The Land Cover/Use Code of the new Istat Census cartography[1]

Stefano Mugnoli, Alberto Sabbi, Fabio Lipizzi

1. Introduction

The renewal process of the Italian National Institute of Statistics (ISTAT) provides the data production through the new Integrated Statistical Registers system (SIR). One of the four SIR registers is the Base Register of the Site (RSLB) that will make it possible to uniquely locate all SIR information. For this reason, ISTAT has planned the implementation of the enumeration areas layer called "microzones". Therefore, the new microzones layer constitutes the base map to realize the new Census Maps, which represents the reference layer to disseminate SIR data and information (Mugnoli et al., 2018). This paper aims to briefly set out the methodology used to realize the new ISTAT microzones and enumeration areas layers; some legend details will be provided in order to better understand the way in which each polygon is represented on the map. The name 'microzones' is related to the fact that the layer is a further subdivision of the ISTAT enumeration areas layer; the latter is divided into very small polygons, homogenous in their Land Cover (LC) and Land Use (LU) aspects; this creates a kind of a plot made up of many micro-areas.

The ISTAT census enumeration areas vector layer, in fact, represents the cornerstone to analyse the Italian territory from a statistical point of view. All the data collected during census surveys are linked to each of about 740.000 enumeration areas drawn on Italy. This dense plot helps us to describe the entire national territory in a very detailed way, particularly in urban areas.

Therefore, in order to improve LC/LU statistics and to better characterize each enumeration area, the ISTAT ATA (Environment Territory) Unit, planned to produce a sort of a microzones mosaic layer described by a land cover/use definition compatible with the LUCAS (Land Use/Cover Area frame Survey) legend. This certainly allows both to define more clearly the homogenous areas contour for the future and to optimize the geo-localization of all census variables. With regard to the above, it is important to remember that again this year ISTAT has just planned a continuous population Census survey. It is thus fundamental to have a very detailed reference cartography for this survey.

Census geographical datasets are essentially used for classifying and characterizing national territory in relation to resident population, buildings, services and industry. Supplementing this information with land cover and land use data, it can be possible not only to produce comprehensive data on land cover/use, but better to calculate some statistical parameters (i.e. population density by masking all the uninhabitable or uninhabited areas) at local and global level too. But not just that: in fact, statistical information at this level of detail can be used to evaluate other important phenomena like soil consumption, urban sprawl (European Environment Agency, 2006), accessibility to territory and the demographic change in population distribution. In short, our product can be considered a sort of "Land Cover/Use (LC/LU) Synthetic Layer", in the sense of getting together geo-statistical information derived from many different geographic datasets. Therefore, its main use is to support statistical surveys since it is the result of integration and harmonization of different kinds of thematic archives such as administrative, demographic, infrastructure (road, railway, ports, airport, etc.), agricultural Census data and environmental maps.

[1] Even if the paper was devised together by the Authors, F. Lipizzi wrote paragraph 1. 'Introduction'; S. Mugnoli wrote paragraphs 2, Microzones and Enumeration areas LC/LU Legend and References and paragraph 5 'Future update'; A. Sabbi wrote paragraph 3. 'Topology rules and accuracy assessments' and paragraph 4. .'Conclusions'

Stefano Mugnoli, ISTAT, Italian National Institute of Statistics, Italy, mugnoli@istat.it
Alberto Sabbi, ISTAT, Italian National Institute of Statistics, Italy, sabbi@istat.it, 0000-0002-3215-6400
Fabio Lipizzi, ISTAT, Italian National Institute of Statistics, Italy, lipizzi@istat.it

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Moreover, the peculiar legend of the map is undoubtedly useful in better understanding the synthesis process. In Italy CISIS[2] (Centro Interregionale per i Sistemi Informatici Geografici e Statistici) has contributed to harmonising geographical and statistical data. One of the most important results is the release of the database "DB Prior 10K" at national level. The database developed by CPSG (Comitato Permanente per i Sistemi Geografici), provides some layers (i.e. streets, railways, hydrography) with the same data structure. Furthermore, in order to implement the INSPIRE[3] directive, the 'Consulta Nazionale per l'Informazione Territoriale e Ambientale (CNITA) was established.

