Transferability between two areas of a data fusion based method for Land Use classification

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

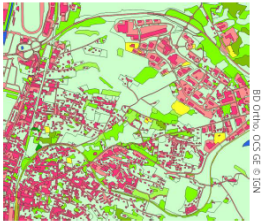
Univ Gustave Eiffel, ENSG, IGN, LASTIG, F-94160 Saint-Mandé, France

March 28, 2024

Layout

Context

Ortho-image



Land Cover

- Built-up areas
- Unbuilt areas
- Areas with mineral materials
- Areas with other composite materials
- Bare ground
- Water surfaces
- Snow and glaciers
- Hardwood stands
- Coniferous stands
- Mixed stands
- Shrub and sub-shrub formations
- Other woody formations
- Herbaceous formations
- Other non-woody formations

Land Use

- Agriculture
- Forestry
- Extractive activities
- Fishing and aquaculture
- Other primary production
- Secondary production
- Secondary and tertiary production and residential use
- Tertiary production
- Road networks
- Rail networks
- Aerial networks
- Inland waterway and maritime transport networks
- Other transport networks
- Logistics and storage services
- Public utility networks
- Residential use
- Areas in transition
- Abandoned areas
- No use
- Unknown use

Context

Current Issues with Land Use Products:

- **Update frequency**
- **Difficulties** in **automating** Land Use determination from Remote Sensing data

Land Cover:
Built-up
Land Use:
Residential

Land Cover:
Built-up
Land Use:
Tertiary production

BD ORTHO © IGN

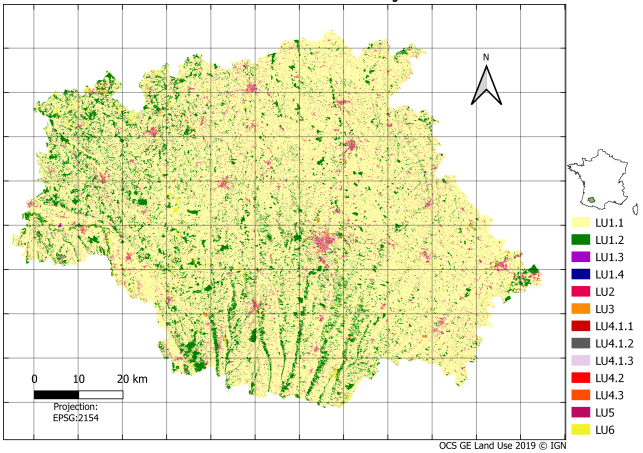
Objective

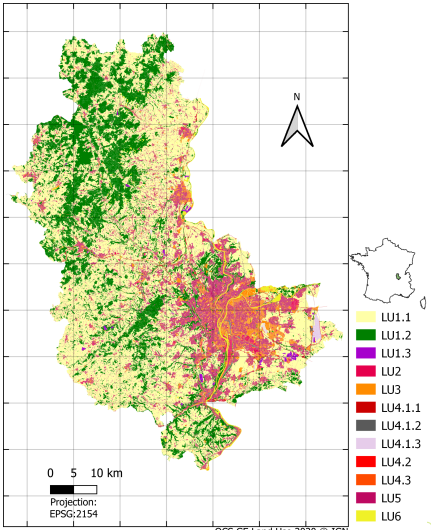
From an **existing** partition of the territory into polygons, using **several input data sources** (*e.g.* aerial images, topographic databases, land cover...), assign to each polygon **a unique Land Use class**.

Assess the **transferability** of the proposed method.

Study datasets: OCS GE LU in Gers and Rhône, France

Semi-manual and time-costly Ground Truth.





Difficulties and challenges

Imperfections of the sources:

- inaccuracy
- incompleteness
- imprecision

Heterogeneity of the sources:

- differences of:
  - actuality
  - scale
  - meaning of the classes
- how informative is the source for our task

