MULTI-SOURCE APPROACH FOR ENHANCED LUCAS STATISTICS: A PILOT STUDY IN PORTUGAL[1]

Name (s) of author(s):

Diana Almeida[1], Filipe Marcelino[2], Francisco Gutierres[3], Pedro Campos[1], Francisco Vala[1], Mário Caetano[2]

[1]Statistics Portugal (PORTUGAL)

diana.almeida@ine.pt, francisco.vala@ine.pt, pedro.campos@ine.pt

[2] Directorate-General for Territory (PORTUGAL)

fmarcelino@dgterritorio.pt, mario.caetano@dgterritorio.pt

[3]Eurecat - Technology Centre of Catalonia, Big Data Analytics Unit (SPAIN)
francisco.sacramento@eurecat.org (formerly in Statistics Portugal)

Organization: **Statistics Portugal, Directorate-General for Territory, Eurecat**

1. Introduction

The European Statistical System (ESS) results from a sustainable partnership between the Community Statistical Authority, composed by the Eurostat, the National Statistic Institutes of the Member States and other national authorities which are responsible for the development, production and dissemination of European statistics. The ESS has a continuous mission of providing comparable statistics at EU level, harmonizing procedures of acquisition and dissemination.

Under the scope of the ESS objectives and mission, the LUCAS GT 2015 project was developed to produce harmonised, quality-assured, land cover land use (LCLU) information according to a predefined classification and with a given precision (NUTS level 3). This project accommodates the feasibility of future updates based on a National Integrated Approach to produce LCLU information to comply with the European Statistical System (ESS) medium term strategy for LCLU statistics. The model of close cooperation and follow-up between Statistics Portugal (INE) and Directorate-General for Territorial Development (DGT), through a

[1] This paper was developed under the Project LUGAS GT 2015, No. 08441.2015.002-2015.724, supported by Eurostat, European Commission.

Memorandum of Understanding-MoU, and articulated with the National Committee for LCLU Mapping (CACTO) will ensure the envisaged National Integrated Approach to produce LCLU information centred on the National LCLU Map (COS) guaranteeing coherence and comparability, namely between statistical and geographical LCLU data. Additionally, another main goal of LUCAS GT 2015 is the development of methodologies for LCLU mapping based on satellite image processing and spatial analysis within GIS, in which this paper will focus on.

This paper aims to present a methodological study centred in LCLU statistics developed in Portugal, following the nomenclature adopted in the LUCAS project, at a scale of NUTS3 level for the year 2013, relying in COS2010 as main reference. It is going to be applied to a testing area, comprising fully three NUTS3 and partially eight sub-regions. The methodological focus is centred on the application of Remote Sensing techniques, satellite imagery and ancillary data to derive LCLU statistics harmonized with LUCAS GT 2015 nomenclature.

This methodology finds a wider application because it is grounded on the assurance of an accurate and efficient process to derive land cover maps and therefore can be disseminated to other Member States. This study reveals to be of great value due to multi-source integration, which ultimately will be a crucial component for developing Small Area Estimation (SAE) methodology. It is expected that the SAE methodology contributes to enhance LCLU desegregation. This process allows a continuous updating and harmonized LCLU data.

2. National Data Sources for ESS

The classification system adopted by each national producer emphasizes different aspects of LC and LU information, related to its specific requirements and scope. These aspects result on different classification systems and nomenclatures, several spatial and temporal resolution and accuracy. This section presents the PT data sources that have a precise LC and/or LU classification scheme. The analysis focused on the main features of the adopted national classification systems, taking into account the purposes of the LUCAS project. The following characteristics were examined: a) the classification system; b) the minimum mapping unit and spatial resolution; c) the time coverage, periodicity and geometry type. Following these characteristics and owing to a well-defined classification system, COS was identified as being the data source which is able to provide improved thematic compliance with LUCAS.

It is important to note that COS is the official LC and LU Map of Portugal Mainland. It is a large-scale product with a high level of thematic and positional accuracy, harmonised with global and European policies related with geographic data quality and standards. Other National Data will complement and support the LC and LU information that is provided by COS.

Other National Data function as complement for the LCLU information provided by COS: the National Forest Inventory (NFI) is the official survey system of Portugal Mainland on Forestry domain, providing information on forests and woodlands; the Land Parcel Identification System (LPIS) comprises an identification system for agricultural parcels, and a data base on cropland. More specific data such as quarries (mineral masses), mining concession, road network and the railway network, the National System of Water Resources Information and the national mapping of burned areas are also a complement of LCLU information.

