERING SPATIAL STRUCTURE CHARACTERISTICS”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-4/W9 (Sept. 2019), pp. 111–118. DOI: 10.5194/isprs-annals-iv-4-w9-111-2019.

[330] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation”. In: *CoRR* (2016). arXiv: 1609.08144 [cs.CL]. URL: http://arxiv.org/abs/1609.08144v2.

[331] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *ArXiv* abs/1609.08144 (2016).

[332] Yuxin Wu and Kaiming He. “Group Normalization”. In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 3–19. DOI: 10.1007/978-3-030-01261-8_1.

[333] Michael Wulder et al. *A Guide to the Estimation of Canada’s National Forest Inventory Attributes from Landsat TM Data*. Jan. 2001.

[334] Michael A. Wulder et al. “Land cover 2.0”. In: *International Journal of Remote Sensing* 39.12 (Mar. 2018), pp. 4254–4284. DOI: 10.1080/01431161.2018.1452075.

[335] B.K. Wyatt et al. *Comparison of land cover definitions.* Research rep. London: Department of the Environment, 1994.

[336] Tong Xiao et al. “Learning from massive noisy labeled data for image classification”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. DOI: 10.1109/cvpr.2015.7298885.

[337] Nan Xing et al. “Zero-Shot Learning via Discriminative Dual Semantic Auto-Encoder”. In: *IEEE Access* 9 (2021), pp. 733–742. DOI: 10.1109/access.2020.3046573.

[338] Gang Xu et al. “Remote Sensing Mapping of Build-Up Land with Noisy Label via Fault-Tolerant Learning”. In: *Remote Sensing* 14.9 (May 2022), p. 2263. DOI: 10.3390/rs14092263.

[339] Qianxiang Xu et al. “Modelling semantic uncertainty of land classification system”. PhD thesis. The Hong Kong Polytechnic University, 2014. URL: http://hdl.handle.net/10397/53693.

[340] Yonghao Xu et al. “Advanced Multi-Sensor Optical Remote Sensing for Urban Land Use and Land Cover Classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.6 (June 2019), pp. 1709–1724. DOI: 10.1109/jstars.2019.2911113.

[341] Chuan Yan et al. “Improved U-Net Remote Sensing Classification Algorithm Based on Multi-Feature Fusion Perception”. In: *Remote Sensing* 14.5 (Feb. 2022), p. 1118. DOI: 10.3390/rs14051118.

[342] Liang Yan et al. “Triplet Adversarial Domain Adaptation for Pixel-Level Classification of VHR Remote Sensing Images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.5 (May 2020), pp. 3558–3573. DOI: 10.1109/tgrs.2019.2958123.

[343] Xiongfeng Yan et al. "A graph convolutional neural network for classification of building patterns using spatial vector data". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (Apr. 2019), pp. 259–273. DOI: 10.1016/j.isprsjprs.2019.02.010.

[344] Hui Yang et al. "The Standardization and Harmonization of Land Cover Classification Systems towards Harmonized Datasets: A Review". In: *International Journal of Geo-Information* 6.5 (May 2017), p. 154. ISSN: 2220-9964. DOI: 10.3390/ijgi6050154.

[345] Wenli Yang and James W. Merchant. "Impacts of upscaling techniques on land cover representation in Nebraska, U.S.A." In: *Geocarto International* 12.1 (1997), pp. 27–39. DOI: 10.1080/10106049709354571. eprint: https://doi.org/10.1080/10106049709354571. URL: https://doi.org/10.1080/10106049709354571.

[346] Yinfei Yang et al. "Improving Multilingual Sentence Embedding using Bi-directional Dual Encoder with Additive Margin Softmax". In: *CoRR* abs/1902.08564 (2019). arXiv: 1902.08564. URL: http://arxiv.org/abs/1902.08564.

[347] Liu Yaolin et al. "Frameworks for generalization constraints and operations based on object-oriented data structure in database generalization". In: *Geo-spatial Information Science* 4.3 (Jan. 2001), pp. 42–49. DOI: 10.1007/bf02826923.

[348] Yifang Yin et al. "GPS2Vec". In: *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, Nov. 2019. DOI: 10.1145/3347146.3359067.

[349] Naoto Yokoya et al. "Open Data for Global Multimodal Land Use Classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.5 (May 2018), pp. 1363–1377. DOI: 10.1109/jstars.2018.2799698.

[350] Chenyu You et al. "CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE)". In: *IEEE Transactions on Medical Imaging* 39.1 (Jan. 2020), pp. 188–203. DOI: 10.1109/tmi.2019.2922960.

[351] Qiuze Yu et al. "Universal SAR and optical image registration via a novel SIFT framework based on nonlinear diffusion and a polar spatial-frequency descriptor". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 171 (2021), pp. 1–17. DOI: https://doi.org/10.1016/j.isprsjprs.2020.10.019.

[352] Wenhao Yu et al. "Crossing Variational Autoencoders for Answer Retrieval". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.498.

[353] Benito Zaragozi et al. “Improving the Usability of the Land Cover and Use Information System of Spain (SIOSE): A Proposal to Distribute New Thematic Layers and Predefined Reclassifications”. In: *Proceedings of the 6th International Conference on Geographical Information Systems Theory, Applications and Management*. SCITEPRESS - Science and Technology Publications, 2020. DOI: 10.5220/0009579502940301.

[354] Benito Zaragozi et al. “Integration of New Data Layers to Support the Land Cover and Use Information System of Spain (SIOSE): An Approach from Object-Oriented Modelling”. In: *Communications in Computer and Information Science*. Springer International Publishing, 2021, pp. 85–101. DOI: 10.1007/978-3-030-76374-9_6.

[355] Rina Zazkis and Stephen Campbell. “Prime decomposition: Understanding uniqueness”. In: *The Journal of Mathematical Behavior* 15.2 (June 1996), pp. 207–218. DOI: 10.1016/s0732-3123(96)90017-6.

[356] Baolei Zhang et al. “Understanding Land Use and Land Cover Dynamics from 1976 to 2014 in Yellow River Delta”. In: *Land* 6.1 (Mar. 2017), p. 20. DOI: 10.3390/land6010020.

[357] Daoqiang Zhang and Dinggang Shen. “Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer’s disease”. In: *NeuroImage*. Vol. 59. 2012, pp. 895–907.

[358] Han Zhang et al. “Self-Attention Generative Adversarial Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 7354–7363. URL: https://proceedings.mlr.press/v97/zhang19d.html.

[359] Pengbin Zhang et al. “Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery”. In: *Sensors* 18.11 (Nov. 2018), p. 3717. DOI: 10.3390/s18113717.

[360] Yihang Zhang et al. “Example-Based Super-Resolution Land Cover Mapping Using Support Vector Regression”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7.4 (Apr. 2014), pp. 1271–1283. DOI: 10.1109/jstars.2014.2305652.

[361] Yu Zhang and Qiang Yang. “An overview of multi-task learning”. In: *National Science Review* 5.1 (Sept. 2017), pp. 30–43. DOI: 10.1093/nsr/nwx105.

[362] Zengxiang Zhang et al. “A 2010 update of National Land Use/Cover Database of China at 1:100000 scale using medium spatial resolution satellite images”. In: *Remote Sensing of Environment* 149 (June 2014), pp. 142–154. DOI: 10.1016/j.rse.2014.04.004.

[363] Ellen D. Zhong et al. “Reconstructing continuous distributions of 3D protein structure from cryo-EM images”. In: *International Conference on Learning Representations*. 2020.

[364] Tianyi Zhou, Shengjie Wang, and Jeff A. Bilmes. "Robust Curriculum Learning: from clean label detection to noisy label self-correction". In: *ICLR*. 2021.

[365] Will Zou et al. "Generic Object Detection with Dense Neural Patterns and Regionlets". In: *Proceedings of the British Machine Vision Conference 2014*. British Machine Vision Association, 2014. DOI: 10.5244/c.28.72.

APPENDIX A

Land cover nomenclatures

color	id	name	Main semantic link				
			C	P	G1	G2	M
	1	Continuous urban fabric	111	9	1111	235	7
	2	Discontinuous urban fabric	112	9	1111	235	6
	3	Industrial or commercial units	121	9	1111	235	8
	4	Road surfaces	122	9	1112	411	10
	5	rapeseeds	211	8	221	11	3
	6	cereals	211	8	221	11	3
	7	protein crops	211	8	221	11	3
	8	soy	211	8	221	11	3
	9	sunflower	211	8	221	11	3
	10	maize	211	8	221	11	3
	11	rice	211	8	221	11	3
	12	tubers	211	8	221	11	3
	13	Intensive grassland	231	4	221	11	3
	14	Orchards	222	3	2111	11	3
	15	Vineyards	221	2	213	11	3
	16	Broad-leaved forest	311	1	2111	12	1
	17	Coniferous forest	312	1	2112	12	1
	18	Natural grasslands	321	4	221	63	2
	19	Woody moorlands	324	3	212	63	2
	20	Bare rock	332	7	121	63	2
	21	Beaches, dunes and sand plains	331	7	1121	63	2
	22	Glaciers and perpetual snow	335	10	123	63	2
	23	Water bodies	523	12	122	14	4

Table A.1: OSO nomenclature. The main semantic link column gives for each OSO class the semantically closest class in the other LULC.

color	id	name	Main semantic link				
			P	O	G1	G2	M
	111	Continuous urban fabric	9	1	1111	235	7
	112	Discontinuous urban fabric	9	2	1111	235	6
	121	Industrial or commercial units	9	3	1111	235	8
	122	Road and rail networks and associated land	9	4	1112	411	10
	123	Port areas	9	3	1112	414	8
	124	Airports	9	3	1112	413	10
	131	Mineral extraction sites	7	20	1121	13	11
	132	Dump sites	9	20	1122	43	11
	133	Construction sites	9	20	1121	61	11
	141	Green urban areas	4	13	221	235	5
	142	Sport and leisure facilities	9	3	1111	235	5
	211	Non-irrigated arable land	8	6	221	11	3
	212	Permanently irrigated land	8	13	221	11	3
	213	Rice fields	8	11	221	11	3
	221	Vineyards	3	15	213	11	3
	222	Fruit trees and berry plantations	2	14	2111	11	3
	223	Olive groves	2	14	2111	11	3
	231	Pastures	4	13	221	11	3
	241	Annual crops associated with permanent crops	8	13	221	11	3
	242	Complex cultivation patterns	8	14	221	11	3
	243	Mainly agriculture but significant areas of natural vegetation	8	14	2111	12	3
	244	Agro-forestry areas	1	16	2111	12	1
	311	Broad-leaved forest	1	16	2111	12	1
	312	Coniferous forest	1	17	2112	12	1
	313	Mixed forest	1	17	2113	12	1
	321	Natural grassland	4	18	221	63	2
	322	Moors and heathland	3	19	212	63	2
	323	Sclerophyllous vegetation	3	19	212	63	1
	324	Transitional woodland/shrub	3	19	212	12	1
	331	Beaches, dunes, sands	7	21	121	63	2
	332	Bare rock	7	20	121	63	2
	333	Sparsely vegetated areas	7	18	221	63	2
	334	Burnt areas	3	19	212	63	2
	335	Glaciers and perpetual snow	10	22	123	63	2
	411	Inland marshes	5	19	212	11	2
	412	Peatbogs	5	19	212	11	2
	421	Salt marshes	7	21	1121	63	2
	422	Salines	11	21	1121	13	11
	423	Intertidal flats	11	21	1121	63	2
	511	Water courses	11	23	122	63	4
	512	Water bodies	11	23	122	63	4
	521	Coastal lagoons	11	23	122	63	4
	522	Estuaries	12	23	122	63	4
	523	Sea and ocean	12	23	122	63	4

Table A.2: CLC nomenclature. The main semantic link column gives for each CLC class the semantically closest class in the other LULC.

color	id	name	Main semantic link				
			C	P	O	G1	G2
	1	Forest	311	1	16	2111	12
	2	Semi-natural areas	321	3	18	221	11
	3	Crops	211	8	6	221	11
	4	Water	511	11	23	122	414
	5	Artificialized green urban areas	142	4	2	221	235
	6	Individual housing	112	9	2	1111	235
	7	Colective housing	111	9	1	1111	235
	8	Activities	121	9	3	1111	235
	9	Facilities	111	9	2	1111	235
	10	Transports	122	9	4	1112	411
	11	Mine/dump/construction	131	9	3	1121	13

Table A.3: MOS nomenclature. The main semantic link column gives for each MOS class the semantically closest class in the other LULC.

color	id	name	Main semantic link				
			C	P	O	G2	M
	1111	Built-up areas	111	9	1	235	6
	1112	Undeveloped areas	122	9	4	411	10
	1121	Mineral material areas	131	9	21	412	11
	1122	Areas with other composite materials	132	9	3	43	11
	121	Bare soils	332	7	20	63	2
	122	Water surfaces	512	11	23	414	4
	123	Snowfields and glaciers	335	10	22	63	2
	2111	Deciduous stands	311	1	16	12	1
	2112	Conifer stands	312	1	17	12	1
	2113	Mixed stands	313	1	16	12	1
	212	Shrub and sub-shrub formations	324	3	19	63	1
	213	Other woody formations	221	3	15	11	3
	221	Herbaceous formations	211	8	6	11	3
	222	Other non-woody formations	334	4	18	63	2

Table A.4: OCS-GEc nomenclature. The main semantic link column gives for each OCS-GEc class the semantically closest class in the other LULC.

color	id	name	Main semantic link				
			C	P	O	G1	M
	11	Agriculture	211	8	6	221	3
	12	Forestry	311	1	16	2111	1
	13	Extraction activities	131	7	20	1121	11
	14	Fisheries and aquaculture	521	11	23	122	4
	235	Secondary or tertiary production and residential usage	112	9	2	1111	6
	411	Road networks	122	9	4	1112	10
	412	Rails networks	122	9	4	1121	10
	413	Overhead networks	124	9	23	1112	10
	414	River and maritime transport networks	123	12	3	122	10
	42	Logistics and storage services	121	9	3	1111	8
	43	Public utility networks	121	9	3	1111	8
	61	Transitionnal Areas	133	9	3	1121	11
	62	Abandoned areas	322	2	3	212	11
	63	Without use	321	2	18	212	2

Table A.5: OCS-GEu nomenclature. The main semantic link column gives for each OCS-GEu class the semantically closest class in the other LULC.

color	id	name	Main semantic link				
			C	O	G1	G2	M
	11	Closed forest	311	16	2111	12	1
	12	Open forest	231	16	212	12	1
	20	Shrubland	221	15	213	11	3
	30	Herbaceous vegetation	321	13	221	11	3
	90	Herbaceous wetland	411	23	122	63	4
	100	Moss and lichen	333	20	222	63	2
	60	Bare / sparse vegetation	332	20	121	63	2
	40	Cropland	211	6	221	11	3
	50	Built-up	112	2	1111	235	6
	70	Snow and ice	335	22	123	63	2
	80	Permanent water bodies	512	23	122	14	4
	200	Ocean	523	23	122	414	4

Table A.6: CGLS-LC100 nomenclature. The main semantic link column gives for each CGLS-LC100 class the semantically closest class in the other LULC.

APPENDIX B

Semantic correpondance analysis between maps of the MLULC dataset

The following figure presents the porposed handcarfted semantic translation from source (in row) to target in column for all couple of source/target maps. Blue square indicates strong semantic correspondance. The codes of classes are provided in appendix. For instance, in Figure B.1 the first row presents the possible semantic association for the CGLS class 1 (Closed Forest) which establish three semantic semantic relation with CLC class 22,23,24,25 (Agro-Forestry, Broad-leaved Forest, Coniferous Forest, Mixed Forest). Note that some possible semantic translations are not represented as they are perceived as weaker. For instance, it could also be translated in CLC class 21 (Land principally occupied by agriculture, with significant areas of natural vegetation) but the semantic link is weaker as class 21 includes other stuffs than Forest.

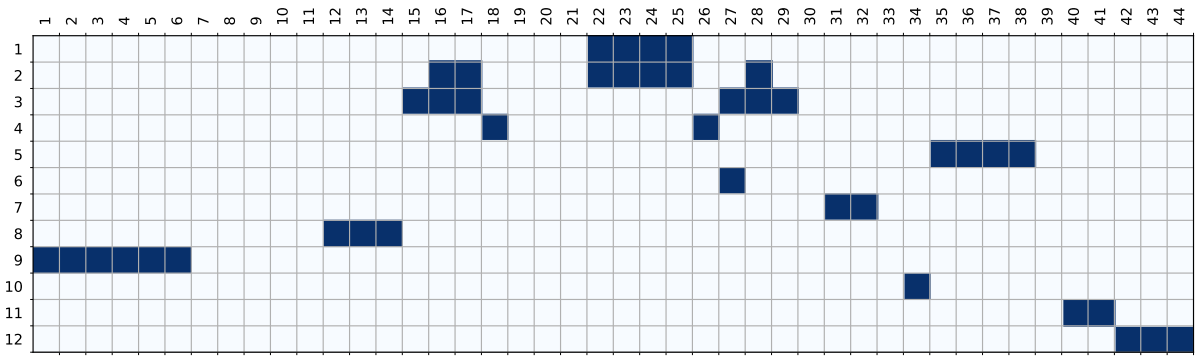

Figure B.1: Semantic translation from CGLS-LC100 (in row) to CLC (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

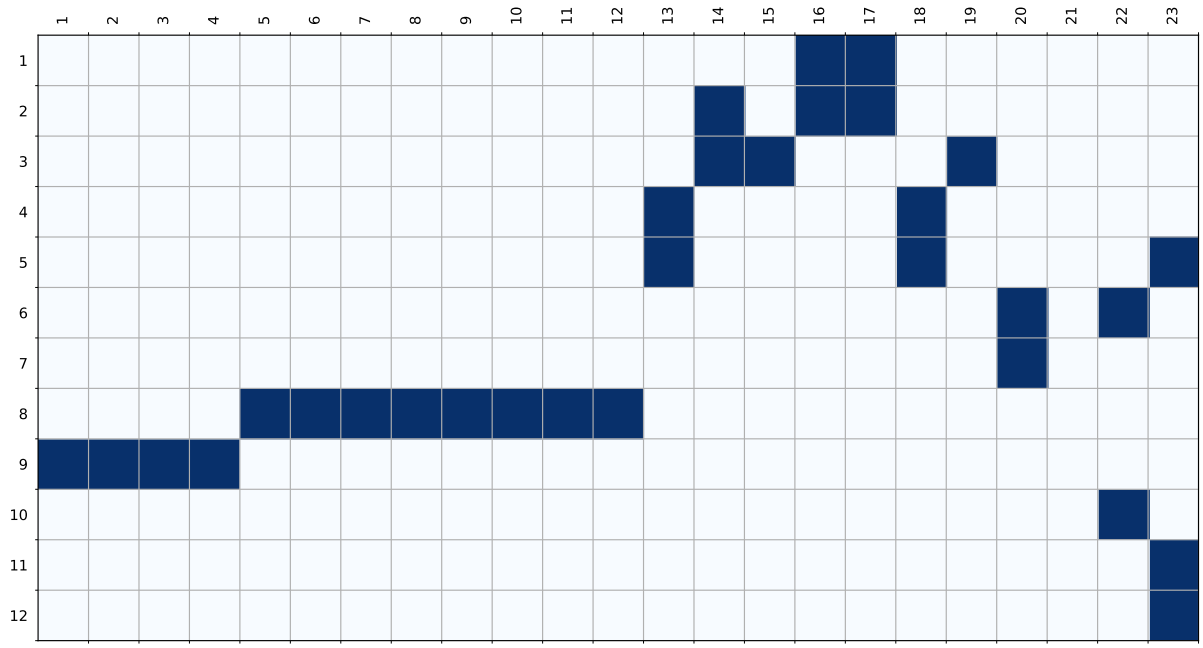

Figure B.2: Semantic translation from CGLS-LC100 (in row) to OSO (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

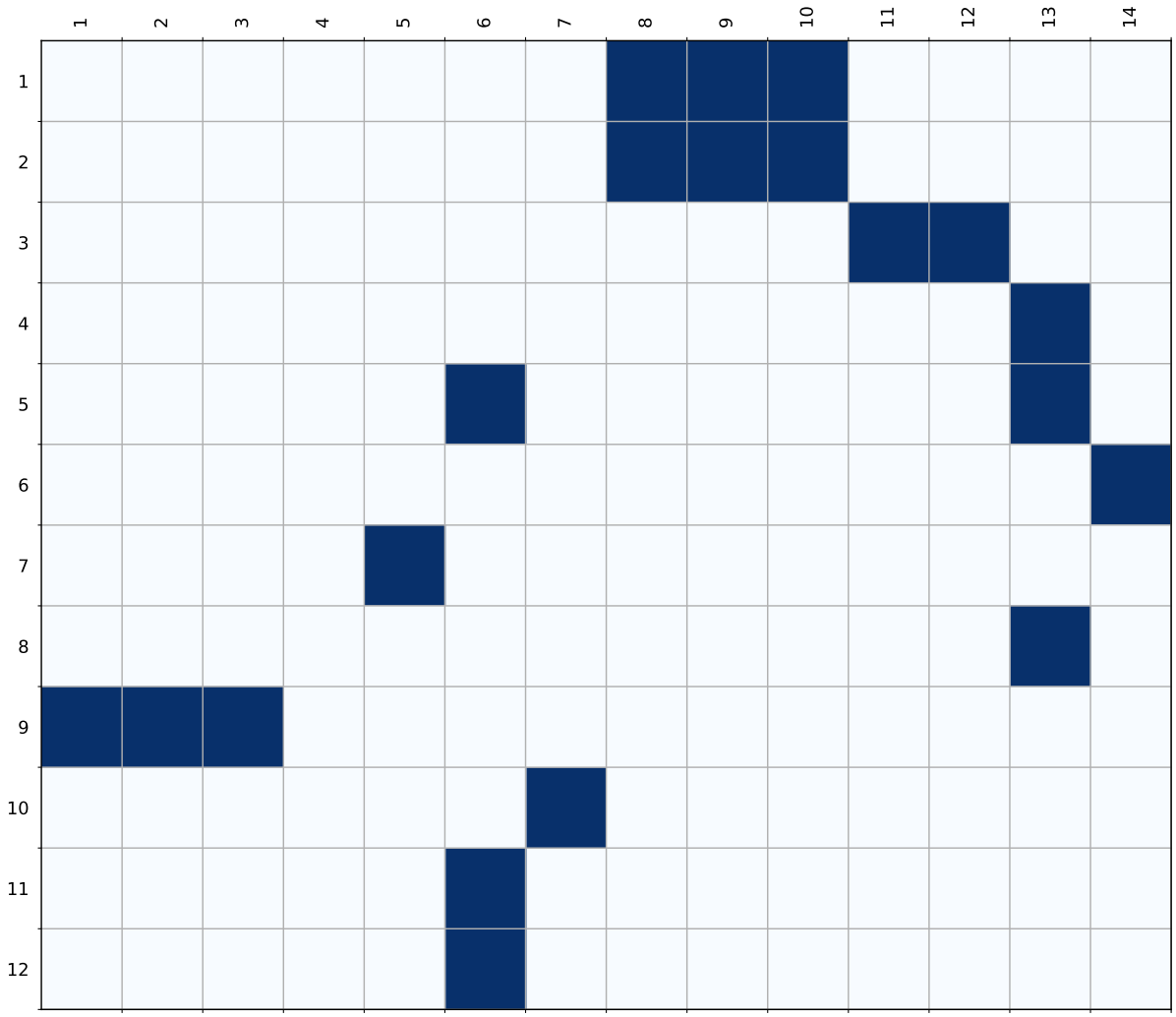

Figure B.3: Semantic translation from CGLS-LC100 (in row) to OCSGE-cover (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

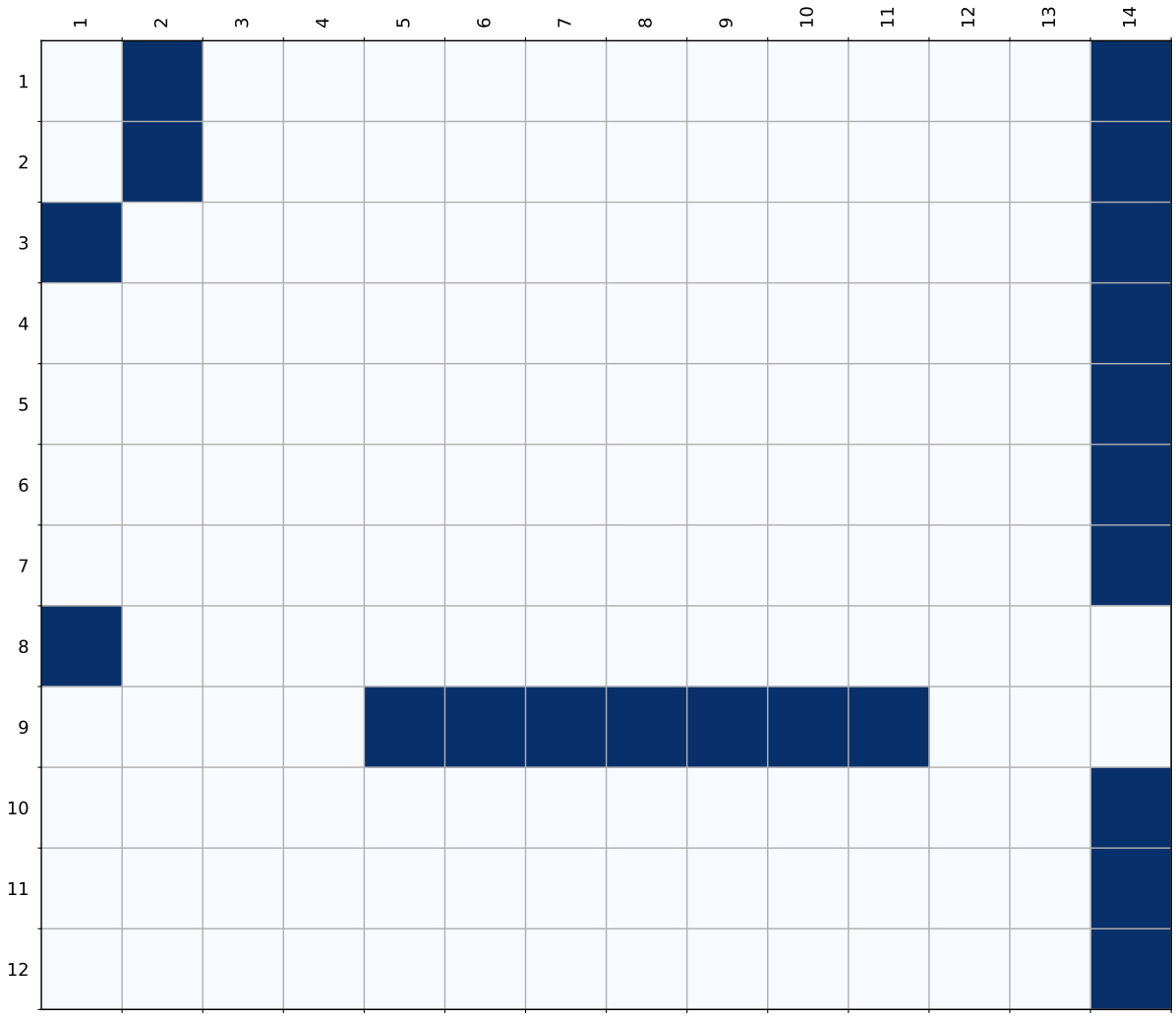

Figure B.4: Semantic translation from CGLS-LC100 (in row) to OCSGE-use (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