Therefore, to align with from the above, every geographical ISTAT data is designed to pursue the same purpose: to provide standardised information for the entire national territory.

The final geo-statistical microzones layer was developed through collaboration of many people and after the review of many different intermediate products. In the end, the activity is the sequel of many ISTAT experimentations (Lipizzi and Mugnoli, 2010; Chiocchini and Mugnoli, 2014; Mugnoli et al., 2011; Lombardo et al., 2017).

2. Microzones and Enumeration areas LC/LU legend

ISTAT enumeration are as are described by a lot of attributes that identify each polygon from an administrative and statistical point of view. There are some codes that can be useful to frame each area in a sort of LC/LU classification. Since 2011 each enumeration area had been identified according to a key related to its main "vocation". This sort of legend was focused especially on human activities, uses or services for the citizen.

Having considered the need to define a clear and useful LC/LU legend to uniquely describe the entire national territory, the choice has fallen upon LUCAS (Land Use and Cover area frame Survey) because this is a "*survey that provides harmonized and comparable statistics on land use and land cover across the whole of the EU's territory*"[4]. And not just for this reason, the microzones and enumeration areas class legend has been based on the LUCAS one because it is based on two LC and LU pure legends; moreover, all the map layers at our disposal make it possible to identify each polygon by a LUCAS class. Upon completion of the two layers description, it is easy to transfer the classification to the microzones layer since the latter is a sort of summary of the former. The first draft provides a 45 LC class, mostly at LUCAS level 1. But classifying each microzone is not always simple, especially in the case in which polygons can be referred to LU rather than to LC. For example, it is very difficult to characterize the "green urban areas" on the basis of the LC pure legend, as LUCAS is. Usually green areas are classified on the basis of their use (i.e. amusement parks, community gardens, etc.). Attempts have been made to separate grasslands and woodlands from "green artificial" ones. So, in these cases a specific code, which comes from the fusion by LUCAS LC and LU codes, was created and named *COD_MZ* for the microzones layer; then each enumeration area has been identified by a single code *COD_TIPO_S* that represents a simplification of the *COD_MZ*.

3. Topology rules and accuracy assessment

When different geographical databases are merged into a unique layer, some overlay errors inevitably occur. It is therefore essential to define very strict topology rules upstream. First of all, you have to decide the overlay order of the layer. In our case, in addition to enumeration areas cartography, the basic layer is represented by water (river, lake, lagoons, etc.) and wetlands; above this, railways, streets and buildings in this order; then, agricultural and natural area layers; finally, and if it's possible, the polygons derived from the vegetation indices calculated starting from

[2] For more information regarding CISIS activities: http://www.cisis.it/

[3] https://www.mite.gov.it/pagina/inspire

[4] For more information regarding LUCAS survey: http://ec.europa.eu/eurostat/statistics-explained/index. php/LUCAS_-_Land_use_and_land_cover_survey.

ortophotos.

Of course, in so doing, it is necessary to deal with the overlay areas (bridge, road crossings, etc.). Using some simple ArcGIS© 10.7.1 by ESRI analysis algorithms (Intersect and Symmetrical difference), (Law and Collins, 2018; Bolstad P., Manson, 2019), different layers can be merged automatically without topology errors.

It is only thanks to the fact that the topology is correct that it is possible to evaluate the land cover of each class. In Table 1 is shown a summary of land cover surfaces for each Italian region (in percentages) related to the LUCAS legend at level 0.

X,Y[5] tolerance is set at 1m, the same as the enumeration area layers.

An additional benefit in using the LUCAS legend is the possibility to assess the accuracy of the microzones layer by LUCAS points themselves. Class accuracy varies from 72.02% for the woodland to 33.33% for the grassland.