The nomenclature of COS2010 is compatible with the nomenclature of CORINE Land Cover (CLC), which constitutes a reference LCLU product at European level. As a result of this matching effort, there are similarities within the first three levels of COS2010 nomenclature and the CLC nomenclature. COS2010 is based on an a priori and 5-level hierarchical system and has 225 classes at the most detailed level, whereas at the most general level there are only five classes, namely: Artificial areas; Agricultural and agro-forestry areas; Forests and natural and semi-natural areas; Wetlands; and Water bodies. The nomenclature of COS2010 is embedded in an a priori and hierarchical classification system. The fact that it is a priori system means that its LCLU classes are abstractions of the reality and that all possible categories and combinations of categories are generated prior to any data collection or photo-interpretation. The fact of being a hierarchical system means that the classification is able to accommodate different levels of information, which is achieved by simultaneously having classes that describe general characteristics, and classes referring to more specific details.

At the more general levels, few classification criteria are used, whereas for the more specific classes, a higher number of classification criteria are applied. These more specific classes are a result of the sub-division of the more general classes, and at each level the classes are exhaustively and mutually exclusive. The nomenclature of COS2010 encompasses five levels of thematic detail.

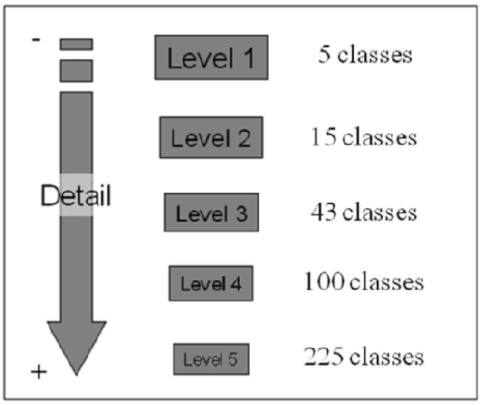

Figure 1 – Hierarchical levels and number of LCLU classes in each level in COS2010 nomenclature

2.1 Potentialities of COS2010 when compared with LUCAS

To evaluate the compliance and limitations of the national classification (COS) when compared with LUCAS2012 attributes and LUCAS GT 2015 nomenclature and thematic scope, a comparability analysis has been performed.

The objectives of the comparison with LUCAS2012 relied on simultaneously compare COS2010 nomenclature with the LUCAS2012 attributes, aiming to establish an equivalence between both nomenclatures, and afterwards, validate the COS polygons classification, using the LUCAS2012 field work (including *in situ* photography's).

Regarding the LUCAS2012 project in mainland Portugal, there are 22 257 points, from which, 7 336 (32%) are field work points. The analysis focused only on corresponding the field work point classification and respective attributes with the COS classifications originating a reclassification of the LUCAS points to a COS classification. This resulted in 47% of points in agreement and 53% of points in disagreement between both products. The work of validating the disagreements (53%) within COS polygons revealed different results: 1) both products were correctly different (49%), based on their nomenclature differences and date; 2) situations where COS2010 was incorrect (29%); 3) situations where LUCAS2012 was incorrect (20%); 4) situations were both products were incorrect (2%). Following this analysis the COS polygons with error were corrected to LUCAS2012 classification.

Regarding the LUCAS GT 2015 LC nomenclature a correspondence with the 225 classes of COS can be identified in Table 1. Only the class Ocean (5.2.3.01.1 in COS) was not comparable, which leads to a correspondence of just 224 classes.

Table 1 – Comparability between LUCAS GT 2015 Land Cover classification and COS 2010

LUCAS GT 2015 LC Precision		**Number of COS 2010 Classes per LUCAS Level**
Level 1	**Level 2 and 3**	
A. ARTIFICIAL LAND	A10. Roofed built-up areas	13
	A20. Artificial non built-up areas	9
	A30. Other built-up areas	6
B. CROPLAND	B. CROPLAND	38
C. WOODLAND	CF10. Broadleaved forest	49
	CF20. Coniferous forest	21
	CF30. Mixed forest	25
	COLT. Other land with tree cover	34
D. SHRUBLAND	S. Shrubland	5
E. GRASSLAND	LCE. Permanent grassland	2
F. BARELAND, LICHENS, GLACIERS AND PERMANENT SNOW	F10. Rocks and Stones	4
	F20. Sand	2
G. WATER	G10. Inland water bodies	7
	G20. Inland running water	3
H. WETLANDS	H. WETLANDS	6
Total classes COS		224

Table 1 show an optimal agreement between the two data sets in the LC data for the more detailed level. However, it should be noted that there are situations where the correspondence could be two ways. There are classes that could be considered not totally correspondent to just one classification – i.e the construction sites (COS class 1.3.3.01.1) are present in every "Artificial Land" (level 1), therefore they could be considered either in roofed built-up areas or in artificial non built-up areas depending if there are buildings under construction or if it's the construction of a road. It has been decided to consider a correspondence with Roofed built-up areas (A10) the most valid based on the argument that it's a more frequent situation of construction (1.3.3.01.1).

