



HAL
open science

Assessing the transferability of a multi-source land use classification workflow across two heterogeneous urban and rural areas

Martin Cubaud, Laurence Jolivet, Arnaud Le Bris, Ana-Maria Olteanu-Raimond

► To cite this version:

Martin Cubaud, Laurence Jolivet, Arnaud Le Bris, Ana-Maria Olteanu-Raimond. Assessing the transferability of a multi-source land use classification workflow across two heterogeneous urban and rural areas. *International Journal of Digital Earth*, 2024, 17 (1), 10.1080/17538947.2024.2376274 . hal-04631529

HAL Id: hal-04631529

<https://hal.science/hal-04631529>

Submitted on 11 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License



Assessing the transferability of a multi-source land use classification workflow across two heterogeneous urban and rural areas

M. Cubaud, A. Le Bris, L. Jolivet & A.-M. Olteanu-Raimond

To cite this article: M. Cubaud, A. Le Bris, L. Jolivet & A.-M. Olteanu-Raimond (2024) Assessing the transferability of a multi-source land use classification workflow across two heterogeneous urban and rural areas, *International Journal of Digital Earth*, 17:1, 2376274, DOI: [10.1080/17538947.2024.2376274](https://doi.org/10.1080/17538947.2024.2376274)

To link to this article: <https://doi.org/10.1080/17538947.2024.2376274>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 10 Jul 2024.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

Assessing the transferability of a multi-source land use classification workflow across two heterogeneous urban and rural areas

M. Cubaud , A. Le Bris , L. Jolivet  and A.-M. Olteanu-Raimond 

Univ Gustave Eiffel, IGN-ENSG, LASTIG, Saint-Mande, France

ABSTRACT

Mapping Land Use (LU) is crucial for monitoring and managing the dynamic evolution of the human activities of a given area and their consequential environmental impacts. In this study, a multimodal machine learning framework, using the XGBoost classifier applied to attributes constructed from heterogeneous spatial data sources, is defined and used to automatically classify LU in the two French departments of Gers and Rhône. It reaches a mean F1 score of 83 and 86% respectively. This research work also assesses the robustness and transferability of the machine learning model between these two diverse study areas and highlights the challenges encountered, arising mainly from the differences of distribution of the attributes and classes between the study areas. Adding a few samples from the test study area allows the model to learn some specificities of the test study area, and thus improves the results. Moreover, the study evaluates the individual contributions of each data source to the accuracy of predictions of the LU classes, providing insights concerning the relevance of each data source in enhancing the overall precision of the Land Use classification. The findings contribute to a validated LU classification workflow, identify valuable data sources, and enhance understanding of model transferability challenges.

ARTICLE HISTORY

Received 26 February 2024
Accepted 30 June 2024


KEYWORDS

Land use classification;
transferability; data-fusion;
machine learning; LULC

1. Introduction

Land Cover (LC) and Land Use (LU) are dual elements to monitor Earth surface over time. Land Cover (LC), i.e. the physical description of the territory, describes the surface cover of the Earth (e.g. building, vegetation, water). Land Use refers to the description of the territory in terms of the human activities, e.g. residential, agricultural or tourism activity. It reflects the socio-economic, cultural, and environmental priorities of a particular area. Land Use maps enable the monitoring of the dynamics of changes in a given area, providing valuable insights into urbanization or land take studies. Indeed, they are used to implement strategies for a net land take of zero by 2050 or to prioritize the preservation of natural habitats (EEA (European Environment Agency) 2023; Li, Ma, and Zhou 2022). They also support spatial planning, in order to understand the potential uses of land and their evolution over time, at large and small spatial scales. Nowadays, in the context of climate change and environmental monitoring, Land Use maps are crucial in assessing the impact of

CONTACT M. Cubaud  martin.cubaud@ign.fr  Univ Gustave Eiffel, IGN-ENSG, LASTIG, F-94160 Saint-Mande, France

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/17538947.2024.2376274>

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

human activities on biodiversity allowing to quantify the threats and degradation caused by them over time, to manage natural resources or to plan sustainable agriculture. In addition, they provide useful information to assess the status and trends of different indicators of sustainable development goals (SDGs) defined by the United Nations, such as SDG 15: Life on Land, SDG 11: Sustainable Cities and Communities, SDG 2: Zero Hunger, to name but a few (Asuquo Enoch, Ebere Njoku, and Chinenye Okeke 2023; Reed et al. 2017). Finally, Land Use maps are vital for public institutions and governments to define public policies and regulations (Reed et al. 2017).

Many different LU and/or LC data are available at different spatial and temporal scales, such as CORINE Land Cover and Urban Atlas on a European scale, or national products, such as the OCS GE produced by the French National Mapping Agency (NMA) (García-Álvarez and Nanu 2022). In recent years, significant progress has been made in mapping and updating LC maps, thanks to the advancement of machine and deep learning techniques, as well as the increased availability of massive new satellite programs such as Copernicus (Vali, Comai, and Matteucci 2020). In addition to the research community, NMAs are also actively engaged in advancing this endeavor by either independently or collaboratively working with the research community to define automatic or semi-automatic solutions to produce accurate and up-to-date LULC data on a national scale (Olteanu-Raimond et al. 2022). For instance, an AI workflow for LC classification based on deep learning methods has been developed by the French NMA, allowing for the pixel based LC classification, following a segmentation method to derive vector LC databases. Moreover, to continually enhance outcomes and facilitate data sharing with various communities, challenges are organized. For example, the FLAIR challenge, which consists in studying the transferability of LC classification between several areas and dates using a very large dataset, received more than 1000 contributions, and allowed to improve the baseline mean intersection over union (mIoU) from 0.55 to 0.65 (Garioud et al. 2023).

However, automatic LU mapping remains a challenge for both research and national data producer communities and still requires significant manual effort. In fact, traditional methods based on photo interpretation are costly. Recognizing the use of a place or a construction might even be impossible due to the limitations of aerial images that mostly convey information about the LC. For example, two buildings with a similar appearance may serve for different activities, e.g. one can be a dwelling and the other work offices. More advanced methodologies based on remote sensing and machine learning encounter the same limitations. For instance, Zhang et al. (2018) used a U-Net network to predict urban Land Use and Land Cover from Word-view remote sensing images. However, they limited themselves to 6 classes, mostly encapsulating LC aspects.

The inherent complexity of addressing this challenge requires not to limit ourselves to remote sensing images. Our research hypothesis is that extracting information from existing heterogeneous spatial data sources describing both the physical features and the use of space can be used as valuable input to map LU. However, it can be difficult to exploit several of these sources using predefined rules.

Thus, Cubaud et al. (2023) developed a new framework based on machine learning that enables the improvement of the accuracy and diversity of LU classification. First, Cubaud et al. (2023) defined attributes from several data sources that describe target LU polygons. Then, they compared pre- and post-classification data-fusion approaches to predict three LU classes: industrial, commercial, and residential. The authors showed that the pre-classification approach using the XGBoost algorithm gave the most accurate prediction outcomes, compared with the other approaches and classifiers tested. Thus, the XGBoost algorithm will be chosen for the current work.

The objectives of the current research work are as follows. The first objective is to extend the methodology of Cubaud et al. (2023) by constructing new attributes on the target LU polygons to be predicted, and to assess its ability to predict the LU classes of a more complete nomenclature of 13 classes. Moreover, in real-life scenarios, it is costly to train a new model for each new area, and it requires a ground truth training set which is not always available. Thus, the second objective is to assess the transferability capacities of the machine learning process between two study areas with large differences.

The contributions of this paper are: (1) to validate and consolidate an existing workflow for LU classification by extending it to 13 LU classes and heterogeneous study areas, (2) to identify which available data sources are valuable for LU classification, (3) to analyze the transferability of the XGBoost machine learning model between two study areas, and (4) to provide open access to the code and a dataset containing 639,066 annotated LU polygons, representing a total of 9516 km², with their constructed attribute values. The dataset is available at <https://doi.org/10.5281/zenodo.10462844> and the code at github.com/mcubaud/multi-source-Land-Use-classification.

The paper is organized as follows. First, Section 2 gives a comprehensive overview of the current state of the art in LU classification and studies several LU nomenclatures. Next, Section 3 proposes a workflow for LU classification using several heterogeneous data sources. Section 4 explains how this workflow is applied to two different study areas and how its transferability is assessed. Section 5 then presents the results of the experiments. Finally, Section 6 discusses the importance of the sources in the different experiments and analyzes why some classes are not adequately predicted.

2. Related work

This section first presents an overview of the current research on LU classification, followed by a comparison of existing LU nomenclatures.

2.1. Land use classification

Some of the proposed approaches directly address LU mapping, while others indirectly contribute to the LU mapping process by enhancing topographic data sources with LU information. Thus, the approaches adopt different types of mapping units, ranging from urban neighborhoods, often referred to as urban functional zones, to meter resolution raster grids, passing through building scale, known as building function. The objective is often to have a geometry in which the activities are as homogeneous as possible.

In the literature, several data sources and algorithms have been proposed to predict Land Use associated with these mapping units.

The remote sensing approach for LU mapping consists of classifying the pixels of satellite or aerial images by exploiting their radiometry, their texture, or their spatio-temporal variations (Yin et al. 2021). However, these pieces of information are mostly related to the Land Cover. Some studies therefore exploit the relationships between imagery, LC and LU. In Zhang et al. (2019), LU and LC are learned together iteratively using a deep neural network based on imagery and class probabilities of LC and LU. Graph Neural Networks can also be used to integrate spatial and contextual relationships between LC segments (Li and Stein 2020; Liu et al. 2022).

Several geographical data sources other than optical imagery have also been used to map LU, such as topographic data, Radar, LiDAR, volunteered geographic information (VGI), or involuntary geographic information (iVGI) (See et al. 2016). The use of several of these sources can be referred to as data fusion, according to the definition given by Hall and Llinas (1997): 'data fusion techniques combine data from multiple sensors and related information from associated databases to achieve improved accuracy and more specific inferences than could be achieved by the use of a single sensor alone'.

To identify building functions, Fonte et al. (2018) used rule-based classifications of OpenStreet-Map (OSM), Facebook and Foursquare VGI data. Meng et al. (2012) identified residential buildings by combining images, a Digital Surface Model extracted from LiDAR data, and distances from major roads sourced from an authoritative database, using a decision tree classifier. Pan et al. (2013) utilized Taxi GPS traces (iVGI) to infer the social function of specific locations employing a Support Vector Machine. Deng et al. (2022) identified building functions from images, Points of Interest (POI) and building footprint (authoritative database) from AMAP and distance to OSM roads using a XGBoost classifier. Zhang et al. (2023) identified building functions applying a

transformer network on building footprints and POI from AMAP, and user density data from the Tencent application (iVGI).

At the scale of the urban neighborhood, Vu et al. (2021) mapped urban functional zones by segmenting Sentinel 2 satellite images, clustering the segments using K-means, and making a rule-based fusion of the clusters and their OSM content. Wu et al. (2023) mapped urban functional zones by combining Google Earth satellite images, impervious surface generated from Landsat 8, Sentinel-1 and nighttime images, parcels extracted from OSM roads, POI from AMAP and Tencent users location, using an auto-encoder network and a AdaBoost classifier. Hu et al. (2023) estimated the LU distribution of traffic analysis zones by applying fuzzy C-Mean on attributes obtained by three unsupervised neural networks on the location and times of taxi flows and on the content of Sina Weibo tweets (iVGI). Su et al. (2023), Xu et al. (2023) and Yang, Bo, and Zhang (2023) mapped urban LU applying a Graph Neural Network on the graph of neighbors of the POI from AMAP (VGI) within a block defined by OSM roads. Mawuenyegah, Li, and Xu (2022) defined attributes describing the spatio-temporal variations of the location of Twitter messages (iVGI) and used Random Forest to map urban functional zones.

LU has also been mapped at other scales. Tu et al. (2020) employed a Random Forest classifier to categorize LU by combining classical optical imagery, night lights intensity, radar imagery, POI from Baidu (VGI) and demographic data from WorldPop. Liu et al. (2021) integrated VGI data from multiple mapathon campaigns and on-site assessments (*in-situ*) using the Dempster-Shafer Theory (DST) to classify Land Use changes. Li et al. (2023) used a knowledge representation learning model and a multilayer perceptron on the knowledge graph linking POI from AMAP to administrative regions, adjacent regions between them, regions to OSM roads and regions to districts, in order to estimate the proportion of each use within the region. He et al. (2021) combined optical images and user density data from the Tencent Web application (iVGI) using a convolutional neural network to classify Land Use areas.