The real problem is due to the number of LUCAS points of the less represented classes. In our case, for example, we have very few points for the "Bare land and lichens/moss" and for the "wetlands". Moreover, it is clear from the error matrix that there are clear overlaps between natural grasslands (pastures) and agricultural ones.

The microzones layer is completed for all Italian regions and it is now in the pipeline to transfer information to the Census 2021 enumeration area layer.

In Figure 1 a focus on the Census 2021 enumeration layer (Municipality of Florence) at the second LUCAS level; different colours represent different LC classes.

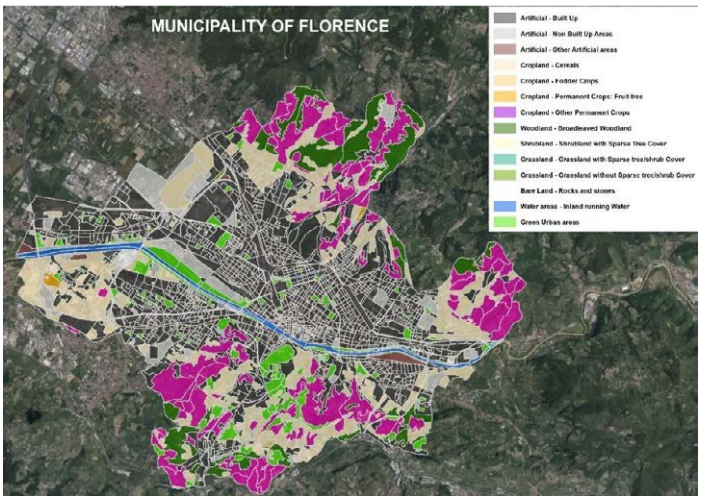
Table 1 – Summary of land cover surface for each Italian region (in percentages)

Regione	Artificial	Cropland	Woodland	Bare Land	Grassland and Shrubland	Water and Wetlands	Green Urban Areas[6]
Piemonte	8,97	40,72	37,55	7,66	3,98	0,99	0,14
Valle d'Aosta	2,11	4,99	31,13	50,74	1,80	9,18	0,04
Lombardia	14,17	42,54	28,16	9,20	1,77	3,82	0,35
Trentino-Alto Adige	2,95	18,84	54,22	15,33	6,79	1,82	0,05
Veneto	14,12	48,68	24,01	2,98	4,33	5,56	0,31
Friuli-Venezia Giulia	9,84	33,10	40,46	6,32	5,56	4,62	0,10
Liguria	10,64	8,55	74,65	1,51	4,55	0,01	0,10
Emilia-Romagna	8,16	58,46	25,86	0,65	3,92	2,70	0,25
Toscana	6,29	39,48	51,69	1,17	0,75	0,53	0,09
Umbria	4,95	47,72	41,25	0,13	3,97	1,95	0,04
Marche	5,93	54,85	33,58	0,41	4,82	0,30	0,13
Lazio	11,67	46,97	29,03	2,14	8,20	1,52	0,47
Abruzzo	5,36	41,92	31,08	3,48	17,84	0,27	0,04
Molise	3,37	50,30	36,91	1,17	7,80	0,40	0,05
Campania	10,83	51,35	28,16	1,01	7,97	0,51	0,18
Puglia	5,90	78,69	5,37	0,95	7,65	1,18	0,26
Basilicata	2,72	55,10	33,30	0,50	7,19	1,18	0,01
Calabria	5,28	35,33	39,43	1,33	17,30	1,28	0,05
Sicilia	6,43	62,45	8,32	3,25	18,96	0,55	0,03
Sardegna	3,48	40,13	23,52	1,69	30,05	1,11	0,02
ITALIA	**7,16**	**43,01**	**33,88**	**5,58**	**8,26**	**1,97**	**0,14**

[5] The x,y tolerance refers to the minimum distance between coordinates before they are considered equal.