3. Methodology and technical solutions

The methodological approach (Figure 2) for statistical results assures full integration with the national LCLU Map (COS2010), acquired through Remote Sensing techniques. A pixel-based and multi-temporal satellite data (Landsat imagery 5 and 7) is being simultaneously used. Pixel-based technique allows improved understanding of the infra and inter annual changes on LCLU and as a result, achieving training areas by overlapping with COS2010. The mapping procedure to derive the Land Cover Map 2010 (LCMap 2010) consists in a supervised classification approach, such as the Support Vector Machine (SVM).

To produce an improved Land Cover Map 2010 (iLCMap 2010), previously produced LCMap 2010 will be integrated with COS2010 at level 5 detail and 1ha MMU. This operation of spatial analysis allows to evaluate the results of the LCMap 2010, and to provide thematic enrichment (detail) to the present classification.

A Landscape Change Detection methodology has been also developed to update iLCMap 2010 to iLCMap 2011 and from year to year till iLCMap 2015. Taking the reference year of 2011 as example, the previous produced iLCMap 2010 is going to be used and combined with the LC change from the images 2010 and 2011.

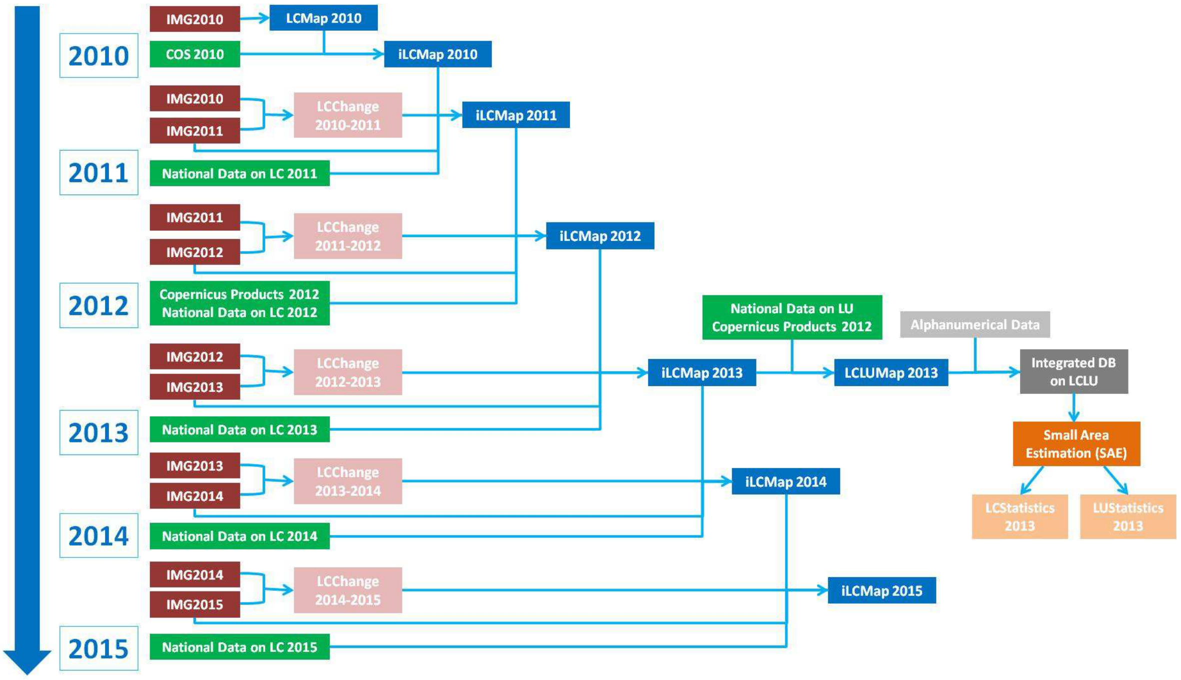

Figure 2– Complete sequence of the methodological workflow

3.1 Testing Area (Pilot study)

The testing area covers central-northern Portugal, including entirely NUTS3 of Region of Aveiro; Viseu-Dão Lafões and Region of Coimbra, and partially the Metropolitan Area of Porto, Médio Tejo, Region of Leiria, Alto Alentejo, Oeste, Tâmega-Sousa, Beira Serra da Estrela and Douro (Figure 3).