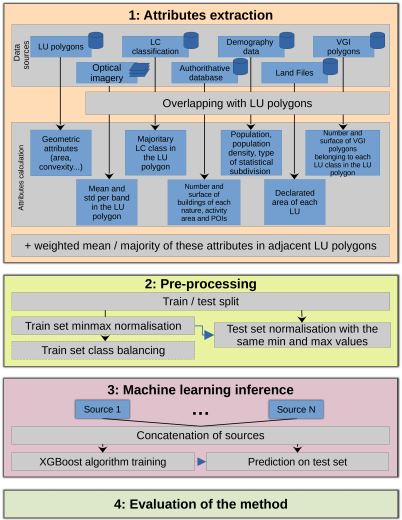
High class imbalance

Differences of distribution between the two study areas

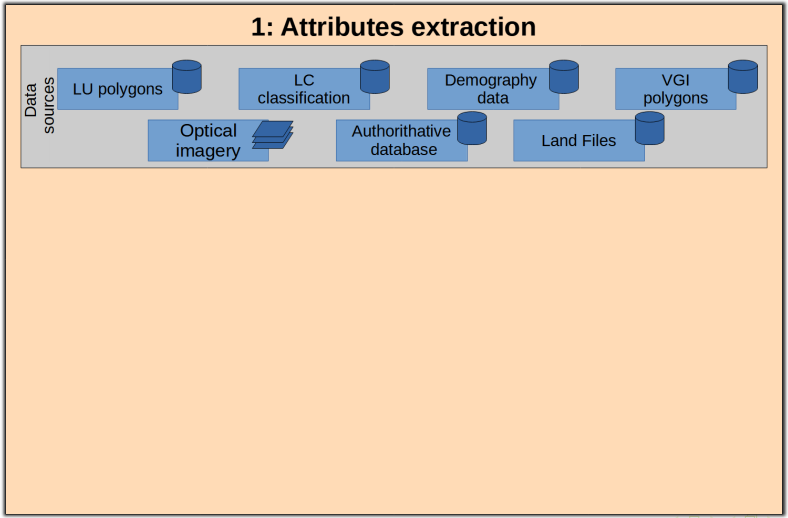
Layout

Characterization of Land Use

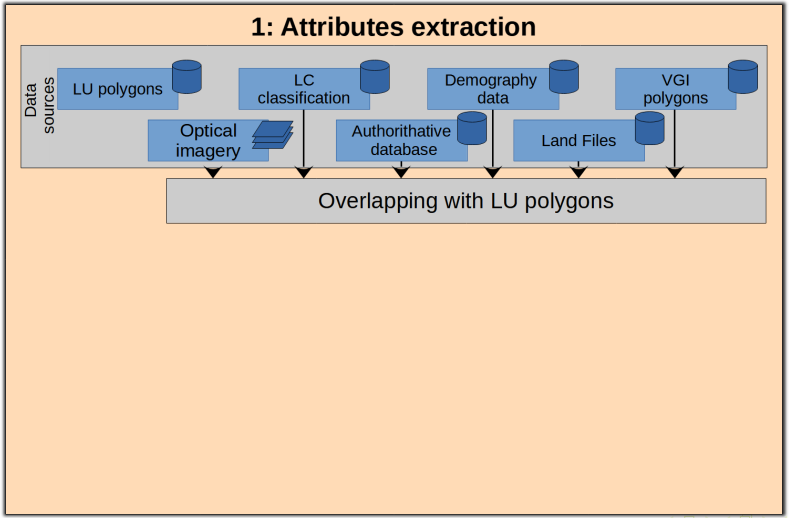
- **Base geometry**: OCS GE polygons
- **Hypothesis**: Learning could enable the understanding of polygon's LU through ground truth and multiple sources.
- Proposal of a **four-step methodology**:
  1. Attribute Extraction,
  2. Pre-processing,
  3. Inference,
  4. Method Evaluation.



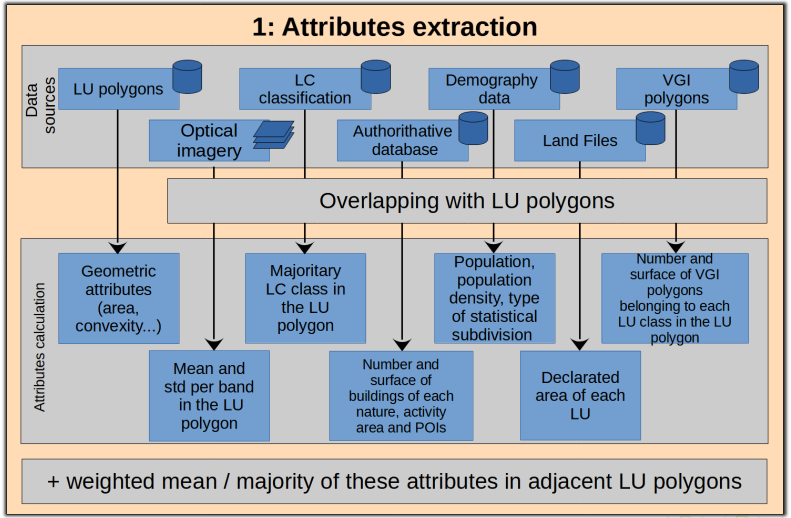
1: Attributes extraction from the sources



