An improved European land use/cover map derived by data integration

Konštantín Rosina
European Commission,
Joint Research Centre,
Via E. Fermi 2749, Ispra, Italy
konstantin.rosina@ec.europa.eu

Filipe Batista e Silva
European Commission,
Joint Research Centre,
Via E. Fermi 2749, Ispra, Italy
filipe.batista@ec.europa.eu

Pilar Vizcaino-Martinez
European Commission,
Joint Research Centre,
Via E. Fermi 2749, Ispra, Italy
pilar.vizcaino-martinez
@ec.europa.eu

Mario Alberto Marín Herrera
European Commission,
Joint Research Centre,
Via E. Fermi 2749, Ispra, Italy
mario-alberto.marin-herrera
@ext.ec.europa.eu

Sérgio Freire
European Commission,
Joint Research Centre,
Via E. Fermi 2749, Ispra, Italy
sergio.freire@ec.europa.eu

Marcello Schiavina
European Commission,
Joint Research Centre,
Via E. Fermi 2749, Ispra, Italy
marcello.schiavina@ec.europa.eu

Abstract

Available Europe-wide data on land use and land cover (LULC) for the year 2012 have been rich and detailed, but fragmented in disparate data sets and data types, each having limitations either in spatial detail, thematic detail or coverage. In this study, we try to overcome this issue by processing and integrating a considerable volume of diverse data to create a single, ready-to-use data layer covering the EU-28 and 11 neighbouring countries at 100 m pixel resolution. Using cartographic (map algebra) and statistical (machine learning) techniques we refined the spatial detail from 25 ha down to one hectare and derived seven new LULC classes, while respecting the original CORINE nomenclature. Importantly, we decomposed the class ‘Industrial and commercial units’ into ‘Production facilities’, ‘Commercial/service facilities’ and ‘Public facilities’. The accuracy of the result is overall satisfactory, although we recognised a significant confusion between Production and Commercial facilities. Other limitations of the map are discussed and future research avenues are proposed. This spatial data fusion exercise fulfilled the objective of creating a dataset better suited to socioeconomic research, such as modelling of the human population, economic activity and land use. Immediately, the results are being applied in EU-wide spatio-temporal population modelling and LULC projection.

Keywords: geospatial data integration, land use, land cover, CORINE, classification, random forests

1 Introduction

Land use and land cover (LULC) data provide information about the biophysical cover of the Earth surface and related human actives. Detailed, harmonised and up-to-date LULC maps are a vital input for environmental and socioeconomic sciences, as well as for evidence based planning and policy targets. In Europe, CORINE Land Cover (CLC) has been widely used as the best available harmonised dataset in Europe, although not without limitations. In the last decade, new Europe-wide geodata sources have surfaced with the potential to alleviate some of CLC’s limitations. These include public authority sources (such as those from the EU Copernicus programme), commercial data providers and volunteered geographic information.

In this article, we aim at improving the knowledge of the European LULC both in the spatial and thematic domains by integrating a multitude of available geodatasets. The motivation is to provide better support for territorial modelling of population distribution, land use, transportation, environment and their interactions. Given the spatial extent and resolution of the data, the work is relevant for models constructed at continental, national or regional scales.

Several efforts to produce refined CLC data by visual intrpretation using expanded nomenclature of classes and at finer scale have been made mostly at the level of individual countires (Hazeu et al. 2016). The HELM project (Harmonised European Land Monitoring) has argued for the need to integrate, harmonise and increase the resolution European land monitoring data. However, the production of new datasets was not the goal of the project (Ben-Asher et al. 2013).

More closely related works have focused on producing spatially (Batista e Silva et al. 2013, Fonte at al. 2017, Pazúr & Bolliger 2017) or thematically enhanced data (Jiang et al. 2015) by fusion of existing LULC datasets with other data sources. Alhough, to our knowledge, our work is the first attempt targeting both domains and, at the same time, the continental extent of the study area.

