LULC segmentation in historical images under domain shift: an empirical study

Swann Briand, Flora Weissgerber, Pierre Fournier, Magali Weissgerber

LULC SEGMENTATION IN HISTORICAL IMAGES UNDER DOMAIN SHIFT: AN EMPIRICAL STUDY

Swann Briand[a] *, Flora Weissgerber*[a] *, Pierre Fournier*[a] *, Magali Weissgerber*[b,c]

[a] DTIS, ONERA, Université Paris Saclay, FR-91123 Palaiseau, France
[b] German Centre for Integrative Biodiversity Research (iDiv), Leipzig, Germany
[c] Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

ABSTRACT

Agricultural abandonment is a global trend leading to vegetation succession and Forests expansion. Manual annotations of 1946 and 2019 aerial surveys images from a peri-urban area in Massif Central shows Land Use and Land Cover (LULC) evolution in this period. We propose to use a convolutional neural network trained on labelled years images to predict LULC maps from 13 intermediate years unlabelled images. However, sensors variety used for acquisition during this time induce variability in ground sampling distances and colorimetry. We have shown using transfer between labelled years that sampling distances have to be the same in the training and testing set, and that coarse scaling offer sufficient performances for the considered LULC classes. Colorimetric data augmentations were individually used after sampling unification to make models more robust to sensors and illumination changes, but proved to be inconsistent in transfer on intermediate years.

***Index Terms*—** Agricultural abandonment, aerial survey, CNN semantic segmentation, domain adaptation

1. INTRODUCTION

Agricultural abandonment is a global trend with 150 million Ha abandoned between 1700 and 1992 [1] generally followed by vegetation succession and Forests expansion [2]. However, it exhibits regional differences [3]. A recent study over an urban hinterland in Auvergne (France) showed that 23 % of agricultural lands were abandoned between 1946 and 2019 [4]. This study was conducted by manually annotating orthomosaics from aerial images acquired in 1946 and in 2019 by the National Institute of Geographical and forestry information (IGN). However, multiple aerial surveys conducted between these two dates were not exploited. The automatic annotation of these images would provide a better temporal resolution of Land Use and Land Cover (LULC) changes.

Neural networks have already been used in remote sensing for semantic segmentation tasks on aerial historical orthophotos for land cover semantic segmentation. Adversarial methods have been used to reduce the amount of annotated historical data needed. In [5], a fully annotated dataset of recent RGB images are translated into historical images using a generative neural network (GAN). These translated images are used to train a CNN to predict LULC classes historical maps. While still using an adversarial approach, [6] aims to construct semantic segmentation features invariant between two datasets of historical images over Central Africa, one with label and one unlabelled.

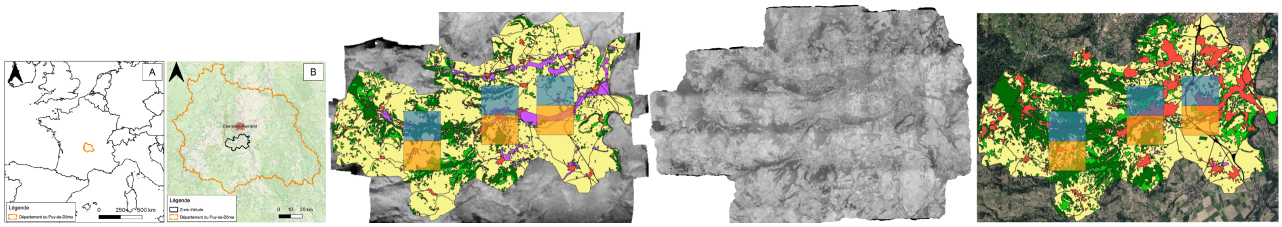
As with these previous datasets, our data include multiple unlabelled images, with the image sampling and colorimetry changing between the years. These colorimetric changes are due to the lighting condition and the parameters such as exposure time set manually for each acquisition campaign as well as the digitisation parameters in the case of analog images. This article aims first at evaluating the performance losses due to these differences. Then, instead of considering a problem with a fixed ensemble of target domains, we evaluate the capacity of data augmentation to train a model robust to these changes.

In section 2, we present the dataset used in this study. Then, in section 3, we present the domain adaption methods studied. Finally, in section 4, we assess the capacity of data augmentation to facilitate transfer between years.