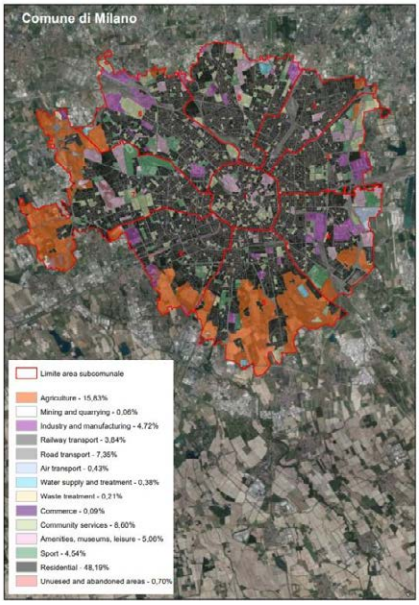
[6] Class not present in LUCAS legend but considered because very important for inhabited localities.

Figure 1 – Municipality of Florence (Enumeration areas 2021) at the second LUCAS level



Another advantage in using LUCAS legend is represented by the opportunity of using a Land Use pure legend too. Below, just as an example, Milan municipality represented on the base of LUCAS LU legend.

Figure 2 – Municipality of Milan (Enumeration areas 2021) according the LUCAS LU legend



4. Future update

From the above, it is clear that the new Census Maps represents a fundamental benchmark for territorial analysis. However, up to now, it is a sort of something static which may lost its original meaning over time.

For this reason, in parallel to the production of the new layer, it is also though to their dynamic update. So, some studies was carried out in this regard.

The principal of these related the inhabitant areas, which are the most important features of the layer was based on the use of deep CCN (Convolutional Neural Network) U-NET.

The U-NET was first used by Olaf Ronneberger O. (Ronneberger O. et al., 2015) in biomedical

image segmentation. The name comes from how the authors arranged their architecture in an image that resembled the letter "U". The model implemented in our project is similar to the original model in architecture but has convolution layers that take in the 8 bands in the tiff files used.

This experimentation was sourced on python with keras. The tiff files contain 8 channels in the ortho data which requires us to define the input layer to accept an input that has dimensions of a patch (2D) times 8. A patch is a spot on the original tiff file that is randomly selected and then undergoes a random transformation to produce an analogous patch which only differs from this original patch by the transformation.

The images are 8-band commercial grade satellite imagery accessed from the SpaceNet dataset. The 8 channels are red, red edge, coastal, blue, green, yellow, near-IR1 and nearIR2. In addition to the training images, there are masks corresponding to these images which contain the true segmentation of these images; they contain information about 5 different classes: buildings, roads, trees, crops and water. The images are 16 bit resolution while the mask files are 8 bit.

The model was trained with a batch size of 10. 400 train images and 100 validation patches were generated from 24 training images with their corresponding masks. While there were only 24 images in the dataset, the code performs six random transformations including mirroring, transpose, and rotation to produce enough patches - this process is called image augmentation and increases the dataset. The validation and training losses are important parameters to understanding the fit of the model. In an ideal situation, in the long run at least, both of these quantities have identical values. If the validation loss is greater than the training loss by a large amount the model overfits; on the contrary, if the reverse happens, it is a case of underfitting.

The output of the test image and its corresponding labelled outputs are presented in Figure 2. The colour coding is as given in table 2. The test files also undergo image augmentation and the final result is the averaged out result of the independent predictions of the transformed images. In addition to the segmented images, the mask of the test image is also returned by the program. In some sense this model can be used as an extension to prepare masks for future training images once it has been perfected to a certain degree of training and validation loss.