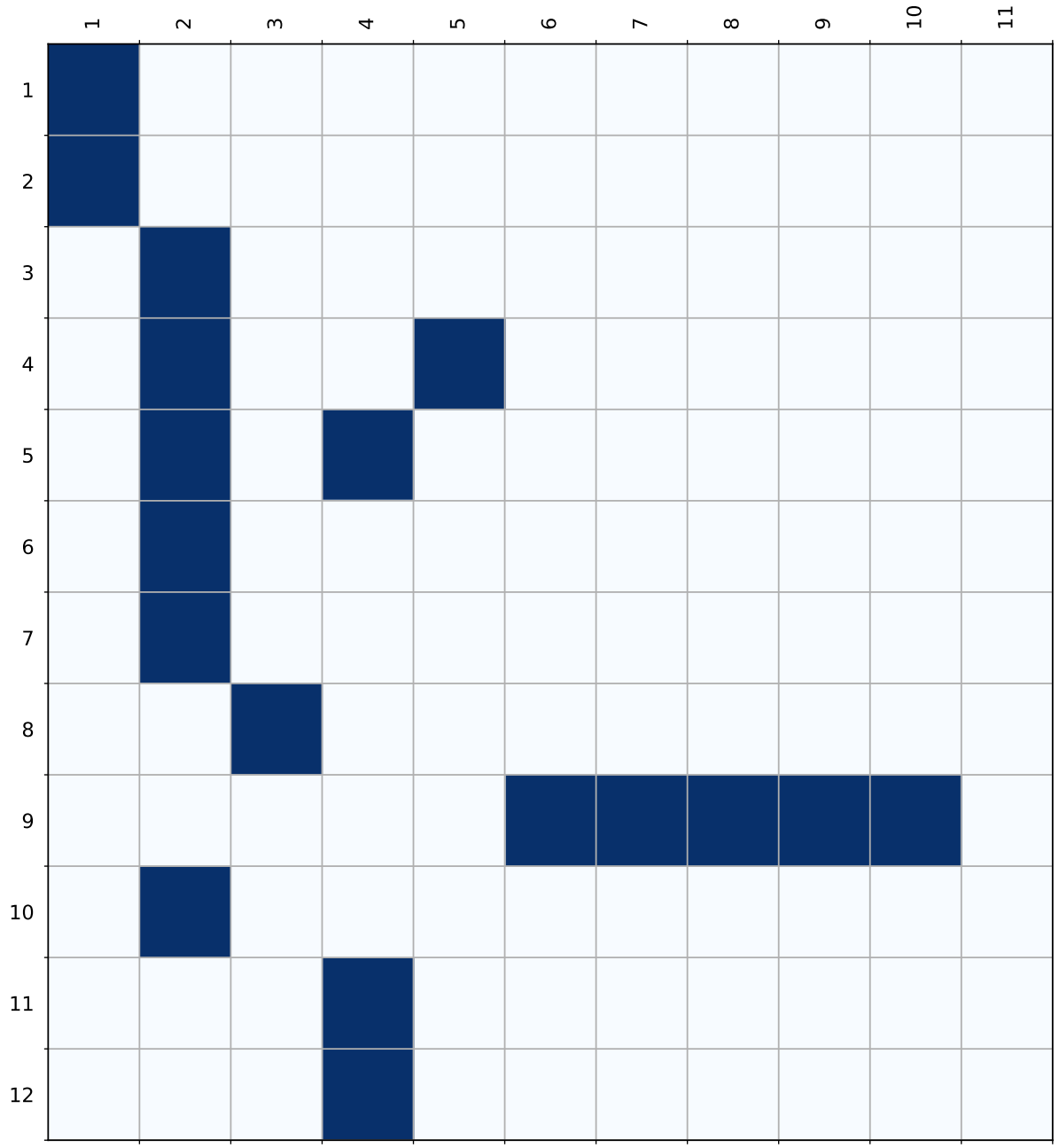

Figure B.5: Semantic translation from CGLS-LC100 (in row) to MOS (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

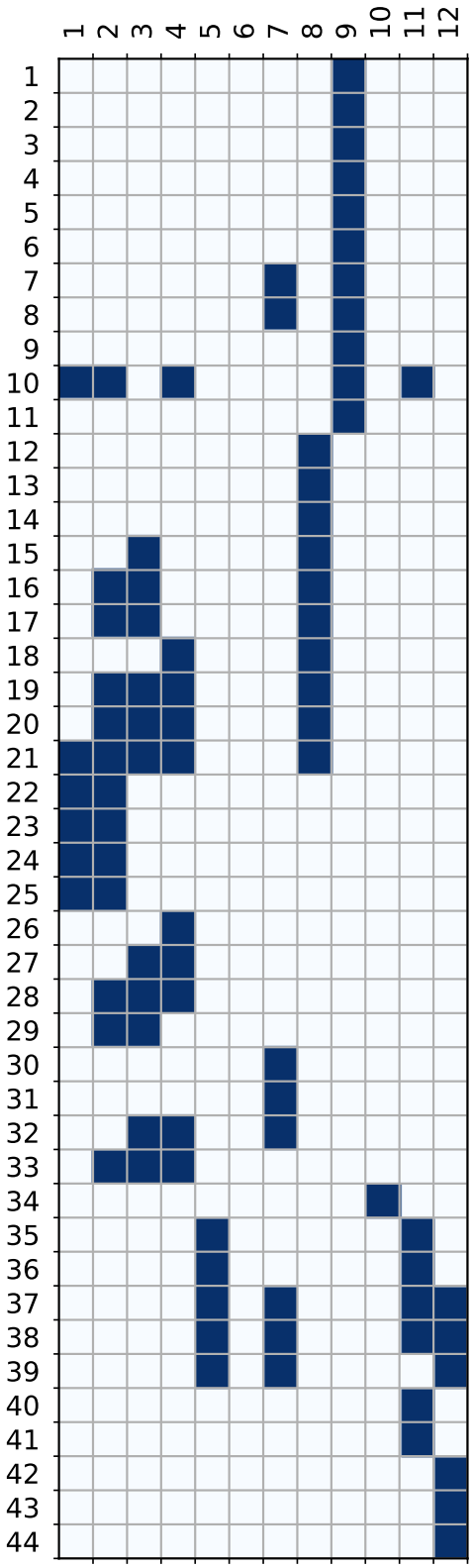

Figure B.6: Semantic translation from CLC (in row) to CGLS-LC100 (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

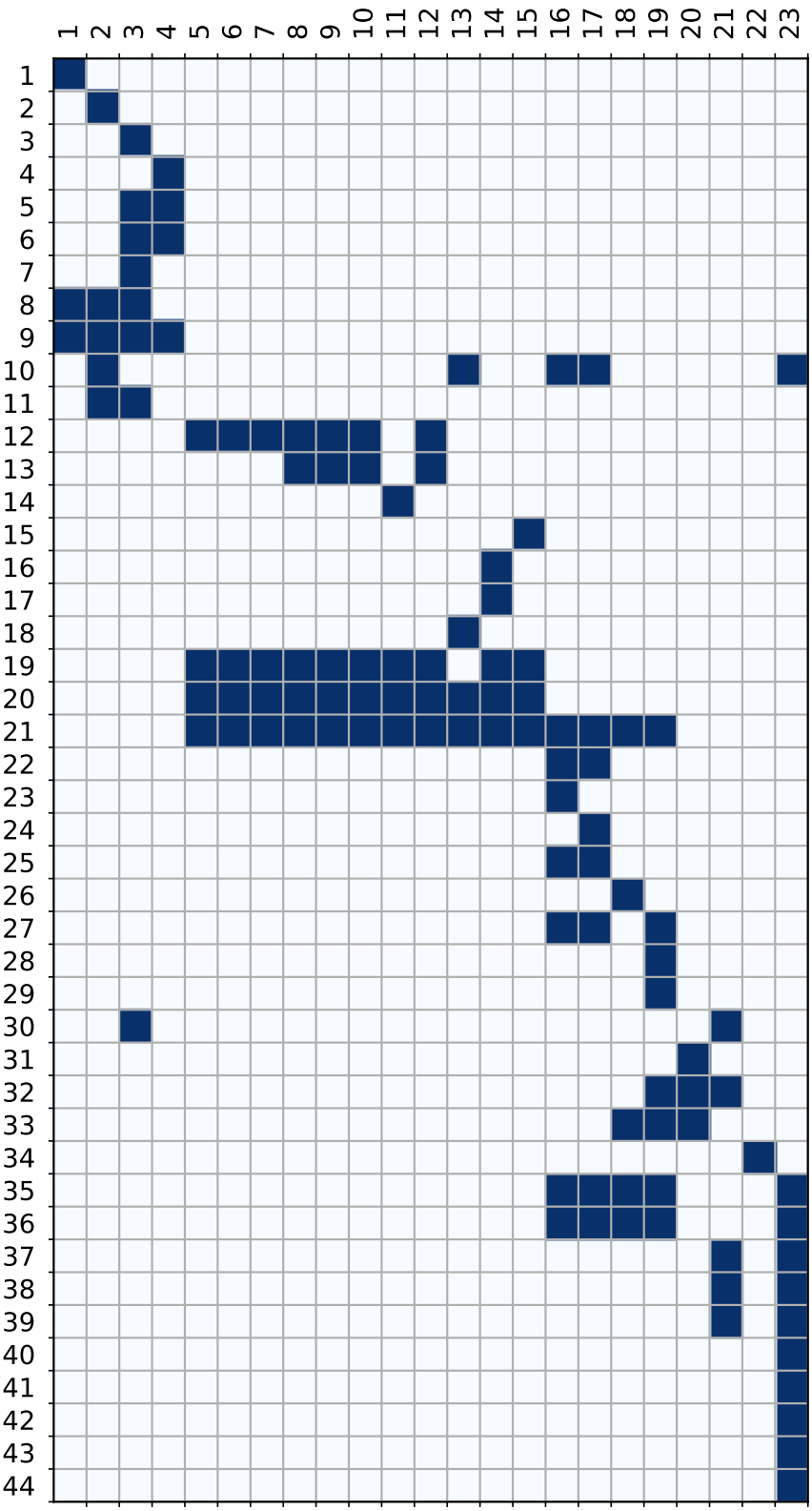

Figure B.7: Semantic translation from CLC (in row) to OSO (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

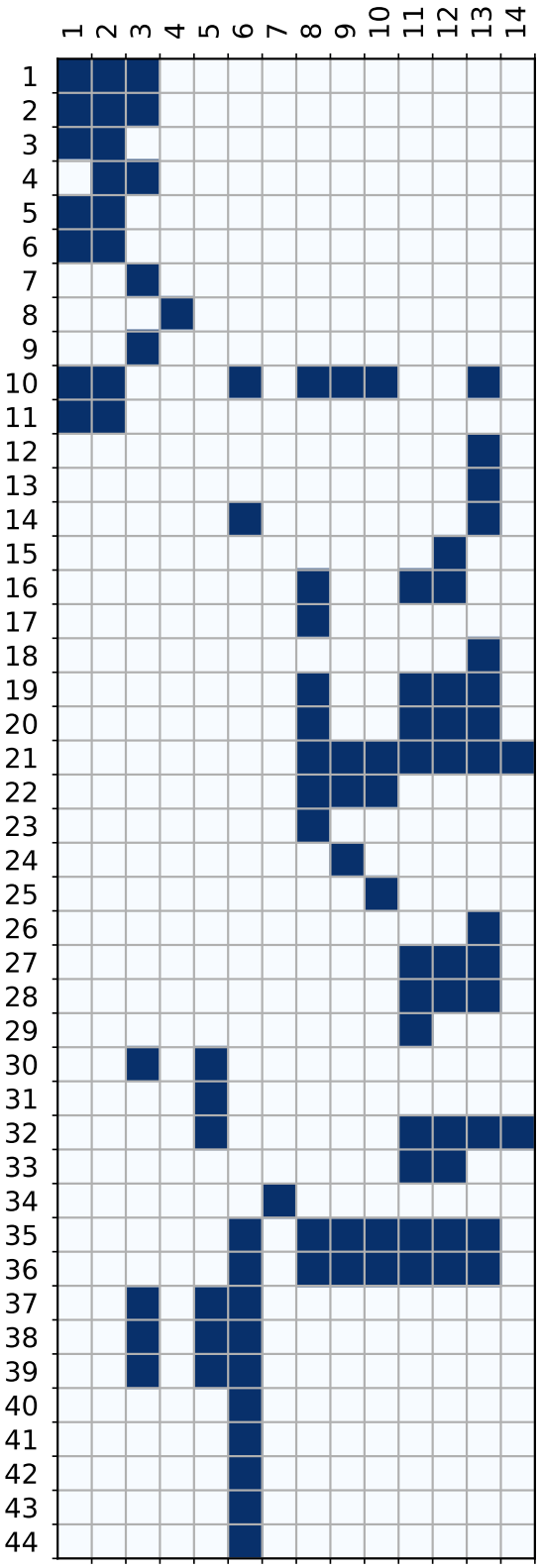

Figure B.8: Semantic translation from CLC (in row) to OCSGE-cover (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

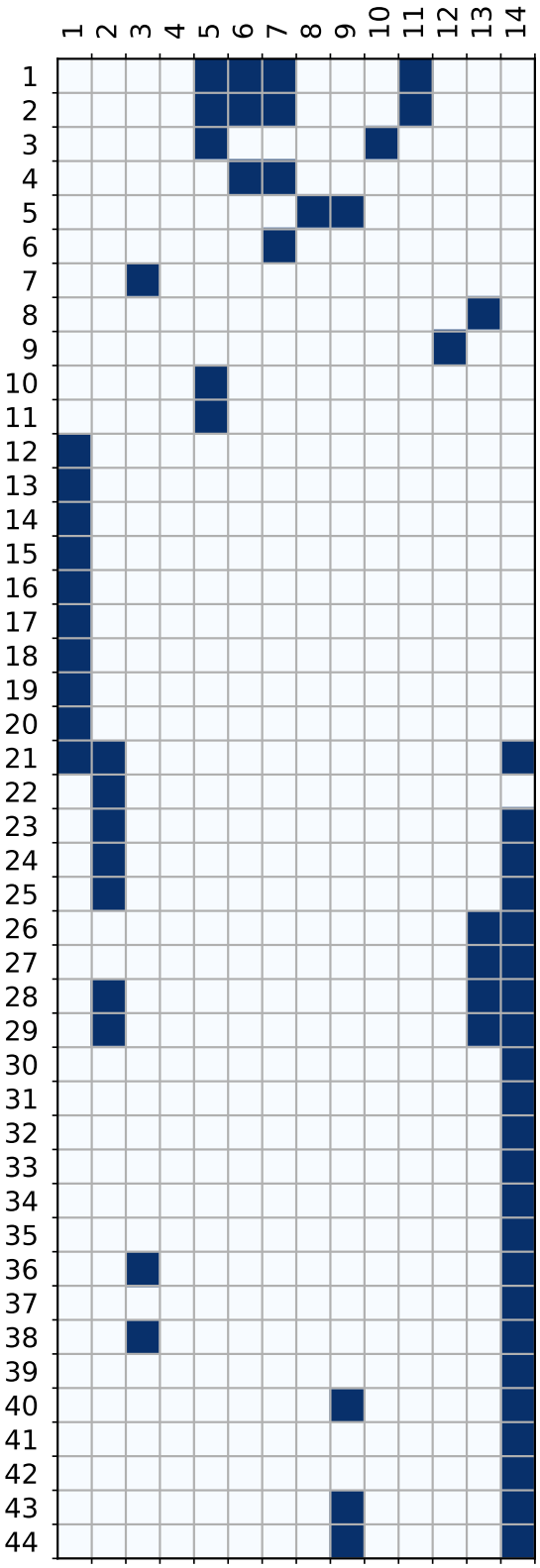

Figure B.9: Semantic translation from CLC (in row) to OCSGE-use (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

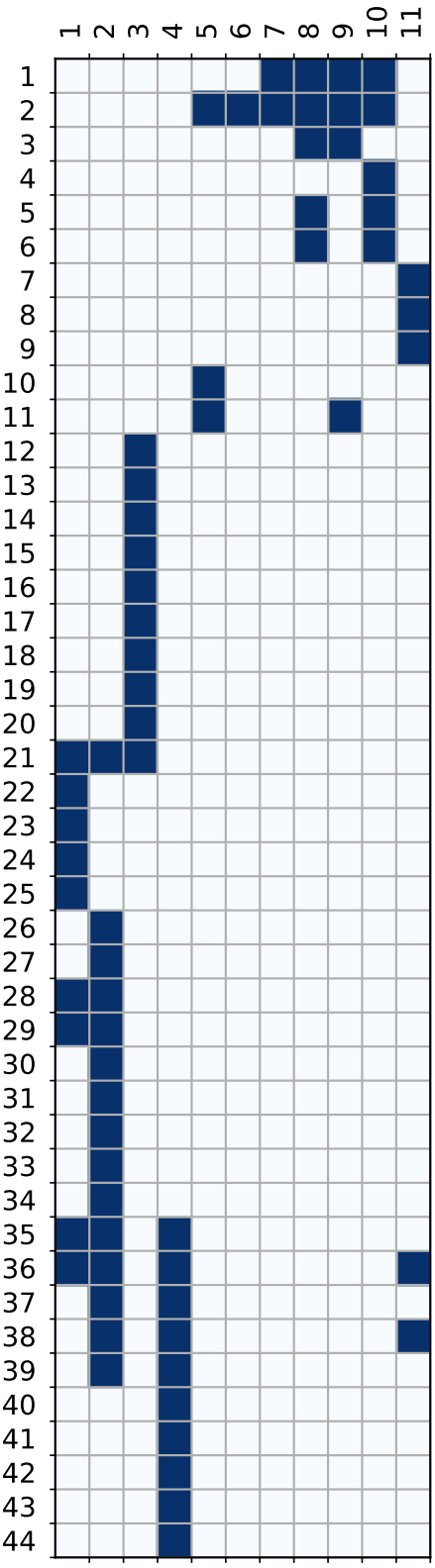

Figure B.10: Semantic translation from CLC (in row) to MOS (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

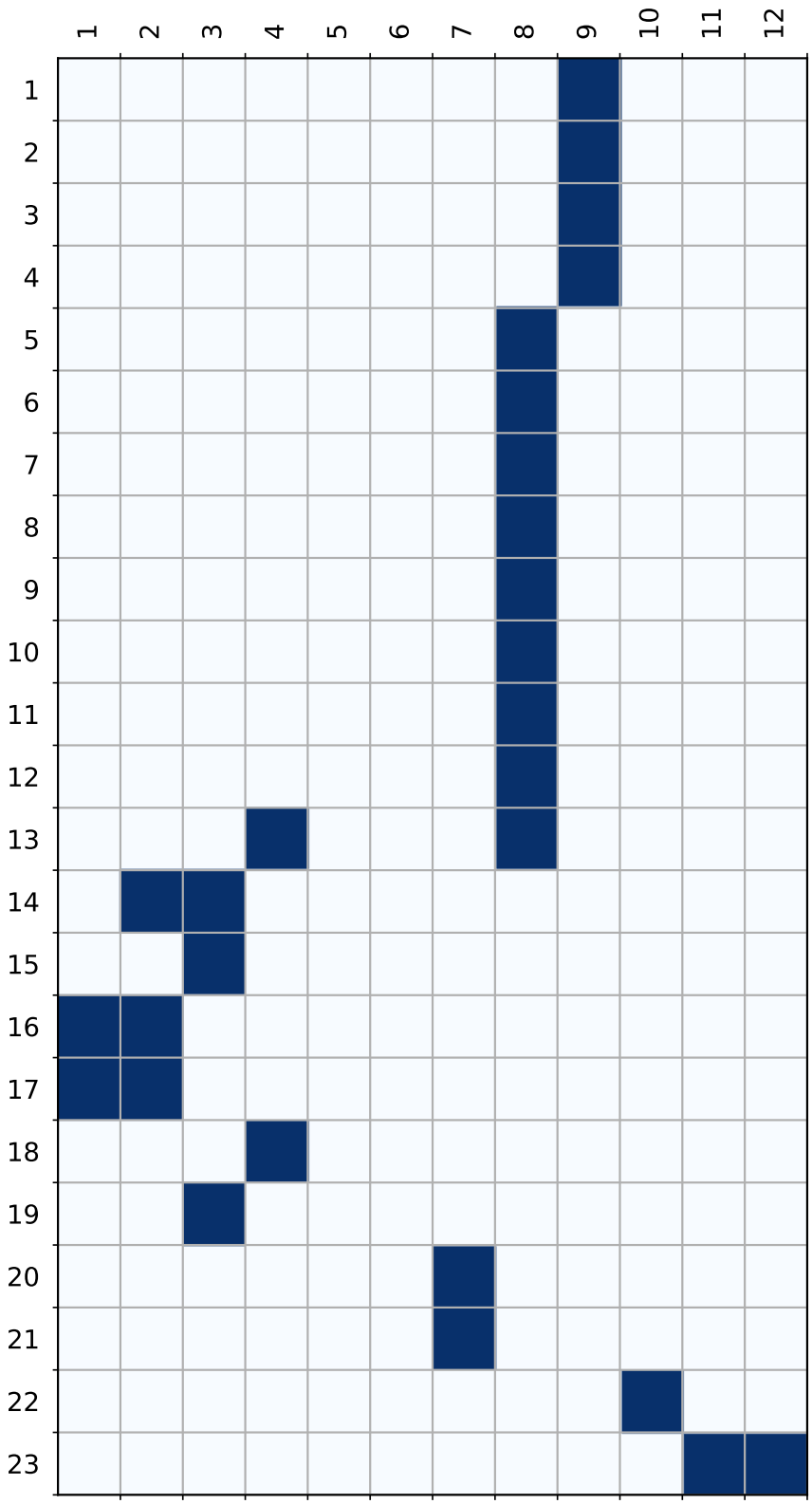

Figure B.11: Semantic translation from OSO (in row) to CGLS-LC100 (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

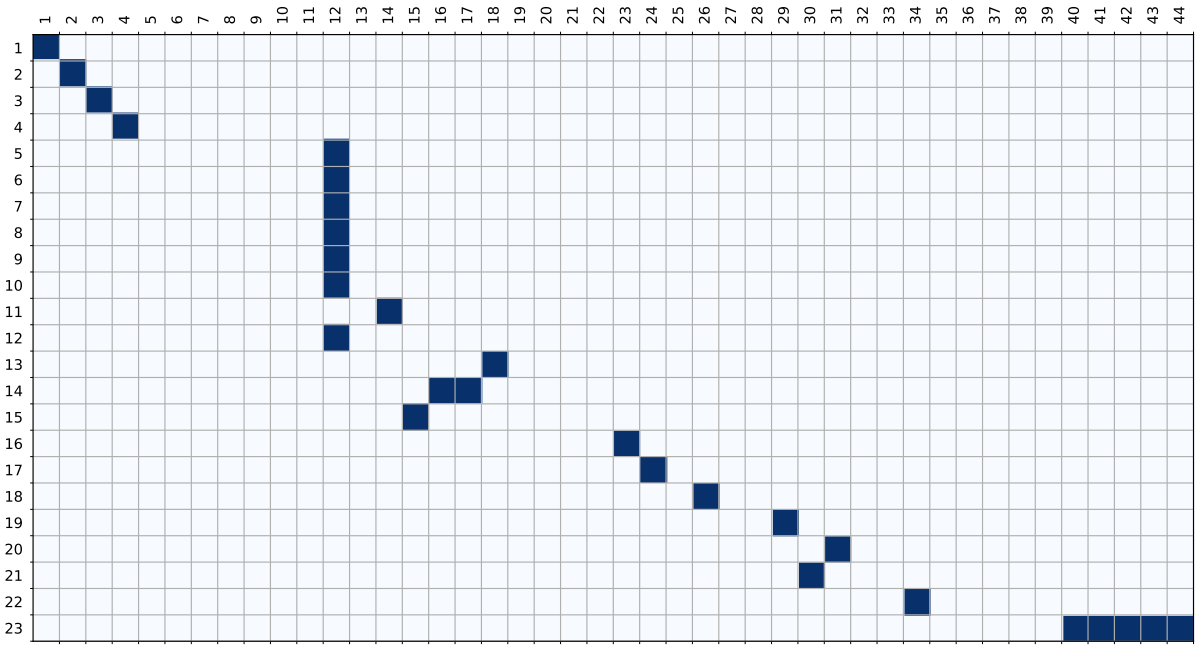

Figure B.12: Semantic translation from OSO (in row) to CLC (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