2 Methodology

The methodology consisted of two main parts (Figure 1). First, we improved the spatial detail of the original CLC2012 map from 25 ha down to 1 ha (5 ha for non-artificial classes outside of urban areas). The increased spatial detail was a prerequisite to subsequent thematic refinement of the artificial surfaces by deriving seven additional classes (Table 2).

2.1 Spatial refinement

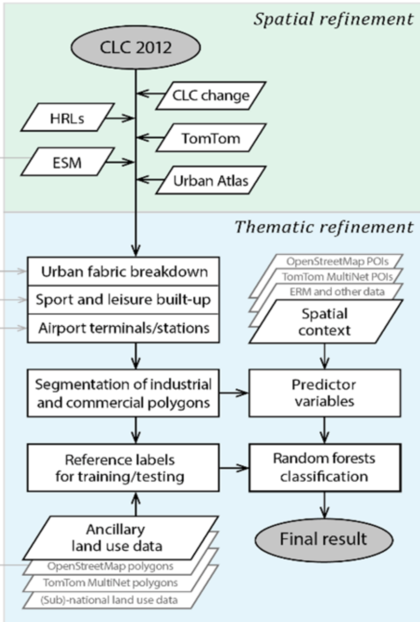
The presented approach elaborated on the methodology proposed by Batista e Silva et al. (2013) for the production of spatially refined 2006 map. However, we introduced several updates to maximise the utility of data available for 2012, such as the EU Copernicus monothematic high-resolution layers (HRL) and an extended set of Urban Atlas (UA) data. Another newly included dataset was the European Settlement Map (ESM, Ferri et al. 2014). ESM is related to the Global Human Settlement Layer (Pesaresi et al. 2016). Both datasets detect the share of built-up area per pixel from high resolution satellite imagery.

The spatial refinement relied on a GIS-based synthesis of categorical raster, interval raster, and polygon vector data. We deployed an automated chain of raster map algebra operations at 100 m pixel size on raw datasets (Table 1); vector data were pre-rasterised using the maximum combined area method to identify the dominant class in each cell. The CLC2012 map served as a seamless background which was sequentially updated. At each step, the cells either remained unchanged or were updated by the overlaid input data layer, following pre-established decision rules.

Table 1: Inputs and steps of the spatial refinement (simplified)
Source: Authors.

	Input data source	Description of procedure	Affected classes
1	CLC change maps	Selected CLC change patches that were not included into CLC2012 map due to generalisation rules were added. Minimum mapping unit (MMU) = 5 ha	All
2	Copernicus HRLs (Forest, Wetlands, Water)	A threshold of 50% was applied to the pixel values (i.e. the respective class must account for the majority of the pixel to be considered). MMU = 5 ha	31x, 41, 51x
3	TomTom MultiNet land use layer	The polygons were rasterized; a look-up table was used to establish the relationship between the TomTom and CLC nomenclatures. MMU = 1 ha	121, 122, 123, 124, 141, 142
4	European Settlement Map	Pixels overlapping non-residential artificial classes were excluded, as were the pixels under minimum building density threshold (empirically derived value of 5%). MMU = 1 ha	11x
5	Urban Atlas (UA) 2012	The polygons were rasterized; decision matrix (CLC class vs UA class) was used to establish the final classification of overlapping pixels. MMU = 1 ha	All
7	Linear features (roads, rivers)	The inclusion of linear features observed less restrictive thresholds of within-pixel cover to preserve the contiguity of these features (given their distinct function and importance in structuring and fragmenting the territory).	122, 511

Figure 1: Data processing chain



Source: Authors.

2.2 Thematic refinement

The largest of the artificial classes, urban fabric (UF), ensued from the spatial refinement as a heterogeneous mixture including settlement types ranging from cities to very low density and isolated rural settlements. It comprised two CLC urban classes, six UA classes and areas extracted from ESM. We split the UF into four consistently defined classes by applying ESM-based density intervals: UF dense (>50% built-up), UF medium density (30-50% built-up), UF low density (10-30% built-up), and UF very low density and isolated (<10% built-up).