To conclude this subsection, an analysis of the current state of the art with respect to our requirements is provided. First, we observed that most studies on Land Use classification and the resulting maps only focus on urban LU. The data sources used differ between the study areas, depending on their availability or completeness. Some are open data, but some have restricted access or have to be paid for. To classify rural areas, using taxi GNSS tracks or social media usage data may be irrelevant in terms of data completeness.

Second, among the related research works, we can note that several of them (He et al. 2021; Hu et al. 2023; Li et al. 2023; Liu et al. 2022; Su et al. 2023; Sun et al. 2023; Xu et al. 2023; Yang, Bo, and Zhang 2023) use deep learning to extract attributes from one or several data sources. Some studies (Liu et al. 2022; Su et al. 2023; Xu et al. 2023; Yang, Bo, and Zhang 2023) use graph neural networks (GNN) that are able to model the spatial relationships between vector objects. Nevertheless, some of them (Su et al. 2023; Yang, Bo, and Zhang 2023) apply the GNN to graphs of POI within the mapping units, which is an inappropriate approach in rural areas, since the majority of mapping units will not have any POI. Others (Liu et al. 2022; Xu et al. 2023) use the graph of adjacent polygons, with the hypothesis that it can identify LU spatial patterns. In addition, the main use of deep learning for multi-source LU classification is to extract information from images using convolutional neural networks (CNN), but it poses practical challenges. The process involves training autoencoder or encoder-decoder networks, extracting latent representations of all images, possibly resampling these representations, computing aggregation at the scale of the polygon, and finally concatenating with other attributes. This multistep procedure significantly increases the complexity of the classification process, making it less feasible. Furthermore, given that images primarily contain information related to Land Cover rather than Land Use, the additional complexity of integrating CNN-based features may not outweigh the potential benefits in improving classification results.

Finally, there are still few studies on model transferability for LU classification, which remains a challenge. To our knowledge, Zhang et al. (2023) made the only related work in which the authors attempted to transfer their model between two study areas. However, this study has the following

limitations: it is focused only on buildings, the input data sources (i.e. POI within the buildings, social media usage) have completeness issues in rural areas, and the transferability is assessed only between two neighborhoods of the same city. Therefore, we will study the transferability of the XGBoost model between two highly heterogeneous study areas.

2.2. Semantic analysis of land use classes

The aim of this subsection is to give an overview of existing LU nomenclatures and to discuss the different ways to organize a LU nomenclature.

Several LULC nomenclatures are composed of Land Use classes and Land Cover classes, mixed together within a single hierarchical system (Cihlar and Jansen 2001). This is for example the case of CORINE Land Cover (CLC), which has classes defined by their LU (e.g. industrial or commercial areas) and classes defined by their LC (e.g. broad-leaved forest). However, a given use can take place in areas with different LC (e.g. for residential use both buildings and the vegetation of the gardens are taken into account). Similarly, a given LC can be found in areas with different LU (e.g. for built-up areas both residential use and commercial use are possible). Thus, as mentioned in the introduction, LU and LC represent two distinct aspects of the territory. On the contrary, some other products strictly differentiate between LU and LC, defining a separate nomenclature for each. Most of them have a hierarchical structure, where the LU classes are organized in several levels. The more general classes are subdivided into more specific classes. The Hierarchical Land Use Classification System (HILUCS), imposed by the European INSPIRE directive, has such a hierarchical structure (Temporary MIWP 2021-2024 sub-Group 2.3.1 2023). HILUCS has a three level hierarchy of 98 classes, based on their socio-economic activity. The UK National Land Use Database (Harrison 2006) and the GB 50137–2011 Chinese code for Classification of Urban Land Use (chinesestandard.net 2014) have a two-level nomenclature also based on socio-economic activity but with a different organization. The Australian Land Use and Management Classification (DAFF (Department of Agriculture and Water Resources) 2016) is composed of 196 classes in three levels, the highest level being the intensity of the use, therefore reflecting its impact on the environment, and the other levels being split according to the socio-economic activity.

In current existing products, such a high level of semantic precision is not always reachable without decreasing the accuracy of the map. This is primarily due to the challenge of balancing fine-grained distinctions in land use classes with the overall reliability of the classification results. The inherent complexity and variability of land use patterns, coupled with limitations in available data sources, contribute to the difficulty in achieving both high semantic precision and map accuracy simultaneously. The LU nomenclature of the OCS GE (Occupation des Sols à Grande Echelle) (IGN (Institut National de l'Information Géographique et Forestière) 2022) is therefore constructed to be compliant with the INSPIRE directive, with a less detailed nomenclature than HILUCS. This ensures interoperability with other European datasets and standards, and allows OCS GE LU to have one of the highest semantic and geometric accuracy in Europe. OCS GE LU nomenclature has a three level hierarchy and is composed at the most detailed level of 20 classes, that are presented in the Table 1. To clarify the meaning of some classes, the class LU1.5 (Other primary production) represents areas dedicated to the exploitation of natural resources not included in the previous classes, e.g. hunting areas. The class LU4.3 (Public utility networks) represents the infrastructures used for energy, wastes or water supply. The class LU6.1 (areas in transition) describes an area under construction. The class LU235 defines a mix use between the classes LU2 (Secondary Production, i.e. industry and manufacturing), LU3 (Tertiary Production, i.e. commercial and services) and LU5 (Residential Use). The LU classification is made by a rule-based approach with intensive manual corrections. The OCS GE product is available in open license at <https://geoservices.ign.fr/ocsge>. It will be used in France as a measure of land take in order to reach a net land take of zero by 2050. It is possible, to a certain extent, to translate an existing LULC map from one nomenclature to another. For instance, the EAGLE matrix (Arnold et al. 2013) is a tool for comparing LULC nomenclatures. It

Table 1. OCS GE's LU Nomenclature compliant to the INSPIRE directive (IGN (Institut National de l'Information Géographique et Forestière) 2022).

Level 1		Level 2		Level 3	
Code	Name	Code	Name	Code	Name
LU1	Primary production	LU1.1	Agriculture		
		LU1.2	Forestry		
		LU1.3	Mining And Quarrying		
		LU1.4	Fisheries and Aquaculture		
		LU1.5	Other primary production		
LU2	Secondary Production (industry and manufacturing)				
LU3	Tertiary Production (commercial and services)				
LU4	Transport Networks, Logistics And Utilities	LU4.1	Transport Networks	LU4.1.1	Road Transport
				LU4.1.2	Railway Transport
				LU4.1.3	Air Transport
				LU4.1.4	Water Transport
				LU4.1.5	Other Transport Networks
LU4.2	Logistical And Storage Services				
LU4.3	Public utility networks				
LU5	Residential Use				
LU6	Other	LU6.1	Areas in transition		
		LU6.2	Abandoned areas		
		LU6.3	No use		
		LU6.6	Unknown use		
LU235	Secondary and tertiary production and residential use				

decomposes each class in a given LULC nomenclature by aligning it with the classes specified in the INSPIRE directive, indicating for each which INSPIRE classes must be included, are compatible, or are otherwise related to the class definition. Rules can therefore be designed for translation thanks to this tool, but with the limitation that one-to-one associations are not always possible. Therefore, Baudoux, Inglada, and Mallet (2023) proposed to learn a common representation of several LULC maps, using CNN-based encoder-decoder networks. Then, at inference time, the models can be used to translate from one map to another. However, this approach supposes that there are areas for which there are LULC maps using both the input and target nomenclature.

Thus, the goal of our research is to assign a land use class for each mapping unit, according to a given nomenclature. In this paper, we assume that the mapping units are already defined, and we adopt the OCS GE LU nomenclature. For that reason, we use the mapping units already defined in the OCS GE. This justifies our decision not to use semantic image segmentation models in our research work.

3. Materials and method

3.1. Description of the workflow for LU classification

In this subsection, we describe our general workflow for LU classification. It is composed of four main steps, illustrated in Figure 1:

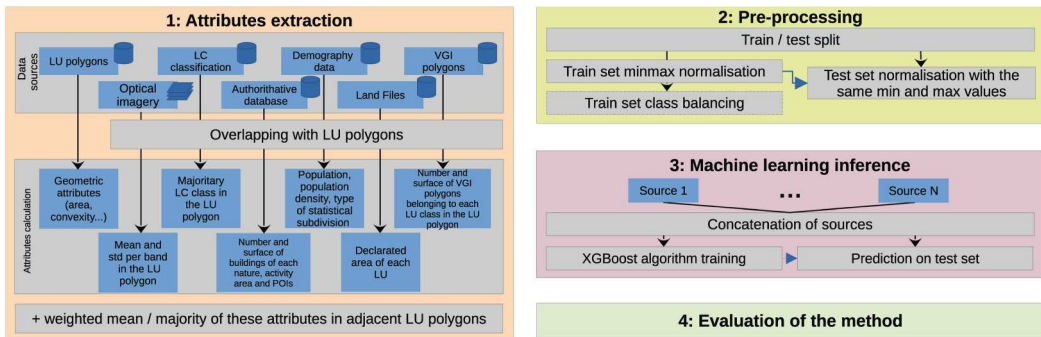


Figure 1. Workflow for LU classification adapted from Cubaud et al. (2023).

- Attributes extraction (Section 3.1.1);
- Pre-processing (Section 3.1.2);
- Machine learning inference (Section 3.1.3);
- Evaluation of the method (Section 3.1.4).

As mentioned before, the assumption is made that the desired output geometry, referred to in the rest of the article as LU polygons, is already defined. This geometry could for instance be obtained by image segmentation or by constructing blocks from the road network. The LU polygons must partition the study area in order to fully map it. Hereafter, the geometry is chosen to be the polygons of an existing LU map, i.e. the OCS GE. These polygons were obtained by combining Land files, a topographic database and an LC segmentation, followed by manual geometric correction.

3.1.1. Attributes extraction

The aim of this first computing step is to characterize each LU polygon by computing attributes from the intersection between the LU polygon and each data source. To consider the spatial context of the LU polygons and the potential spatial relationships between different land uses, we additionally designed neighboring attributes. For example, a vegetated area in front of a building is likely to be a garden, and thus considered as residential use. These attributes are constructed as the average of each attribute in the adjacent LU polygons, weighted by the length of the shared borders. The weights are based on the hypothesis that the more two adjacent polygons share borders, the more they influence each other. The list of the data sources used and the details about the attributes are provided in Section 3.2.

3.1.2. Pre-processing

First, the dataset containing the computed attributes and their ground truth LU is randomly split into an 80% training set and a 20% test set. A minmax normalization is then applied to both, based on the values computed from the training set. Finally, to address the challenge of strong class imbalance for machine learning algorithms, the classes in the training set are rebalanced. In Cubaud et al. (2023), the SMOTE-NC (Synthetic Minority Oversampling Technique for Nominal and Continuous) (Chawla et al. 2011) algorithm was identified as the best strategy. However, since the smaller classes only contain a few samples, there were stability issues with the algorithm. Moreover, setting all classes to the size of the larger class leads to a massive training dataset, which results in longer training times. Therefore, to rebalance the training set, the larger class is randomly under-sampled to half its size, all classes are randomly over-sampled to have at least 1000 samples, and finally, all classes are over-sampled using SMOTE-NC to reach the new size of the largest class. The threshold

of 1000 was chosen empirically to ensure stability in SMOTE-NC oversampling, addressing potential exceptions in the process. This was implemented using the Python library imblearn.¹

3.1.3. Machine learning inference

For this third step, all the attributes from all sources are concatenated. A machine learning algorithm is trained on the training set, and then applied to predict the samples in the test set. In Cubaud et al. (2023), the Gradient Boosted Trees (XGBoost) (Chen and Guestrin 2016) algorithm obtained the best results for predicting three LU classes. Thus, this algorithm was chosen for this work, with the same hyperparameter values, i.e. $\text{learning_rate} = 0.1$ and $\text{n_estimators} = 1000$, and the other left to their default value in the XGBoost² Python library.