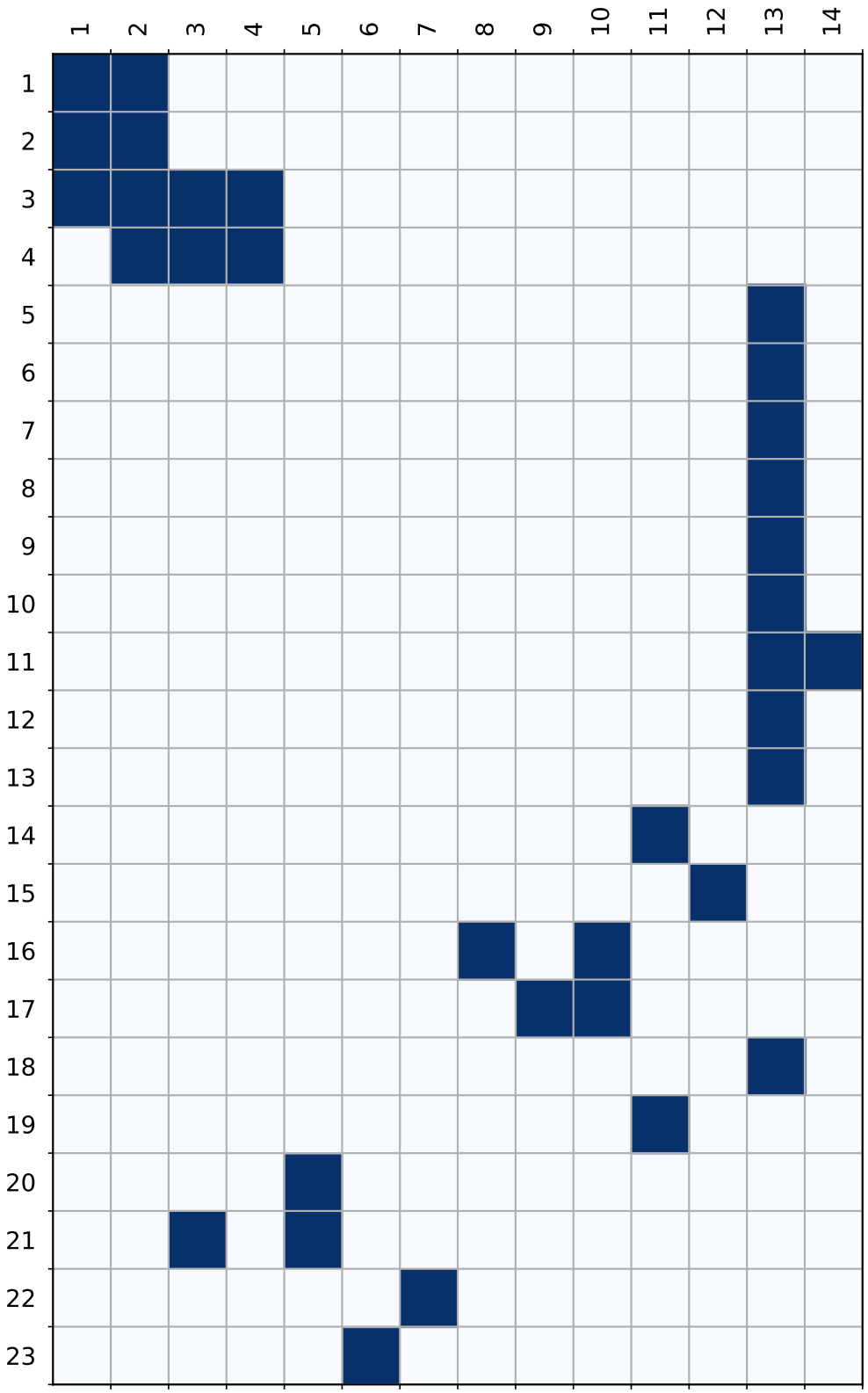

Figure B.13: Semantic translation from OSO (in row) to OCSGE-cover (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

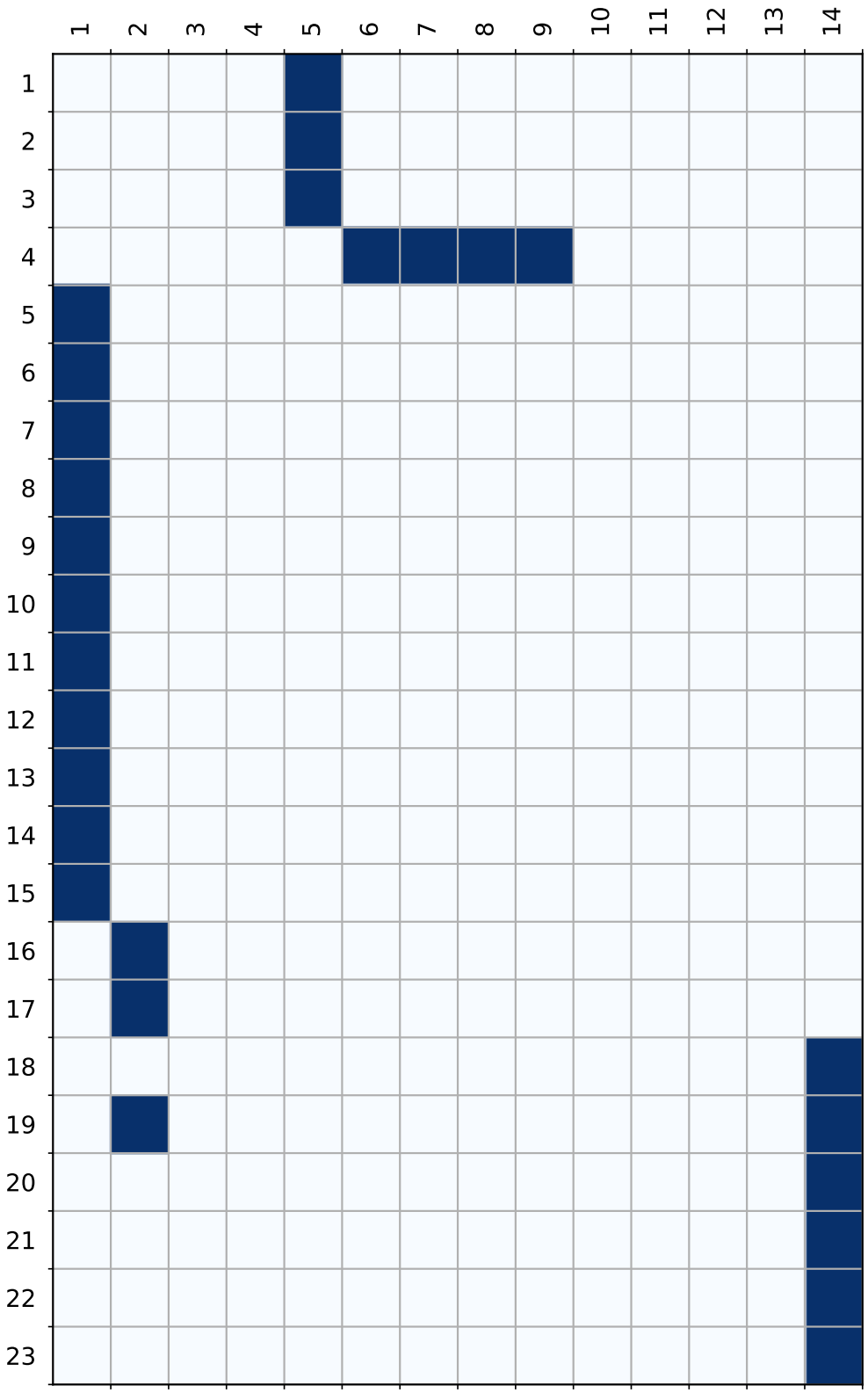

Figure B.14: Semantic translation from OSO (in row) to OCSGE-use (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

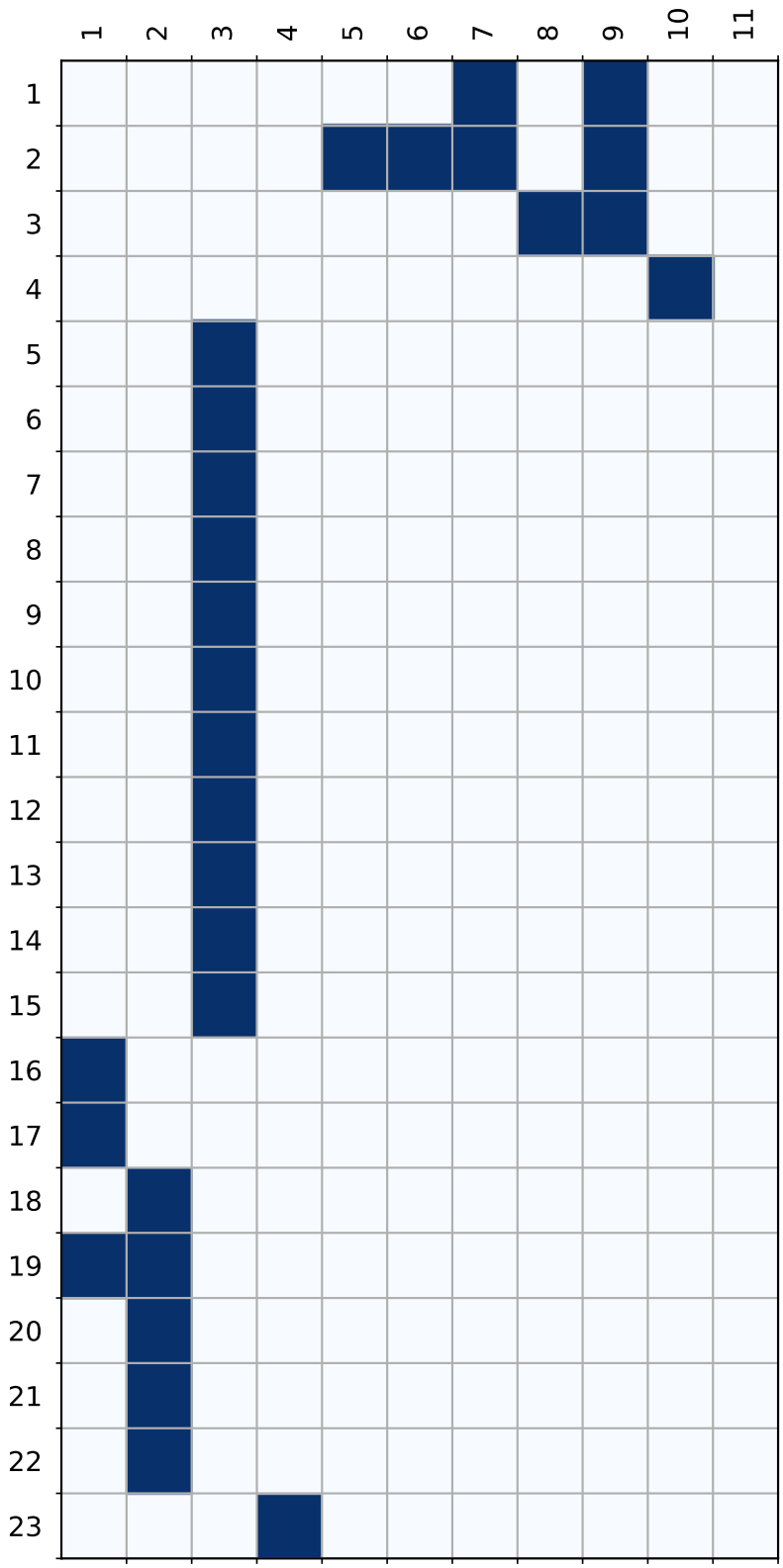

Figure B.15: Semantic translation from OSO (in row) to MOS (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

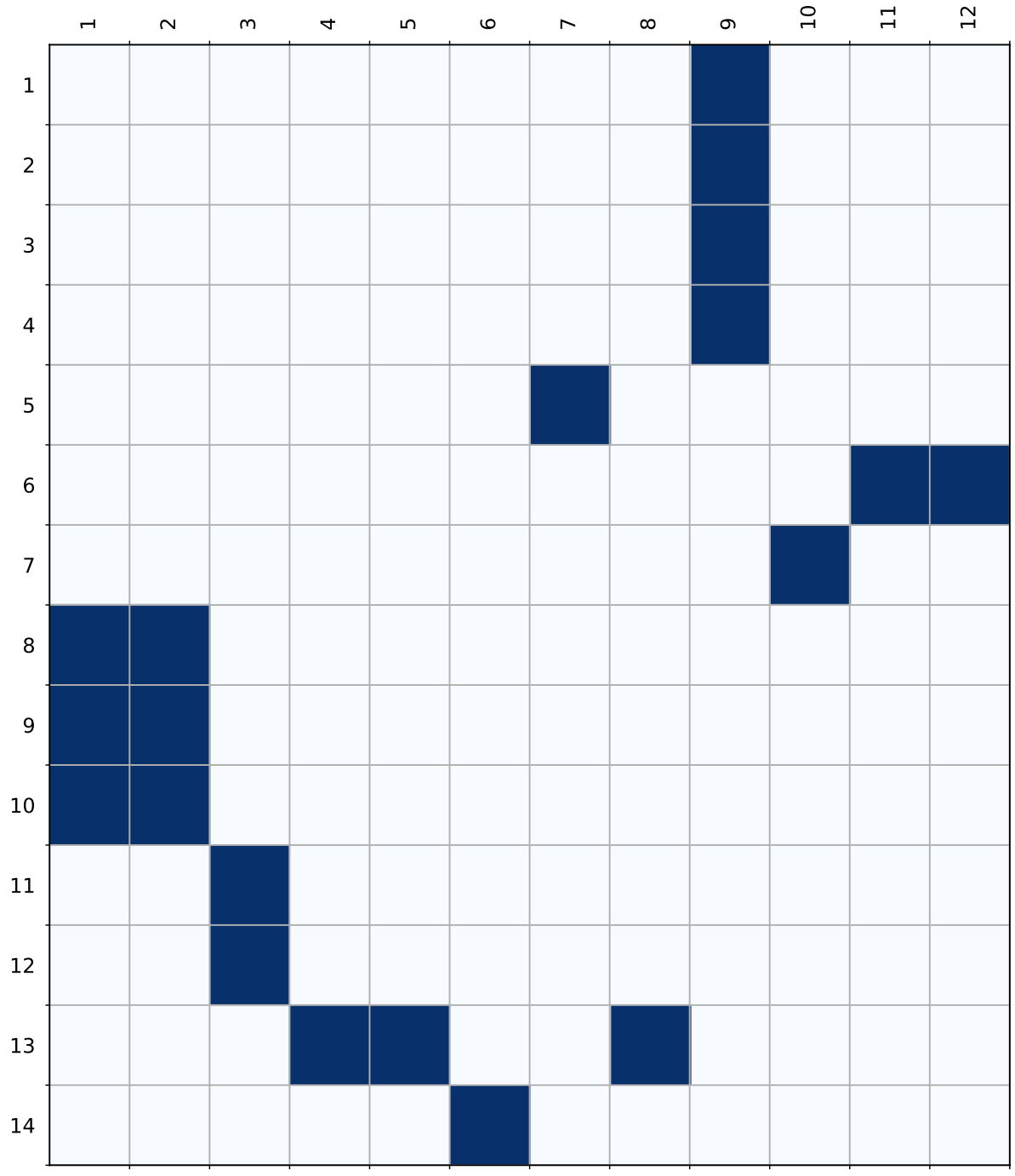

Figure B.16: Semantic translation from OCSGE-cover (in row) to CGLS-LC100 (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

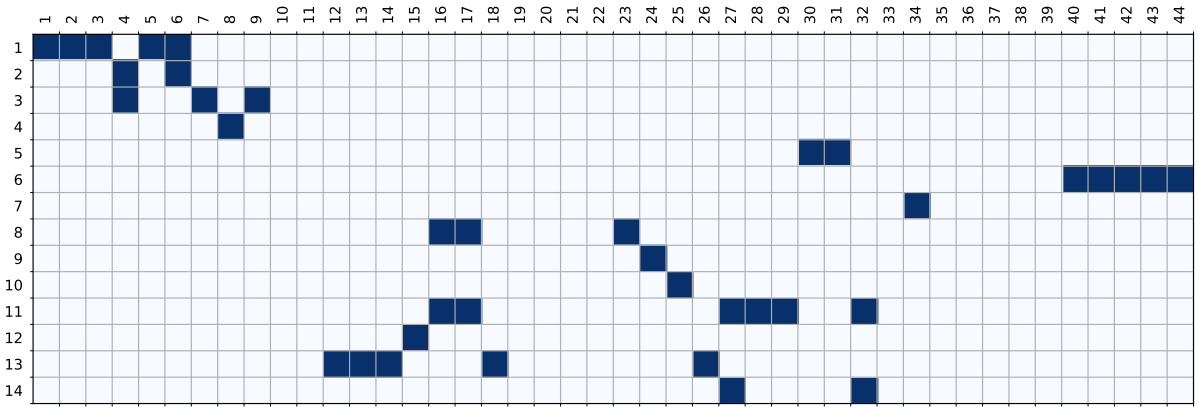

Figure B.17: Semantic translation from OCSGE-cover (in row) to CLC (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

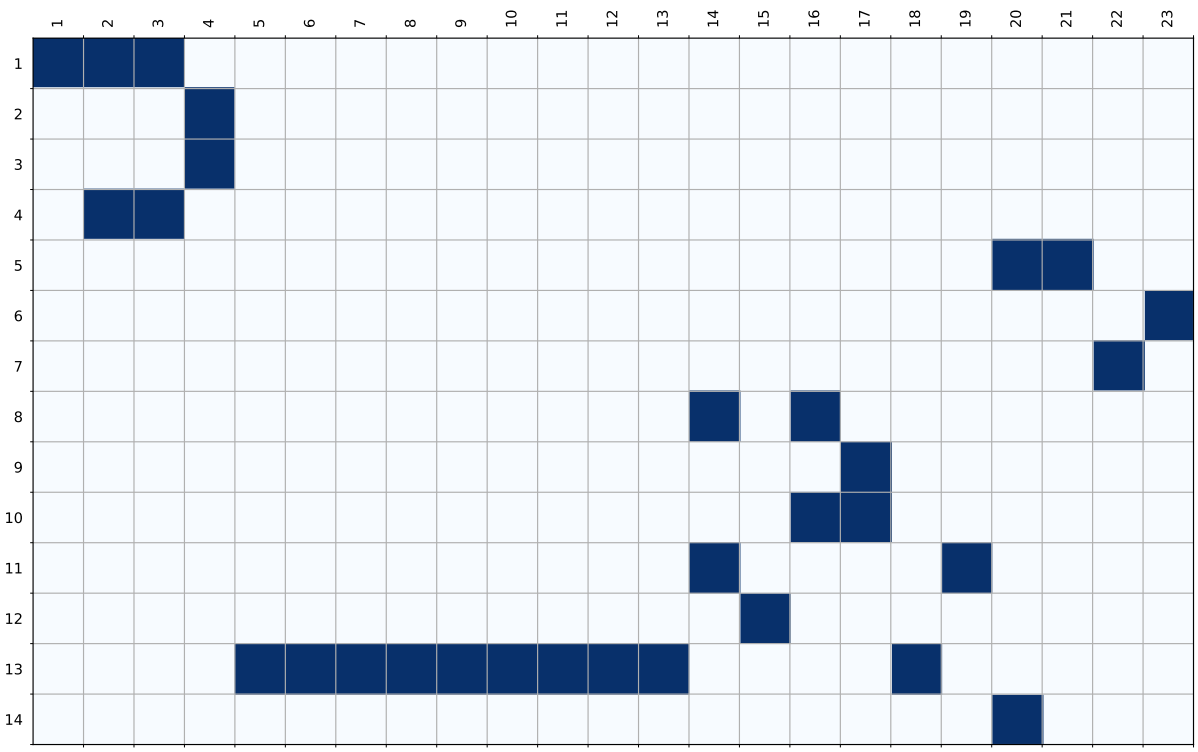

Figure B.18: Semantic translation from OCSGE-cover (in row) to OSO (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

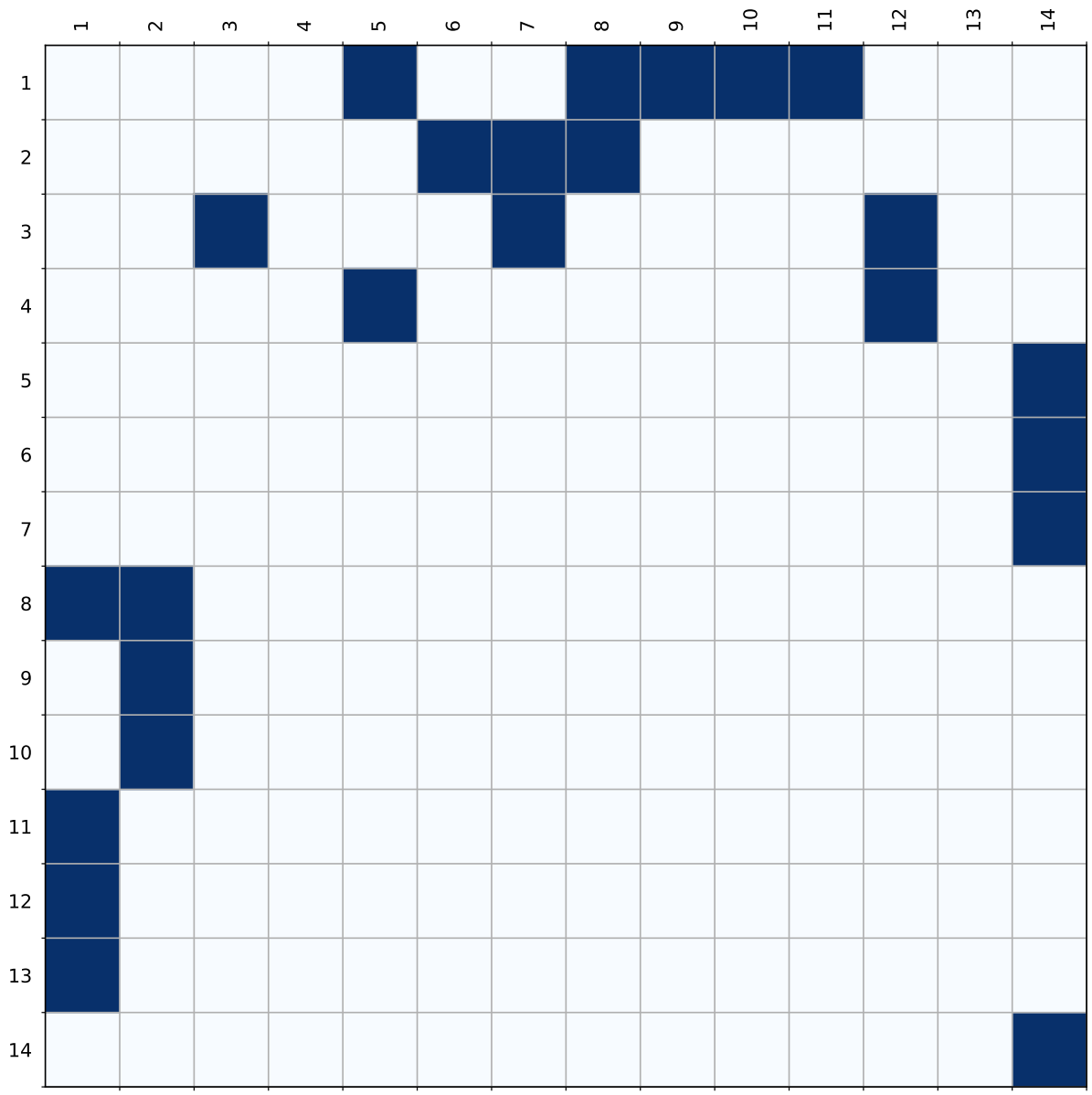

Figure B.19: Semantic translation from OCSGE-cover (in row) to OCSGE-use (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

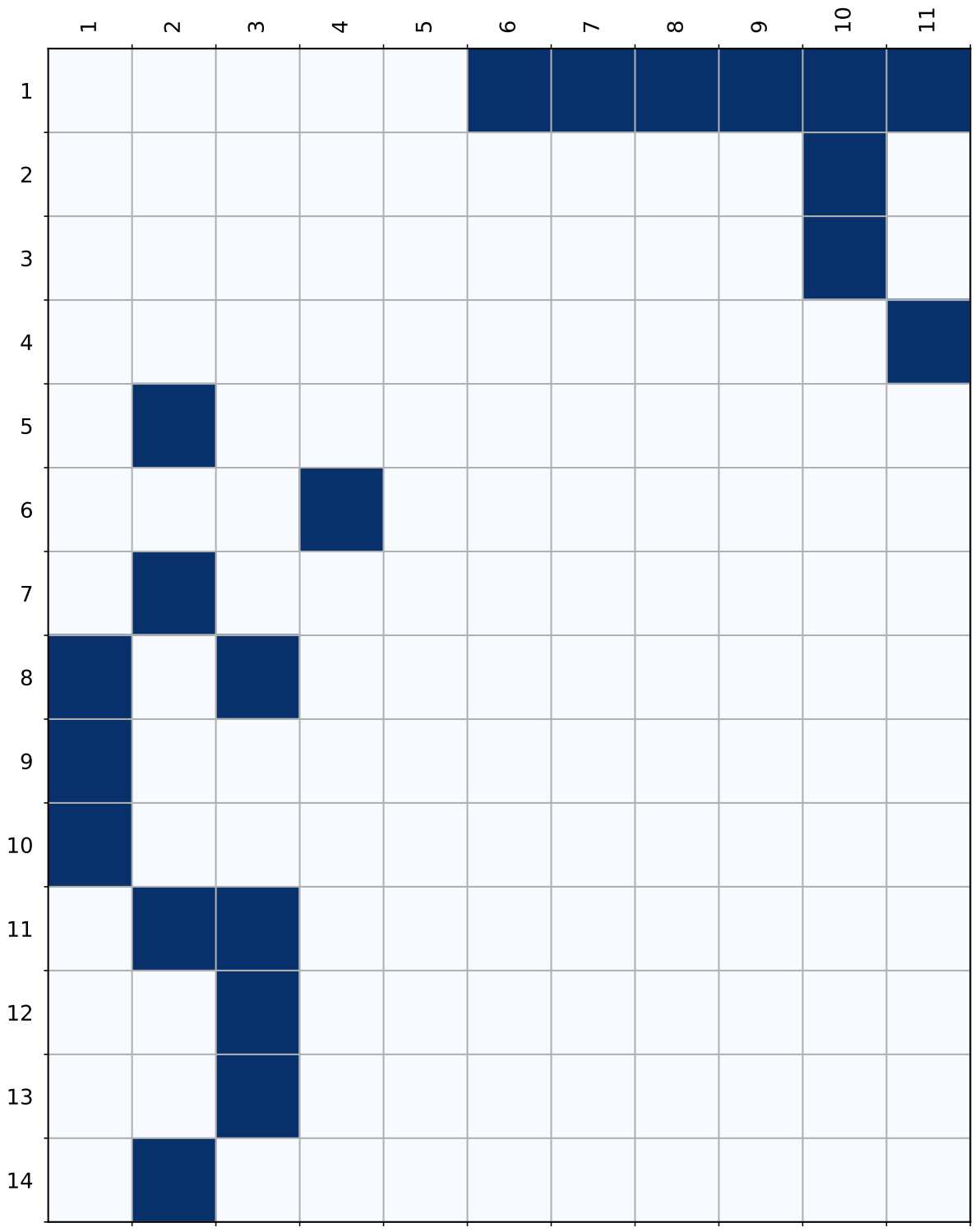

Figure B.20: Semantic translation from OCSGE-cover (in row) to MOS (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