Further classes were defined representing sport and leisure built-up facilities (using ESM) as well as for airport and ground transport terminals (using OpenStreetMap polygons).

The key task was, however, to break down the CLC class 'Industrial and commercial units' (ICU) that also comprises public and other facilities. We aimed at untangling it into three subclasses, matching broad economic sectors (the NACE classification, Eurostat 2008): 'Production facilities' (sectors ABCDE), 'Commercial/service facilities' (GHIJKLMN) and 'Public facilities' (OPQ). This breakdown enables to link the classes to employment and other sectoral statistics.

First, we segmented the ICU pixel-clusters into smaller, more homogenous segments by road network. Second, we labelled a subset of the segments based on intersection with semantically matched ancillary land use data where available. For example, the 'Commercial/service facilities' were matched with OSM polygons tagged as landuse=commercial, landuse=retail, building=office, shop=*, amenity=bank office=*, and so forth. TomTom land use polygons and a

Table 2: The target legend of the artificial classes expanded into fourth level (grey – thematically enhanced classes) Source: Authors.

CLC1	CLC2	Level 2 Label	CLC3	Level 3 Label	CLC4	Level 4 Label
1 Artificial surfaces	11	Urban fabric	111	Continuous urban fabric	1111	Urban fabric dense
			112	Discontinuous urban fabric	1121	Urban fabric medium density
					1122	Urban fabric low density
					1123	Urban fabric very low density / isolated
	12	Industrial, commercial and transport units	121	Industrial and commercial units	1211	Production facilities
					1212	Commercial/service facilities
					1213	Public facilities
			122	Road or rail networks and associated land	1221	Road/rail networks and associated land
					1222	Major stations
			123	Port areas	1231	Port areas
			124	Airport areas	1241	Airport areas
					1242	Airport terminals
	13	Mines, dumps and construction sites	131	Mineral extraction sites	1311	Mineral extraction sites
			132	Dump sites	1321	Dump sites
			133	Construction sites	1331	Construction sites
	14	Artificial vegetated non-agricultural areas	141	Green urban areas	1411	Green urban areas
			142	Sport and leisure facilities	1421	Sport and leisure green
					1422	Sport and leisure built-up

compilation of (sub)-national land use data for Spain, Portugal, Lombardy and Wallonia were similarly used. In the end, around one third of the total 740,000 segments obtained a label. We used 70% of the labelled segments for training while keeping 30% aside for testing of the model.

Next, we characterised the segments by features (predictor variables) obtained through spatial analyses performed on a large volume of geodata: millions of points of interest (POI), transport infrastructure, population and building density. The most important features were those based on the POI, separated into three categories corresponding to the target ICU subclasses. We measured per category the frequency of POIs in the interior, as well as kernel POI density in the neighbourhood of each segment.

We ran multiple instances of random forests classification algorithm, while exploratively tuning several hyperparameters, including the size of training sample. The best fitting model was used to predict the missing labels.

Finally, we compared the result to an independent, random sample of 600 segments, interpreted using street level photos. In many cases, there were two or three classes mixed in a single segment. Therefore we sharpened the definition of ground truth as the perceived *dominant* group of NACE sectors.

3 Results and discussion

Figure 2 displays examples of CLC2012 (A, C) and the integrated map (B, D) in full thematic resolution. The selected areas comprise various settlement types from metropolitan to dispersed rural ones. The highest level of refinement is evident in the urban area. The LULC texture of cities is, in reality, more fine-grained compared to exurban/rural areas, and UA data captures well this detail. Though, the UA covers only the larger urban zones (LUZ) that comprise only ~15% of the study area.

For comparison, west half (approximately) of each transect is located outside of LUZ; where only the artificial classes, forest, wetlands and water were refined (the latter three only at the inferior spatial detail of 5 ha). Despite that, the LUZ/non-LUZ dichotomy is not markedly visible, which is an important aspect of the map's cartographic quality

The employed random forest classification was able to correctly predict the reference land use class in over 87% of the testing sample (Kappa 0.72). A comparison with the independent validation sample has shown that the reference labels were noisier than expected (78.5% accuracy, 0.64 Kappa), when compared to more strictly defined ground truth. As a result, the accuracy of the ICU classification in the final map was lower too (74%, Kappa 0.53, Table 3), with a high omission rate of 'Commercial/service facilities' (61% of the validation plots from this class were misclassified as 'Production facilities').