The selection of the study area has been based on the prior knowledge of the LCLU class characteristics and Landscape diversity of the region. This study applied supervised classification-SVM algorithm in ERDAS imagine/ENVI to produce a LC map statistics and to detect LCLU changes, observed in the selected study area using multispectral satellite data (Spring and Summer) obtained from Landsat 5 for the year 2010 and Landsat 7 for 2013 respectively. The images are composed by 12 bands, resulted from a spatiotemporal image-fusion model to enhance the temporal resolution of Landsat. This part of the process is essential to the classification´s accuracy.

The goal of this investigation is to use Remote Sensing data in a GIS framework in order to integrate LULC changes, through the analysis of normalized time-series, in the LC map statistics process.

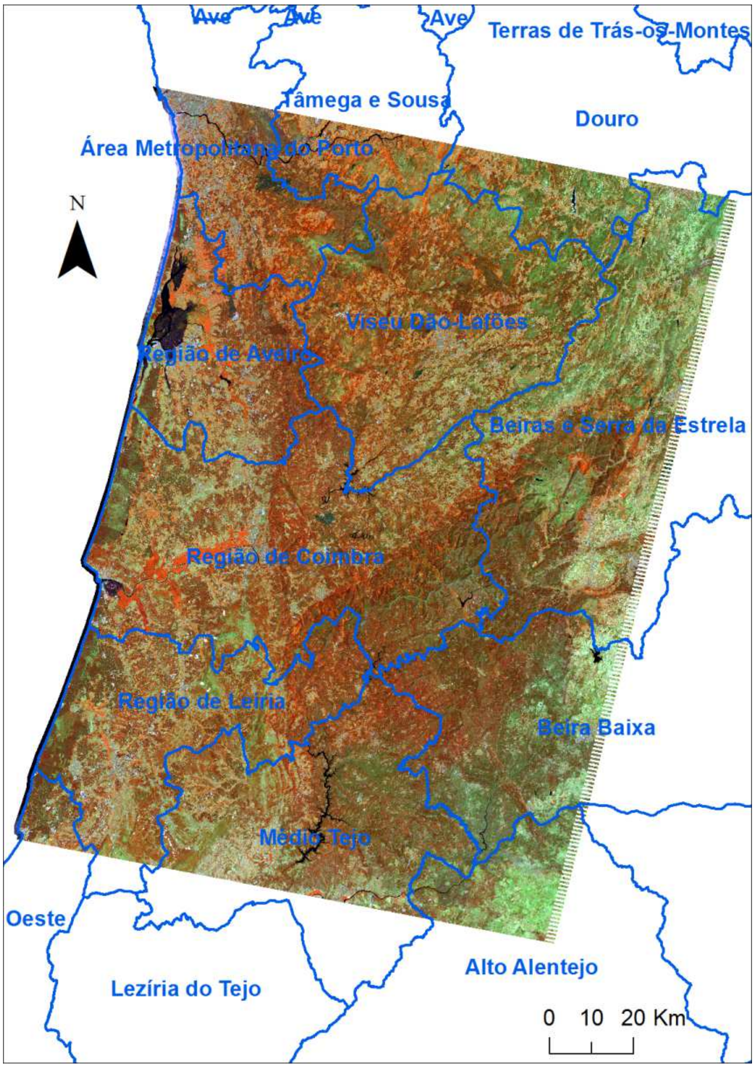

Figure 3 – Location of the study area.
Landsat 5 image (Spring 25th April; Summer 30th July) 2010.

3.2 Remote Sensing

Remote Sensing (RS) is the technique of information acquisition about an object or phenomenon when physical contact with the object is not possible. The Remote Sensing and GIS analytical procedures are characterized for its high universality, from the perspective of data and possible application (described in Gutierres, 2014 and Gutierres *et al.*, 2016). Designers of the method focused on solving the scale problem of the analyzed phenomenon and repeatability of results and emphasized the optimization of calculations, which is significant in cases of large image sets (Baatz & Schäpe, 2000; Lewiński, 2006).

RS provides useful information and tools to identify long term trends and short-term variations, such as impact of rising sea levels and LULC changes (Gutierres, 2014; Bustamante *et al.*, 2013), and can supply complementary information on LC location, limits and extent.