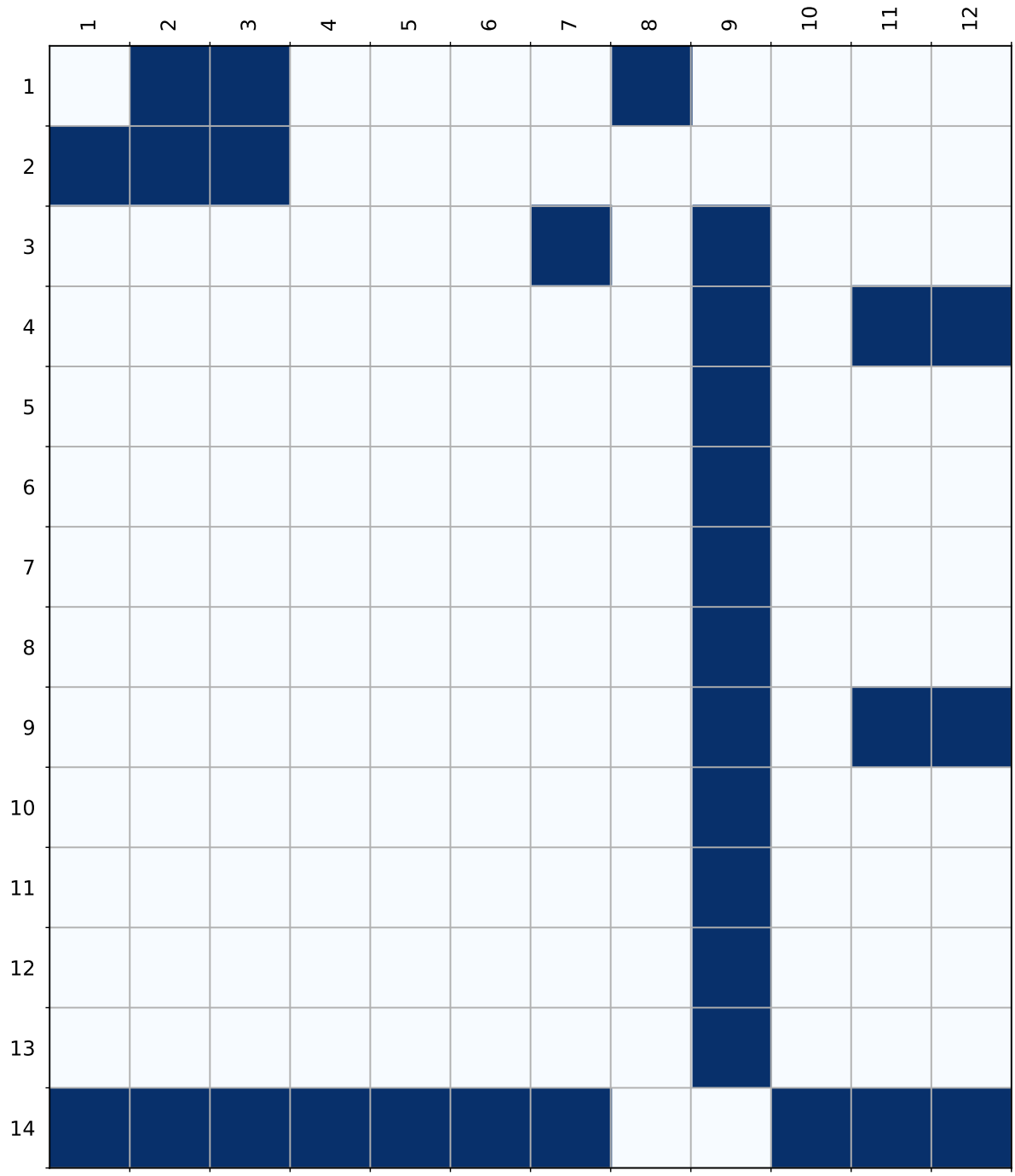

Figure B.21: Semantic translation from OCSGE-use (in row) to CGLS-LC100 (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

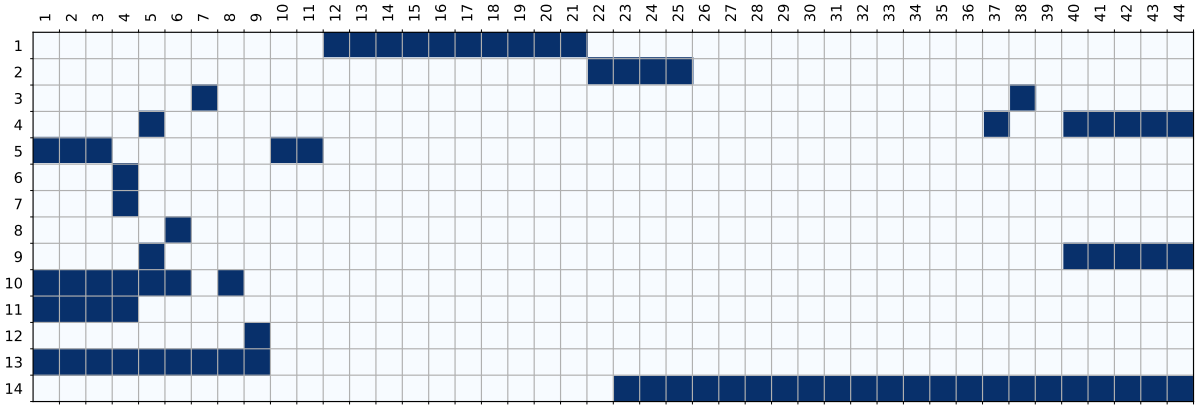

Figure B.22: Semantic translation from OCSGE-use (in row) to CLC (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

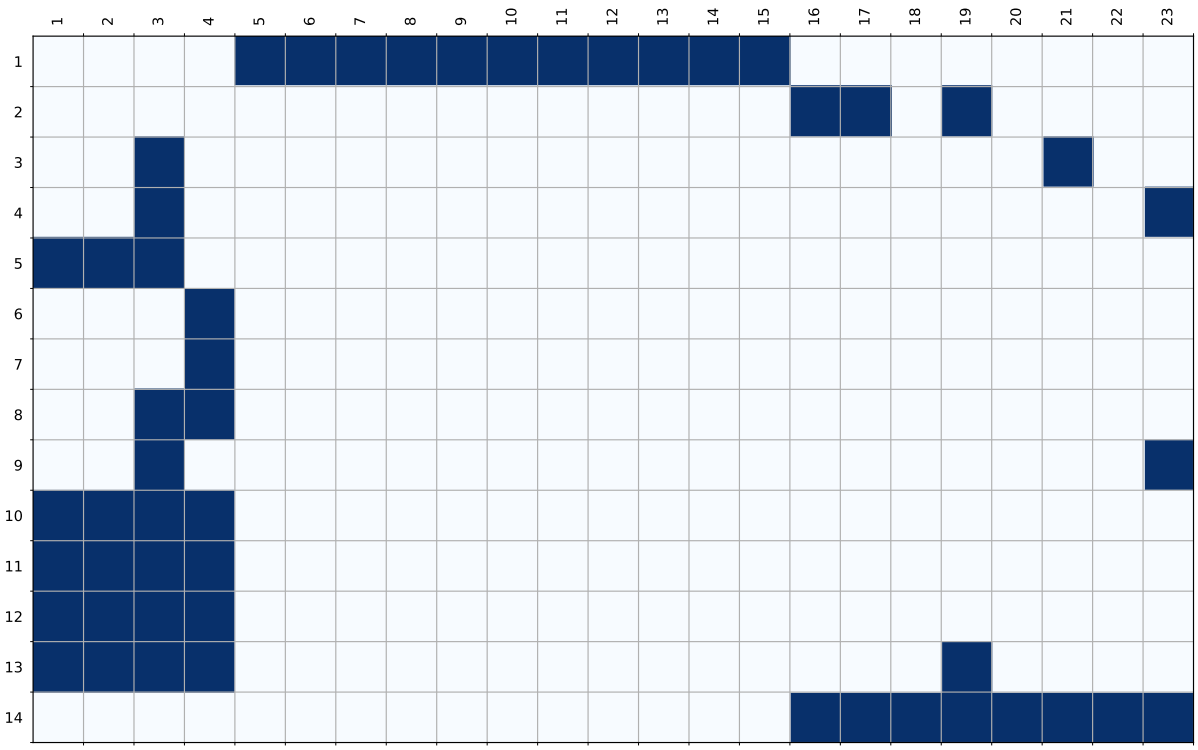

Figure B.23: Semantic translation from OCSGE-use (in row) to OSO (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

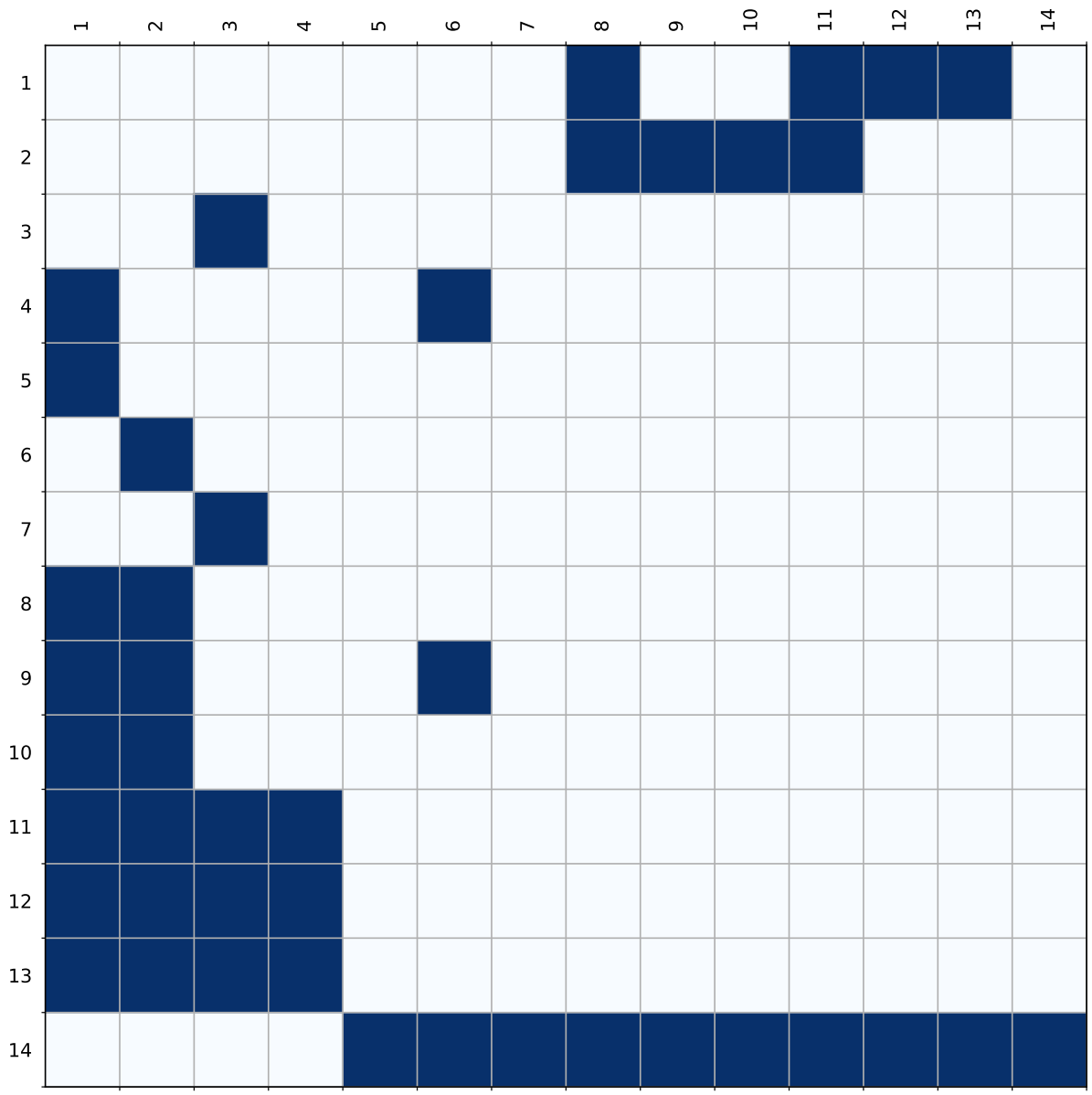

Figure B.24: Semantic translation from OCSGE-use (in row) to OCSGE-cover (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

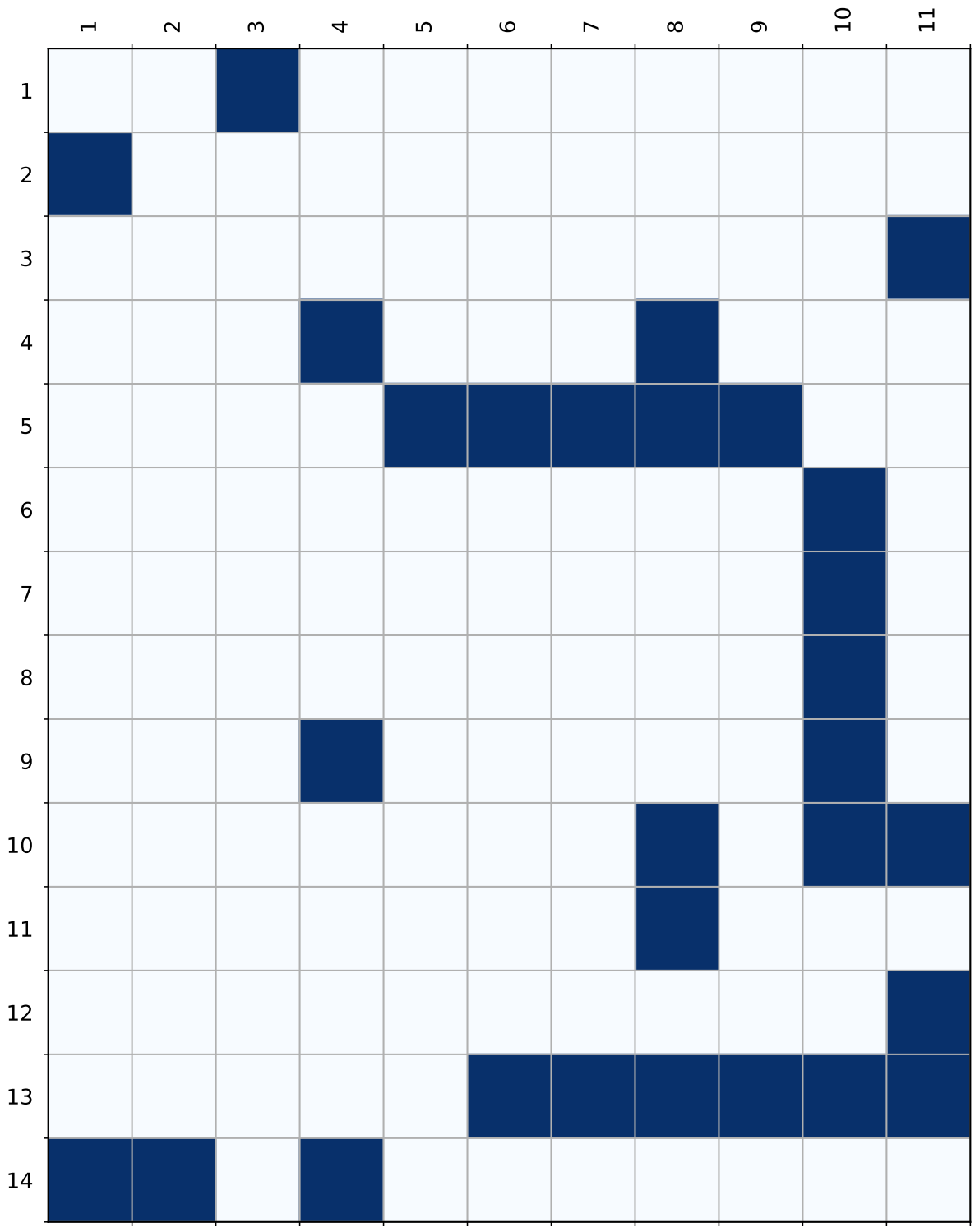

Figure B.25: Semantic translation from OCSGE-use (in row) to MOS (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

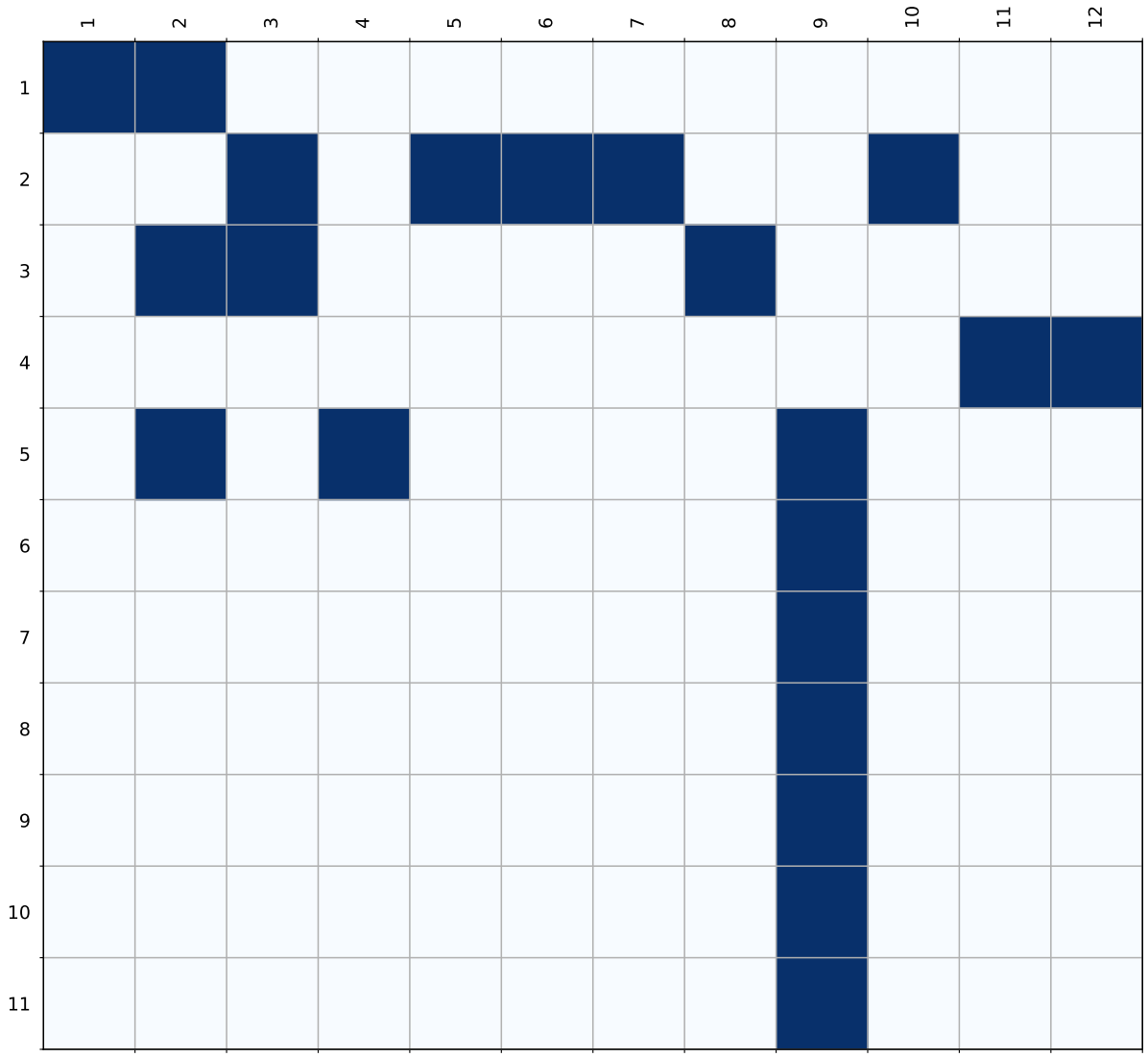

Figure B.26: Semantic translation from MOS (in row) to CGLS-LC100 (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

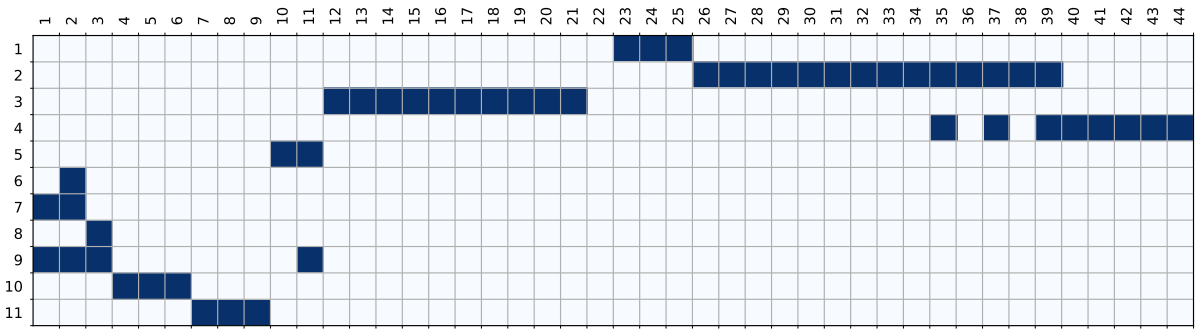

Figure B.27: Semantic translation from MOS (in row) to CLC (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

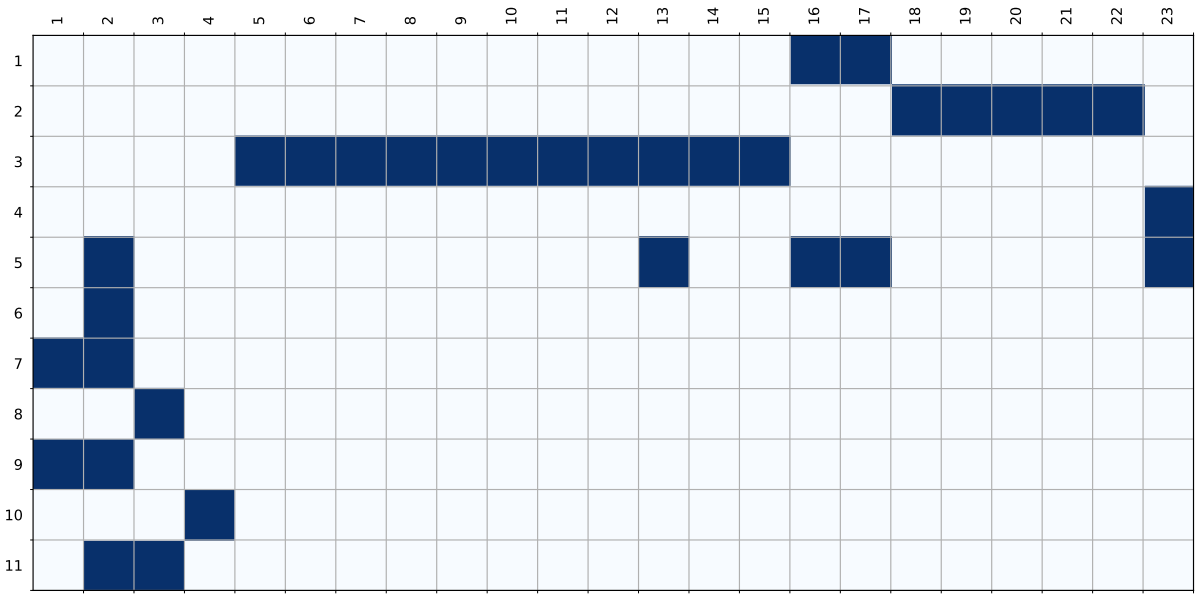

Figure B.28: Semantic translation from MOS (in row) to OSO (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