We attribute the high confusion of this class pair to two underlying factors: a) discrepancy between the morphological and functional (sectoral) notions of the term industrial b) actual co-occurrence of industry ('production facilities') and storage, distribution, logistics and wholesale ('Commercial/service facilities'). On the other hand, the 'Public facilities' attained better accuracy, perhaps thanks to less ambiguous semantics (commission error 16%, omission error 20%). Importantly, the thematic detail attained at the CLC level 4 is always nested in the respective ICU class, thus preventing classification errors from propagation to upper levels of the nomenclature.

Table 3. Error matrix of the ICU classification included in the map, compared to an independent sample.
Source: Authors.

Overall accuracy: 74.0% Cohen's Kappa: 0.53		Prediction 1211	1212	1213	Total reference	Omission error
Reference	1211	232	16	9	257	9.7%
	1212	73	40	5	118	66.6%
	1213	13	5	73	91	19.8%
	Total predicted	318	61	87	466	
	Commission error	27.0%	34.4%	16.1%		

3.1 Limitations

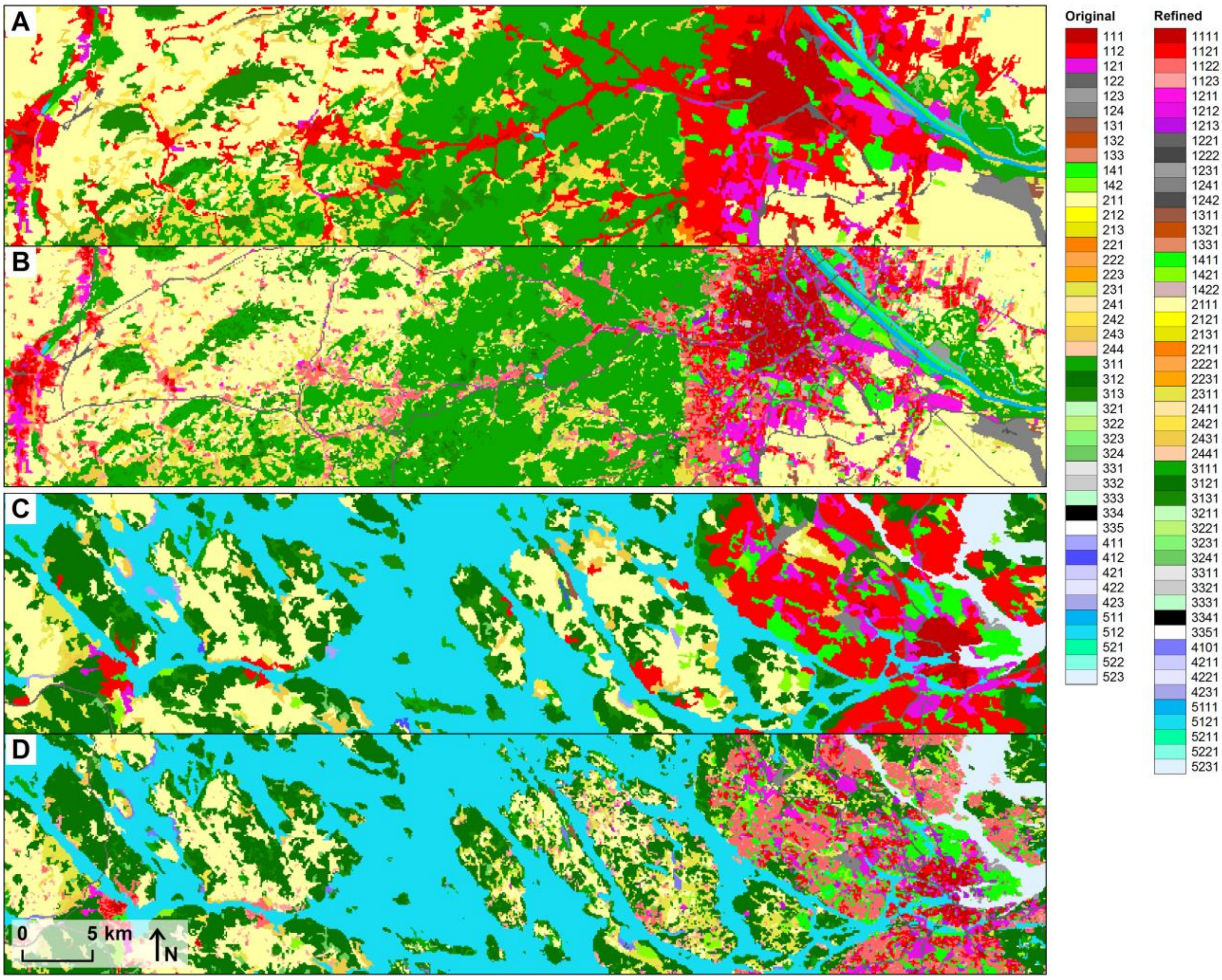
Based on visual examination of the results, we recognise some limitations that pertain mostly to the spatial refinement phase. We combined the input datasets as provided, based on mostly automatized processing and decision rules. Apart from the ICU class, we did not evaluate the accuracy of the map systematically. Here it needs to be noted that the increase of spatial and thematic detail does not imply that increase in classification accuracy compared to CLC2012 was achieved.