1: Attributes extraction from the sources



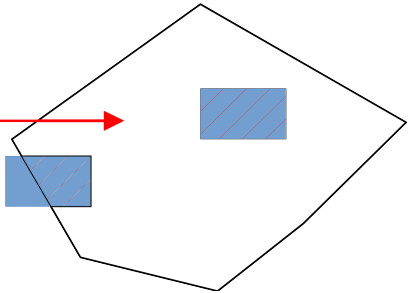
1: Attributes extraction from the sources



1: Attributes extraction from the sources

- **Attribute** : Characterization of the OCS GE LU polygon from its intersection with the data sources.

Residential number : 2
Residential area :
(1+0.66) x S m²



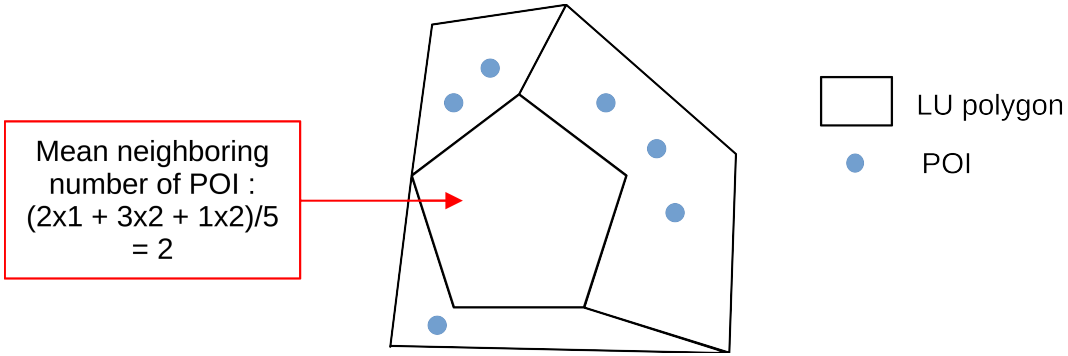
LU polygon

Residential buildings database

with S the area of one of these buildings.

1: Attributes extraction from the sources

- **Neighboring attributes** computed as the mean (or majority) value of each attribute in the neighboring OCS GE LU polygons, weighted by the length of the adjacent side.

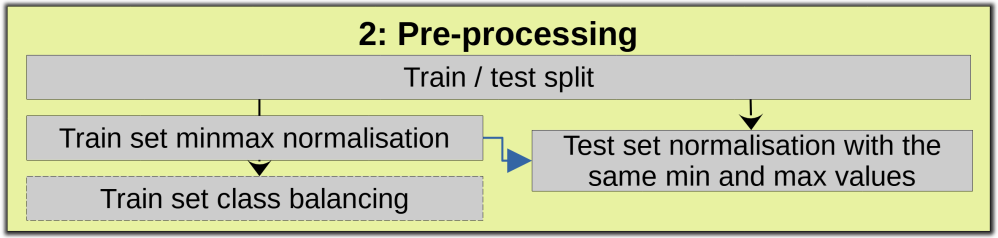


Sources and extracted attributes

Type	Source	Number of attributes	Neighboring attributes
LU segmentation	OCS GE LU Geometry (Authority)	25	25
Optical imagery	BD ORTHO (Authority)	8	8
LC classification	CLC (Authority)	1	1
	OSO (Auto. Research Product)	1	1
Topographic database	BD TOPO building (Authority)	17	17
	BD TOPO other (Authority)	13	13
	RPG (Authority)	1	1
Demographic data	INSEE (Authority)	6	6
Land Files	DGFIP (Authority)	18	18
Topographic database	OpenStreetMap (VGI)	36	36
Total		**254**	

Table: Number of attributes constructed for each source, including the number of neighboring attributes.