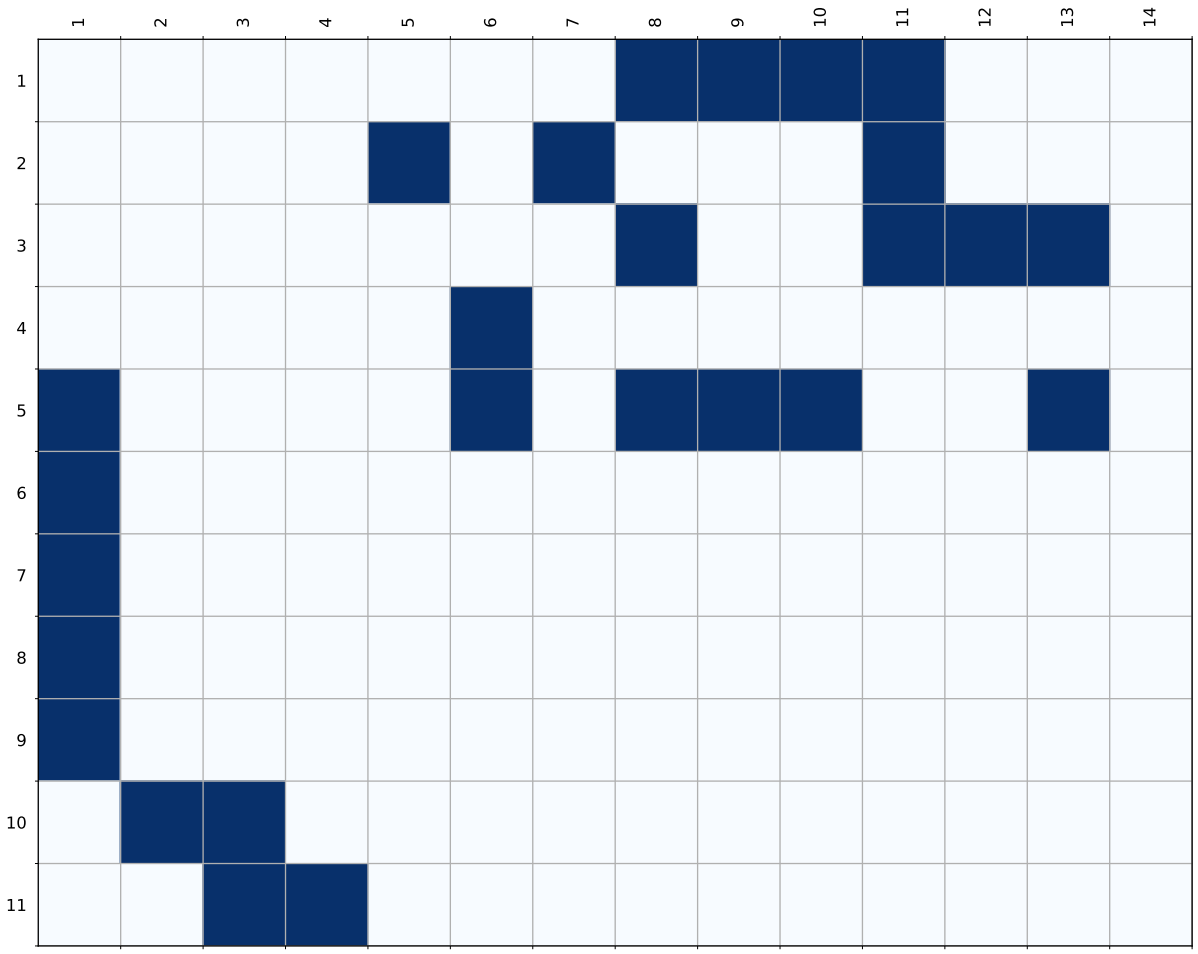

Figure B.29: Semantic translation from MOS (in row) to OCSGE-cover (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

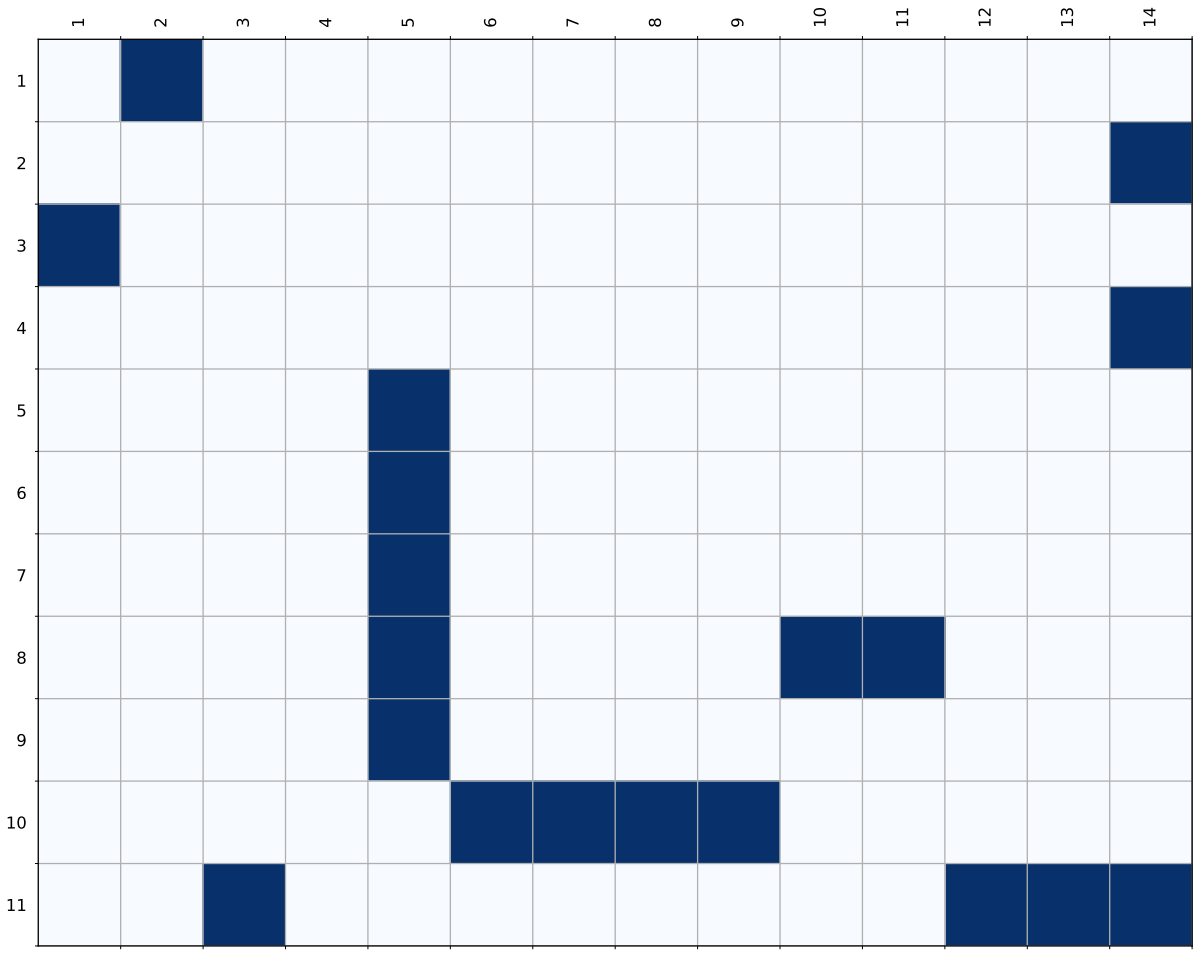

Figure B.30: Semantic translation from MOS (in row) to OCSGE-use (in column). For more explanation on how to read the Figure refer to the presentation of Appendix B.

APPENDIX C

Semantic vs statistic translation

Table C.1 gives the proportion of target classes that can be obtained from the translation of one single source to one single target (1-to-1), one source into multiple target (1-to-n) and no source class corresponds to the target (0-to-1) estimated using the matrices in Appendix B. High 1-to-1 values denotes easy translation as all pixels with a class establishing a 1-to-1 translation shall be translated the same-way. High 1-to-n values denotes complicated translation as pixels with the same class shall be translated into multiple targets which requires context analysis or additional data. High 0-to-1 denotes ill defined translations in which multiple target class have no clear source class correspondent. We estimate those proportion using to different approaches, symmetric and asymmetric. The asymmetric method only consider the source to target translation matrix while the symmetric one combine the source to target with the transposed target to source translation matrix. This asymmetric translation considers that a source class can not only be translated into its main semantic correspondents (asymmetric) but also in all target classes that can be translated into the source. For instance, OSO *Wheat crops* have a clear main semantic correspondent in the CLC nomenclature : *Non-irrigated arable crops*, thus all *Wheat crops* shall be translated into *Non-irrigated arable crops* in a "1-to-1" way. However, this results in the fact that classes such as *Mainly agriculture but significant areas of natural vegetation* which might partially include *Wheat crops* have no corresponding source class in the OSO to CLC translation matrix resulting in a high "0-to-1" proportion. Conversely, when one considers the reverse CLC to OSO translation both *Non-irrigated arable crops* and *Mainly agriculture but significant areas of natural vegetation* can both be translated into OSO *Wheat crops*. In practice this implies that the asymmetric (symmetric) measurement tends to underestimate (overestimate) the proportion of 1-to-n translation and overestimate (underestimate) 1-to-1 and 0-to-n translations. Therefore one should consider that the correct proportion of classes establishing each of the configuration is comprised between the symmetric and asymmetric values.

source		CGLS					CLC					OSO					OCSGEc					OCSGEu					MOS					Average
target		CLC	OSO	OCSGEc	OCSGEu	MOS	CGLS	OSO	OCSGEc	OCSGEu	MOS	CGLS	CLC	OCSGEc	OCSGEu	MOS	CGLS	CLC	OSO	OCSGEu	MOS	CGLS	CLC	OSO	OCSGEc	MOS	CGLS	CLC	OSO	OCSGEc	OCSGEu	
Assymetric	1-to-1 (%)	4	13	35	14	36	**58**	**65**	**78**	**57**	**63**	**50**	34	42	28	45	41	11	26	21	**54**	8	6	4	21	45	8	4	17	14	28	31
	1-to-n (%)	**71**	**83**	**58**	**58**	**55**	34	35	22	36	37	34	16	**58**	30	**46**	**59**	**62**	**74**	**65**	37	**92**	**94**	**96**	**79**	**55**	**92**	**94**	**83**	**79**	**65**	**60**
	0-to-1 (%)	25	4	7	28	9	8	0	0	7	0	16	**50**	0	**42**	9	0	27	0	14	9	0	0	0	0	0	0	2	0	7	7	9
Symetric	1-to-1 (%)	4	8	35	14	36	**58**	**65**	**78**	28	**63**	33	2	35	14	27	41	4	17	7	36	8	2	4	14	45	8	2	17	14	21	24
	1-to-n (%)	**71**	**92**	**58**	**72**	**55**	34	35	22	**72**	37	**67**	**98**	**65**	**86**	**64**	**59**	**87**	**83**	**86**	**64**	**92**	**98**	**96**	**86**	**55**	**92**	**96**	**83**	**79**	**79**	**72**
	0-to-1 (%)	25	0	7	14	9	8	0	0	0	0	0	0	0	0	9	0	9	0	7	0	0	0	0	0	0	0	2	0	7	0	3

Table C.1: Estimation of the proportion of target classes obtainable from translation where one source class is translated into a single target class (1-to-1), into multiple target classes (1-to-n), or where no source classes is found for the the target (0-to-1) for each source and target maps of the MLULC dataset. 2 different estimation technique are used. Symmetric use the source to target and the reverse target to source translation matrix (given in Appendix B) to estimate all possible translation from each source class to each target class, i.e. possible source to target class translation also includes target classes that are translated into source. Asymmetric only considers the source to target translation matrix.

APPENDIX D

Class definition

label	title	text
1	Continuous urban fabric	The continuous urban fabric class is assigned when urban structures and transport networks are dominating the surface area. More than 80% of the land surface is covered by impermeable features like buildings, roads and artificially surfaced areas.
2	Discontinuous urban fabric	The discontinuous urban fabric class is assigned when urban structures and transport networks associated with vegetated areas and bare surfaces are present and occupy significant surfaces in a discontinuous spatial pattern. The impermeable features like buildings, roads and artificially surfaced areas range from 30 to 80 % land coverage.
3	Industrial or commercial units	Buildings, other built-up structures and artificial surfaces (with concrete, asphalt, tarmacadam, or stabilised like e.g. beaten earth) occupy most of the area. It can also contain vegetation (most likely grass) or other non-sealed surfaces. This class is as-signed for land units that are under industrial or commercial use or serve for public service facilities.
4	Road and rail networks and associated land	Motorways and railways, including associated installations.
5	Port areas	Infrastructure of port areas (land and water surface), including quays, dockyards and marinas.
6	Airports	Airports installations: runways, buildings and associated land. This class is assigned for any kind of ground facilities that serve airborne transportation.
7	Mineral extraction sites	Open-pit extraction sites of construction materials (sandpits, quarries) or other minerals (open-cast mines). Includes flooded mining pits.
8	Dump sites	Public, industrial or mine dump sites. Public, industrial or mine dump sites.
9	Construction sites	Spaces under construction development, soil or bedrock excavations, earthworks. This class is assigned for areas where landscape is affected by human activities, changed or modified into artificial surfaces, being in a state of anthropogenic transition.
10	Green urban areas	Areas with vegetation within or partly embraced by urban fabric. This class is assigned for urban greenery, which usually has recreational or ornamental character and is usually accessible for the public.
11	Sport and leisure facilities	This class is assigned for areas used for sports, leisure and recreation purposes. Camping grounds, sports grounds, leisure parks, golf courses, racecourses etc. belong to this class, as well as formal parks not surrounded by urban areas.
12	Non-irrigated arable land	Cultivated land parcels under rainfed agricultural use for annually harvested non-permanent crops, normally under a crop rotation system, including fallow lands within such crop rotation. Fields with sporadic sprinkler-irrigation with non-permanent devices to support dominant rainfed cultivation are included.
13	Permanently irrigated arable land	Cultivated land parcels under agricultural use for arable crops that are permanently or periodically irrigated, using a permanent infrastructure (irrigation channels, drainage network and additional irrigation facilities). Most of these crops cannot be cultivated without artificial water supply. Does not include sporadically irrigated land.
14	Rice fields	Cultivated land parcels prepared for rice production, consisting of periodically flooded flat surfaces with irrigation channels.
15	Vineyards	Areas planted with vines, vineyard parcels covering >50% and determining the land use of the area.
16	Fruit tree and berry plantations	Cultivated parcels planted with fruit trees and shrubs, intended for fruit production, including nuts. The planting pattern can be by single or mixed fruit species, both in association with permanently grassy surfaces.
17	Olive groves	Cultivated areas planted with olive trees.
18	Pastures, meadows and other permanent grasslands under agricultural use	Permanent grassland characterized by agricultural use or strong human disturbance. Floral composition dominated by graminacea and influenced by human activity. Typically used for grazing-pastures, or mechanical harvesting of grass–meadows.
19	Annual crops associated with permanent crops	Cultivated land parcels with non-permanent crops (mostly arable land) associated with permanent crops (fruit trees or olive trees or vines) on the same parcel.
20	Complex cultivation patterns	Mosaic of small cultivated land parcels with different cultivation types -annual crops, pasture and/or permanent crops-, eventually with scattered houses or gardens.
21	Land principally occupied by agriculture, with significant areas of natural vegetation	Areas principally occupied by agriculture, interspersed with significant natural or semi-natural areas (including forests, shrubs, wetlands, water bodies, mineral outcrops) in a mosaic pattern.
22	Agro-forestry areas	Annual crops or grazing land under the wooded cover of forestry species.
23	Broad-leaved forest	Vegetation formation composed principally of trees, including shrub and bush understorey, where broad-leaved species predominate.
24	Coniferous forest	Vegetation formation composed principally of trees, including shrub and bush understorey, where coniferous species predominate.
25	Mixed forest	Vegetation formation composed principally of trees, including shrub and bush understorey, where neither broad-leaved nor coniferous species predominate.
26	Natural grassland	Grasslands under no or moderate human influence. Low productivity grasslands. Often situated in areas of rough, uneven ground, steep slopes, frequently including rocky areas or patches of other (semi-)natural vegetation.
27	Moors and heathland	Vegetation with low and closed cover, dominated by bushes, shrubs, dwarf shrubs (heather, briars, broom, gorse, laburnum etc.) and herbaceous plants, forming a climax stage of development.
28	Sclerophyllous vegetation	Bushy sclerophyllous vegetation in a climax stage of development, including maquis, matorral and garrigue.
29	Transitional woodland/shrub	Transitional bushy and herbaceous vegetation with occasional scattered trees. Can represent woodland degradation, forest regeneration / recolonization or natural succession.
30	Beaches, dunes, and sand plains	Natural non-vegetated expanses of sand or pebble/gravel, in coastal or continental locations, like beaches, dunes, gravel pads, including beds of stream channels with torrential regime. Vegetation covers maximum 10%.
31	Bare rock	Scree, cliffs, rock outcrops, including areas of active erosion, rocks and reef flats situated above the high-water mark, inland salt planes.
32	Sparsely vegetated areas	Areas with sparse vegetation, covering 10-50% of surface. Includes steppes, tundra, lichen heath, badlands, karstic areas and scattered high-altitude vegetation.
33	Burnt areas	Natural woody vegetation affected by recent fires.
34	Glaciers and perpetual snow	Land covered by glaciers or permanent snowfields. Permanent snow and ice can be captured by finding the patches' smallest extent during the year. This can be captured when they shrink to minimum due to summer warmth, but before the first snowfall after summer occurs. Such ideal date is between end July and late September.
35	Inland marshes	Low-lying land usually flooded in winter, and with ground more or less saturated by fresh water all year round.
36	Peatbogs	Wetlands with accumulation of considerable amount of decomposed moss (mostly Sphagnum) and vegetation matter. Both natural and exploited peat bogs.
37	Coastal salt marshes	Vegetated low-lying areas in the coastal zone, above the high-tide line, susceptible to flooding by seawater. Often in the process of being filled in by coastal mud and sand sediments, gradually being colonized by halophilic plants.
38	Salines	Salt-pans for extraction of salt from salt water by evaporation, active or in process of abandonment. Sections of salt marsh exploited for the production of salt, clearly distinguishable from the rest of the marsh by their parcellation and embankment systems.
39	Intertidal flats.	Coastal zone under tidal influence between open sea and land, which is flooded by sea water regularly twice a day in a 12 hours cycle. Area between the average lowest and highest sea water level at low tide and high tide. Generally non-vegetated expanses of mud, sand or rock lying between high and low water marks.
40	Water courses	Natural or artificial water-courses serving as water drainage channels. Includes canals. Minimum width for inclusion: 100 m.
41	Water bodies.	Natural or artificial water bodies with presence of standing water surface during most of the year.
42	Coastal lagoons	Stretches of salt or brackish water in coastal areas which are separated from the sea by a tongue of land or other similar topography. These water bodies can be connected to the sea at limited points, either permanently or for parts of the year only.
43	Estuaries	The mouth of a river under tidal influence within which the tide ebbs and flows.
44	Sea and ocean	Zone seaward of the lowest tide limit.

Table D.1: CLC Class definition

label	title	text
1	Continuous urban fabric	Most of the land is covered by structures and the transport network. Building, roads and artificially surfaced areas cover more than 80% of the total surface. Non-linear areas of vegetation and bare soil are exceptional.
2	Discontinuous urban fabric	Most of the land is covered by structures. Building, roads and artificially surfaced areas associated with vegetated areas and bare soil, which occupy discontinuous but significant surfaces.
3	Industrial or commercial units	Artificially surfaced areas (with concrete, asphalt, tarmacadam, or stabilised, e.g., beaten earth) without vegetation occupy most of the area, which also contain buildings and/or vegetation.
4	Road surfaces	motorway rest areas, parking areas, motorway networks, larger than 50 m.
5	Rapeseed	Rapeseed usually seeded between November and February and harvested at the beginning of the summer (mid June to the end of July).
6	cereals	Straw cereal crops usually seeded between November and February and harvested at the beginning of the summer (mid June to the end of July).
7	protein crops	Protein crops usually seeded between November and February and harvested at the beginning of the summer (mid June to the end of July).
8	soy	Soy crops usualluy seeded from March to mid June and harvested at the end of the summer (mid August to mid September).
9	turnsole	Sunflower usualluy seeded from March to mid June and harvested at the end of the summer (mid August to mid September).
10	corn	Maize usualluy seeded from March to mid June and harvested at the end of the summer (mid August to mid September).
11	rice	Rice usualluy seeded from March to mid June and harvested at the end of the summer (mid August to mid September).
12	tubers	Tubers or roots usualluy seeded from March to mid June and harvested at the end of the summer (mid August to mid September).
13	Intensive grassland	Dense grass cover, of floral composition characterized by agricultural use or strong human disturbance not under a rotation system used.
14	Orchards	Agricultural lands parcels planted with fruit trees or shrubs.
15	Vineyards	Agricultural lands areas planted with vines.
16	Broad-leaved forest	Area populated by forest trees covering at least 40% of the soil, and composed of more than 75% of hardwoods (relative coverage rate). Young hardwood plantations, natural reforestation and clearcuts of these species are also entered.
17	Coniferous forest	Area populated by forest trees covering at least 40% of the soil and composed mainly of conifers (i.e. a relative cover rate of more than 75%). Young coniferous plantations, natural reforestation and clearcutting of these species are also entered.
18	Natural grassland	Grasslands under no or moderate human influence. Low productivity grasslands. Often situated in areas of rough, uneven ground, steep slopes, frequently including rocky areas, briars and heathland.
19	Woody moorlands	Spontaneous vegetation dominated by woody plants (heather, briar, broom, etc.) and semi-woody plants (fern, phragmites, etc.) shorter than 5 m.
20	Bare rock	Scree, cliffs, rock outcrops, including areas of active erosion, rocks and reef flats situated above the high-water mark.
21	Beaches, dunes, and sand plains	beaches, dunes and expanses of sand or pebbles in coastal or continental locations, including beds of stream channels with torrential regime.
22	Glaciers and perpetual snow	land covered by glaciers or permanent snowfields.
23	Water bodies	All water bodies longer than 20 m and all water courses larger than 7.5 m

Table D.2: OSO Class definition

label	title	text
1	Forest	Vegetation of trees, shrubs, bushes that may result from regeneration or shrub recolonization. Area composed of at least 40% of trees 5m high (except orchards), including tree heaths. Glades, regeneration cuts, clear cuts, seedlings, including poplar forest cuts.
2	Semi-natural areas	Wetlands, marshes, non-treed moors, sparse or herbaceous vegetation, agricultural wastelands including fallows and multi-year frosts, abandoned quarries with vegetation, maneuvering grounds, deforestation rights-of-way for power lines or aqueducts and undeveloped banks of waterways.
3	Agricultural areas	Areas occupied by annual crops or grass areas except lawns mainly grazed but from which fodder can be harvested, including those of equestrian centers. Also includes Nurseries and fruit crops of more than 1000m^2 homogeneous or mixed and of commercial production, vine as well as abandoned or fallow orchards, vegetable crops (salads, etc.), market gardening, flower crops and greenhouse crops.
4	Water	Water surfaces of at least 500m^2, including park ponds, groundwater layers of gravel pits and retention basins. Permanent water courses without maximum width restriction, including canals.
5	Artificial open spaces	Parks. Pleasure gardens, vegetable gardens or orchards gardens, for family use. Outdoor courts. Outdoor bathing areas. Roller-skating and cross-country tracks. Open shooting ranges. Evolution parks for the practice of golf, including buildings. Equipment for horse racing. Campgrounds and caravans, including mobile homes. Zoos, amusement parks, recreation centers without accommodation. Non-agricultural grass surfaces: military maneuver grounds, near airfield runways, grassed areas in business and commercial areas, castles. Vacant or undeveloped land, located within the urban grid.
6	Individual housing	Housing estates and individual buildings. Groups of buildings spaced less than 100m apart, mostly rural in form from 1 to 2 levels, exceptionally 3, built contiguously forming a built core, comprising in its central part a point of convergence or a particular point (monument, church), including buildings farm, including a road structure whose narrow width and layout testify to a road of village origin. Castles will be classified in low continuous habitat for the building itself, and in village for the outbuildings.
7	Collective housing	They includes continuous and dicontinous collective or individual housings in cities. The areas concerned are in the old suburbs, centers and in the new "townhouse" districts. Includes castles, car parks, green spaces, shops, play areas that are an integral part of the whole, prisons, hotels (excluding activity zones), reception centers, holiday and leisure centers, homes for workers and students, convents, seminars, retirement homes, precarious or mobile housing (caravans or mobile isolated homes).
8	Activity zone	Industrial plants .Industrial activities (in business premises, laboratories, warehouses, workshops, etc.) dispersed in residential areas, thus forming a mixed fabric, but which are individualized in relation to the habitat. Animal production activities: kennels, stud farms, poultry facilities, veterinarians, etc. Parking lots and large vacant spaces. Storage areas for new vehicles, caravans, construction materials, sawmills, vehicle scrap yards, including port areas. Logistics warehouses. Commercial areas with surface greater than 400m^2. Gas stations. Offices of more than 5000m^2.
9	amenities	Indoor and outdoor sports facilities. Outdoor and indoor swimming pool including biological pools. Autodromes including a speed or road track.All schools public or private.Hospital. Large convention and exhibition centers. Museums, some libraries, castles open to the public. All administrations, police stations, fire stations. Large public installations including military, radio installations. Permanent markets. Places of worship. Nurseries, leisure activity buildings, post offices, motorway tolls and locks.
10	transport network	Yard yards, stations, equipment maintenance facilities, rail tracks including fill and cut. Tracks$>$ 25m wide, including access ramps, embankment and cuttings embankments, entire interchanges. Car parks with their own right of way, excluding underground parks. This item includes car parks associated with equipment and housing. Bus, bus and coach stations for travelers. Freight transport facilities are identified in storage activities. This item includes RATP bus depots. Air terminals, aircraft parking areas, technical installations (hangars, etc.) and runways only are included in this station.
11	Quarries, landfills and construction sites	Quarries, sand pits in activity or abandoned, without traces of vegetation. Authorized landfills, waste reception centers and landfill areas. Construction and demolition sites.