2. DATASET

The area of interest is a 166 km^2 peri-urban area located in the Massif Central region of France, near Clermont-Ferrand (see Figure 1). It is composed of different topographic profiles, with the west of the area being located in the Chaîne des Puys mountain range and the east of the area being located in the Limagne alluvial plains. The altitude ranges from 321m to 1025m. Due to the nature of the terrain, the validation and test sets were divided into three 10 km^2 sub-areas, selected to be representative of the terrain's topography: one mountainous area in the West, one plain area in the East and a transition area in the center. The southern part of those areas is used for model validation, the northern one as a test set after model training as represented in Figure 1.

This study is based on 15 orthomosaics constructed from aerial surveys conducted by IGN between 1946 and 2019. Be-



(a) (b) (c) (d)

Fig. 1: **(a)** Study area (*B*, black line) in Puy-de-Dôme department (*A, B*, orange line) (figure from [4], p.39). 1946 **(b)** and 2019 **(d)** orthomosaics and their corresponding annotations, and 1974 **(c)** unannotated orthomosaic. Background, Forests, Shrubs, Urban and artificial, Orchards, Transportation routes. The square is the validation set and the square is the test set.

fore 1989, the images are collected in the *remonter le temps* IGN service [7] The orthomosaics have been constructed using Agisoft Metashape. From 2000 onward, RGB images are available and orthomosaics were directly downloaded from the CRAIG database [8]. In this study, the RGB images were converted to grayscale to keep the number of channels consistent in the timeserie. Sampling varies from 1m for the 1946 image to 0.20m for the 2019 image.

The 1946 and 2019 orthomosaics are hand-labelled, the other 13 remain unlabelled. Eight Land Use and Land Cover (LULC) classes were selected for the agricultural abandonment study [4], including Urban and artificial areas, Water and Transportation routes and 5 vegetation classes capturing the vegetation succession: Agricultural lands, Orchards, Shrubs, Planted forests and Forests. Due to the geographic constraints in the choice of the validation and test sets, some classes were missing from these sets. The classes Agricultural lands and Water were merged in a Background class and the class Planted Forests was merged to the Forests class.

3. METHOD

We train a U-Net architecture with a EfficientNet encoder pretrained on Imagenet. During training, random 256x256 px^2 patches with at least 80% of labelled pixels and that don't overlap validation and test sets are selected on the fly. To train all our models, we used 120 epochs of 250 batches of 8 patches. Pixel labels prediction are highly dependent on context due to the nature of CNNs, especially in our case where we want to predict LULC classes, where a same object can have different labels based on context. As our dataset is small and a fixed tiling for patch selection limits the variability of training data, we increase the number of possible patches by randomly selecting them to ensure the same pixels are seen with different context. To further increase our models' robustness, we use spatial data augmentations techniques by applying one of several operations with a 50% probability: $\pm$ 90° or 180° rotations and horizontal or vertical flip.

To train the models, we minimise the weighted multiclass cross entropy. Because our dataset class distribution is imbalanced and some classes, like Orchards or Roads, are almost absent, we use weights for each term of the loss function $p_k = \frac{1}{f_k}$ with f_k the kth class frequency in the dataset, so that classification mistakes for the least occurring classes are penalised regardless of their occurrences.

During the study, we only consider two samplings: the 1m sampling of the 1946 orthomosaic and the 0.20m sampling of the 2019 orthomosaic. When undersampling images to the 1m grid of the 1946 orthomosaic, a low-pass filter is applied to avoid aliasing. Label maps are undersampled using nearest neighbour interpolation.

The sensor changes do not only modify the resolution, they also lead to changes in the image appearance, that can be combined with changes in acquisition condition, to create a dataset shift from one year to another. We analyse the impact of the following four data augmentation to mitigate these distribution shifts :

- $\mathcal{M}^*$: Modifying the brightness and the contrast
- $\mathcal{M}^\gamma$: Applying a power law
- $\mathcal{M}^{\text{EQ}}$: Histogram equalisation

We also consider Fourier Domain Adaptation [9], $\mathcal{M}^{\text{FDA}}$, a domain adaptation technique that replaces the source domain low frequencies with the target domain low frequencies, to assess how introducing information from the target domain compares to the colorimetric data augmentation. In FDA, we applied a Hamming window instead of a Rectangular window to avoid spectrum discontinuities.

For quantitative evaluation, we use the Intersection over Union (IoU) metric, averaged over the classes without weighting. In this study, we conducted two types of experiments. The first one is a transfer between the two annotated orthomosaics of 1946 and 2019. It enables to measure the loss in performances compared to models without transfer. The performances are evaluated on the test set presented in Figure 1. A second type of experiment is the inference on unlabelled

intermediate years. We evaluated the model on the labels that did not change between 1946 and 2019 hypothesising that these labels were stable for the whole period. They represent 60% of the 1946 labels. For the test set and the intermediate year images inference, we used a 64 pixel stride for our tiling so that each pixel could be predicted 16 times with different context like during training. These different network outputs were then averaged, weighted by the distance d between the considered pixel and the centre of the patch during the inference.