Table 2 - Colour Coding of output

Label	R, G, B	Colour
Buildings	150, 150, 150	Gray
Roads	223, 194, 225	Pale Yellow
Trees	27, 120 55	Dark Green
Crops	166, 219, 160	Pale Green
Water	116, 173, 209	Sky Blue

5. Conclusions

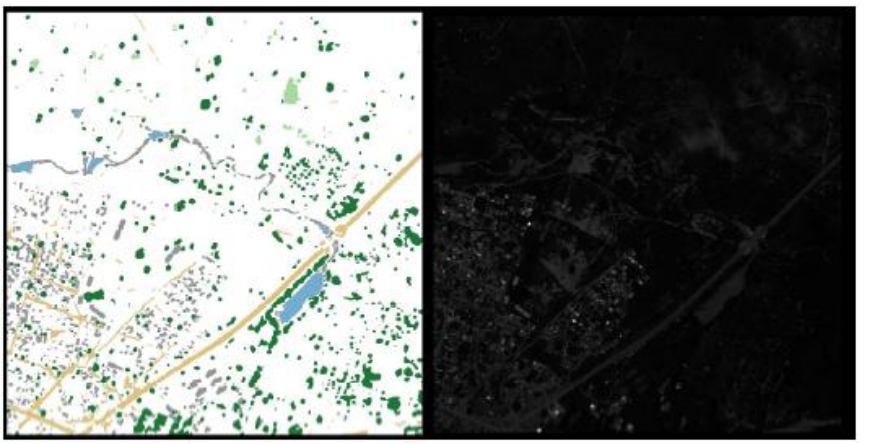
The need to have a homogeneous statistical cartography for the entire national territory is a priority not only for ISTAT but for national and local administrative authorities too. Enumeration areas layer have played a crucial role until now in describing statistical indicators in their territorial and environmental aspects.

However, until now, old enumeration areas layers was not suitable to describe LC and get territorial parameters to some important ISTAT surveys (i.e. agricultural census, transport and services surveys, etc.). So, the new ISTAT microzones and enumeration areas 2021 layer has to be seen as the base map phenomena.

Image processing activities are planned for the future to update all the ISTAT geographic databases, especially by deep learning technics.

The Authors thank all the people of the ISTAT ATA unit who daily works to implement ISTAT geodatabases.

Figure 3 - Sample Output with test image (right)



References

Mugnoli S., Lipizzi F. and Esposto A. (2018). New ISTAT 'microzones' layer: a new way to read land cover statistics. *Journal of Research and Didactics in Geography* (J-READING). 2: 95-104.

European Environment Agency, (2006). *Urban Sprawl in Europe – the ignored challenge*, EEA report n. 10/2006.

Lipizzi F. and Mugnoli S. (2010). Le statistiche agricole verso il Censimento del 2010: valutazione e prospettive. In: *Proceedings of the conference: Le Statistiche agricole verso il Censimento del 2010: valutazioni e prospettive.* Cassino, Università di Cassino, 26-27 October 2006, Cassino, 2010, pp. 381-394.

Chiocchini R. and Mugnoli S. (2014). Land Cover and Census integration geographic datasets to realize a statistics synthetic map. In: *Proceedings of European Forum for Geography and Statistics (Krakow, 22-24 October 2014), European Forum for Geography and Statistics Conference, 2014,* https://www.efgs.info/conferences/efgs/2014 -krakow/.

Mugnoli S., Chiocchini R., Cruciani S., Esposto A. and Lipizzi F. (2011). Integrazione di dataset geografici di copertura del Suolo e Censuari per la realizzazione di una mappa statistica sintetica. In: *Proceedings of the XV National Conference ASITA 2011* (Colorno, 15-18 November 2011), Parma, 2011, pp. 1633-1640.

Lombardo G., Esposto A., Minguzzi R. and Mugnoli S. (2017). La CSS ISTAT un nuovo strumento per le statistiche territoriali. *Geomedia, XXI, 2, pp. 26-30.*

Law M. and Collins A. (2018). *Getting to Know ArcGIS Desktop, fifth edition*, Esri Press.

Bolstad P., Manson S. (2019). *GIS Fundamentals: A First Text on Geographic Information Systems, Sixth Edition*, XanEdu Publishing, Inc.

Ronneberger O., Fischer P, and Brox T. (2015). U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention.* Springer, pages 234–241.