Table D.3: MOS classes definitions

label	title	text
1	Built-up areas	Areas covered with buildings or other types of construction. These areas include buildings of a permanent nature, covered with a roof and the associated spaces in compliance with the entry thresholds. They are intended to shelter, house or place people, animals, material, goods. Includes glass greenhouses or plastic tunnels capable of accommodating a standing man.
2	Undeveloped areas	Areas covered with patrially or fully waterproofed materials, in particular asphalt, concrete, paved or slab floors. The paved road network, squares, car parks and tennis courts regardless of the surface are impermeable non-built areas.
3	Mineral material areas	Areas covered with stabilized and compacted soils, partially or totally permeable, and covered with mineral materials. Includes railroad networks, stone paths, forest tracks or services, non-vegetated firewalls, transport track sites (roads, motorways, railways, etc.) quarries, salt works and construction sites (groynes and rip-rap).
4	Areas with other composite materials	Areas covered with heterogeneous and artificial materials with a mixture of non-mineral materials. Includes, in particular, dumpsites.
5	Bare soils	Areas covered with natural bare soil. Includes soils covered with sand, pebbles, rocks, stony surfaces or any other mineral material.
6	Water surfaces	Areas permanently covered with water. Includes submerged surfaces (Areas permanently covered with water). The limits are the banks or vegetation without the rate of coverage exceeding 25%.
7	Snowfields and glaciers	Areas permanently covered with snow or ice. Includes surfaces covered mainly by glaciers and snowfields.
8	Broad-leaved trees	Areas only covered with deciduous trees. This class includes pure stands of the same species of hardwoods or a mixture of hardwoods (oak, beech, poplars, fruit trees...). This class excludes woodland pastures. Absolute tree cover rate greater than or equal to 25% with a free coverage rate relative hardwood greater than or equal to 75%.
9	Coniferous trees	Areas only covered with coniferous trees. This class includes pure stands of the same species of conifers or a mixture of conifers (pines, firs...). Absolute tree cover rate greater than or equal to 25% with a free coverage rate relative hardwood greater than or equal to 75%.
10	Mixed trees	Areas covered with a mixed of deciduous and coniferous trees. This class includes mixed stands of hardwoods and conifers (oak-beech forest, beech forest pinewood, etc.) in massifs or linearly (wooded cordons). Absolute tree cover rate greater than or equal to 25% with a free coverage rate relative hardwood greater than or equal to 75%.
11	Shrub and sub-shrub formations	Areas covered with shrubs and sub-shrubs. This class includes alpine moors, moors mountain areas, non-wooded scrubland or maquis, uncultivated or fallow land, the moors on salty land, phragmites heaths. This class also includes plantations of small fruits, horticultural plants, medicinal plants and Aromatic...
12	Liana-like vegetation	Areas covered with liana-like vegetation. Includes plantations of vines, hops, kiwi ...
13	Herbaceous formations	Areas covered with herbaceous vegetation. This class includes permanent and temporary meadows, natural lawns, arable land, ornamental lawns, collective or individual market gardens in the vicinity of dwellings, vegetable lawns in sports complexes, etc.
14	Other non-woody formations	Areas covered with other non-woody formations. Includes Areas covered with lichen or moss. Includes areas of banana, palm or bamboo.

Table D.4: OCS-GEc classes definitions

label	title	text
1	Agriculture	Production of crops (plants, mushrooms, etc.) and animal products for food, sale, home consumption or industrial uses. This category includes crops intended for the production of biofuels as well as field cultivation and cultivation under cover. Land fallow under rotation also falls into this category. The preparation of products for their primary marketing is included, as well as land reclamation (earthworks, drainage, preparation of rice fields, etc.) as well as landscaping and maintenance. This class includes agricultural infrastructure.
2	Forestry	Production of roundwood and other primary wood products. In addition to wood production, forestry activities generate products that undergo limited processing, such as firewood, charcoal, and roundwood used in unprocessed form (e.g., minewood, pulpwood, etc.). Nurseries as well as storage and transport areas related to the exploitation of timber, trees and woody plants intended for the production of biofuels are also covered. These activities can be carried out in natural forests or in plantations.
3	Extraction activities	Extractive industries consisting of the extraction of minerals and materials occurring naturally in the form of solids (coal, minerals, gravel, sand, salt), liquids (petroleum), gas (natural gas) or biomass (peat). Extraction can be done in different ways: underground, on the surface, by digging wells, etc. This class includes buildings and infrastructure related to these activities.
4	Fisheries and Aquaculture	This category includes professional fishing and inland or marine aquaculture (fish farming, seaweed farming, shellfish farming, mussel farming, carcinoculture, echinoculture). This class includes buildings and infrastructure related to these activities.
5	Commercial, industrial and residential areas	Industrial and manufacturing activities consisting in manufacturing, from the output of the primary sector, manufactured goods and intermediate products for other sectors. Services constituting products for other businesses and consumers, both private and public. This category includes commercial services, financial services, specialist services and information services, public services, cultural services, leisure services and recreational services. Areas mainly used for housing people.
6	Road networks	Areas used for road transport as well as associated infrastructure, for example, roads, car parks or service stations, rest area ...
7	Rail networks	Areas used for rail transport as well as associated infrastructure, for example, tracks, railway stations and marshalling yards, parking ...
8	Air transport network	Areas used for air transport as well as associated infrastructure, e.g. airports and related services, ...
9	River and maritime transport networks	Areas used for inland waterway transport and associated infrastructure, for example ports, rivers, docks and related services, ... This class includes canals dedicated to navigation.
10	Logistics and storage services	Areas used for separate storage and logistics services (not directly related to industries)
11	Public utility networks	Infrastructures linked to public utility networks grouping the areas used for the distribution of electricity, gas and thermal energy, the areas used for the collection, collection, purification, storage and distribution of water as well as only for the collection and treatment of wastewater, areas used for the collection and recycling of waste.This class includes areas regularly maintained (shredding) under high-voltage lines (interruption across the width of the tree cover).
12	Transitional areas	Areas under construction.
13	Abandoned areas	Agricultural, residential or industrial areas and areas devoted to transport and basic infrastructure in a derelict state. An area belongs to the category of abandoned areas if it is no longer in use or can no longer be used for its original purposes without major repair or renovation work.
14	Unused	Areas which occur in their natural state and are not subject to other economic use.

Table D.5: OCS-GEu classes definition

label	title	text
1	Closed forest.	Forest with tree canopy covering more than 70% of the area.
2	Open forest	Forest with top layer-trees covering 15 % to 70% of the area and second layer-mixed of shrubs and grassland,
3	Shrubs	These are woody perennial plants with persistent and woody stems and without any defined main stem being less than 5 m tall. The shrub foliage can be either evergreen or deciduous.
4	Herbaceous vegetation	Plants without persistent stem or shoots above ground and lacking definite firm structure. Tree and shrub cover is less than 10 %.
5	Herbaceous wetland	Lands with a permanent mixture of water and herbaceous or woody vegetation. The vegetation can be present in either salt, brackish, or fresh water.
6	Moss and lichen	Moss and lichen
7	Bare / sparse vegetation	Lands with exposed soil, sand, or rocks and never has more than 10 % vegetated cover during any time of the year
8	Cultivated and managed vegetation/agriculture (cropland)	Lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type.
9	Urban / built up	Land covered by buildings and other man-made artificial structures
10	Snow and Ice	Lands under snow or ice cover throughout the year.
11	Permanent water bodies	lakes, reservoirs, and rivers. Can be either fresh or salt-water bodies.
12	Open sea	Oceans, seas. Can be either fresh or salt-water bodies.

Table D.6: CGLS classes defintion

label	title	text
1	Water	Areas where water was predominantly present throughout the year. may not cover areas with sporadic or ephemeral water. contains little to no sparse vegetation, no rock outcrop nor built up features like docks. examples: rivers, ponds, lakes, oceans, flooded salt plains.
2	Trees	Any significant clustering of tall (15-m or higher) dense vegetation, typically with a closed or dense canopy. examples: wooded vegetation, clusters of dense tall vegetation within savannas, plantations, swamp or mangroves (dense/tall vegetation with ephemeral water or canopy too thick to detect water underneath).
3	Flooded vegetation	Areas of any type of vegetation with obvious intermixing of water throughout a majority of the year. seasonally flooded area that is a mix of grass/shrub/trees/bare ground. examples: flooded mangroves, emergent vegetation, rice paddies and other heavily irrigated and inundated agriculture.
4	Crops	Human planted/plotted cereals, grasses, and crops not at tree height. examples: corn, wheat, soy, fallow plots of structured land.
5	Built Area	Human made structures. major road and rail networks. large homogenous impervious surfaces including parking structures, office buildings and residential housing. examples: houses, dense villages / towns / cities, paved roads, asphalt.
6	Bare ground	Areas of rock or soil with very sparse to no vegetation for the entire year. large areas of sand and deserts with no to little vegetation. examples: exposed rock or soil, desert and sand dunes, dry salt flats/pans, dried lake beds, mines.
7	Snow/Ice	Large homogenous areas of permanent snow or ice, typically only in mountain areas or highest latitudes. examples: glaciers, permanent snowpack, snow fields.
8	Clouds	No land cover information due to persistent cloud cover.
9	Rangeland	Open areas covered in homogenous grasses with little to no taller vegetation. wild cereals and grasses with no obvious human plotting (i.e., not a plotted field). examples: natural meadows and fields with sparse to no tree cover, open savanna with few to no trees, parks/golf courses/lawns, pastures. Mix of small clusters of plants or single plants dispersed on a landscape that shows exposed soil or rock. scrub-filled clearings within dense forests that are clearly not taller than trees. examples: moderate to sparse cover of bushes, shrubs and tufts of grass, savannas with very sparse grasses, trees or other plants.

Table D.7: ESRI land-cover map [a]

[a] https://www.arcgis.com/home/item.html?id=d6642f8a4f6d4685a24ae2dc0c73d4ac

label	title	text
1	Evergreen Needleleaf Forests	Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%.
2	Evergreen Broadleaf Forests	Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%.
3	Deciduous Needleleaf Forests	Dominated by deciduous needleleaf (larch) trees (canopy >2m). Tree cover >60%.
4	Deciduous Broadleaf Forests	Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%.
5	Mixed Forests	Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2m). Tree cover >60%.
6	Closed Shrublands	Dominated by woody perennials (1-2m height) >60% cover.
7	Open Shrublands	Dominated by woody perennials (1-2m height) 10-60% cover.
8	Woody Savannas	Tree cover 30-60% (canopy >2m).
9	Savannas	Tree cover 10-30% (canopy >2m).
10	Grasslands	Dominated by herbaceous annuals (<2m).
11	Permanent Wetlands	Permanently inundated lands with 30-60% water cover and >10% vegetated cover.
12	Croplands	At least 60% of area is cultivated cropland.
13	Urban and Built-up Lands	At least 30% impervious surface area including building materials, asphalt, and vehicles.
14	Cropland/Natural Vegetation Mosaics	Mosaics of small-scale cultivation 40-60% with natural tree, shrub, or herbaceous vegetation.
15	Permanent Snow and Ice	At least 60% of area is covered by snow and ice for at least 10 months of the year.
16	Barren	At least 60% of area is non-vegetated barren (sand, rock, soil) areas with less than 10% vegetation.
17	Water Bodies	At least 60% of area is covered by permanent water bodies.

Table D.8: MCD12Q1: IGBP nomenclature

label	title	text
0	Water bodies	At least 60% of area is covered by permanent water bodies.
1	Evergreen Needleleaf Forests	Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%.
2	Evergreen Broadleaf Forests	Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%.
3	Deciduous Needleleaf Forests	Dominated by deciduous needleleaf (larch) trees (canopy >2m). Tree cover >60%.
4	Deciduous Broadleaf Forests	Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%.
5	Mixed Forests	Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2m). Tree cover >60%.
6	Closed Shrublands	Dominated by woody perennials (1-2m height) >60% cover.
7	Open Shrublands	Dominated by woody perennials (1-2m height) 10-60% cover.
8	Woody Savannas	Tree cover 30-60% (canopy >2m).
9	Savannas	Tree cover 10-30% (canopy >2m).
10	Grasslands	Dominated by herbaceous annuals (<2m).
11	Permanent Wetlands	Permanently inundated lands with 30-60% water cover and >10% vegetated cover.
12	Croplands	At least 60% of area is cultivated cropland.
13	Urban and Built-up Lands	At least 30% impervious surface area including building materials, asphalt, and vehicles.
14	Cropland/Natural Vegetation Mosaics	Mosaics of small-scale cultivation 40-60% with natural tree, shrub, or herbaceous vegetation.
15	Non-Vegetated Lands	At least 60% of area is non-vegetated barren (sand, rock, soil) or permanent snow and ice with less than 10% vegetation.

Table D.9: MCD12Q1: UMD nomenclature

label	title	text
0	Water bodies	At least 60% of area is covered by permanent water bodies.
1	Grasslands	Dominated by herbaceous annuals (<2m) including cereal croplands.
2	Shrublands	Shrub (1-2m) cover >10%.
3	Broadleaf Croplands	Dominated by herbaceous annuals (<2m) that are cultivated with broadleaf crops.
4	Savannas	Between 10-60% tree cover (>2m).
5	Evergreen Broadleaf Forests	Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%.
6	Deciduous Broadleaf Forests	Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%.
7	Evergreen Needleleaf Forests	Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%.
8	Deciduous Needleleaf Forests	Dominated by deciduous needleleaf (larch) trees (canopy >2m). Tree cover >60%.
9	Non-Vegetated Lands	At least 60% of area is non-vegetated barren (sand, rock, soil) or permanent snow and ice with less than 10% vegetation.
10	Urban and Built-up Lands	At least 30% impervious surface area including building materials, asphalt, and vehicles.

Table D.10: MCD12Q1: LAI nomenclature

APPENDIX E

Confusion matrix of CGLS-LC100, CLC, OSO on our ground truth

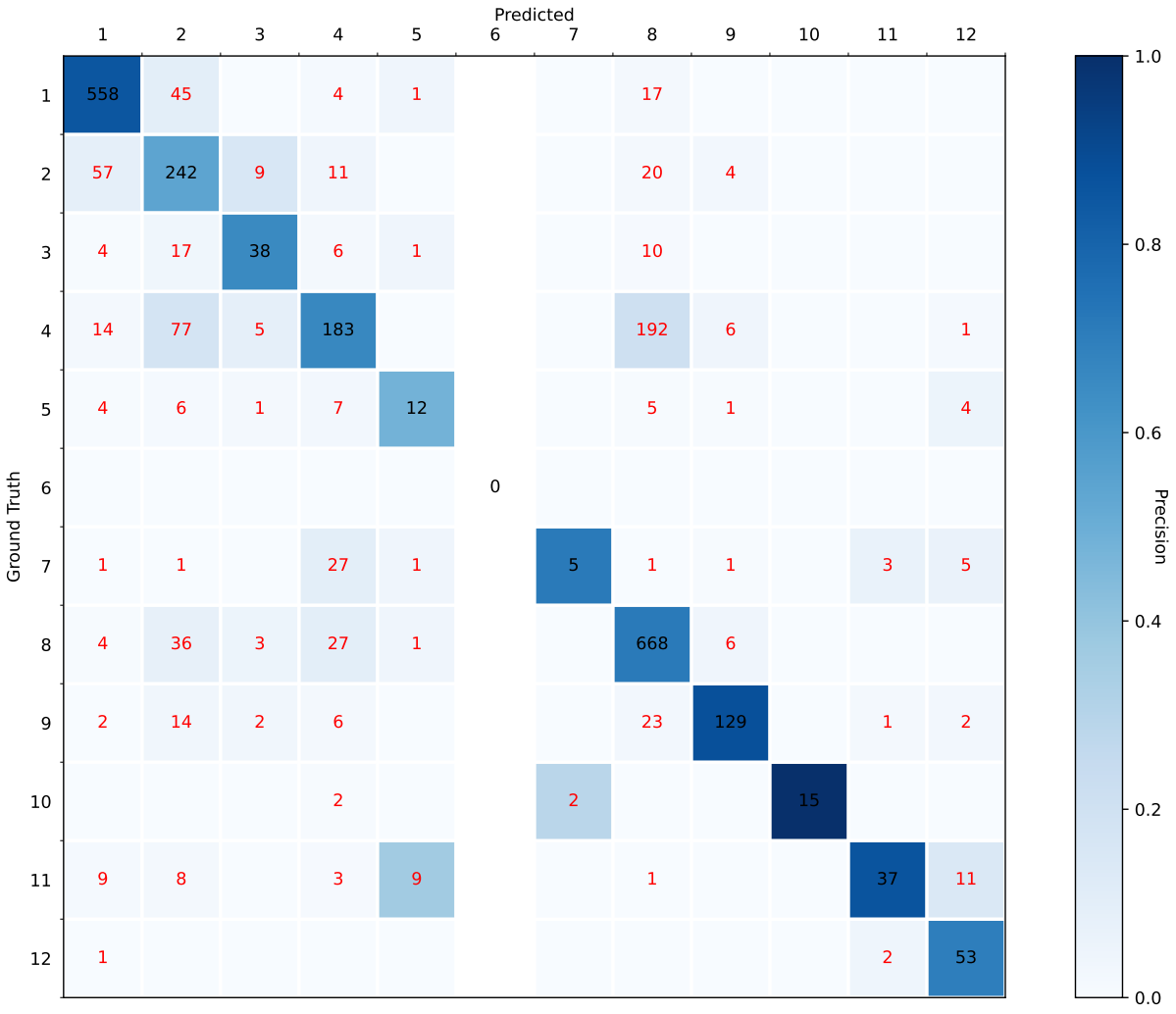

Figure E.1: Confusion matrix of CGLS-LC100 computed on our ground truth normalized by number of elements predicted, i.e. Color on the diagonal reflects per-class precision.

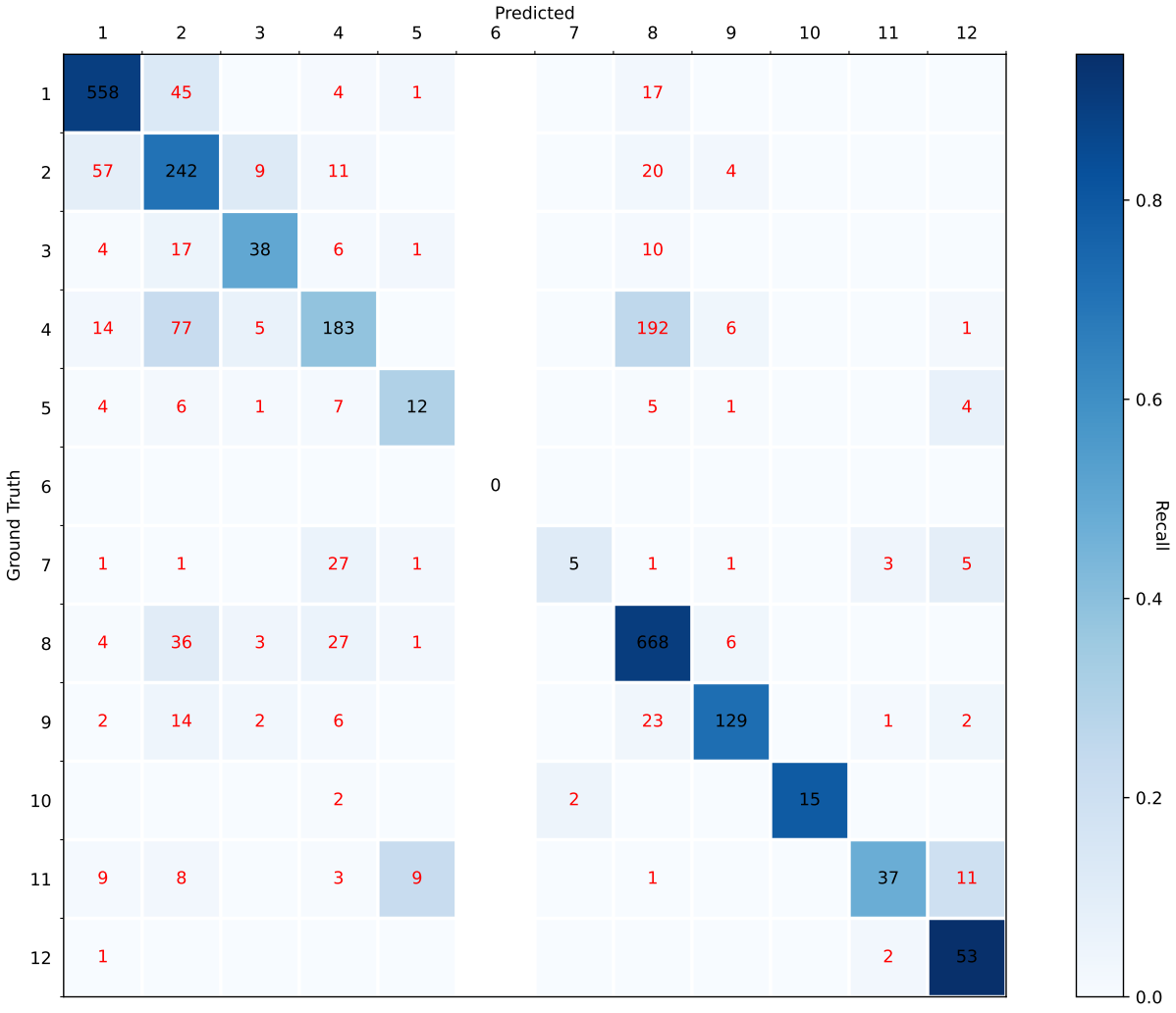

Figure E.2: Confusion matrix of CGLS-LC100 computed on our ground truth normalized by number of elements in the ground truth, i.e. Color on the diagonal reflects per-class recall.

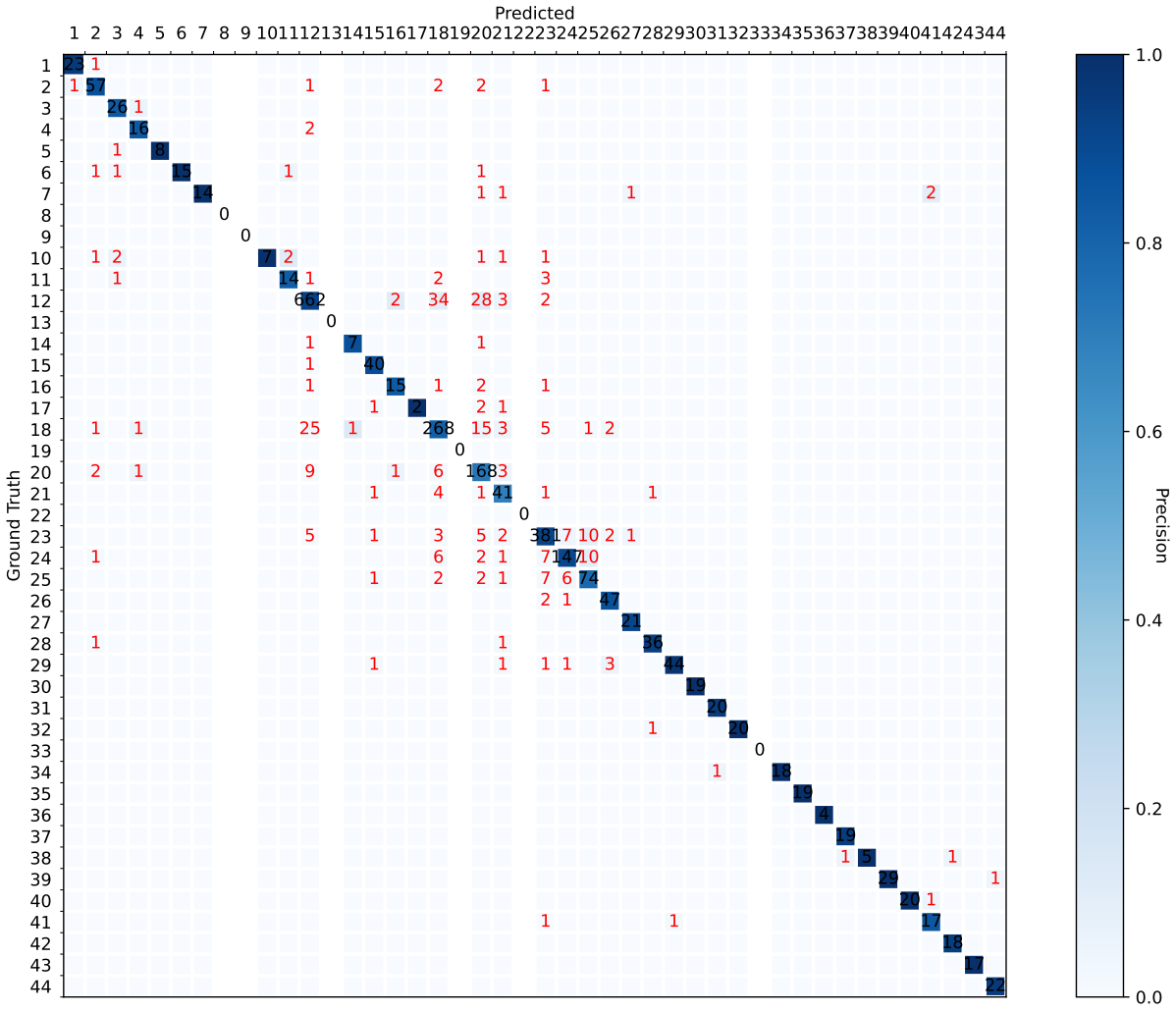

Figure E.3: Confusion matrix of CLC computed on our ground truth normalized by number of elements predicted, i.e. Color on the diagonal reflects per-class precision.