4. RESULTS

4.1. Without transfer

Table 1 shows models performance with resolution adaptation on labelled years. Best IoU scores are obtained when the year and sampling remain unchanged between training and testing.

These good performances can be observed in Figure 2c and 2i for 1946 and 2019 respectively. However, in Figure 2 (i) we can see two majors confusions that highlights the difficulties of considering LULC classes :

- A park area in the south east of the city was classified as a mix of forest, shrubs and background
- A road in the north east of the city was classified as urban areas.

Indeed, since our labels are constituted of both LU and LC classes, thus the same object, such as a tree or a road, can belong to different classes depending on its surrounding.

For the 2019 image, training and testing with a 1m sampling leads to more homogeneous prediction map, that yield better IoU than with a 0.2m sampling.

4.2. Transfer between 1946 and 2019

4.2.1. Sampling adaptation

When changing the year between training and testing while keeping their original sampling, models performs badly. The prediction of $\mathcal{M}^{1m}_{1946}(\mathcal{I}^{0.2m}_{2019})$ are very scattered. As it can be seen in Figure 2j, instead of detecting one homogeneous Urban area, houses are being individually detect as Urban area, road between them as Transportation routes and their gardens as Orchards. On the contrary, the prediction of $\mathcal{M}^{0.2m}_{2019}(\mathcal{I}^{1m}_{1946})$ are too homogeneous with Shrubs being over-represented as shown in Figure 2d.

When transferring between years with a common a 1m spacing, the performances of $\mathcal{M}^{1m}_{1946}$ and $\mathcal{M}^{1m}_{2019}$ increases both of more than 0.1 of IoU. $\mathcal{M}^{1m}_{1946}$, illustrated in Figure 2k performs better than $\mathcal{M}^{1m}_{2019}$, that over-predicts Background as illustrated in Figure 2e. The model trained on $\mathcal{I}^{1m}_{1946}$ seems thus more robust to illumination change than the model train on $\mathcal{I}^{1m}_{2019}$.

	$\mathcal{I}^{1}_{1946}$	$\mathcal{I}^{1}_{2019}$	$\mathcal{I}^{0.2}_{2019}$
$\mathcal{M}^{1\text{m}}_{1946}$	0.56	0.40	0.24
$\mathcal{M}^{1\text{m}}_{2019}$	0.23	0.71	0.28
$\mathcal{M}^{0.2\text{m}}_{2019}$	0.11	0.27	0.66

Table 1: Mean IoU for sampling change.

	$\mathcal{M}^{\text{EQ}}_{1946}$	$\mathcal{M}^{\text{FDA}}_{1946}$	$\mathcal{M}^{\gamma}_{1946}$	$\mathcal{M}^{*}_{1946}$	$\mathcal{M}^{\text{EQ}}_{2019}$	$\mathcal{M}^{\text{FDA}}_{2019}$	$\mathcal{M}^{\gamma}_{2019}$	$\mathcal{M}^{*}_{2019}$
$\mathcal{I}^{1}_{1946}$	0.57	0.57	0.58	0.58	0.37	0.38	0.38	0.42
$\mathcal{I}^{1}_{2019}$	0.41	0.50	0.34	0.37	0.70	0.70	0.71	0.70

Table 2: Mean IoU for colorimetric data augmentation.

4.2.2. Colorimetric data augmentation

Table 2 shows the model performances with different colorimetric data augmentations when evaluated on the same year or in transfer to another year. For same year prediction, there is very few modifications of the IoU when using colorimetric data augmentation. Even the FDA, that takes information from the target year, didn't degrade the results on the source year.

When transferring between years, data augmentation impacts the IoU, especially for the 2019 to 1946 transfer where the brightness and contrast data augmentation improve the IoU up to 0.19. This increase in IoU is reflected in the larger presence of Shrubs and Orchards areas as represented figure 2l. For the 1946 to 2019 transfer, only FDA increases the IoU. Despite a large increase in the precision and recall of the Urban areas prediction the gain in mean IoU is only of 0.1 due to an overprediction of the Shrubs class in Forest areas it can been seen in figure 2f.