2: Pre-processing

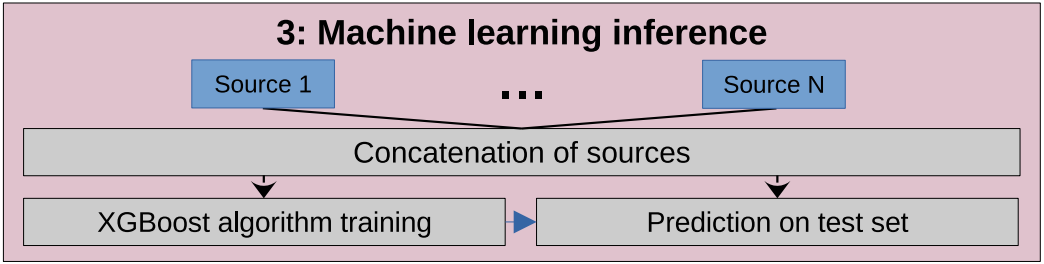


Train set (80 %) / test set (20 %) split.

Minmax normalization of attributes from the values of the train set.

Balancing the classes of the train set using the **SMOTE-NC** algorithm: oversampling of the minority classes by synthesizing new data.
⇒ Goal: To compensate for the significant class imbalance.

3: Machine learning inference



Learning is made using **all attributes** from **all sources** together.

The **XGBoost** machine learning algorithm (Chen and Guestrin, 2016) was identified as the most suited in a previous related work (Cubaud et al., 2023).

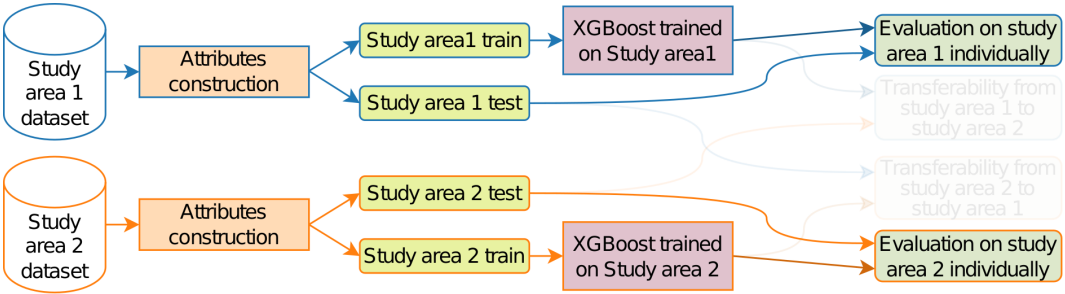
Experiences

3 types of experiences:

- **For each study area individually:** model trained and evaluated on the same study area.
- **Transferability:** model trained from one study area and evaluated on the other study area.
- **Mixed train set:** model trained from a mix of both study area and evaluated on one of the two study areas.

Layout

For each individual study area

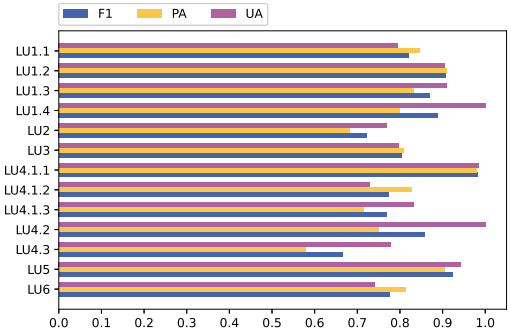


Study area 1: Gers, Study area 2: Rhône

Results for each individual study area

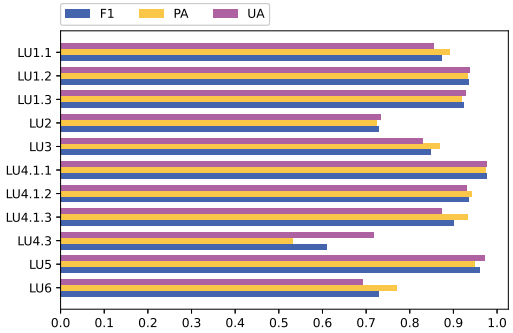
Improvement over the automatic version of the OCS GE.