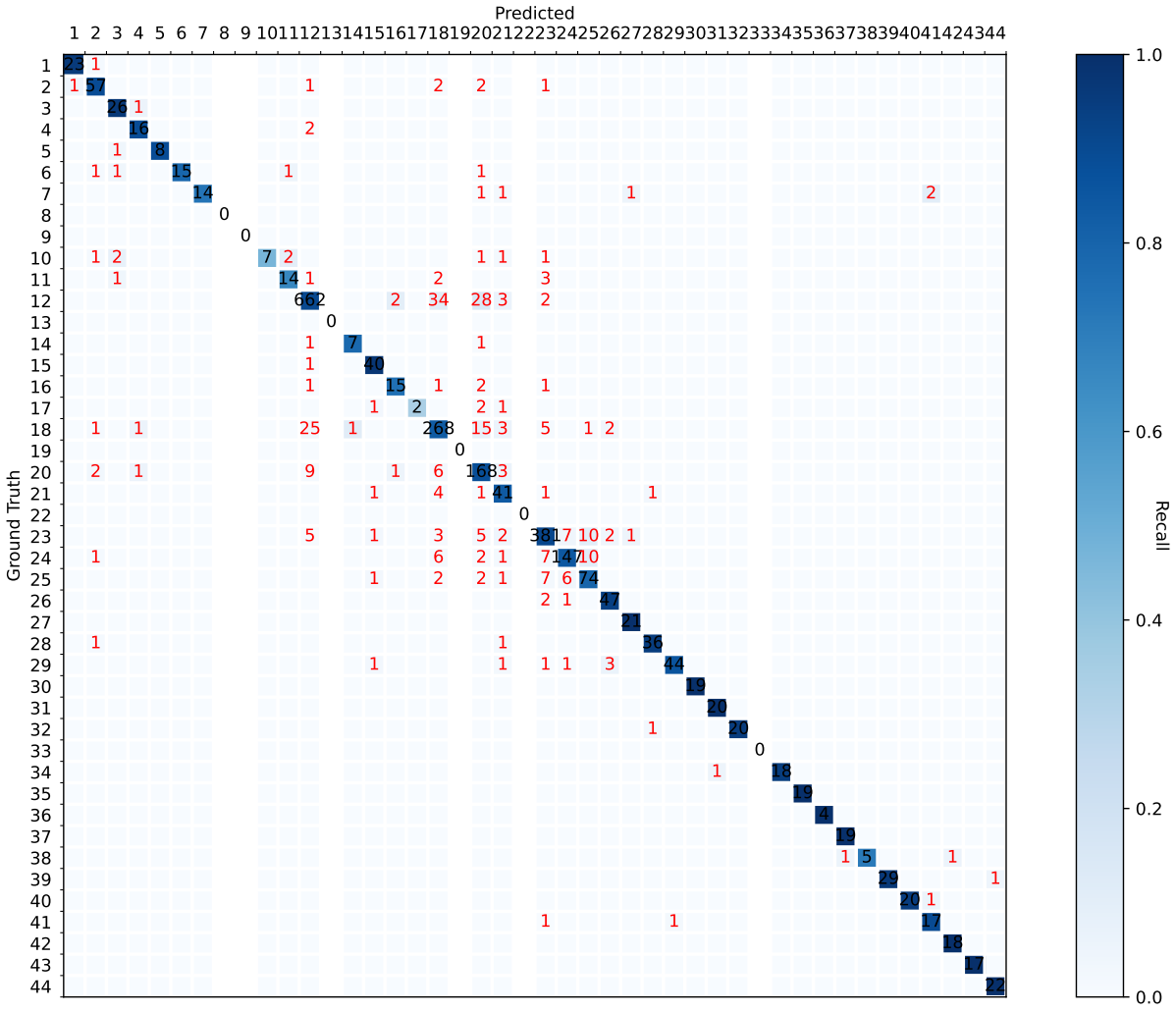

Figure E.4: Confusion matrix of CLC computed on our ground truth normalized by number of elements in the ground truth, i.e. Color on the diagonal reflects per-class recall.

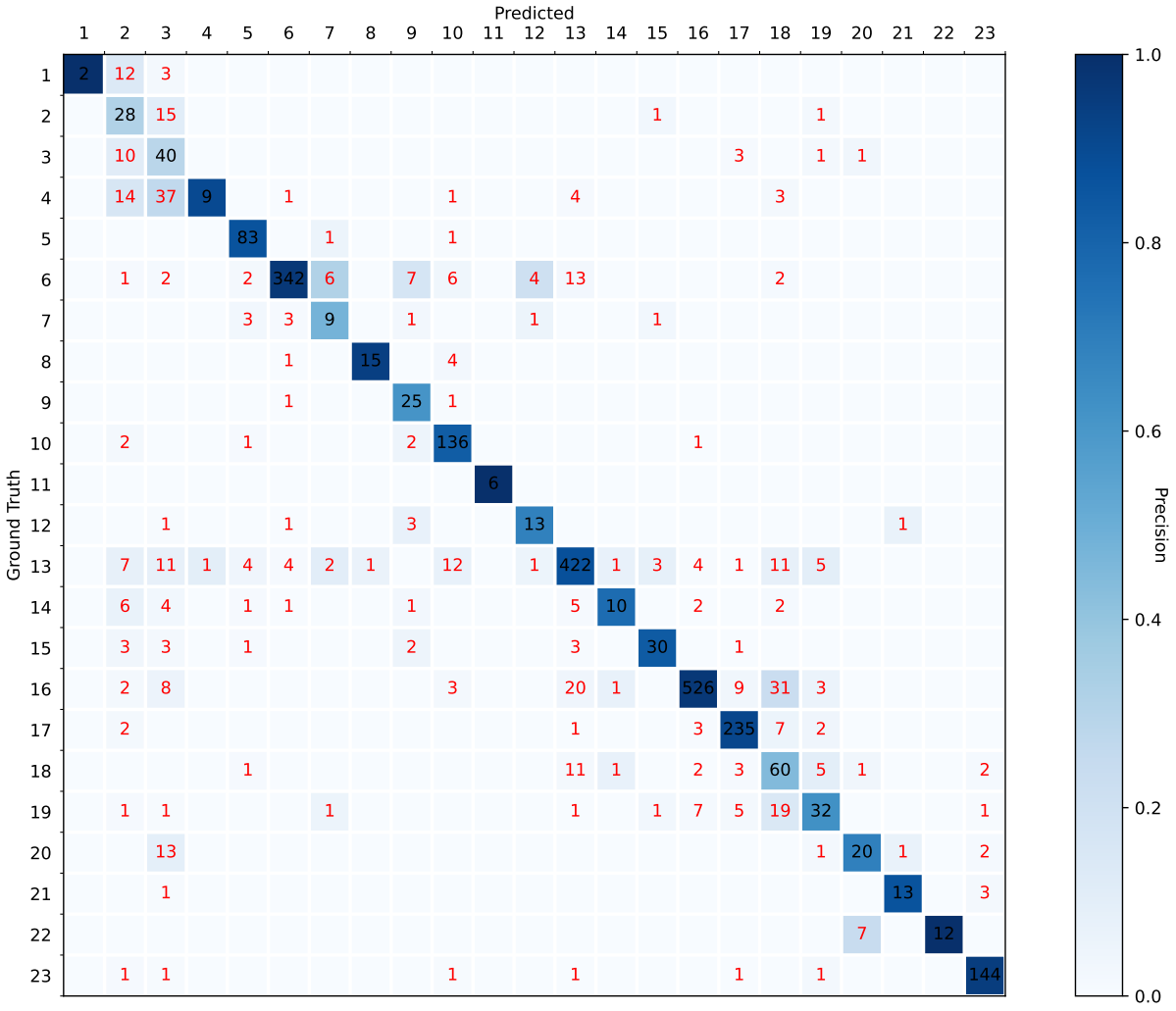

Figure E.5: Confusion matrix of OSO computed on our ground truth normalized by number of elements predicted, i.e. Color on the diagonal reflects per-class precision.

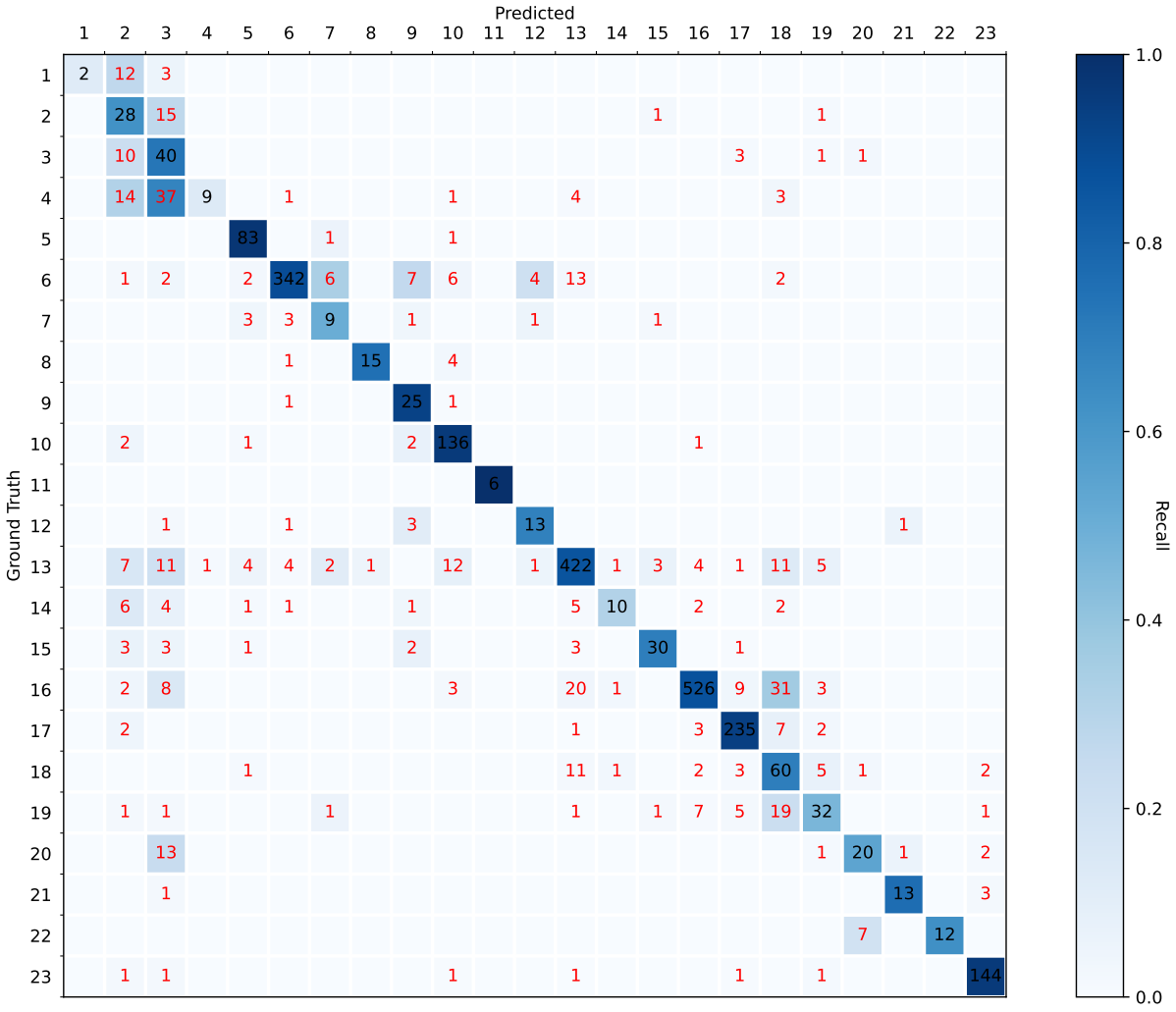

Figure E.6: Confusion matrix of OSO computed on our ground truth normalized by number of elements in the ground truth, i.e. Color on the diagonal reflects per-class recall.

APPENDIX F

Assessing sampling error uncertainty

As we did not find published work on estimating per-class Fscore margin of error depending on the ground truth size, this appendix provide a simple statistical formula. We use the notation introduce in Figure F.1 in our formulas.

		Predicted class	
	Total population = P + N	Positive (PP)	Negative (PN)
Ground truth class	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Table F.1: Confusion Matrix illustration

We define the following values:

$$TPR = recall = \frac{TP}{P} = \frac{TP}{TP + FN}. \tag{F.1}$$

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}. \tag{F.2}$$

$$PPV = precision = \frac{TP}{TP + FP}. \tag{F.3}$$

$$F1_{score} = 2\frac{precision \times recall}{precision + recall} = 2\frac{PPV \times TPR}{PPV + TPR} = 2\frac{A}{B} \tag{F.4}$$

Margin for TPR and TNR values can be computed [52, 230]. Assuming a confidence level of 95% ($z = 1.96$) and n the per class sample size, TPR (same for TNR) is computed using Equation F.5 :

$$\delta_{TPR} = z\sqrt{\frac{TPR(1 - TPR)}{n}}. \tag{F.5}$$

To estimate PPV uncertainty we rewrite Equation F.3 using F.1 and F.2. We obtain Equation F.6 :

$$PPV = \frac{TP}{TP + FP} = \frac{P \times TPR}{P \times TPR + N(1 - TNR)} = \frac{C}{D}. \tag{F.6}$$

As P and N corresponds to ground truth values, they are considered exact (with no uncertainties). Then we can compute uncertainties for PPV by merging Equation F.6 and F.5. We obtain Equation F.9 :

$$\delta_C = P\delta_{TPR}. \tag{F.7}$$

$$\delta_D = P\delta_{TPR} + N\delta_{TNR}. \tag{F.8}$$

$$\delta_{PPV} = PPV(\frac{\delta_C}{C} + \frac{\delta_D}{D}). \tag{F.9}$$

Finally, by merging Equation F.4 with F.5 and F.9 we obtain Equation F.14 :

$$A = PPV \times TPR. \tag{F.10}$$

$$B = PPV + TPR. \tag{F.11}$$

$$\delta_A = A(\frac{\delta_{PPV}}{PPV} + \frac{\delta_{TPR}}{TPR}). \tag{F.12}$$

$$\delta_B = \delta_{PPV} + \delta_{TPR}. \tag{F.13}$$

$$\delta_{F1} = 2\frac{A}{B}(\frac{\delta_A}{A} + \frac{\delta_B}{B}). \tag{F.14}$$

APPENDIX G

Machine Learning and Deep Learning: short introduction

This section concisely introduces the main principle behind machine learning and convolution neural networks to facilitate reading for users with limited knowledge of those techniques. For an in-depth review of all the underlying concepts, refer to [27, 101, 155].

G.1 Machine Learning

Machine learning englobes a set of methods (including deep learning) designed to solve a problem by implicitly modelling a set of discriminative rules using real data. An illustration of the difference between solving a problem with machine learning and imperative approaches can be given using the example of water areas detection on optical satellite images. An imperative approach would consist in assuming a prior knowledge of water characteristics (e.g. water is blue) and writing an algorithm that will annotate all blue areas as water areas. Conversely, in a machine learning-based approach, one provides the algorithm with a set of areas annotated as water/not water and lets it determine a distinctive criterium to distinguish the water areas.

The phase allowing the method to produce the model to respond to the problem from the data provided is referred to as the learning or training phase. It is systematically followed by a validation phase carried out on another set of data to ensure the validity of the model obtained and its ability to generalize (i.e. its transferability to other data). If the network performs better on seen during training data than on unseen data (validation), the network is said to overfit.

Machine learning methods differ from each other both by the algorithms used for building the model and their training methods, that is to say, by the method making it possible to obtain the model. Thus, we distinguish supervised training, for which the data is labelled (we compare the expected output and the output for each input) from unsupervised training. In our land-cover scenario, we will always assume that we have an existing sample

of the desired target. Only supervised training methods are explored in our manuscript.

Deep Learning is a subset of the Machine learning algorithms for which the model is built using neural networks algorithm.

G.2 Neural network

One of the multiple representations of the basic brick of a neural network is the perceptron [258].

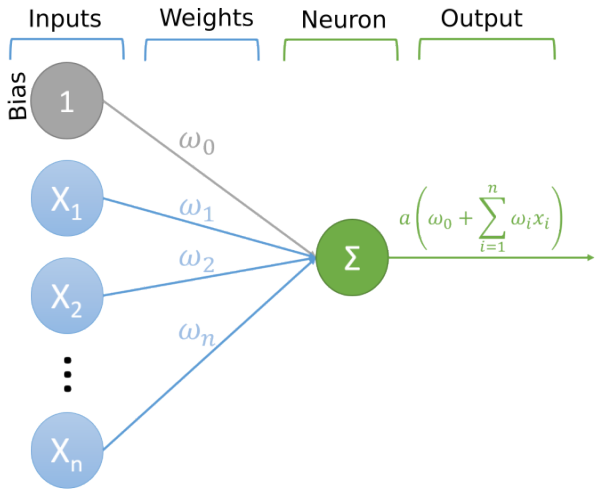

Figure G.1: Perceptron[258]

The following explanations are based on the figure G.1. Let $(X_1, X_2, ..., X_n)$ be a set of input variables corresponding to data provided to the network. The neuron Σ performs a weighted summation of these different input variables with per input variables weight denoted $(\omega_1, \omega_2, ..., \omega_n)$. Finally, it applies an activation function denoted a to this sum and returns the result. The activation function allows, among other things, the resolution of non-linear problems. A classic example of a is the Rectified Linear Unit (ReLU) function defined as follows:

$$ReLU(x) = \begin{cases} 0 \text{ if } x < 0. \\ x \text{ if } x > 0. \end{cases} \tag{G.1}$$

As explained above, machine learning methods are based on two phases (training and inference). In the case of neural networks, the training determines the ideal values of the ω_i to answer the stated problem. Training a neural network is an iterative process in which ω_i are initialised with random values that are progressively adjusted as training progresses. At each iteration, the network process a set of values $(X_1, X_2, ..., X_n)$ input for which the expected output y is known. A comparison between the output of the neurone and the expected value y is then computed using a loss function. Multiple error functions have

been defined depending on the problem type (classification, regression) and the data to be processed. The choice of the loss function greatly conditions the result obtained.

Once the difference between obtained and expected results has been measured, we determine how to update the weights ω_i using the gradient descent algorithm. We set, $x^{(j)} = (X_1^{(j)}, X_2^{(j)}, ..., X_n^{(j)})$ the j$^{\text{th}}$ set of variables corresponding to a statistical individual and $x_i^{(j)}$ the i$^{\text{th}}$ variable of the j$^{\text{th}}$ set. We then call $f(x^{(j)})$ the output obtained by the perceptron for the j$^{\text{th}}$ set and $y^{(j)}$ the expected real value for this set. We can then write:

$$\omega_i \leftarrow \omega_i - lr \frac{\partial Loss(f(x^{(j)}), y^{(j)})}{\partial \omega_i}. \tag{G.2}$$

lr is called the learning rate and is a network hyperparameter. If this parameter is too high, the network weights vary significantly at each iteration to provide a convergence towards a minimum for the considered $x^{(j)}$ and partly forget the learning carried out during the previous iterations. Conversely, if it is too weak, the network will need many iterations before obtaining a relevant result. If the error function used is the function Mean square error ou Ecart Moyen Quadratique (MSE), we can write based on the previous equation:

$$\omega_i \leftarrow \omega_i - lr(y^{(j)} - f(x^{(j)}))x_i^{(j)}. \tag{G.3}$$

We then iterate these operations on the $x^{(j)}$ at least until the error function converges. The resulting weight is then fixed, forming the model used for inference and marking the end of the training phase.

Deep learning or deep learning is based on stacking these building blocks in layers, the output of one of these layers constituting the input of the next (see figure G.2). We then update the w_i (one per arrow on our diagram) by calculating the error on the last layer and back-propagating it. A challenge is then to ensure the convergence of the error function, which does not necessarily converge to the solution, but potentially to one of the many local minima. The choice of specific parameters (learning rate, number of layers, etc.) dramatically conditions the convergence capacity of the network. One of the many possible techniques to facilitate and accelerate the convergence of the network is to randomly group the observations into packets before submitting them to the network to limit the risk that the network over-learns at each iteration. These packets are called "batches".

Through this short introduction to the principles of neural networks, we identify two main points of interest that have been addressed during this internship:

- Deep Learning consists of a set of methods based on neural networks performing learning that can be supervised or unsupervised. The supervised or unsupervised approach often depends on the availability of ground truth. If possible, these two forms of learning should be studied in this study.
- Many hyper-parameters (parameters often chosen empirically) are involved in such networks: learning speed, number of layers, number of neurons in each layer, cost

function... Their choices significantly affect the results provided by the network. It is therefore essential to study their influence.

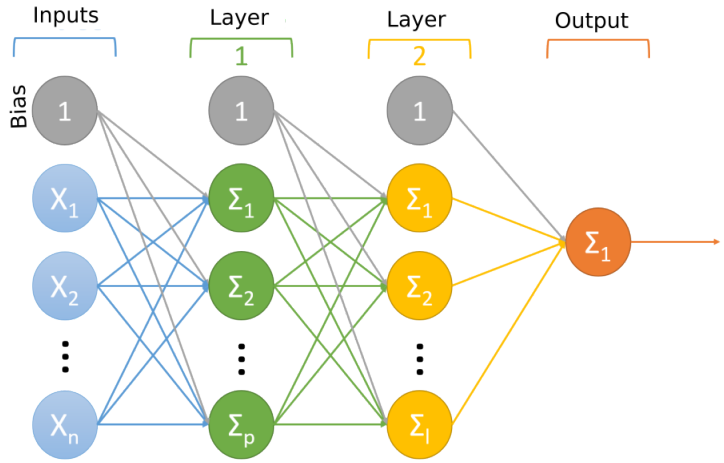

Figure G.2: 3 layers Multi-Layer Perceptron ou Perceptron multi-couches (MLP)

G.3 Convolution neural networks

As we seek to perform land-cover translation on rasterized land-cover, it is necessary to rely on deep learning architectures efficient in image processing. Such architectures mainly rely on the spatial filter convolution of an image. This type of network is arranged in the same way as on the MLP (see G.2). The input X_i is then replaced by a matrix (an image) and the ω_i by convolution filters. The parameters of each filter are learned during the training phase and correspond to feature detectors (elements of interest) present in the image. The idea is that the network learns to recognize through these filters the important elements of an image to perform the expected task. The result of the convolution of an image by one of these filters is called an activation map (or feature map) and corresponds to the highlighting of the places where the feature is present. An example of convolution is presented in figure G.3.

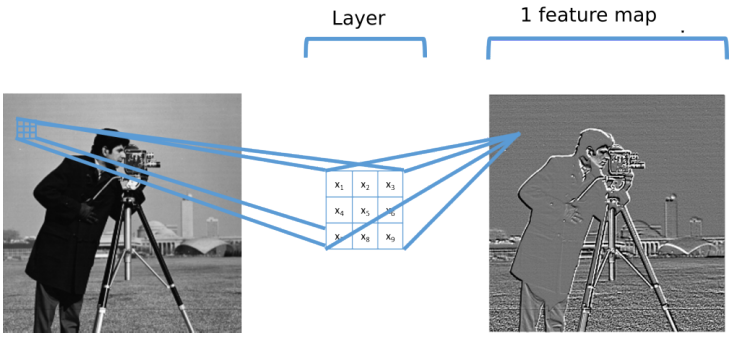

Figure G.3: Convolution by a 3×3 filters

A brief illustration could be given by taking the example of a simple 3-layer network of a neuron aiming to determine "green urban areas" on an image. A conceptual view of what could be learned by the network is that the first layer carries out a convolution by a filter aiming to detect grasslands. The second then seeks to determine on the activation map obtained if an element resembling a city is convolved by a grassland detector (since the city present in the activation map has also been convolved by the grassland filter). We then obtain a second activation map corresponding to the places where a green urban area is possible in a city. Finally, the last neuron sums the elements present on the feature map and returns 0 or 1 depending on the absence or presence of grassland in a city. This example illustrates, in particular, that the more the number of layers increases, the more the concepts learned become abstract, making it resolve the complex task.

Each convolution layer is characterized by its number of filters and the stride at which it moves on the image. Filters are characterized by their size. For each convolution layer, the number of parameters to be learned by the network is equal to the product between the size of the filter and the number of filters present in the layer. We can classically break down a CNN into four structuring elements that are successively stacked (see Figure G.4):

- A convolution layer: convolution between the features and the image to see if (and where) the feature is present. This is always at least the first layer of the network. Features are learned gradually during training. The number of feature maps, as well as the size of the features, are part of the hyperparameters of the network.
- A pooling layer (typically ReLU): we reduce the size of the feature maps. This considerably speeds up the calculation times (by reducing the number of pixels).
- A layer of normalisation (typically Batch Normalization (BN)): we normalise the input values to speed up learning and reduce the risk of over-fitting.
- An activation layer (typically ReLU).

Those four structuring elements are stacked multiple times to increase the depth of the network into what is referred to as a layer. For instance, a two-layer network can be composed of the first layer with a convolution followed by an activation layer and a second layer, taking the output of the first layer and convolution and an activation layer as input. An important observation is that at each convolution, the value of one pixel is influenced by its neighbourhood. The convolution by a 3×3 filter modifies the value of one pixel using the values of its eight neighbouring pixels. Stacked convolution increases the size of the neighbourhood at each layer. The maximum neighbourhood size is termed receptive field and is a key element in leveraging spatial context. A small receptive field only enables to analyse of far-range influence. Various solutions exist to increase the receptive field. The first strategy is to increase the network depth, as each convolution will increase the receptive field. However, the bigger the network, the more complex the training is due to vanishing gradients [176] and potential overfitting due to an increased

number of learnt parameters. The second strategy is to modify the convolution parameters by either using bigger filters (which is memory-consuming as it increases the number of learnt parameters) or dilated filters (instead of processing the nearest neighbours, the filter process the neighbours at a given distance). The last one is to modify the feature maps by reducing their size through, for example, a pooling layer.

Figure **??** shows the effect of different layers on the histogram of a feature map.