3.1.4. Evaluation of the method

To evaluate the quality of the method, the predicted LU class of the polygons in the test set are compared to their LU class in the ground truth. The confusion matrix M is constructed, where M_{ij} represent the number of elements of class i in the ground truth that are predicted in class j . A per class quality assessment is given by the Producer Accuracy (PA, also referred to as recall) and the User Accuracy (UA, also referred to as precision). These two accuracies allow the per class F1-score to be constructed as follows:

$$PA_i = \frac{M_{ii}}{\sum_{j=1}^c M_{ij}}, \quad UA_i = \frac{M_{ii}}{\sum_{j=1}^c M_{ji}}, \quad F1_i = 2 \frac{PA_i \times UA_i}{PA_i + UA_i}, \quad (1)$$

with c the number of classes. Global quality assessment is given by the Overall Accuracy (OA) and the macro-mean F1 score (mF1), defined by the following formulas:

$$OA = \frac{\sum_{i=1}^c M_{ii}}{\sum_{i,j=1}^c M_{ij}}, \quad mF1 = \frac{1}{c} \sum_{i=1}^c F1_i \quad (2)$$

OA shows the proportion of well predicted samples independently of their classes, but due to the strong class imbalance, it may still give a high score even if the model always predicts the majority class. On the other side, mF1 assesses if each class is accurately predicted.

3.2. Data sources and constructed attributes

Ten data sources are used and 127 defined attributes are derived from them. This number of attributes is doubled by the neighboring attributes, thus, a total of 254 attributes. Links to the data sources are given in the Data availability statement section. Some attributes convey explicit information about LU, such as building functions, while others may convey implicit information that indirectly correlates with LU, like the area of the LU polygon. Tables 2–8 list the constructed attributes.

3.2.1. LU polygons geometry

First, 25 attributes are defined to characterize the shape of the LU polygons, thus giving implicit information about LU.

These attributes, listed in Table 2 are the area of the polygon, its convexity, its compactness, its elongation, the number of holes inside the polygon, and its polygonal signature. The polygonal signature is a function that allows to compare the shape of two polygons by assigning for each point of the polygon its distance to the center (Méneroux et al. 2022). To ensure scale invariance, this signature is normalized by the polygon's perimeter. The signature is then sampled into 20 points to construct 20 attributes. The sampling is started by the closest point to the center.

Table 2. The 25 attributes derived from the geometry of the LU polygon.

Name	Description	Formula
Area	Area of the polygon.	
Compactness	Compares the shape to a circle.	$\frac{4\pi \text{area}(\text{polygon})}{\text{perimeter}(\text{polygon})^2}$
Convexity	Measures the regularity of the polygon.	$\frac{\text{area}(\text{polygon})}{\text{area}(\text{convex_hull}(\text{polygon}))}$
Elongation	Measures how stretched the polygon is.	$\frac{\text{length}(\text{OrientedBoundingBox}(\text{polygon}))}{\text{width}(\text{OrientedBoundingBox}(\text{polygon}))}$
Number of holes	Number of holes inside the LU polygon.	
Signature _n (for n in [0, . . . , 19])	Regular sampling of the polygonal signature.	

Table 3. The 8 attributes derived from the radiometry (BD ORTHO).

Name	For x in	Description
Mean _x	{Blue, Green, Red, Near-infrared}	Mean value of the x channel.
Std _x	{Blue, Green, Red, Near-infrared}	Standard deviation of the x channel.

Table 4. The 2 attributes derived from Land Cover maps (CLC and OSO).

Name	Description
CLC	Majority CLC Land Cover class within the LU polygon.
OSO	Majority OSO Land Cover class within the LU polygon.

Table 5. The 31 attributes derived from BD TOPO and RPG.

Name	For x in:	Description
BD TOPO Buildings (17 attributes)		
x buildings area	{agricultural, annex, commercial and services, industrial, religious, residential, sports, undifferentiated}	Area of x buildings intersected with the LU polygon.
x buildings number	{agricultural, annex, commercial and services, industrial, religious, residential, sports, undifferentiated}	Number of x buildings intersected with the LU polygon.
Buildings height		Average height of buildings intersected with the LU polygon.
BD TOPO Other (13 attributes)		
x activity area fraction	{LU1.1, LU1.3, LU1.4, LU2, LU3, LU4.3}	Fraction of the LU polygon covered by x activity areas.
x polygons fraction	{Hydrography, Aerodrome, Cemetery}	Fraction of the LU polygon covered by x polygons.
x length	{roads, railways}	Length of x inside the LU polygon
Number of ERP		Number of Establishments Receiving the Public within the LU polygon.
Nature of Hydrographic surface		Ordinal encoded NATURE class of 'Hydrography' polygons.
RPG (1 attribute)		
RPG area		Area of agricultural terrain from the RPG intersected with the LU polygon.

Table 6. The 6 attributes derived from statistical data (INSEE).

Name	Description
Municipality population	Population of the municipality in which the LU polygon is located.
IRIS density	Population density of the IRIS in which the LU polygon is located.
Type of IRIS	Type of the IRIS in which the LU polygon is located.
Number of inhabitant	Sum of the number of individuals from INSEE 200 m ² gridded data intersected with the LU polygon.
Number of social housing	Sum of the number of social housing from INSEE 200 m ² gridded data intersected with the LU polygon.
Estimated living standards	Sum of the living standards from INSEE 200 m ² gridded data intersected with the LU polygon.

Table 7. The 18 attributes derived from Land Files.

Name	For x in:	Description
x Area	{LU1.1, LU1.2, LU1.3, LU2, LU3, LU4.1.1, LU4.1.2, LU4.1.3, LU4.1.4, LU4.1.5, LU4.2, LU4.3, LU5, LU6.1, LU6.2, LU6.3, LU6.6}	Area of x Land Use within the LU polygon according to Land Files.
Main LU		Majority Land Use according to Land files within the LU polygon.

Table 8. The 36 attributes derived from OSM.

Name	For x i:	Description
OSM Landuse polygons x area	{LU1.1, LU2, LU3, LU5}	Area of x OSM Landuse polygons intersecting the LU polygon.
OSM other polygons x area	{LU1.1, LU1.2, LU2, LU3, LU4.1.1, LU4.1.2, LU4.1.3, LU4.3, LU5, LU6.1, LU6.2}	Area of x OSM other polygons intersecting the LU polygon.
OSM other polygons x number	{LU1.1, LU1.2, LU2, LU3, LU4.1.1, LU4.1.2, LU4.1.3, LU4.3, LU5, LU6.1, LU6.2}	Number of x OSM other polygons intersecting the LU polygon.
OSM points x number	{LU1.1, LU2, LU3, LU4.1.1, LU4.1.2, LU4.1.3, LU5}	Number of x OSM points within the LU polygon.
OSM lines x length	{LU4.1.1, LU4.1.2, LU4.1.3}	Number of x OSM lines (ways) intersecting the LU polygon.

3.2.2. Optical imagery

The BD ORTHO is the 20 cm resolution open-license French national reference ortho-image database, with a planimetric accuracy of 80 cm. Eight attributes, providing implicit information, are computed from BD ORTHO for each LU polygon: mean and standard deviation of the blue, green, red and near infrared channels, as shown in [Table 3](#).

3.2.3. Land cover

The majority Land Cover class within the LU polygon of two LC maps are used as two attributes, shown in [Table 4](#), providing implicit information about LU. These two maps are CORINE Land Cover (CLC) and OSO (Inglada, Vincent, and Thierion 2019), a yearly open-license Land Cover map of France generated using Sentinel data. CLC is a vector map, while OSO is raster. OSO has a minimal mapping unit of 100 m² while CLC's is 10,000 m². According to their specifications, both maps have a thematic accuracy of approximately 85%. In Cubaud et al. (2023), the OCS GE LC map was also considered, but it has been identified as detrimental by the study.

3.2.4. Authoritative database

The BD TOPO is the open-license French reference vector topographic database, with a planimetric accuracy of 2.5 m. The attributes are separated into two categories: those defined from its 'building' layer (hereafter referred to as BD TOPO Building) and several other layers (hereafter referred to as BD TOPO Other). BD TOPO Building indicates the main function for each building, classified into eight defined functions. The area and the number for each of the eight building functions of the buildings intersected by the LU polygon are used as attributes (explicit information). In addition, the building height is also used as an attribute (implicit information).

BD TOPO Other contains the following layers:

- The 'Area of activity or interest' layer describes places with a specific economic activity. The categories have been mapped to six LU classes, and for each class the intersected area between the activity area and the LU polygon is used as an attribute.
- The 'Establishment Receiving the Public' (ERP) layer contains Point Of Interest (POI) locations. The number of ERP within the LU polygon is used as an attribute.

- The ‘Hydrography’ layer represents water areas in the form of polygons. After a filtering step based on the area (minimum area equal to 200 m²), the persistency (set to true) and the nature (not in {Hyporheic flow, Glacier, névé, Phreatic flow, Karst flow}) of the water area, the proportion of the area of the LU polygon covered with water is computed as an attribute.
- The ‘Aerodrome’ layer is represented by polygons. The proportion of the area of the LU polygon covered by an airfield is computed as an attribute.
- The ‘Cemetery’ layer is represented by polygons. The proportion of the area of the LU polygon covered by a cemetery is computed as an attribute.
- The ‘Road section’ layer is described by polylines. After removing roads under tunnels and roads under construction, the length of roads within the LU polygon is computed as an attribute.
- The ‘Railway section’ layer is described by polylines. After removing railways abandoned or under tunnels, the length of railways within the LU polygon is computed as an attribute.

The Graphic Parcel Register (RPG) is the reference database of agricultural parcels for the European Common Agricultural Policy. The area of agricultural parcels intersected by the LU polygon is used as an attribute. [Table 5](#) lists the 31 attributes derived from an authoritative database.

3.2.5. Statistical data

The French National Institute of Statistics and Economic Studies (INSEE) provides several statistical products. They are used to derive attributes with implicit information about LU. Six attributes are constructed at different scales and presented in [Table 6](#). At the municipal scale, the population of the municipality in which the LU polygon belongs is used as an attribute. At the IRIS scale (statistical subdivision of the municipality), the population density of IRIS and its type (i.e. housing, business and other) give two attributes. At the scale of a 200 m square grid, the population, the estimated living standards and the number of social housing units are used as attributes.

3.2.6. Land files

For each cadastral parcel, the method described in Rutkowski et al. (2017) was applied to obtain an area per Land Use class by combining several indicators from the Land Files. We derived 18 attributes: the areas for each LU class in the Land Files, as well as the LU class name with the highest area (explicit information). This data source has some imperfections because it relies on declarative data and doesn’t include public institutions which are not paying taxes. The 18 attributes are presented in [Table 7](#).

3.2.7. VGI databases

OpenStreetMap (OSM)³ is the largest project of collaborative mapping worldwide. Users can create polygons, lines, and points and assign tags to them, which are constructed as pairs of (key = value). Based on the pairs found in the study areas, those described in the OSM wiki⁴ and the most common combinations found in the tag info website,⁵ we created a mapping from the OSM tags with explicit LU information to the 20 LU classes of our nomenclature. Some conditions may be required. For example, if an entity has the key building:use, it means that the value of the building key represents the original use for which the building was created. We then defined as attributes for each LU class the number and area of OSM polygons intersected by the LU polygon, the length of OSM lines intersected by the LU polygon, and the number of points within the LU polygon. After removing the attributes that are always empty, 36 attributes are left using this data source and are presented in [Table 8](#).

4. Experimental protocol

This section describes the study areas on which the method described in the Section 3 is applied, and the experiments made to assess the transferability of the method.

4.1. Study areas

4.1.1. Description

In order to assess the transferability of our methodology, we chose to work on two French departments. The department of Gers (6257 km²), shown in [Figure 2](#), is located in the South-West of France, and is mostly rural (with a population density of 31 inhabitants/km²) with some small towns (only 23,000 inhabitants in the main town Auch).

The department of Rhône (3259 km²), shown in [Figure 3](#), is located in the East of France, and includes the second-largest urban area in France (1.4 M inhabitants in Grand Lyon), but also encompasses some rural regions, with low-altitude mountains (Monts du Beaujolais et du Lyonnais) in the western part. Population density for the whole department is 583 inhabitants/km², but reaches 10,910 inhabitants/km² in Lyon. For both departments, rural spaces can be visualized in [Figures 2 and 3](#) as the areas dominated by LU1.1 (Agriculture) or LU1.2 (Forestry) activities. Overall, these two study areas present a great diversity of landscapes and are highly different from one another.