Nevertheless, the resulting map inherits classification accuracy of the respective input dataset and a large majority of the map relies on input data with documented quality (UA, CLC, HRLs, and ESM).

Some errors might have arisen from the applied decision rules. Extracting urban fabric from the ESM at 5% threshold performed well in selected test-sites, but there might be no single universal optimum. For instance, arid ecoregions may have positively offset ESM values due to sparse vegetation (leading to prevalent commission errors at the given threshold), while in different ecoregions the opposite might hold true.

In multiple stages of the methodology, we assumed a semantic link between categories in input data and CLC nomenclature. The CLC definitions (Bossard et al. 2000) are very specific and oriented towards visual interpretation of texture and context in medium resolution imagery. Such

Figure 2: Comparison of original CLC (A, C) and the final result (B, D). A and B show Vienna with hinterland, C and D show Stockholm with hinterland).



Source: Authors.

definitions are difficult to be closely imitated using other methods. For instance, it contains a group of heterogeneous agricultural classes, whose meaning starts to erode with the increasing spatial resolution. Therefore the link between various notions of LULC classes is not always exact, but a plausible solution can be achieved. Generally, the likelihood of confusion is higher among semantically similar classes (e.g. subclasses of the same level 2 class).

3.2 Further research

Introducing the UA and TomTom data significantly increased the average levels of spatial detail, but also reduced the spatial consistency. UA covers only the LUZ areas, and TM completeness varies across countries. OpenStreetMap (OSM) polygon data could be leveraged to reduce these gaps. Literature suggests that despite being collected by volunteers, OSM data have a great value for LULC analysis (Dorn, Törnros & Zipf 2015) and that producing LULC maps from OSM is worthwhile (Fonte et al. 2017, Schultz et al. 2017). Our experience was similar – the data proved to be valuable for training of the machine learning model and the mapping of transport terminals.

Additionally, deriving a more detailed classification of economic activities would be useful for addressing several policy relevant topics. Due to the mixed patterns, delineation of narrower sectors might require either finer segmentation of the target of the ICU areas, per-pixel classification or data on the dominance ratio per each sector using an alternative data model instead of the categorical raster dataset.