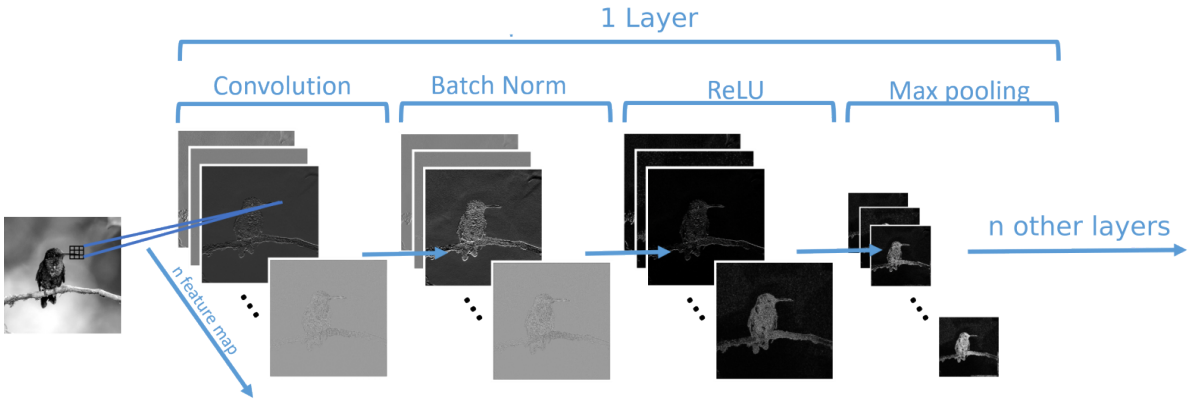

Figure G.4: Convolutional Neural network with a single layers

APPENDIX H

EPI interpretation key

The Edge Preservation Index (see Equation 3.4) is highly correlated to the edge proportion in the reference. To distinguish 6 levels of EPI value (poor, slight, fair, moderate, substantial, almost perfect) we built an interpretation key based on a comparison with other traditional metrics previously (precision, recall, fscore, and Kappa). Figure H.1 presents the EPI, fscore and kappa value for different precision, recall values. As EPI and kappa value depends on the reference's edge proportion, those plots varies depending on the considered land-cover map. We display here the plot for the maps with the smaller (OCSGE use) and bigger (CGLS) edge proportion. The principal take away are listed bellow:

- Kappa and EPI value exhibits 0 value when the precision on edge prediction is identical to the edge proprtion in the dataset. A value bellow 0 indicates that a random classifier replicating edge proportion would perform better than the evaluated classifier.
- When the reference's proportion of edges is high (right figure), the fscore is a bad indicator when its value is below 0.4 as its only takes into account variation in recall and is relatively insensible to broad precision variation.
- When the reference's proportion of edges is low (left figure), the fscore and the Kappa are bad indicators when their value are below 0.4 as they exhibits both a precision and recall asymptotic behavior when the precision and recall widely differ, i.e. they become insensible to wide precision or recall variation, only focusing on the lowest of the two values.
- Kappa, EPI and fscore exhibits almost the same pattern twoard precision and recall when they reach a 0.8 value making them all usable.

To sum up, the main advantage of the EPI compared to fscore and kappa is that it enable studying quality variation when the precision and recall even when the edges are poorly detected. Based on those plots we identify the 6 levels (poor, slight, fair, moderate, substantial, almost perfect) by assuming that each levels must represents the same overall

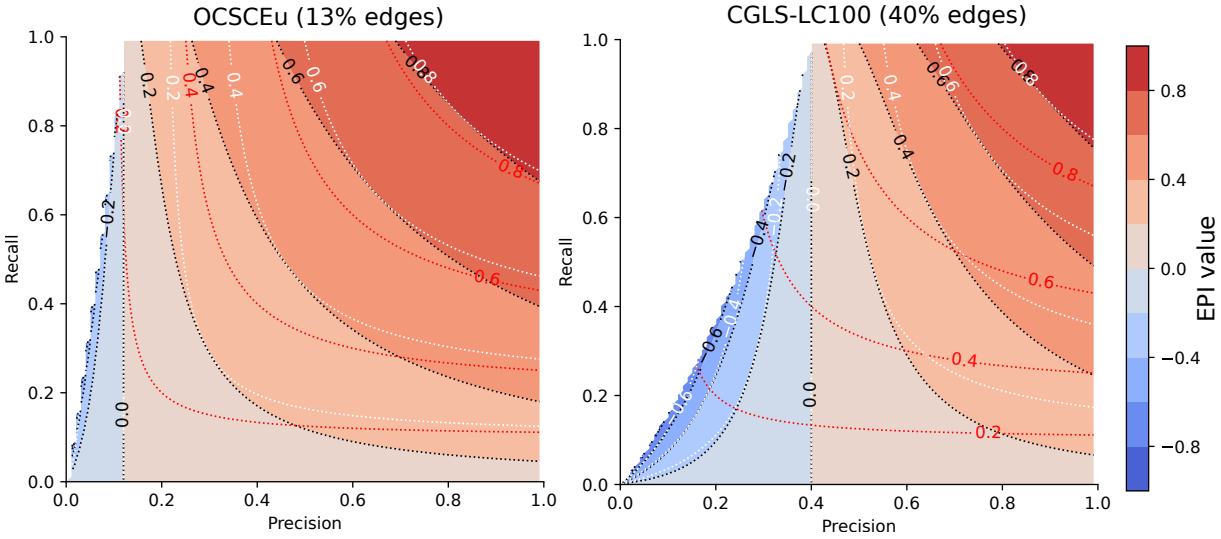

Figure H.1: EPI (black line and colormap), fscore (red lines) and kappa (white line), for different precision recall value. The left plot is computed assuming 13% edges (OCSGE use) while the right is computed assuming 40% edges (CGLS). We underline that since we assume the target proportion of edges in a binary classification setup (edge/non edge) some recall/precision couple can not be observed, in which case they are represented in white.

Target	CGLS	CLC	OSO	MOS	OCSGE cover	OCSGE use	Average
Edge proportion	40	33	23	19	15	12	
Poor	0	0	0	0	0	0	0
Slight (percentile 20)	13	14	16	17	18	18	16
Fair (percentile 40)	26	28	30	31	32	32	30
moderate (percentile 60%)	40	42	45	46	47	47	45
Substantial (percentile 80%)	58	60	62	63	63	63	62

Table H.1: EPI interpretation key, based on percentile of EPI values above 0 computed on diagrams such as the those presented in Figure H.1.. Value above the substantial threshold are termed almost perfect

area on the graph except for the poor value that represent EPI value below 0. As those values depends on the proportion of edges we review the per land-cover results in Table H.1.

APPENDIX I

Map and S1/DEM fusion results

Source		CGLS (P)					CLC (C)					OSO (O)					OCSGEc (G1)				OCSGEu (G2)				MOS (M)			Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	
Early	SourceSpe	64	60	73	82	82	75	62	75	83	83	74	64	78	85	85	70	57	59	89	69		55	56	79	86	84	66
Mid-1	Unique	65	63	75	84	84	76	64	77	85	**85**	73	65	79	86	86	70	59	61	91	70	57	59	82	87	84	67	74
	SourceSpe	65	62	76	83	84	77	65	77	85	84	73	66	80	87	86	71	59	60	90	71	57	59	82	86	84	66	74
	TargetSpe	64	64	75	83	84	**77**	65	76	84	**85**	74	64	80	87	86	69	59	61	92	71	58	58	81	**87**	83	67	74
Mid-2	Unique	65	62	75	84	83	75	64	77	84	84	74	66	79	86	86	71	59	61	90	70	56	58	80	87	**85**	67	74
	SourceSpe	66	63	74	83	83	76	64	77	83	84	73	67	79	85	86	70	60	61	90	71	57	58	80	**87**	84	67	74
	TargetSpe	66	62	75	84	82	75	63	78	83	**85**	74	65	80	86	86	72	60	62	91	70	55	57	79	86	**85**	67	74
Mid-3	Unique	66	64	75	84	84	76	65	77	85	**85**	73	66	79	87	86	71	59	62	92	71	57	60	82	**87**	**85**	67	75
	SourceSpe	66	63	76	85	83	76	64	77	85	86	74	67	81	86	86	69	59	61	92	70	57	59	81	**87**	**85**	66	75
	TargetSpe	64	62	76	85	84	76	64	78	84	84	75	65	81	86	86	69	59	62	92	70	58	58	83	86	**85**	66	75
Late	Unique	**70**	68	**79**	**86**	**85**	**78**	**69**	**79**	**86**	**85**	**77**	**70**	**82**	**88**	**87**	**73**	**63**	**65**	**93**	**73**	62	**63**	**85**	**87**	**85**	**69**	**77**
	SourceSpe	**70**	**69**	**79**	**86**	**85**	**78**	**69**	**79**	**86**	**85**	**76**	**70**	**82**	**88**	86	72	**63**	64	92	72	62	62	84	**87**	84	68	**77**
	TargetSpe	69	67	78	**86**	**85**	**78**	68	78	**86**	**85**	76	**70**	**82**	**88**	86	72	62	**65**	**93**	72	62	62	84	86	**85**	68	**77**

Table I.1: Comparison of various Fusion strategies using Sentinel-1, our 6 land cover maps and the MLCT-Net in terms of OA_{ag}.

Source		CGLS (P)					CLC (C)					OSO (O)					OCSGEc (G1)				OCSGEu (G2)				MOS (M)			Average
Target		C	O	G1	G2	M	P	O	G1	G2	M	P	C	G1	G2	M	P	C	O	G2	P	C	O	G1	P	C	O	
Early	SourceSpe	59	54	69	78	78	71	57	72	80	80	75	65	78	85	85	70	57	57	91	70	54	52	80	86	84	64	71
Mid-1	Unique	60	56	70	80	80	72	57	73	81	81	75	65	78	85	85	71	58	56	91	70	54	52	79	86	83	64	72
	SourceSpe	60	57	70	80	80	73	58	73	81	81	75	65	78	85	85	71	58	56	91	70	54	52	79	86	83	64	72
	TargetSpe	60	57	70	80	80	73	58	73	81	81	75	65	78	85	85	71	58	56	91	70	54	52	79	86	83	64	72
Mid-2	Unique	60	56	70	80	80	72	57	73	81	81	75	66	78	85	85	71	58	56	91	70	54	52	79	86	83	64	72
	SourceSpe	60	57	70	80	80	73	58	73	81	81	75	65	78	85	85	71	59	56	91	70	54	52	79	86	83	64	72
	TargetSpe	60	57	70	80	80	73	58	73	81	81	75	65	78	85	85	71	58	56	91	70	54	52	79	86	83	64	72
Mid-3	Unique	61	58	71	80	80	73	59	73	82	82	75	65	79	85	85	71	58	58	92	71	56	54	81	86	84	64	72
	SourceSpe	61	57	71	80	80	72	59	73	82	82	75	66	79	85	85	71	59	58	92	71	56	54	81	86	84	64	72
	TargetSpe	60	57	71	80	80	73	59	73	82	82	75	65	79	85	85	71	58	58	92	71	56	54	81	86	84	64	72
Late	Unique	67	60	72	80	81	75	60	74	81	82	79	69	80	86	86	74	61	60	93	73	59	55	81	87	84	63	74
	SourceSpe	67	59	72	80	80	75	60	74	81	82	79	68	81	86	86	74	61	60	93	73	59	55	81	87	84	63	**74**
	TargetSpe	67	60	72	80	81	75	60	74	81	82	79	68	80	86	86	74	61	60	93	73	60	55	81	87	84	63	**74**

Table I.2: Comparison of various Fusion strategies using DEM+Aspect, our 6 land cover maps and the MLCT-Net in terms of OA_{ag}.

APPENDIX J

Viual representation of SRS obtained with BoW and Word2Vec

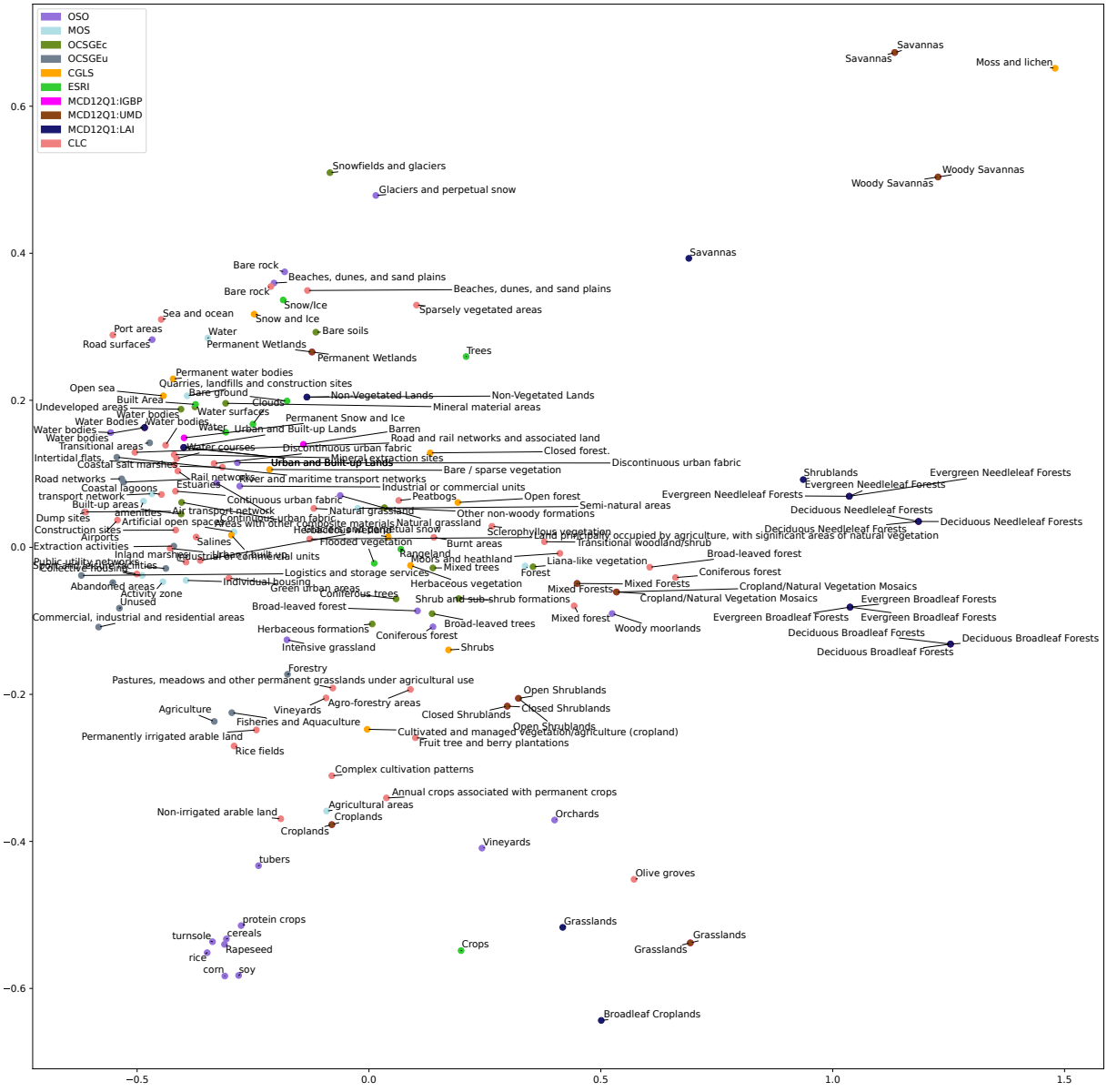

Figure J.1: PCA representation of the HDSRS obtained with a trained on generic text corpora Word2Vec.

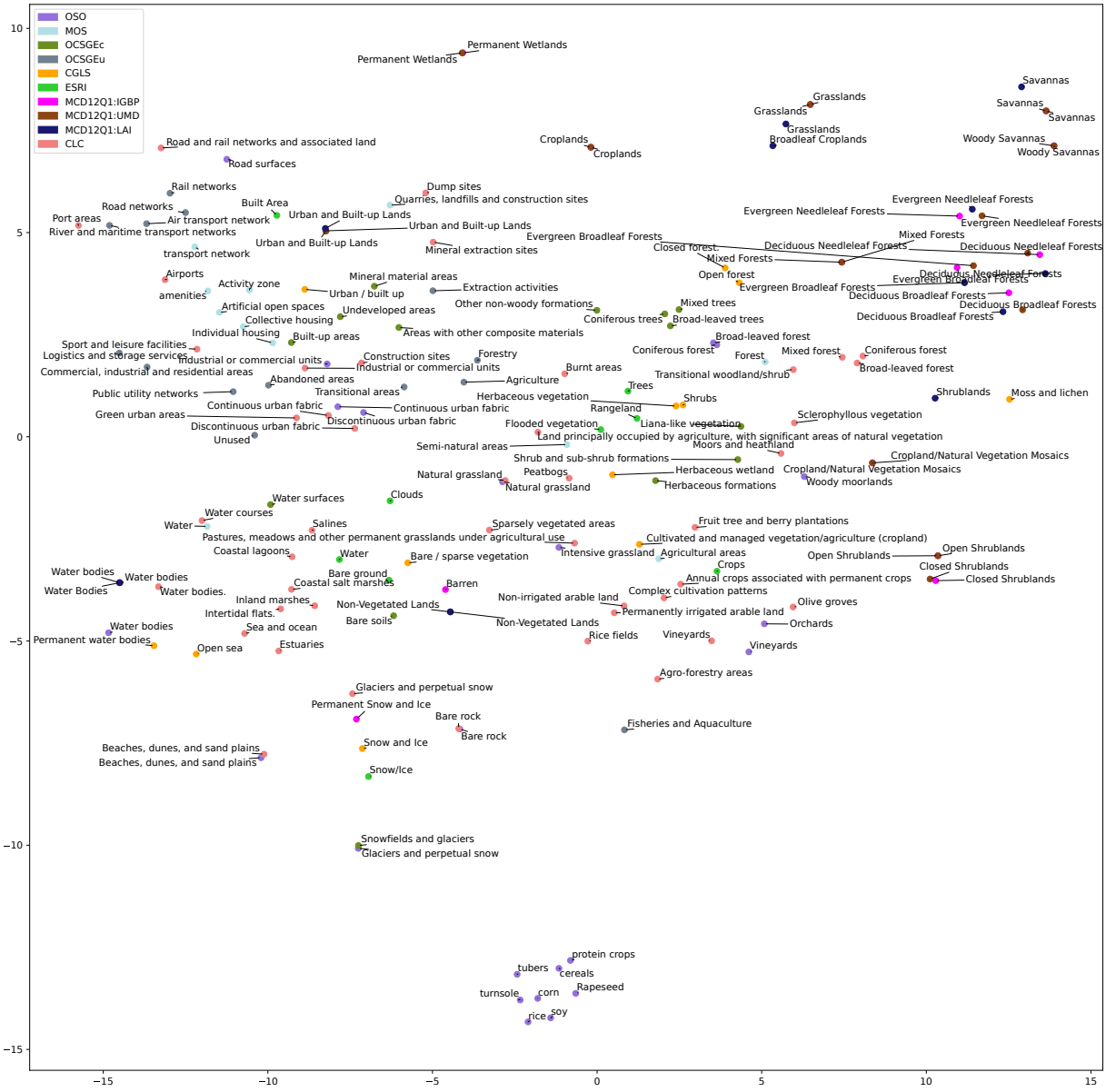

Figure J.2: TSNE representation of the HDSRS obtained with a trained on generic text corpora Word2Vec.

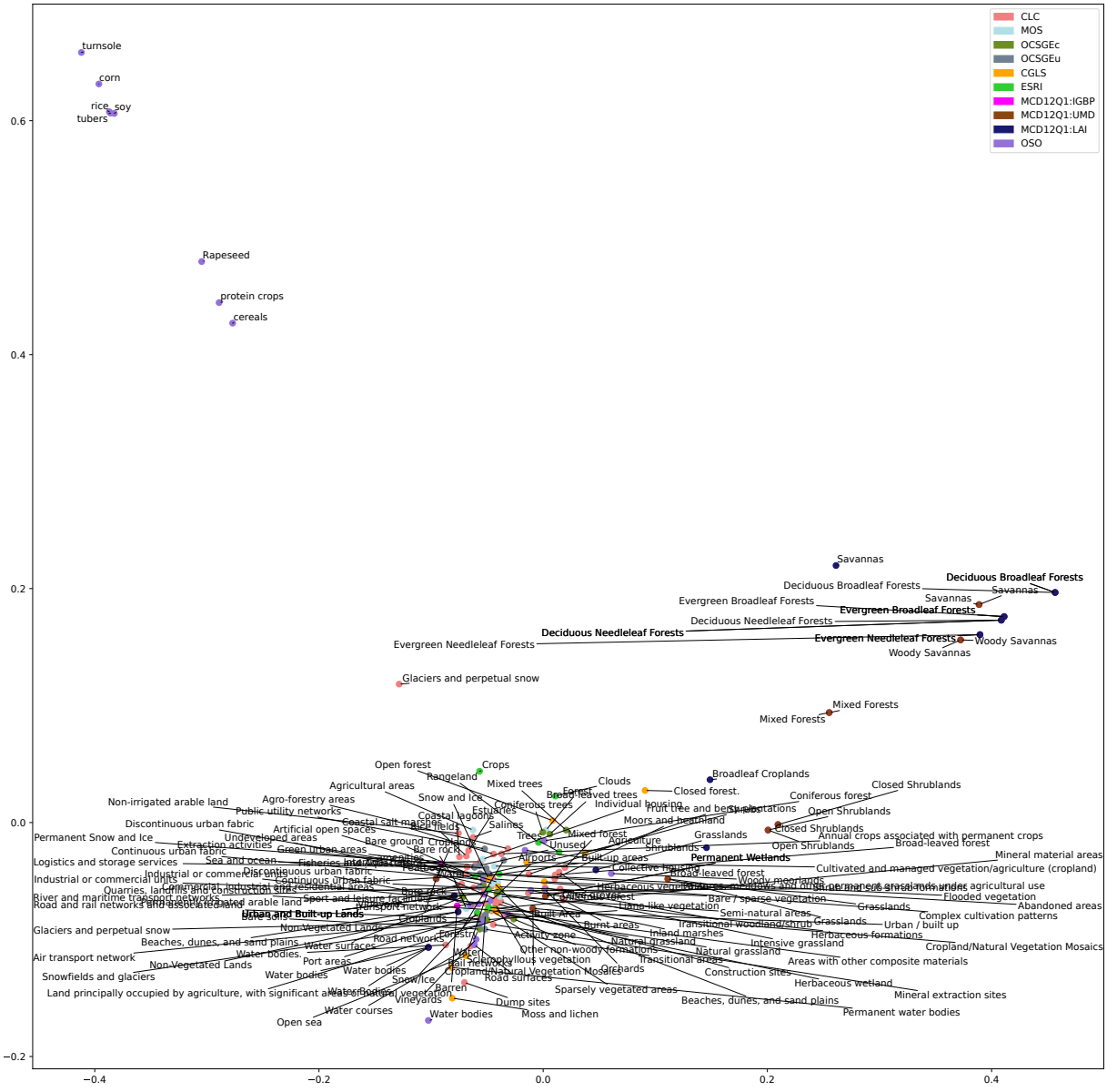

Figure J.3: PCA representation of the HDSRS obtained with a trained on LCDD BoW. OSO is here the test map, unseen when building the dictionnary.

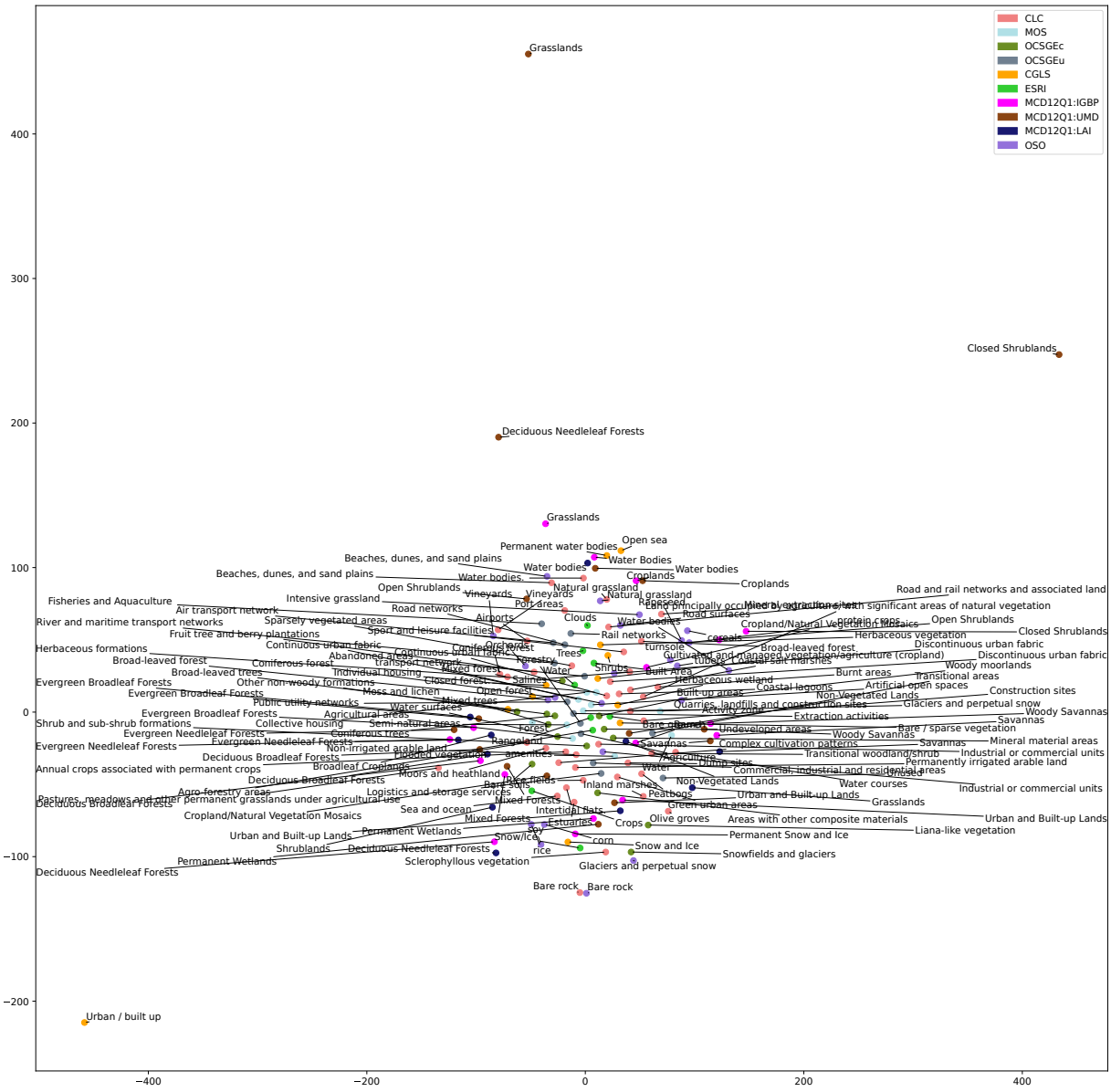

Figure J.4: TSNE representation of the HDSRS obtained with a trained on LCDD BoW. OSO is here the test map, unseen when building the dictionnary.