The OCS GE (OCcupation des Sols à Grande Echelle) dataset is used as the ground truth. It is an authoritative LULC map produced by IGN (The French National Mapping Agency), with separate LU and LC nomenclature. Non-overlapping polygons are derived from the Land Files, the BD TOPO and image segmentation. A rule-based method automatically assigns to each polygon a unique LU class and a unique LC class. It is followed by an intensive manual correction stage because the rule-based method gives low accuracy. Indeed, when the rule-based version is evaluated against the corrected version, it obtains an OA of 67% and an mF1 of 49%. The edition of the OCS GE used for the Gers dates from 2019 and that for the Rhône from 2020. [Figures 2 and 3](#) show this ground truth. The Gers dataset is composed of 230,637 LU polygons and the Rhône dataset of 386,802 LU polygons.

As some classes have a very small amount of samples, or even no samples at all, we chose to merge them into the same class LU6 (Others). We also removed from the training and test sets

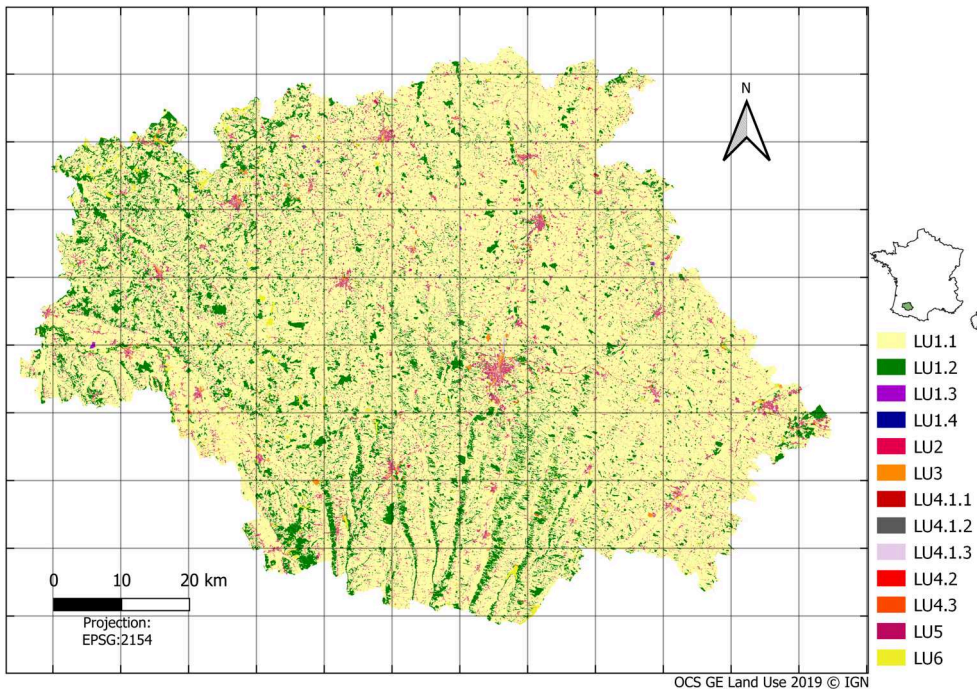


Figure 2. Ground truth for the department of Gers: agricultural and farming Landscapes with a network of small towns.

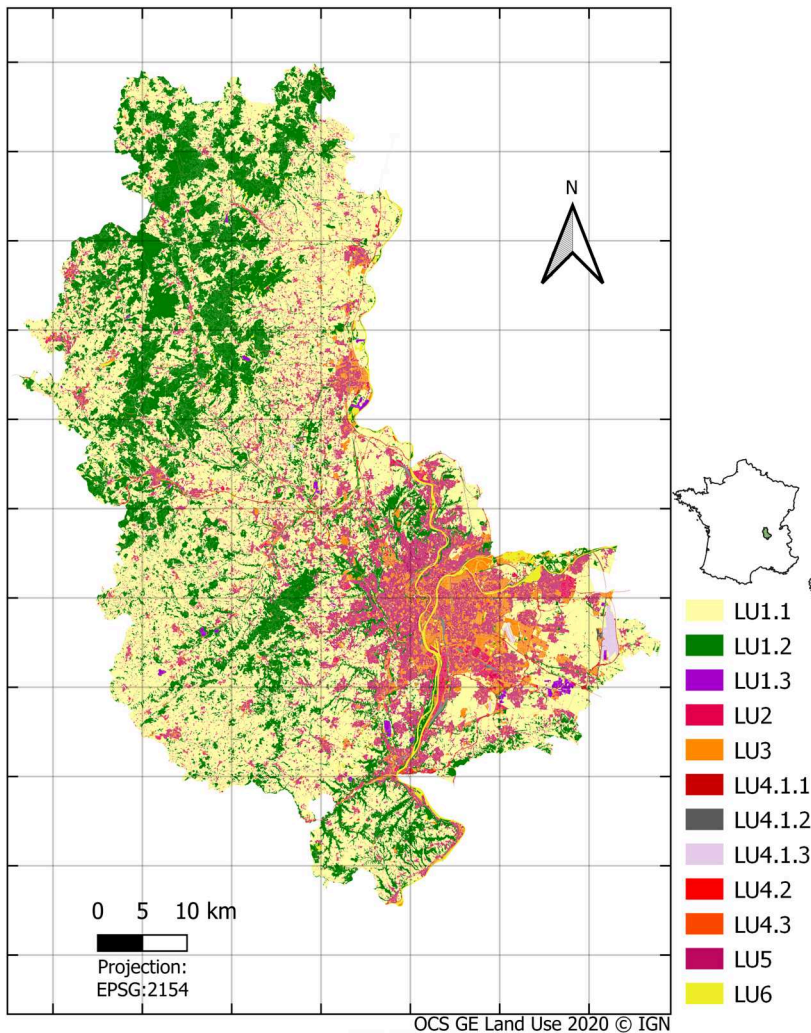


Figure 3. Ground truth for the department of Rhône: a large urban area along the Rhône valley and low-altitude mountains in the west part.

all samples representing the class LU235, which is supposed to represent mixed LU but rather represents an unknown use among industrial, commercial or residential uses. The classes removed or merged with LU6 are written in gray in [Table 1](#).

4.1.2. Differences of distribution of the data sources between the two areas

The attributes from the data sources have different distributions between the two areas, which is a challenge for the transferability process. These differences are due to several causes. First, as the two areas are very different, so is their distribution of LU classes, as shown in the [Figure 4](#). This has obviously an impact on the distribution of the computed attributes. Second, geographical differences may also exist, for instance regarding the plant species and crop types, or the material used in construction. For example, bricks are commonly used in Gers and rarely found in Rhône. Third, differences arise from the varying degree of completeness of the data sources. It is the case for OSM data, where the number of OSM contributors is known to vary depending on how populated and urban the area is (Zhou et al. [2022](#)). Fourth, there are

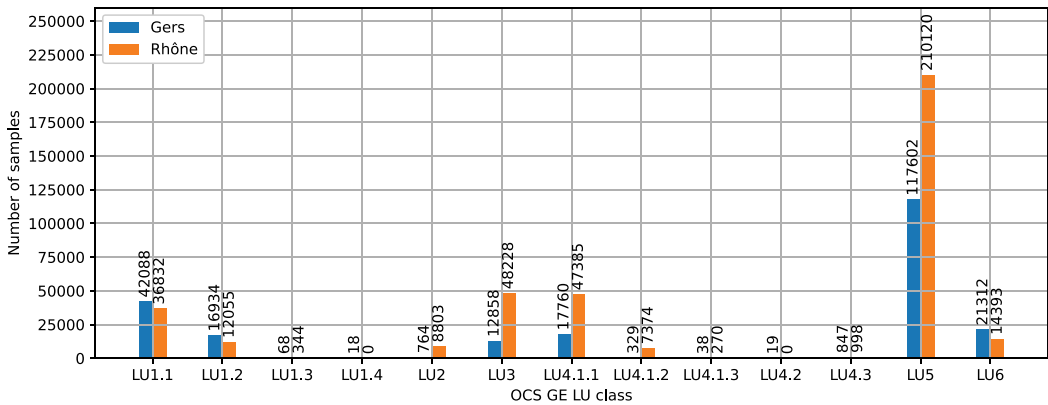


Figure 4. Distribution of the LU classes in the ground truth of each study area.

temporal gaps in the data collection of the data sources. For instance, seasonal differences in aerial images (e.g. changes in vegetation cover between winter and summer), or changes in the specification for a data source between the two dates (e.g. a change in the minimal mapping unit, or in the class definitions) could have affected the attribute distributions. However, this is not the case here, and temporal differences have very little responsibility for the difference in distributions.

As the distributions of most attributes do not follow a normal law, we chose to use the Wasserstein distance to quantify these differences, since it is frequently employed to compare datasets and distributions. It indeed considers the shape of the distributions rather than assuming a specific form. In a first step, we standardized each non-categorical attribute with the same values of mean and standard deviation for the two study areas. Then we computed the Wasserstein distance between the two distributions. The Wasserstein distance between two distributions X and Y can be expressed as:

$$W(X, Y) = \inf_{\gamma \in \Gamma(X, Y)} \int_{X \times Y} \|x - y\| d\gamma(x, y), \quad (3)$$

where $\Gamma(X, Y)$ is the set of all joint distributions with marginals X and Y . It thus represents how distant the two distributions are, and the values are expressed as the number of multiples of the initial standard deviation due to the normalization. A Wasserstein distance of 0 indicates a perfect similarity between the distributions, whereas biggest values indicate dissimilarity between them. Finally, the distances are averaged by data source. Thanks to the normalization, the obtained mean distances can be meaningfully compared. The results are presented in Figure 5. As explained, the statistical data from INSEE has the most distant distributions, because one study area is mostly rural and the other mostly urban. The Wasserstein distance value of 2.5 can be considered as quite high in this context. It is followed by OSM due to the difference in completeness between the two areas. In BD Ortho, radiometric corrections, homogenization, and processing were performed independently for each department, which explains the difference in the distributions of the radiometry data source. The differences in RPG are due to the fact that there are more fields in the Gers department, and they tend to have a larger surface area (on average 58,430 m²) than in the Rhône department (on average 24,923 m²).

4.2. Experiment

Three main types of experiments are described in this paper. First, the methodology is applied to each study area independently (Section 5.1). Second, the transferability of the model is considered (Section 5.2), i.e. the model is trained on one study area and evaluated on the other. Figure 6 illustrates these two types of experiments. However, for this transferability, the model must be able to deal with the differences of distribution between the two study areas, as illustrated in Section 4.1.2.

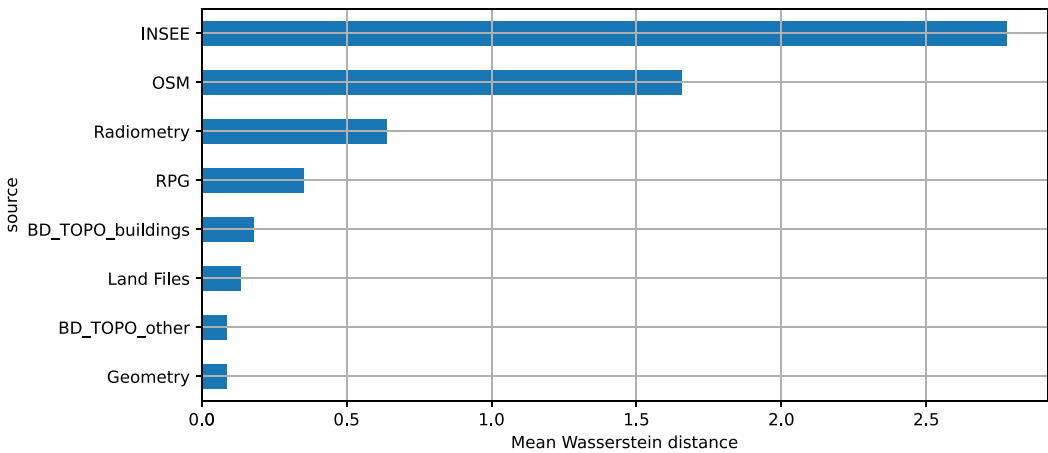


Figure 5. Mean Wasserstein distance of the distribution of non-categorical attributes between Rhône and Gers.