The mapping procedure consists in a supervised classification approach. Here a training sample has to be provided to the classification algorithm that then produces the classification model (Bishop, 2006). The training sample consists in a set of pixels, commonly known as training cases, which are then used as representatives of each land cover class (Hastie *et al.*, 2009). The collection of these training examples is typically done by image interpreters that utilize their image analysis experience and ancillary data, such as ortho-imagery and field work, to define a consistent and representative set of training cases (Foody, 2004; Foody & Mathur, 2004). The proposed approach to minimize the training sampling effort is based on the automatic selection of training pixels using previous land cover information in the format of a vector land cover map produced manually by an official institution (e.g. DGT), here designated as base map. The principle is to use this land cover data to inform a sampling process of pixels from the image. This is implemented by using the polygons present in the base map as strata in a stratified random sampling. This methodology aims to be an accurate and efficient process to derive land cover maps that can be applied in other European Union countries.

3.3 Small Area Estimation (SAE)

The model-based approach of Small Area Estimation (SAE) is as an alternative to design-based approaches for the domains where the sample size is unable to deliver reliable results. Model-based approaches in SAE are obtained by fitting a model to the data, frequently there is used a regression model, in which covariates are used as auxiliary information. This auxiliary information comprises the several sources already available for NUTS3 (i.e. administrative data, national surveys). This procedure usually provides good results even in small samples, because estimation is based on regressions between the variables underlying the model. This is the case of indirect estimators that borrow strength from another areas and/or time periods, in order to increase effective sample size. In addition, mixed models can be used to combine different sources of information and explain different sources of errors.

The NUTS3 data derived from LUCAS estimates, resort of covariates as auxiliary information, mainly using estimated areas from National Data Sources (NDS). To accommodate these processes, a feasibility study is being applied to evaluate data accuracy. Completing these disaggregation processes, it will be possible to obtain a detailed Land Cover classes at NUT3 level.

The application of this method allows collecting information at different levels of detail and combining upper and lower levels of the hierarchical system.

4. Final Remarks

The implementation of this methodology allowed selecting a valid testing area, in order to apply the pilot study. The application of remote sensing and ancillary data to derive LCLU statistics harmonized with LUCAS GT 2015 nomenclature will allow to retrieve LCLU for NUTS3 level. This multi-source integration methodology, focusing RS, satellite imagery and SAE can be applied elsewhere and supplant spatial desegregation, using transversal methods, that comply with INSPIRE Directive, and be harmonized with the procedures of data acquisition and dissemination of the ESS, in the sense that can be continually updated.

It is recommended that this experimental approach may be used for different data set from various new sensor platforms (e.g. Sentinel-2), and provide a standard guideline for specific applications by the Member States.

References

Baatz, M. & Schäpe, A. (2000). *Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation*. In: Strobl J et al (eds) Angewandte Geographische Informationsverarbeitung, XII. Wichmann, Heidelberg, pp 12–23.

Bishop, C. M. (2006). *Pattern recognition and machine learning, information science and statistics*. Springer, Berlin.

Bustamante, J.; Díaz-Delgado, R.; Aragonés, D.; García Murillo, P. & Castellanos E. M. *et al* (2013). *Proyecto HYDRA: aplicación de la teledetección al estudio de la dinámica hídrica y de la vegetación acuática en las marismas de Doñana*. In: Fernández-Renau González-Anleo A, de Miguel Llanes E (eds) Teledetección: Sistemas Operacionales de Observación de la Tierra. XV Congreso de la Asociación Española de Teledetección (AET). Torrejón de Ardoz, Madrid, España, 22–24 Oct 2013.

Foody, G. M. (2004). Supervised image classification by MLP and RBF neural networks with and without an exhaustively defined set of classes. *International Journal of Remote Sensing*, 25 (15), 3091–3104.

Foody, G. M. & Mathur, A. (2004). Toward intelligent training of super- vised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment*, 93 (1-2), 107–117.

Gutierres, F. (2014). *Structure and dynamics of habitats and landscape of Sado Estuary and Comporta/Galé Natura 2000 Sites – A contribution to sustainable land management and ecological restoration*. Ph.D. dissertation, Institute of Geography and Territorial Planning, University of Lisbon.

Gutierres, F.; Teodoro, A. C.; Reis, E. & Neto, C. (2016). *Remote Sensing Technologies for the Assessment of Marine and Coastal Ecosystems*. In: Seafloor Mapping along Continental Shelves: Research and Techniques for Visualizing Benthic Environments. Charlie Finkl & Chris Makowski (Ed.), Coastal Research Library (CRL), 13, Springer (Dordrecht, The Netherlands), 293 pp. DOI 10.1007/978-3-319-25121-9.

Hastie, T.; Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd edition*. Springer-Verlag. 763 pages.

Lewiński, S. (2006). *Applying fused multispectral and panchromatic data of Landsat ETM+ to object oriented classification*. In: EARSeL (ed) New Developments and Challenges in Remote Sensing 26th EARSeL Symposium.Poland, Warsaw, pp 233-240.