4.3. Transfer for intermediate years

For the model trained without colorimetric data augmentation, table 3 shows that the models trained on 1946 have better results up to the year 2000, underlying again the robustness of the training on lower quality data. However, for the model trained on the 1946 image, colorimetric data augmentation does not significantly improve models. The model trained without augmentations even perform better for some intermediate years. On the other hand, models trained on the 2019 image gain from colorimetric data augmentation. Apart from 1954 and 1962 that are the two years the closest to 1946, the models trained on the 2019 image outperform the models trained on the 1946 image. However, no colorimetric data augmentation has consistent performances between years.

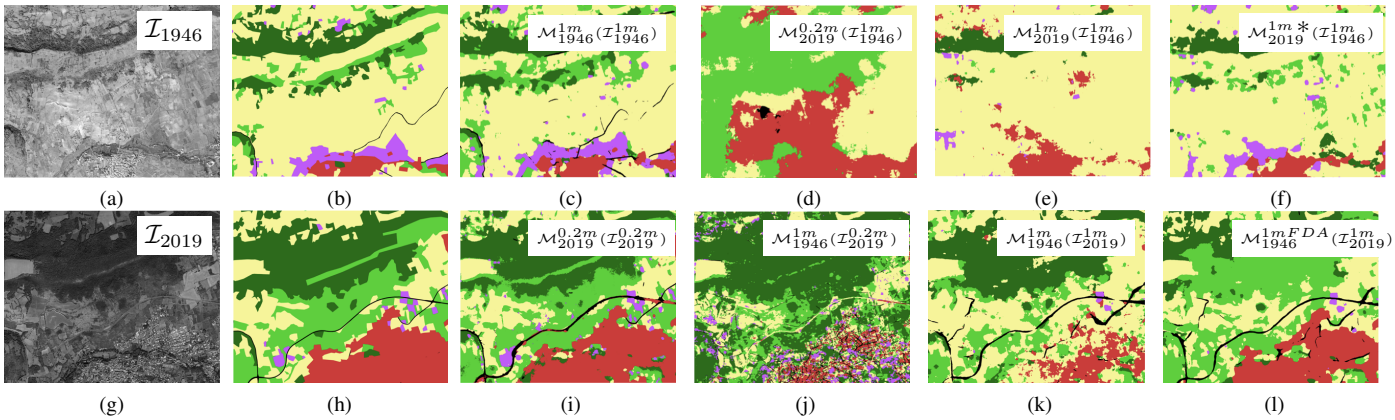


Fig. 2: **(a)**,**(g)** 1946 and 2019 central test zones image (see Fig. 1) ; **(b)**,**(h)** ground truth ; **(c)**,**(i)** LULC map predictions from models trained in the same year with same sampling ; **(d)**,**(j)** transfer without sampling changes; **(e)**, **(k)** with sampling changes ; **(f)**,**(l)** transfer with sampling changes and colorimetric data augmentation from best models , $\mathcal{M}^{*}_{2019}$ and $\mathcal{M}^{FDA}_{1946}$ respectively (see Tab. 2).

	1946	1954	1962	1974	1978	1981	1984	1986	1989	2000	2004	2009	2013	2016	2019
$\mathcal{M}_{1946}$		0.31	**0.33**	0.29	**0.39**	0.33	0.32	**0.32**	0.31	**0.36**	0.32	0.27	**0.31**	0.36	**0.37**
$\mathcal{M}^{*}_{1946}$		**0.33**	0.3	0.26	0.37	0.35	0.32	0.3	0.31	0.34	**0.35**	0.15	0.29	0.34	0.36
$\mathcal{M}^{\gamma}_{1946}$		0.31	0.31	**0.32**	0.37	**0.39**	0.31	0.27	0.3	0.3	0.31	0.23	0.23	0.32	0.33
$\mathcal{M}^{\mathrm{EQ}}_{1946}$		0.32	0.27	0.28	0.35	0.35	**0.33**	0.30	**0.32**	0.32	0.3	0.24	0.26	0.30	0.30
$\mathcal{M}^{\mathrm{FDA}}_{1946}$		0.30	0.29	0.30	0.35	0.31	0.32	0.30	0.31	0.35	**0.35**	**0.34**	**0.31**	**0.39**	0.3
$\mathcal{M}_{2019}$	0.23	0.13	0.26	0.27	0.28	0.29	0.23	0.30	0.29	0.41	0.40	0.35	0.48	0.54	
$\mathcal{M}^{*}_{2019}$	**0.36**	0.26	0.27	0.37	0.42	0.35	0.31	0.42	0.34	**0.48**	**0.49**	**0.45**	0.47	0.55	
$\mathcal{M}^{\gamma}_{2019}$	0.33	0.22	**0.29**	0.37	**0.44**	0.23	**0.36**	**0.45**	**0.39**	**0.48**	0.44	0.30	0.47	0.50	
$\mathcal{M}^{\mathrm{EQ}}_{2019}$	**0.36**	**0.27**	0.24	**0.39**	0.43	**0.40**	0.33	0.44	0.35	0.45	0.48	0.40	**0.50**	0.53	
$\mathcal{M}^{\mathrm{FDA}}_{2019}$	0.33	0.15	0.27	0.36	0.42	0.18	0.34	0.44	0.37	0.5	0.46	0.38	**0.50**	**0.57**	