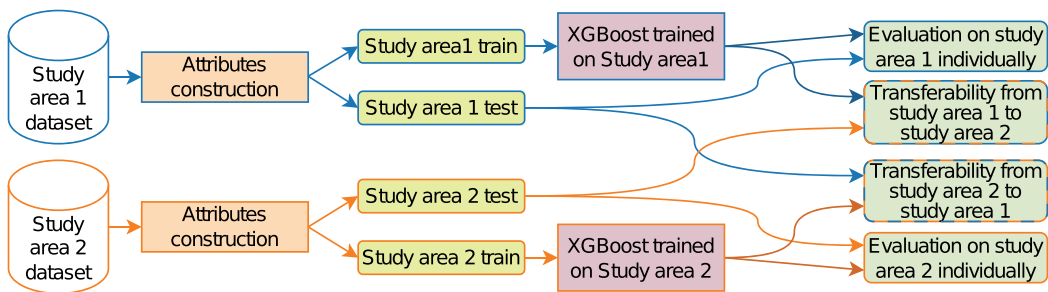


Figure 6. Evaluation on each study area and transferability assessment process.

Another difficulty is that some classes are present only in one study area. Thus, a model trained on the other study area is not able to predict them, and inversely a model trained on this study area may predict them, although they are not present in the test study area. Third, to overcome these difficulties, we applied the approach of Zhang et al. (2023) by integrating samples of the target area in the training set and study the influence of it in Section 5.3. This last experiment is described in Figure 7. The use of a domain adaptation model has not been attempted because existing models would have encountered difficulties. In fact, some attributes are continuous, while others are categorical. Moreover, depending on the attribute, the difference in distributions between the two study areas may have different origins. Each origin would therefore require a different approach.

5. Results

First, this section will present the results obtained from training and evaluating in each study area separately. Then, the transferability results from one study area to the other will be considered. Finally, the results when the training set combines both study areas in varying proportions will be examined.

5.1. Results for each individual study area

In the Gers study area, an OA of 88% and a mF1 of 83% were obtained, while in Rhône, the OA was equal to 92% and the mF1 to 86%. These scores are better than the original rule-based process, justifying the use of a machine learning process. Figures 8 and 9 show the per class metrics obtained for

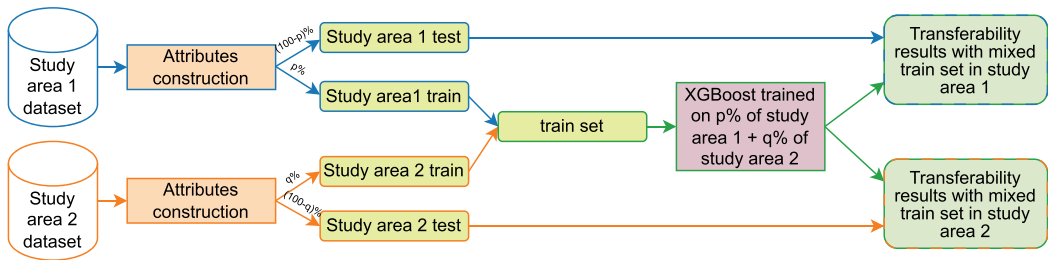


Figure 7. Transferability with a part of the test study area included in the training dataset.

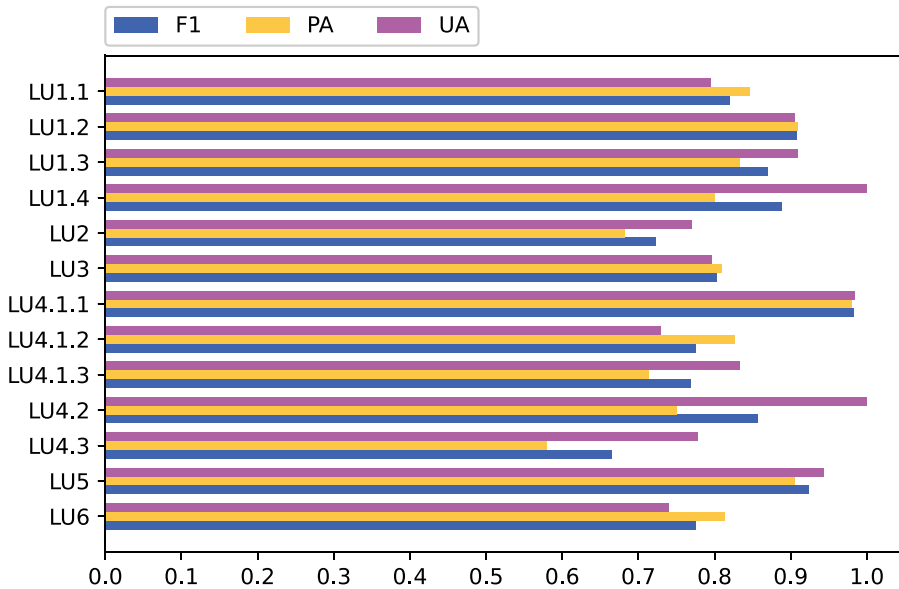


Figure 8. Per class metrics for the individual prediction of Gers.

Gers and Rhône respectively. Most of the classes are well predicted, as the lowest F1 score is 61% for LU4.3 (public utility networks) in Rhône. The classes LU4.1.1 (road network), with a F1 score of 98% in Gers and in Rhône, obtained the best classification results. It is followed by LU5 (residential), with a F1 score of 92% in Gers and 96% in Rhône, and LU1.2 (forestry), with a F1 score of 91% in Gers and 94% in Rhône.

5.2. Results with the model trained on a study area and applied to the other

When the model is trained on Gers and evaluated on Rhône, the OA is 79% and the mF1 score is 50%, whereas when training on Rhône and evaluating on Gers, the OA is 76% and the mF1 is 51%. Figures 10 and 11 show the per class results for Gers to Rhône and Rhône to Gers respectively. In Figure 10, the classes LU1.4 and LU4.2 are not present in the Rhône dataset, but the model learned them and predicted them. Inversely, in Figure 11, the model did not encounter these classes during training, resulting in its inability to predict them. This explains the low mF1 obtained.

5.3. Influence of including a part of the test study area on the train dataset

This subsection presents the results when the training set includes both study areas in different proportions. Tables 9 and 10 give the mF1 score when the model is evaluated on the rest of the Gers and

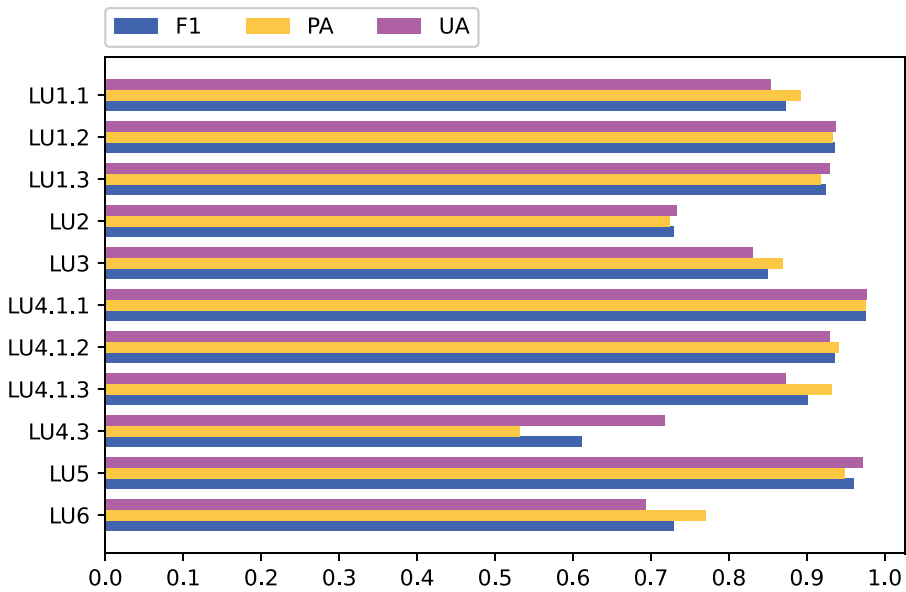


Figure 9. Per class metrics for the individual prediction of Rhône.

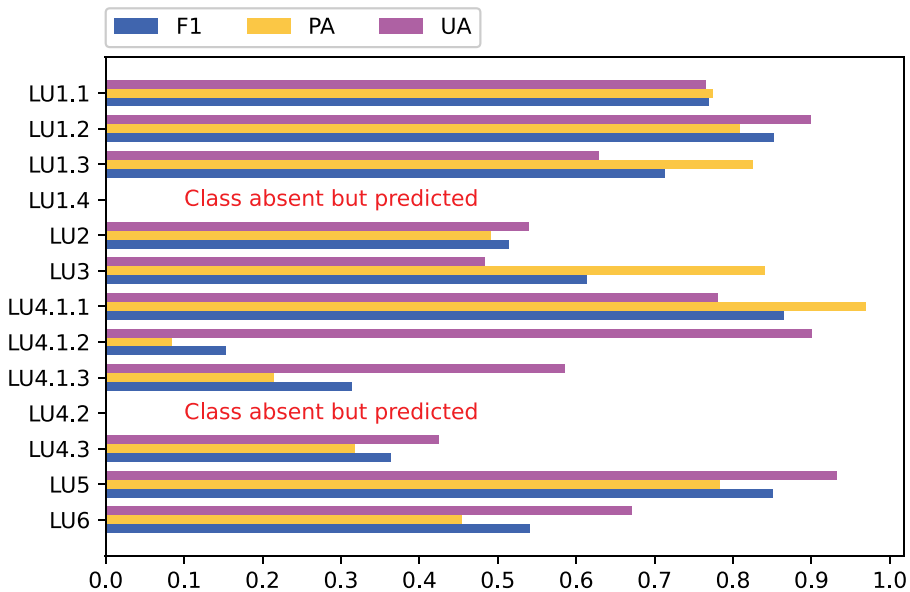


Figure 10. Per class metrics for the transferability from Gers to Rhône.

Rhône dataset, respectively. For instance, the cell (0.05, 0.10) of Table 9 represents the experiment where the training set is composed of 5% of the total Rhône dataset (i.e. 19,340 LU polygons) and 10% of the total Gers dataset (i.e. 23,064 LU polygons), and the training set of the remaining polygons from the Gers dataset. For both tables, the first row corresponds to the results using only one study area, with a varying size of the training set. The results are notable, compared to the transferability results, even with a small training dataset, and the performances globally improve as the size of the training set increases. The first column in both tables corresponds to the transferability

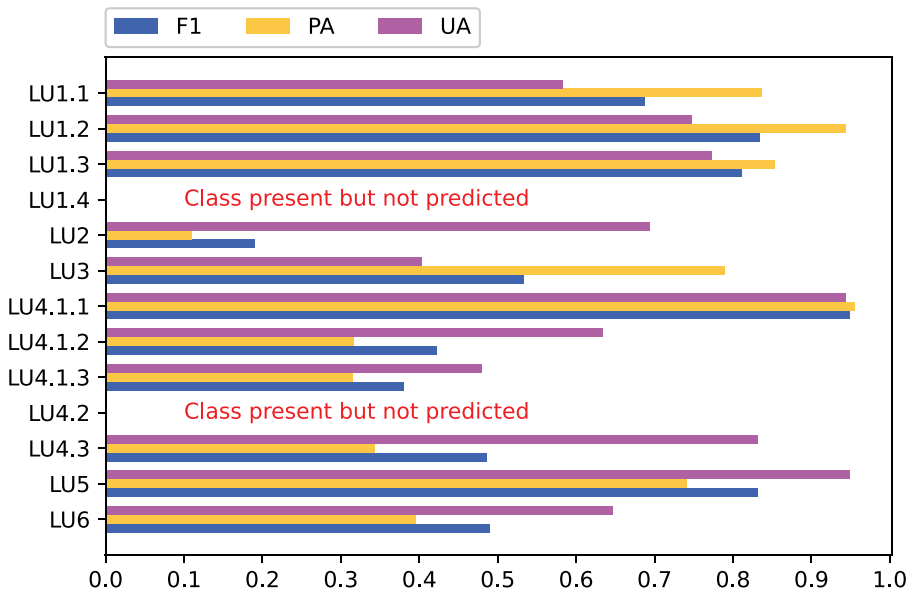


Figure 11. Per class metrics for the transferability from Rhône to Gers.

assessment. The variation of mF1 is smaller here, and the best values are obtained for 10% of the Rhône dataset in the training set in Table 9 and for 20% of the Gers dataset in the training set in Table 10. It can be explained by the fact that with more training data, the model becomes too specific to one study area which may hinder its ability to generalize effectively. Finally, by adding samples from the test study area, even a small amount, the results systematically improve. Overall, the performance consistently improves as more test study area samples are added.