OA: 88%, mF1: 83%



Per class metrics for the individual prediction of Gers

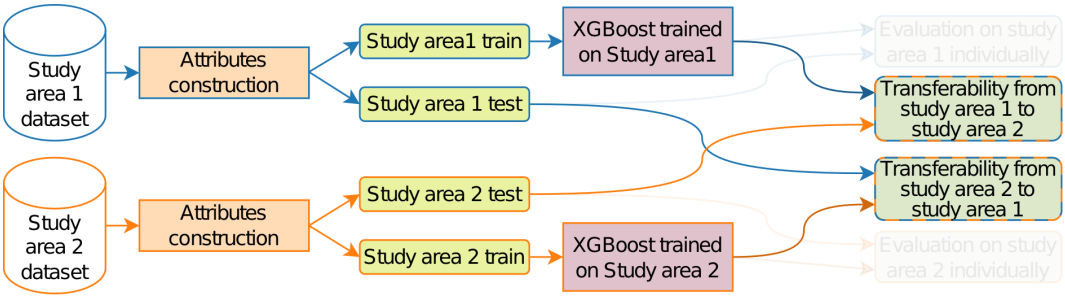
OA: 92%, mF1: 86%



Per class metrics for the individual prediction of Rhône

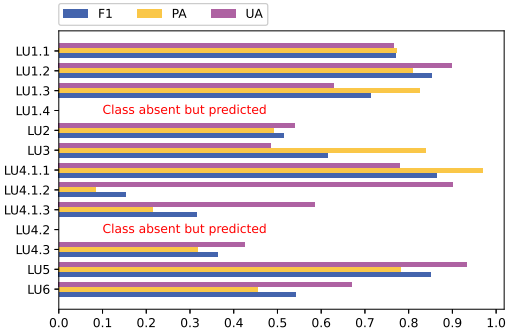
LU1.1: Agriculture, LU1.2: Forestry, LU1.3: Mining, LU1.4: Fishing, LU2: Secondary production, LU3: Tertiary production, LU4.1.1: Road networks, LU4.1.2: Rail networks, LU4.1.3: Aerial networks, LU4.2: Logistics, LU4.3: Public utility networks, LU5: Residential, LU6: Other

Transferability



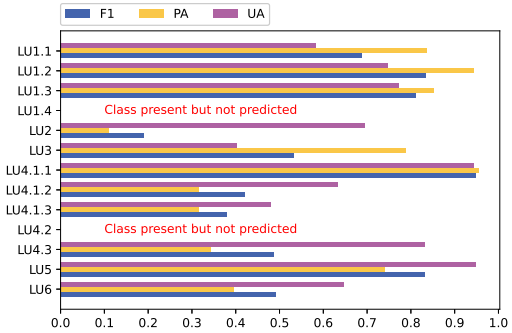
Transferability results

OA: 79%, mF1: 50%



Per class metrics for the transferability from Gers to Rhône

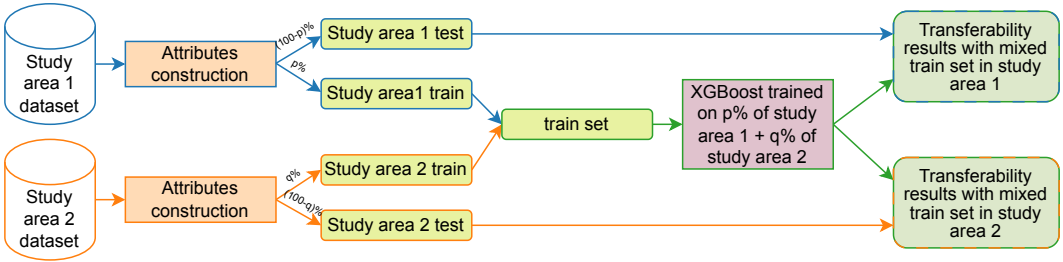
OA: 76%, mF1: 51%



Per class metrics for the transferability from Rhône to Gers

LU1.1: Agriculture, LU1.2: Forestry, LU1.3: Mining, LU1.4: Fishing, LU2: Secondary production, LU3: Tertiary production, LU4.1.1: Road networks, LU4.1.2: Rail networks, LU4.1.3: Aerial networks, LU4.2: Logistics, LU4.3: Public utility networks, LU5: Residential, LU6: Other