Table 3: Table of the mean IoU on common labels for intermediate years. Best score is underlined and best score within each training year is in bold. Top rows are models trained with 1946 dataset, bottom rows with 2019 dataset. All orthomosaics are undersampled to the 1m 1946 grid.

5. CONCLUSION

In this work, we investigate the possibility to automatically annotate LULC on a temporal stack of 15 orthomosaics constructed from aerial survey, with only the two extreme images, 1946 and 2019, annotated by hand. These orthomosaics exhibit changes in sampling between 1m and 0.2m and changes in colorimetry.

We first investigate the impact of sampling changes. We showed that the features learned by our model are sensitive to scaling, but that the model trained on the 2019 image undersampled to 1m yields slightly better result than the model trained on the 2019 image at a 0.2m sampling. Its prediction maps are more homogeneous and closer to the ground truth composed of LULC classes and not object classes.

When transferring between years with a constant 1m sampling, colorimetric data augmentation improve the transfer from 2019 to 1946. Only the FDA, that mixes low frequencies of 2019 image and high frequencies of the 1946 image increase the transfer between 1946 and 2019. When testing on intermediate years, the transfer from 2019 with colorimetric data augmentation outperform the transfer from 1946 that has stable IoU between 0.3 and 0.4. However, no single colorimetric data augmentation could compensate for all the appearance change in the intermediate images. Using all the colorimetric data augmentations at once could improve results consistency for intermediate years.

Moreover, due to memory limitations, we always predicted 256x256 px^2 patches, which limits the field of view for network trained on images with 0.2m sampling. Given the nature of the LULC classes of this dataset, we would like to investigate using larger patches could help to improve the homogeneity of the output map of network trained on images at 0.2m sampling.

6. REFERENCES

[1] N. Ramankutty and J. A. Foley, "Estimating historical changes in global land cover: Croplands from 1700 to 1992," *Global Biogeochemical Cycles*, vol. 13, pp. 997–1027, Dec. 1999.

[2] V. A. Cramer, R. J. Hobbs, and S. for Ecological Restoration International, eds., *Old fields: dynamics and restoration of abandoned farmland*. The science and practice of ecological restoration, Washington: Island Press, 2007. OCLC: ocn145379806.

[3] T. Lasanta, J. Arnáez, N. Pascual, P. Ruiz-Flaño, M. Errea, and N. Lana-Renault, "Space–time process and drivers of land abandonment in Europe," *CATENA*, vol. 149, pp. 810–823, Feb. 2017.

[4] M. Weissgerber, *La succession végétale post-agricole : approche par les services écosystémiques et les perceptions des acteurs dans un territoire entre Chaîne des Puys et plaine de la Limagne*. PhD thesis, Université Clermont Auvergne, Jan. 2023.

[5] W. A. J. Van Den Broeck, T. Goedemé, and M. Loopmans, "Multiclass Land Cover Mapping from Historical Orthophotos Using Domain Adaptation and Spatio-Temporal Transfer Learning," *Remote Sensing*, vol. 14, p. 5911, Nov. 2022.

[6] N. Mboga, S. D'Aronco, T. Grippa, C. Pelletier, S. Georganos, S. Vanhuysse, E. Wolff, B. Smets, O. Dewitte, M. Lennert, and J. D. Wegner, "Domain adaptation for semantic segmentation of historical panchromatic orthomosaics in central africa," *ISPRS International Journal of Geo-Information*, vol. 10, no. 8, 2021.

[7] "Remonter le temps." Available at https://remonterletemps.ign.fr.

[8] "Flux WMS/WFS | Craig." Available at https://www.craig.fr/contenu/361-flux-wmswfs.

[9] Y. Yang and S. Soatto, "FDA: Fourier Domain Adaptation for Semantic Segmentation," Apr. 2020. arXiv:2004.05498 [cs].