6. Discussion

In this section, we discuss the results obtained, with particular attention to three relevant aspects: the contributions and the relevance of the data sources and the challenges encountered in LU classification.

Table 9. mF1 score obtained with mixed trained set when evaluated on the rest of the Gers dataset.

		Proportion of Gers dataset in training set					0.80
		0.00	0.05	0.10	0.20	0.50	
Proportion of Rhône	0.00		0.60	0.66	0.73	0.81	0.83
	0.05	0.47	0.71	0.71	0.73	0.80	0.80
	0.10	0.52	0.68	0.70	0.73	0.80	0.82
	0.20	0.49	0.69	0.71	0.77	0.78	0.81
	0.50	0.48	0.71	0.65	0.73	0.79	0.81
	0.80	0.51	0.70	0.70	0.73	0.82	0.84

Table 10. mF1 score obtained with mixed trained set when evaluated on the rest of the Rhône dataset.

		Proportion of the Rhône dataset in training set					0.80
		0.00	0.05	0.10	0.20	0.50	
Proportion of Gers	0.00		0.77	0.75	0.77	0.82	0.86
	0.05	0.49	0.72	0.80	0.82	0.84	0.85
	0.10	0.49	0.78	0.80	0.82	0.85	0.86
	0.20	0.52	0.79	0.81	0.82	0.84	0.86
	0.50	0.51	0.72	0.74	0.82	0.84	0.86
	0.80	0.50	0.79	0.80	0.83	0.84	0.83

6.1. Contributions of the sources for each individual study area

Studying the contributions of the data sources enable to quantify their impact on the classification results and to assert that these sources are indeed useful to take as input for the model. This identification of valuable sources is likely to be relevant for other study areas, as the following results are consistent on these 2 heterogeneous study areas. It is also likely to be relevant for similar nomenclatures, but the importance of each source may vary depending on the classes. Some specific data sources may not be available in other places, but equivalent local sources may exist (e.g. National topographic databases). The used sources are however available on the whole French territory.

The contribution of each source is measured by two indicators: first the score obtained when the workflow is applied to this source only, and second the score obtained when the workflow is applied to all the sources except this one, which corresponds to the LOCO (Leave-One-Covariate-Out) metrics (Lei et al. 2018).

Figures 12 and 13 show the mF1 score obtained when the XGBoost algorithm is trained using only the attributes of one source for Gers and Rhône respectively. For instance, when using BD TOPO building the mF1 scores are 0.35 and 0.44, for Gers and Rhône, whereas for Land Files, they are equal to 0.49 and 0.55. In comparison to the mF1 achieved when all sources are used, the difference is significant (around 0.32) and is the same whatever the study area. As each source obtained a lower mF1 than the mF1 obtained by all the sources, it justifies the use of multiple sources and the data fusion process. The score using OCS GE LC data source is indicated in the red row (OCS GE LC is not in All sources), given that this source has been previously demonstrated to be inefficient, as mentioned in Section 3.2.3.

Figures 14 and 15 show the mF1 score of all the sources minus the mF1 obtained when the XGBoost algorithm is trained using all the attributes except the attributes of one source, for Gers and Rhône respectively. It thus corresponds to the score lost without this source. In both study areas, the most important source is BD TOPO Other, which especially improves the results for the classes LU1.4, LU4.1.3, LU1.3, LU4.3 and LU2. It is noteworthy that when this source is used independently (see Figures 12 and 13) the mF1 scores are low (0.07 for Gers and 0.44 for Rhône). However, its integration helps enhancing the performance when it is combined with others, showing once more the relevance of multi-source leveraging. Furthermore, the difference

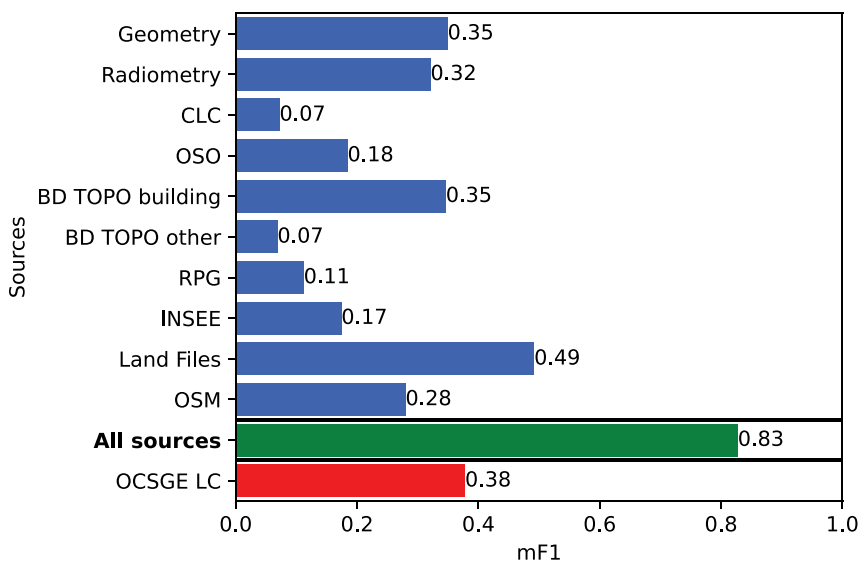


Figure 12. mF1 score obtained when the XGBoost algorithm is trained using only one source in Gers.

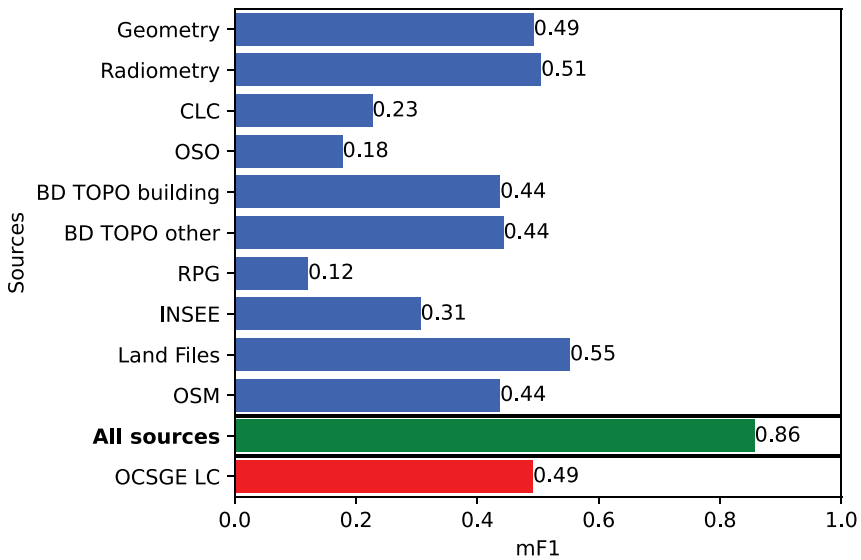


Figure 13. mF1 score obtained when the XGBoost algorithm is trained using only one source in Rhône.

of magnitude of the mF1 loss between the two study areas (in range (0–0.2) in Gers and (0–0.05) in Rhône) is intriguing, especially since the mF1 scores with all sources are similar. It could mean that the sources are more complementary in Gers and more redundant in Rhône. Moreover, [Figures 14](#) and [15](#) also reveal that adding the source OCS GE LC is again detrimental to the classification results. Indeed, adding this source results in a loss of score. Finally, removing neighboring attributes results in a loss of mF1 of 0.04 in Gers and Rhône, demonstrating that these attributes are contributing to the quality of the classification.

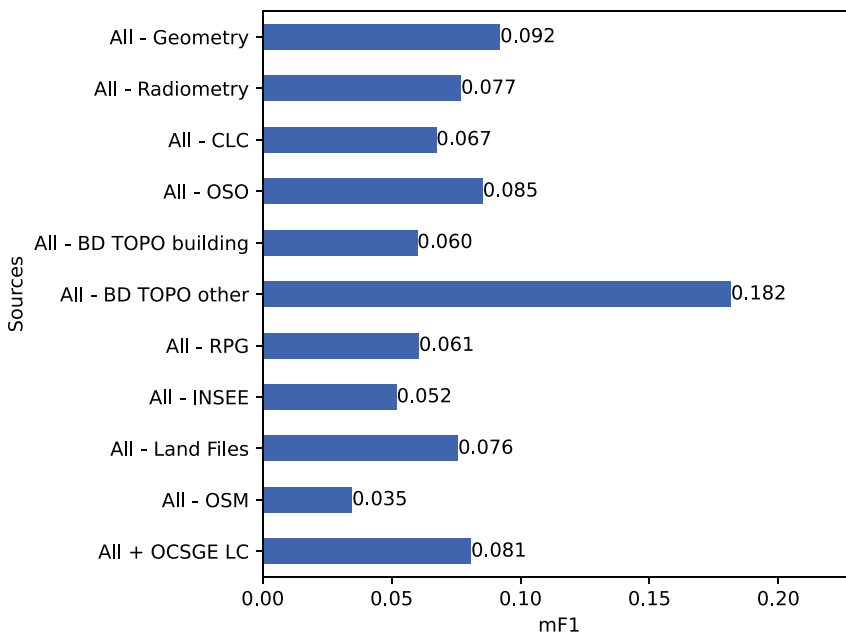


Figure 14. mF1 score lost when the XGBoost algorithm is trained using all sources except one in Gers. Adding the OCSGE LC source is detrimental as it results in a score loss.

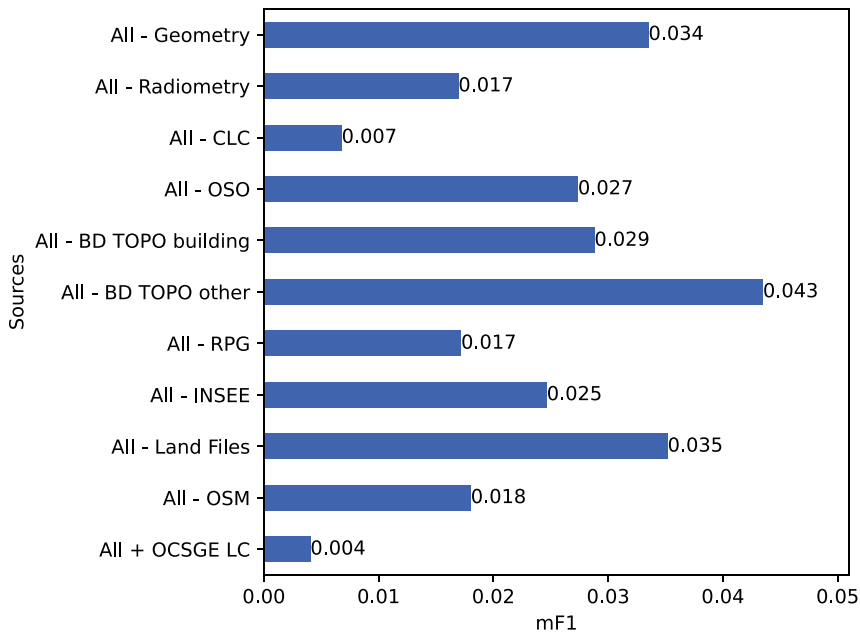


Figure 15. mF1 score lost when the XGBoost algorithm is trained using all sources except one in Rhône. Adding the OCSGE LC source is detrimental as it results in a score loss.

6.2. Importance of the sources, with the model trained on a study area and applied on the other

The comparison of the importance of the sources between the direct and the transferability processes highlights how differences in distributions affect transferability. While one source may be useful for the prediction of an individual study area, it could prove detrimental for transferability. Figures 16 and 17 present the mF1 score lost when using all attributes except those from one source, compared to using all sources. It reveals that several sources obtain a negative score, meaning that transferability results are better without them. These sources vary between the two case studies. Still, BD TOPO Buildings and Other are consistently beneficial in both study areas. However, the reported differences in performance for a specific data source are subject to variability, and the small magnitude of the importance score might be within the range of variability.