Mixed train set



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

		Proportion of Gers dataset in training set					
		0.00	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
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: mF1 score obtained with mixed trained set when evaluated on the rest of Gers dataset

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

		Proportion of Gers dataset in training set					
		0.00	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
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: mF1 score obtained with mixed trained set when evaluated on the rest of Gers dataset

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

		Proportion of Gers dataset in training set					
		0.00	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
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: mF1 score obtained with mixed trained set when evaluated on the rest of Gers dataset

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

		Proportion of Gers dataset in training set					
		0.00	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
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: mF1 score obtained with mixed trained set when evaluated on the rest of Gers dataset

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

		Proportion of Gers dataset in training set					
		0.00	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
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: mF1 score obtained with mixed trained set when evaluated on the rest of Gers dataset

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

		Proportion of Rhône dataset in training set					
		0.00	**0.05**	**0.10**	**0.20**	**0.50**	**0.80**
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

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

Relevance of the different sources

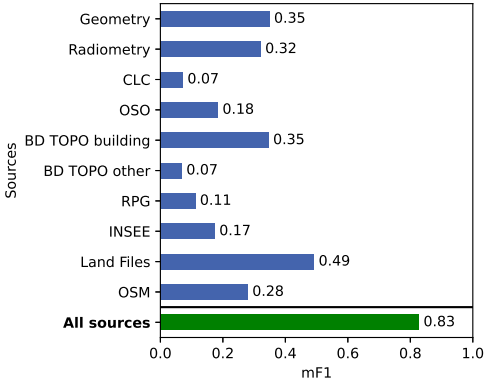


Figure: Metrics for XGBoost trained only with one source in Gers.

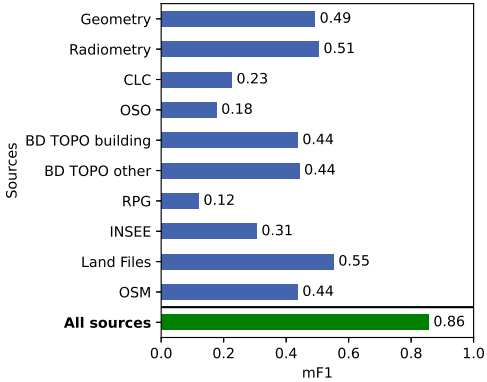


Figure: Metrics for XGBoost trained only with one source in Rhône.

Relevance of the different sources

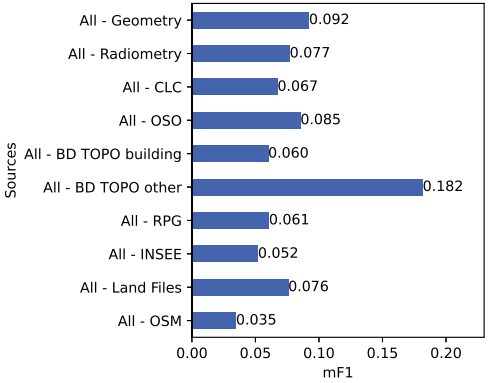


Figure: Score lost when trained without all the attributes from one source in Gers.

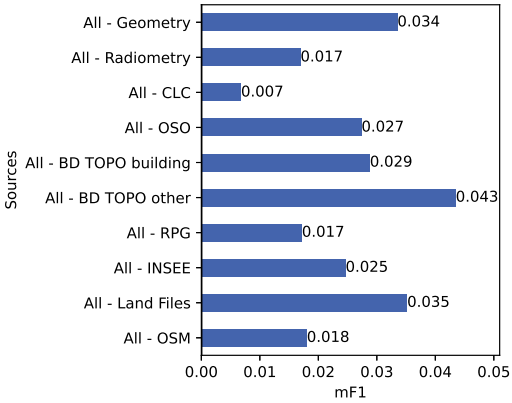


Figure: Score lost when trained without all the attributes from one source in Rhône.

Layout

Conclusion

- Proposed a general **workflow for LU classification** from several heterogeneous sources, applied to 13 LU classes in Gers and Rhône.
- **Individual areas:** Good performances in the two study areas.
  (OA: 88%, 92% respectively, mF1: 83%, 86% respectively)
- **Crisp transferability:** results are lower.
- **Mixed transferability:** Including a small amount of the target study area is identified as a solution to improve these results.
- Crossing **multiple data sources is essential** for accurate LU classification, but imperfections in each source impact the process.