4 Conclusion

European data on land use/land cover for the year 2012 have been either insufficiently detailed (CLC), patchy in coverage (UA, TomTom), or thematically restricted (ESM, HRLs). To increase the spatial resolution, we integrated several more detailed datasets with CLC using the refinement approach of Batista e Silva et al. (2013) adapted for the currently available data. The thematic refinement of ICU presents an experimental method, supported by a comprehensive data on economic activities in POI form (for explanatory variables) and ancillary land use polygons (for model training and testing). The employed random forests model performed with good overall accuracy, although per-class accuracy varied. Both user's and producer's accuracy for 'Public facilities' exceeded 80%, on the other hand there was high omission rate of 'Commercial/service facilities' that were more often than not classified as 'Production facilities'. Nevertheless, the resulting integrated continental-scale layer overcomes several shortcomings of its constituent datasets. It will become a publicly available asset with a potential to underpin diverse research in topics and at scales that are relevant for policy makers. Currently, it is being used as one of ancillary layers supporting fine-resolution spatiotemporal modelling of the EU's population distribution in the ENACT project (Batista e Silva et al. 2018) as well as in European-scale territorial modelling platform LUISA (Jacobs-Crisioni et al. 2017).

Disclaimer

The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

References

Batista e Silva, F., Lavalle, C. & Koomen, E. (2013). A procedure to obtain a refined European land use/cover map. *Journal of Land Use Science*, 8(3), 255–283.

Batista e Silva, F., Rosina, K., Schiavina, M., Marin, M., Freire, S., Ziemba, L., Craglia, M. & Lavalle, C. (2018). From place of residence to place of activity: towards spatiotemporal mapping of population density in Europe. In *Proceedings of the 21st AGILE conference*.

Ben-Asher, Z., Gilbert, H., Haubould, H., Smith, G., & Strand, G. (2013). HELM-Harmonised European Land Monitoring: Findings and Recommendations of the HELM Project. Tel-Aviv: The HELM Project.

Bossard, M., Feranec, J. & Oťaheľ, J. (2000). *CORINE land cover technical guide – Addendum 2000*. Technical report 40. European Environment Agency, Copenhagen.

Dorn, H., Törnros, T. & Zipf, A. (2015). Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. *ISPRS International Journal of Geo-Information*. 4(3), 1657–1671.

Eurostat. (2008) *NACE Rev. 2 – statistical classification of economic activities in the European Community*. Luxembourg: Publications Office of the European Union.

Ferri, S., Syrris, V., Florczyk, A., Scavazzon, M., Halkia, M. & Pesaresi, M. (2014). A new map of the European settlements by automatic classification of 2.5m resolution SPOT data. In *2014 IEEE Geoscience and Remote Sensing Symposium*, Quebec, 2014, pp. 1160–1163.

Fonte, C., Minghini, M., Patriarca, J., Antoniou, V., See, L. & Skopeliti, A. (2017) Generating Up-to-Date and Detailed Land Use and Land Cover Maps Using OpenStreetMap and GlobeLand30. *ISPRS International Journal of Geo-Information*, 6(4), 125.

Hazeu, G., Büttner, G., Arozarena, A., Valcárcel, N., Feranec, J. & Smith, G. (2016). Detailed CLC Data: Member States with CLC Level 4/Level 5 and (Semi-)Automated Solutions. In Feranec, J., Soukup, T., Hazeu, G., Jaffrain, G. (Eds.). European Landscape Dynamics. Boca Raton: CRC Press.)

Jacobs-Crisioni, C., Diogo, V., Perpiña Castillo, C., Baranzelli, C., Batista e Silva, F., Rosina, K., Kavalov, B. & Lavalle, C. (2017). The LUISA Territorial Reference Scenario 2017: A technical description, Publications Office of the European Union, Luxembourg, 2017.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J. & Pereira, F. C. (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53, 36-46.

Pazúr, R. & Bolliger, J. (2017) Enhanced land use datasets and future scenarios of land change for Slovakia. *Data in Brief*, 14, pp. 483–488.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., ... & Syrris, V. (2016) Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. Luxembourg: Publications Office of the EU.

Schultz, M., Voss, J., Auer, M., Carter, S. & Zipf, A. (2017) Open land cover from OpenStreetMap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 63(July), pp. 206–213.