6.3. Analysis of the difficulty of classification of the LU classes

To investigate why some classes received lower classification scores, several aspects are considered. First, the influence of the class sizes on the performance is studied using the Pearson correlation between the size of each class and the metric value (F1, UA or PA) for this class. For each study area individually, this correlation is not significant. For instance, the p -value of the Pearson correlation between F1 and class size is 0.42 in Gers and 0.4 in Rhône. This is due to the balancing process using the SMOTE NC algorithm that allows smaller classes to be also well classified. However, in terms of transferability, i.e. when the model is trained on one study area and evaluated on the other, it appears that class size has a significant influence on the results. Thus, from Gers to Rhône, the Pearson correlation between F1 and class size is 0.78 with a p -value of $1.5e-3$, and 0.72 with a p -value of $5.1e-3$ for Rhône to Gers. It can be explained by the fact that for smaller classes, the model precisely learns their characteristics in one study area. It is thus not able to generalize, leading to poor classification performances for these classes.

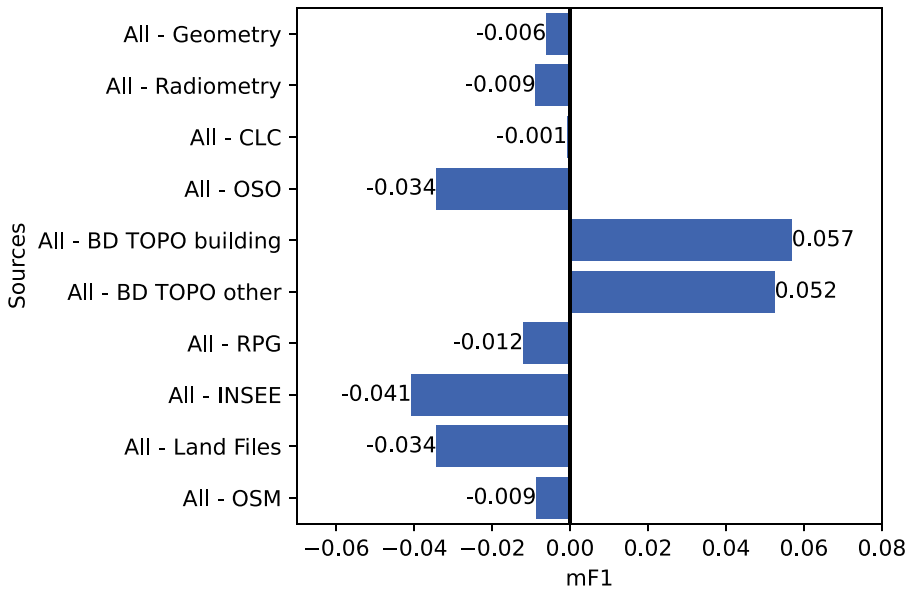


Figure 16. mF1 score lost when the XGBoost algorithm is trained using all sources except one, from Gers to Rhône.

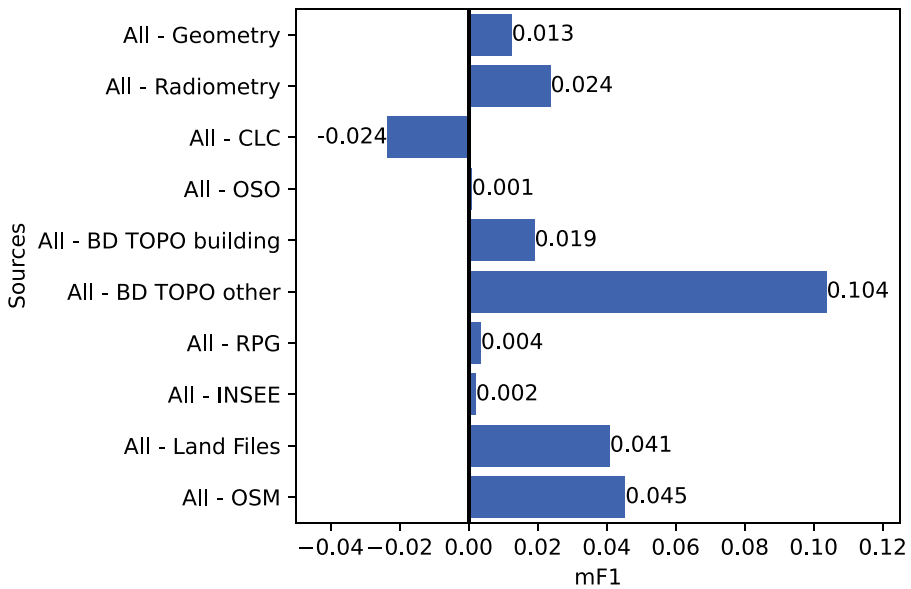


Figure 17. mF1 score lost when the XGBoost algorithm is trained using all sources except one, from Rhône to Gers.

Besides class size, the capacity of the model to accurately predict the classes is also closely tied to the attributes used as input by the model. On the one hand, some explicit attributes give hints that allow some classes to be identified most of the time. It is for instance the case for LU polygons from the class LU4.1.1 (Road Transport) that can usually be identified by the information given by the 'length of roads' attribute. On the other hand, some classes are harder to predict. They may encapsulate different aspects, for instance the class LU6 (Other). They may also be confused with other classes because both classes may have similar values for some attributes. For instance, a non-null 'number of ERP' attribute most often indicates a LU3 (commercial and services) polygon, but it

can sometimes be a workshop building considered as LU2, or be in a building that mostly has a residential (LU5) function. Moreover, sometimes the information required to identify some LU polygons belonging to a class may be missing. For instance, it can be the case in the OSM data source in which some classes that contributors are less interested in, such as agricultural, or have less access to may be less mapped (Zhou et al. 2022).

7. Conclusion

This paper addresses the complexities of land use (LU) classification and the associated challenges of model transferability. Our approach involves leveraging multi-source data coming from 10 data sources, to compute 254 attributes describing the land characteristics. In this paper, we use a National LU nomenclature defined in France (INSPIRE compliant) containing 13 LU classes and consider that the mapping units are already defined.

Several conclusions can be drawn from this study. First, the proposed data fusion machine learning framework shows its ability to map Land Use in two very different study areas, with a mF1 score of 83 and 86% for Gers and Rhône, respectively. However, when it comes to transferring the model from one study area to the other, the quality of the results degrades to a mF1 of 50 and 51% for Gers to Rhône and Rhône to Gers, respectively. Including a small amount of the target study area and limiting the overfitting is identified as a solution to improve these results. Moreover, studying the importance of the sources justifies the use of multiple sources. Nevertheless, some sources that provide useful information for predicting LU in one study area may vary too much from one study area to the other, and thus be detrimental to the transferability process. Therefore, the quality and heterogeneity of data sources play a significant role in performing transferability.

In terms of future work, many directions can be defined. First, the transferability results may be improved by exploiting domain adaptation learning techniques, such as mitigating the difference of distribution or learning a common representation between the two study areas. Transfer learning methodologies, where a trained model is fine-tuned for another task, can also be explored to improve transferability.

Second, LU change and update remains an ongoing research area. In our paper, the transferability has been studied between two different areas, but applying a model from one date to another is also very interesting, as it allows LU maps to be updated in order to follow the evolution of the territory. Additional difficulties may arise from the necessity to modify the geometry between the two dates, the difference of update frequency of the data sources, the possible evolution of their specifications and some changes that can be corrections instead of real world use change.

Third, we focused in this paper on a specific LU nomenclature (OCS GE LU). The proposed method is likely to be adaptable to other LU nomenclatures with little to no modification, but with the necessity to train again the model with training data of this nomenclature. It may also be needed to modify or adjust some attributes. Another solution to obtain a map in another LU nomenclature is to translate the OCS GE LU output of the trained model, as discussed in Section 2.2.

Finally, future work will concentrate on automating the generation of mapping units. Currently, our work assumes that the mapping units are predefined. To address this, our first step would involve starting from cadastral data as a backbone and learning from the data sources how it should be cut to obtain a geometry with as few mix use area as possible.

Notes

1. <https://imbalanced-learn.org>
2. <https://xgboost.readthedocs.io>
3. <https://www.openstreetmap.org>
4. <https://wiki.openstreetmap.org/>
5. <https://taginfo.openstreetmap.org/>

Acknowledgments

The authors are grateful to their colleagues from IGN for providing access to the ground truth data, their insights into the initial Land Use and Land Cover (LULC) process proposed by IGN, and their valuable feedbacks on this work. We would like to thank the reviewers for their interesting comments and recommendations, which helped us to improve the article. We are also grateful to S. Chandora for checking the English spelling of the paper.

Data availability statement

The dataset including the LU polygons of both study areas and the constructed attribute is available at <https://doi.org/10.5281/zenodo.10462844> (Cubaud, Le Bris, and Jolivet 2024). BD ORTHO is available at <https://geoservices.ign.fr/bdortho>. CLC is available at <https://land.copernicus.eu/en/products/corine-land-cover>. OSO is available at <https://www.theia-land.fr/product/carte-doccupation-des-sols-de-la-france-metropolitaine/>. BD TOPO is available at <https://geoservices.ign.fr/bdtopo>. RPG is available at <https://geoservices.ign.fr/rpg>. INSEE's IRIS polygons are available at <https://geoservices.ign.fr/irisge>, while their statistical surveys are at <https://www.insee.fr/fr/statistiques/7704076>. INSEE grid data is available at <https://www.insee.fr/fr/statistiques/6215138?sommaire=6215217>. OSM data was downloaded using <https://download.openstreetmap.fr/extracts/europe/france/>. Land Files are not directly available due to protection of the privacy reasons.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

M. Cubaud  <http://orcid.org/0009-0007-3751-1946>

A. Le Bris  <http://orcid.org/0000-0002-1450-9161>

L. Jolivet  <http://orcid.org/0000-0002-3090-5076>

A.-M. Olteanu-Raimond  <http://orcid.org/0000-0002-1101-1333>

References

- Arnold, Stephan, Barbara Kosztra, Gebhard Banko, Geoff Smith, G. W. Hazeu, Michael Bock, and Nuria Valcarcel. 2013. "The EAGLE Concept – A Vision of a Future European Land Monitoring Framework." In *EARSeL Symposium Proceedings "Towards Horizon 2020"*. June 2013. at Matera, Italy
- Asuquo Enoch, Mfoniso, Richard Ebere Njoku, and Uzoma Chinenye Okeke. 2023. "Modeling and Mapping the Spatial-temporal Changes in Land Use and Land Cover in Lagos: A Dynamics for Building a Sustainable Urban City." *Advances in Space Research* 72 (3): 694–710. <https://doi.org/10.1016/j.asr.2022.07.042>.
- Baudoux, Luc, Jordi Inglada, and Clément Mallet. 2023. "Multi-Nomenclature, Multi-Resolution Joint Translation: An Application to Land-Cover Mapping." *International Journal of Geographical Information Science* 37 (2): 403–437. <https://doi.org/10.1080/13658816.2022.2120996>.
- Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2011. "SMOTE: Synthetic Minority Over-Sampling Technique." <https://doi.org/10.48550/ARXIV.1106.1813>.
- Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 2016, 785–794. ACM. ISBN: 978-1-4503-4232-2. <https://doi.org/10.1145/2939672.2939785>.
- chinesestandard.net. 2014. "GB 50137-2011 Translated English of Chinese Standard. GB50137-2011: Code for Classification of Urban Land Use and Planning Standards of Development Land", March 2014. <https://www.chinesestandard.net>.
- Cihlar, Josef, and Louisa J. M. Jansen. 2001. "From Land Cover to Land Use: A Methodology for Efficient Land Use Mapping Over Large Areas." *The Professional Geographer* 53 (2): 275–289. <https://doi.org/10.1080/00330124.2001.9628460>.
- Cubaud, M., A. Le Bris, L. Jolivet, and A.-M. Olteanu-Raimond. 2023. "Comparison of Two Data Fusion Approaches for Land Use Classification." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-1-W2-2023:699–706. <https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-699-2023>.