Perspectives

Dealing with mixed Land Use.

Automating generation of mapping units from cadastral data.

LU change and update using the model across different dates.

Study the adaptability to other LU nomenclatures.

Improving transferability through domain adaptation techniques and transfer learning methodologies.

Thanks for your attention!

Any question?

References I

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco California USA, 785–794.

Comber, A. J., Wadsworth, R. A., Fisher, P. F., 2008. Using Semantics to Clarify the Conceptual Confusion between Land Cover and Land Use: The Example of 'Forest'. *Journal of Land Use Science*, 3(2-3), 185–198.

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

Deng, Y., Chen, R., Yang, J., Li, Y., Jiang, H., Liao, W., Sun, M., 2022. Identify Urban Building Functions with Multisource Data: A Case Study in Guangzhou, China. *International Journal of Geographical Information Science*, 36(10), 2060–2085.

Fonte, C. C., Minghini, M., Antoniou, V., Patriarca, J., See, L., 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.

References II

He, J., Li, X., Liu, P., Wu, X., Zhang, J., Zhang, D., Liu, X., Yao, Y., 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.

Li, M., Stein, A., 2020. Mapping Land Use from High Resolution Satellite Images by Exploiting the Spatial Arrangement of Land Cover Objects. *Remote Sensing*, 12(24), 4158.

Liu, L., Olteanu-Raimond, A.-M., Jolivet, L., Le Bris, A., See, L., 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.

Liu, Z.-Q., Tang, P., Zhang, W., Zhang, Z., 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.

Meng, X. L., Currit, N., Wang, L., Yang, X. J., 2012. Detect Residential Buildings from Lidar and Aerial Photographs through Object-Oriented Land-Use Classification. *PE&RS, Photogrammetric Engineering & Remote Sensing*, 78(1), 35–44.

References III

Pan, G., Qi, G., Wu, Z., Zhang, D., Li, S., 2013. Land-Use Classification Using Taxi GPS Traces. *IEEE Transactions on Intelligent Transportation Systems*, 14(1), 113–123.

Su, Y., Zhong, Y., Liu, Y., Zheng, Z., 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*, 0(0), 1–33.

Tu, Y., Chen, B., Zhang, T., Xu, B., 2020. Regional Mapping of Essential Urban Land Use Categories in China: A Segmentation-Based Approach. *Remote Sensing*, 12(7), 1058.

Wang, J., Gao, C., Wang, M., Zhang, Y., 2023. Identification of Urban Functional Areas and Urban Spatial Structure Analysis by Fusing Multi-Source Data Features: A Case Study of Zhengzhou, China. *Sustainability*, 15(8), 6505.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P. M., 2019. Joint Deep Learning for Land Cover and Land Use Classification. *Remote Sensing of Environment*, 221, 173–187.

Difficulty of the classes

Class size is not correlated to classification results for each study area individually, but for transferability, it is significantly correlated.

"Easy" classes

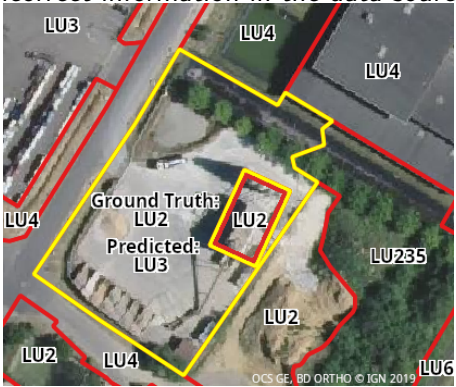
Easily identified by few attributes (e.g. LU4.1.1 Road Network)

"Difficult" classes

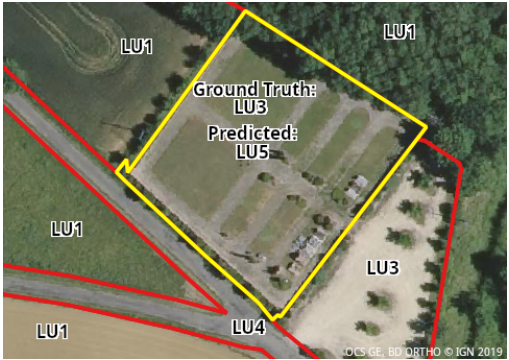
- May encapsulate several aspects (e.g. LU6 Other)
- May share the values of attributes with other classes

Identified Error Sources

Incorrect information in the data sources



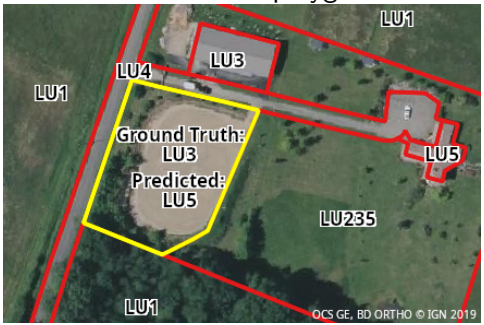
No explicit information in the data sources



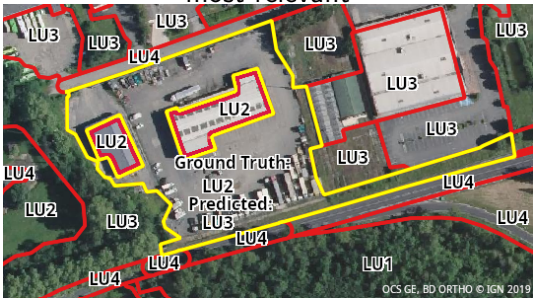
Identified Error Sources - Neighborhood Attributes

Neighborhood attributes help to learn the spatial context and to compensate the lack of explicit information, but:

The relevant information can be separated with the LU polygon



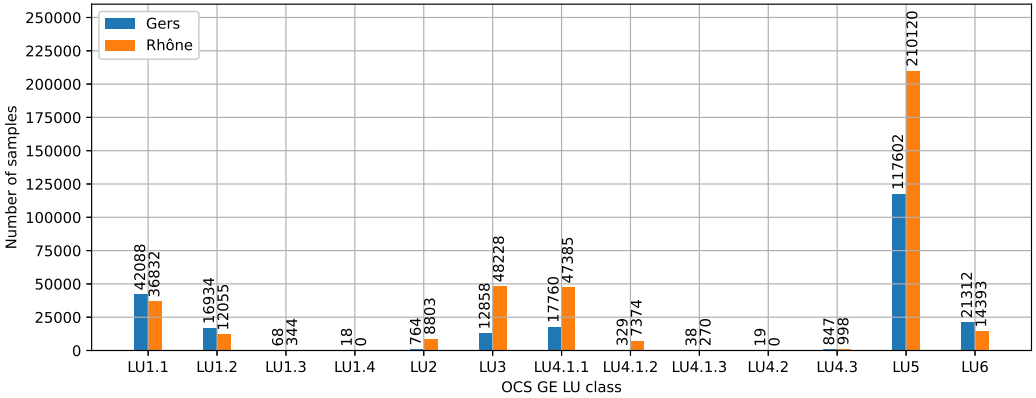
The majority neighborhood can be not the most relevant



Identified Error Sources - Errors due to the quality of Ground Truth

- Generalization of the OCS GE.
- Systematic errors in the ground truth $\rightarrow$ learned by the model (e.g., mechanics).
- Other errors in the ground truth $\rightarrow$ not adequately considered in the evaluation.

High class imbalance



Total: Gers: 230 637 polygons, Rhône: 386 802 polygons.

Positioning in relation to the state of the art

- Semantic confusion between land cover and land use (Comber et al., 2008)
- Remote sensing approaches (Zhang et al., 2019; Li and Stein, 2020; Liu et al., 2022)
- Multi-source approaches:

Multiple sensor imagery	Tu et al. (2020) (daytime and nighttime imagery, Radar), Meng et al. (2012), Deng et al. (2022), He et al. (2021), Wang et al. (2023) (daytime and nighttime imagery, surface temperature)
LIDAR data	Meng et al. (2012) (DSM)
Authoritative datasets	Tu et al. (2020), Meng et al. (2012), Deng et al. (2022), Wang et al. (2023)
VGI	Fonte et al. (2018) (OSM, Facebook, Foursquare), Liu et al. (2021) (in-situ campaigns and mappathon), Deng et al. (2022) (distance to OSM roads), Su et al. (2023)
iVGI	Pan et al. (2013) (GPS traces from taxis), He et al. (2021) (density of Tencent web application users)