- Cubaud, M., A. Le Bris, L. Jolivet, and A.-M. Olteanu-Raimond. 2024. "OCS GE Land Use Dataset including the French Departments of Gers and Rhône." (dataset). Zenodo, January 2024. <https://doi.org/10.5281/zenodo.10462844>
- DAFF (Department of Agriculture and Water Resources). 2016. "Australian Land Use and Management Classification Version 8." Accessed October 2016. <https://www.agriculture.gov.au/abares/aclump/land-use/alum-classification>.
- Deng, Yingbin, Renrong Chen, Ji Yang, Yong Li, Hao Jiang, Wenyue Liao, and Meiwei Sun. 2022. "Identify Urban Building Functions with Multisource Data: A Case Study in Guangzhou, China." *International Journal of Geographical Information Science* 36 (10): 2060–2085. <https://doi.org/10.1080/13658816.2022.2046756>.
- EEA (European Environment Agency). 2023. "Net Land Take in Cities and Commuting Zones in Europe." March 2023. <https://www.eea.europa.eu/en/analysis/indicators/net-land-take-in-cities?activeAccordion=546a7c35-9188-4d23-94ee-005d97c26f2b>.
- Fonte, C. C., M. Minghini, V. Antoniou, J. Patriarca, and L. See. 2018. "Classification of Building Function Using Available Sources of VGI." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-4:209–215. <https://doi.org/10.5194/isprs-archives-XLII-4-209-2018>.
- García-Álvarez, David, and Sabina Florina Nanu. 2022. "Land Use Cover Datasets: A Review." In *Land Use Cover Datasets and Validation Tools: Validation Practices with QGIS*, edited by D. García-Álvarez, M. T. Camacho Olmedo, M. Paegelow and J. F. Mas, 47–66. Cham: Springer International Publishing. ISBN: 978-3-030-90998-7. https://doi.org/10.1007/978-3-030-90998-7_4.
- Garioud, Anatol, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sebastien Giordano, and Boris Wattlelos. 2023. "FLAIR : A Country-Scale Land Cover Semantic Segmentation Dataset from Multi-Source Optical Imagery." In *Advances in Neural Information Processing Systems*, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, Vol. 36, 16456–16482. New Orleans, USA: Curran Associates, Inc.
- Hall, D. L., and J. Llinas. 1997. "An Introduction to Multisensor Data Fusion." *Proceedings of the IEEE* 85 (1): 6–23. <https://doi.org/10.1109/5.554205>.
- Harrison, Andrew R. 2006. "National Land Use Database: Land Use and Land Cover Classification." Technical Report 05 CSR 03696. Office of the Deputy Prime Minister.
- He, Jialyu, Xia Li, Penghua Liu, Xinxin Wu, Jinbao Zhang, Dachuan Zhang, Xiaojuan Liu, and Yao Yao. 2021. "Accurate Estimation of the Proportion of Mixed Land Use at the Street-Block Level by Integrating High Spatial Resolution Images and Geospatial Big Data." *IEEE Transactions on Geoscience and Remote Sensing* 59 (8): 6357–6370. <https://doi.org/10.1109/TGRS.2020.3028622>.
- Hu, Junjie, Yong Gao, Xuechen Wang, and Yu Liu. 2023. "Recognizing Mixed Urban Functions From Human Activities Using Representation Learning Methods." *International Journal of Digital Earth* 16 (1): 289–307. <https://doi.org/10.1080/17538947.2023.2170482>.
- IGN (Institut National de l'Information Géographique et Forestière). 2022. "OCS GE Version 1.1 Occupation du Sol à Grande Échelle Descriptif de contenu." géoservices. November 2022. https://geoservices.ign.fr/sites/default/files/2022-11/DC_OCS_GE_1-1.pdf
- Inglada, Jordi, Arthur Vincent, and Vincent Thierion. 2019. "Theia OSO Land Cover Map 2019." March 2019. (dataset). Zenodo. <https://doi.org/10.5281/ZENODO.6538321>.
- Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113 (523): 1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>.
- Li, Jing, Haiyan Liu, Jia Li, Xiaohui Chen, and Zekun Tao. 2023. "A Knowledge-Based Approach for Estimating the Distribution of Urban Mixed Land Use." *International Journal of Digital Earth* 16 (1): 965–987. <https://doi.org/10.1080/17538947.2023.2184512>.
- Li, Zhijie, Ziyi Ma, and Guoyan Zhou. 2022. "Impact of Land Use Change on Habitat Quality and Regional Biodiversity Capacity: Temporal and Spatial Evolution and Prediction Analysis." *Frontiers in Environmental Science* 10, 1041573. <https://doi.org/10.3389/fenvs.2022.1041573>.
- Li, Mengmeng, and Alfred Stein. 2020. "Mapping Land Use From High Resolution Satellite Images by Exploiting the Spatial Arrangement of Land Cover Objects." *Remote Sensing* 12 (24): 4158. <https://doi.org/10.3390/rs12244158>.
- Liu, Lanfa, Ana-Maria Olteanu-Raimond, Laurence Jolivet, Arnaud Le Bris, and Linda See. 2021. "A Data Fusion-Based Framework to Integrate Multi-Source VGI in An Authoritative Land Use Database." *International Journal of Digital Earth* 14 (4): 480–509. <https://doi.org/10.1080/17538947.2020.1842524>.
- Liu, Zhi-Qiang, Ping Tang, Weixiong Zhang, and Zheng Zhang. 2022. "CNN-Enhanced Heterogeneous Graph Convolutional Network: Inferring Land Use From Land Cover with a Case Study of Park Segmentation." *Remote Sensing* 14 (19): 5027. <https://doi.org/10.3390/rs14195027>.
- Mawuenyegah, Aleta, Songnian Li, and Shishuo Xu. 2022. "Exploring Spatiotemporal Patterns of Geosocial Media Data for Urban Functional Zone Identification." *International Journal of Digital Earth* 15 (1): 1305–1325. <https://doi.org/10.1080/17538947.2022.2107099>.

- Ménéroux, Yann, Ibrahim Maidaneh Abdi, Arnaud Le Guilcher, and Ana-Maria Olteanu-Raimond. 2022. "Is the *radial Distance* Really a Distance? An Analysis of Its Properties and Interest for the Matching of Polygon Features." *International Journal of Geographical Information Science* 37 (2): 1–38. <https://doi.org/10.1080/13658816.2022.2123487>.
- Meng, X. L., N. Currit, L. Wang, and X. J. Yang. 2012. "Detect Residential Buildings From Lidar and Aerial Photographs Through Object-Oriented Land-Use Classification." *Photogrammetric Engineering & Remote Sensing* 78 (1): 35–44. <https://doi.org/10.14358/PERS.78.1.35>.
- Olteanu-Raimond, Ana-Maria, Joep Crompvoets, Inian Moorthy, Clément Mallet, and Bénédicte Bucher. 2022. "The Use of Volunteer Geographic Information for Producing and Maintaining Authoritative Land Use and Land Cover Data." Technical Report, Paris, France, February 2022.
- Pan, Gang, Guande Qi, Zhaohui Wu, Daqing Zhang, and Shijian Li. 2013. "Land-Use Classification Using Taxi GPS Traces." *IEEE Transactions on Intelligent Transportation Systems* 14 (1): 113–123. <https://doi.org/10.1109/TITS.2012.2209201>.
- Reed, James, Josh van Vianen, Jos Barlow, and Terry Sunderland. 2017. "Have Integrated Landscape Approaches Reconciled Societal and Environmental Issues in the Tropics?" *Land Use Policy* 63:481–492. <https://doi.org/10.1016/j.landusepol.2017.02.021>.
- Rutkowski, Perrine, Martin Bocquet, Jérôme Douché, and Laure Chandelier. 2017. "Qualification de l'usage des zones US 235 de l'OCS GE par les Fichiers fonciers." https://datafoncier.cerema.fr/sites/datafoncier/files/fichiers/2019/02/Enrichissement_OCSGE_FF_methode_appfondie_0.pdf.
- See, Linda, Peter Mooney, Giles Foody, Lucy Bastin, Alexis Comber, Jacinto Estima, Steffen Fritz, et al. 2016. "Crowdsourcing, Citizen Science Or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information." *ISPRS International Journal of Geo-Information* 5 (5): 55. <https://doi.org/10.3390/ijgi5050055>.
- Su, Yu, Yanfei Zhong, Yinhe Liu, and Zhendong Zheng. 2023. "A Graph-Based Framework to Integrate Semantic Object/Land-Use Relationships for Urban Land-Use Mapping with Case Studies of Chinese Cities." *International Journal of Geographical Information Science* 37 (7): 1582–1614. <https://doi.org/10.1080/13658816.2023.2203199>.
- Sun, Zhenhui, Peihang Li, Dongchuan Wang, Qingyan Meng, Yunxiao Sun, and Weifeng Zhai. 2023. "Recognizing Urban Functional Zones by GF-7 Satellite Stereo Imagery and POI Data." *Applied Sciences* 13 (10): 6300. <https://doi.org/10.3390/app13106300>.
- Temporary MIWP 2021-2024 sub-Group 2.3.1. 2023. "D2.8.III.4 Data Specification on Land Use – Technical Guidelines." D2.8.III.4_v3.1.0, January 2024. INSPIRE Maintenance and Implementation Group (MIG) <https://inspire-mif.github.io/technical-guidelines/data/lu/>.
- Tu, Ying, Bin Chen, Tao Zhang, and Bing Xu. 2020. "Regional Mapping of Essential Urban Land Use Categories in China: A Segmentation-Based Approach." *Remote Sensing* 12 (7): 1058. <https://doi.org/10.3390/rs12071058>.
- Vali, Ava, Sara Comai, and Matteo Matteucci. 2020. "Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review." *Remote Sensing* 12 (15): 2495. <https://doi.org/10.3390/rs12152495>.
- Vu, T. T., N. V. A. Vu, H. P. Phung, and L. D. Nguyen. 2021. "Enhanced Urban Functional Land Use Map with Free and Open-Source Data." *International Journal of Digital Earth* 14 (11): 1744–1757. <https://doi.org/10.1080/17538947.2021.1970262>.
- Wu, Hao, Wenting Luo, Anqi Lin, Fanghua Hao, Ana-Maria Olteanu-Raimond, Lanfa Liu, and Yan Li. 2023. "SALT: A Multifeature Ensemble Learning Framework for Mapping Urban Functional Zones from VGI Data and VHR Images." *Computers, Environment and Urban Systems* 100:101921. <https://doi.org/10.1016/j.compenvurbysys.2022.101921>.
- Xu, Rungpeng, Zhenjie Chen, Feixue Li, and Chen Zhou. 2023. "Identification of Urban Functional Zones Based on POI Density and Marginalized Graph Autoencoder." *ISPRS International Journal of Geo-Information* 12 (8): 343. <https://doi.org/10.3390/ijgi12080343>.
- Yang, Xin, Shuaishuai Bo, and Zhaojie Zhang. 2023. "Classifying Urban Functional Zones Based on Modeling POIs by Deepwalk." *Sustainability* 15 (10): 7995. <https://doi.org/10.3390/su15107995>.
- Yin, Jiadi, Jinwei Dong, Nicholas A. S. Hamm, Zhichao Li, Jianghao Wang, Hanfa Xing, and Ping Fu. 2021. "Integrating Remote Sensing and Geospatial Big Data for Urban Land Use Mapping: A Review." *International Journal of Applied Earth Observation and Geoinformation* 103:102514. <https://doi.org/10.1016/j.jag.2021.102514>.
- Zhang, Pengbin, Yinghai Ke, Zhenxin Zhang, Mingli Wang, Peng Li, and Shuangyue Zhang. 2018. "Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery." *Sensors* 18 (11): 3717. <https://doi.org/10.3390/s18113717>.
- Zhang, Xucai, Xiaoping Liu, Kai Chen, Fangli Guan, Miao Luo, and Haosheng Huang. 2023. "Inferring Building Function: A Novel Geo-Aware Neural Network Supporting Building-Level Function Classification." *Sustainable Cities and Society* 89:104349. <https://doi.org/10.1016/j.scs.2022.104349>.

- Zhang, Ce, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, and Peter M. Atkinson. 2019. "Joint Deep Learning for Land Cover and Land Use Classification." *Remote Sensing of Environment* 221:173–187. <https://doi.org/10.1016/j.rse.2018.11.014>.
- Zhou, Qi, Yuheng Zhang, Ke Chang, and Maria Antonia Brovelli. 2022. "Assessing OSM Building Completeness for Almost 13,000 Cities Globally." *International Journal of Digital Earth* 15 (1): 2400–2421. <https://doi.org/10.1080/17538947.2022.2159550>.