Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

# Data science pipelines @ Istat: challenges and solutions

by Monica Scannapieco, Massimo De Cubellis and Fabrizio De Fausti

Release date: November 5, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

# Data science pipelines @ Istat: challenges and solutions

Monica Scannapieco, Massimo De Cubellis, and Fabrizio De Fausti[1]

## Abstract

In line with the path taken by the European Statistical System, Istat is investing on innovative methods to harness Big Data sources and to use them for the production of new and enriched Official Statistics products. Big Data sources are not, in general, directly tractable with traditional statistical techniques, just think of specific data types such as images and texts that are examples of the Variety dimension of Big Data. This motivates and justifies the growing interest of National Statistical Institutes in data science techniques.

Istat is currently using data science techniques, including machine learning techniques, in innovation projects and for the publication of experimental statistics. This paper will provide an overview of the main current projects by Istat and will focus on two specific Big Data-based production pipelines, related to the processing of respectively text sources and imagery sources. The paper will highlight the main challenges these two pipelines and the solutions put in place to solve them.

Key Words: Machine Learning; Text Processing; Image Processing; Big Data

## 1. Introduction

The Italian National Institute of Statistics (Istat) has been investigating the potential of Big Data sources for Official Statistics since 2013. As part of the European Statistical System, Istat followed the strategic indications reported in two reference documents, namely: (i) the Scheveningen memorandum[2], with which European NSOs where called to investigate the possible use of Big Data sources to support the production of Official Statistics and (ii) the Bucharest memorandum[3], which indicated the investments necessary to produce Big Data-based statistics as part of Official Statistics to all effects and purposes.

Within such a strategic framework, Istat is investing on the use of several Big Data sources, including:

- Web available data, under the *Web Intelligence* umbrella concept. The main projects using such data are related to enterprise characteristics, price statistics, job vacancies and skills statistics, and traffic statistics.
- Satellite imagery for land cover and green areas statistical products.
- Social media data, and in particular Twitter data for sentiment indexes.
- Scanner data for price statistics.
- AIS (Automatic Identification System) data for vessels traffic.
- Mobile Network Operator data for mobility and tourism statistics.
- Mobile devices sensor data for supporting surveys like Time Use and Households Budget surveys.

From the methodological perspective, the investments have addressed first the issue of the heterogeneity of Big Data sources, i.e. the Variety dimension of Big Data. Indeed, dealing with types of data like texts or images required us to

[1]Monica Scannapieco, ISTAT, Via C. Balbo 16 – 00184 Rome (Italy); [2]Massimo De Cubellis, ISTAT, Via C. Balbo 16 – 00184 Rome (Italy); [3]Fabrizio De Fausti, ISTAT, Via C. Balbo 16 – 00184 Rome (Italy)
[2]Scheveningen Memorandum (2013), https://ec.europa.eu/eurostat/documents/7330775/7339482/SCHEVENINGEN_MEMORANDUM+Final+version.pdf/6d4b84a3-dc92-4110-a621-c10ac06dc487
[3]Bucharest Memorandum (2018), https://ec.europa.eu/eurostat/documents/7330775/7339482/The+Bucharest+Memorandum+on+Trusted+Smart+Statistics+FINAL.pdf

design and develop proper data preparation pipelines, as well as introducing machine learning as the principal inference paradigm. In addition to that, the treatment of external (private) sources demanded for investments on new access techniques, like e.g. Web crawling and scraping, and on privacy-preserving methods on the input side.

In this paper, we will describe two pipelines that are becoming the reference ones for projects dealing respectively with textual data and with imagery. We will exemplify such pipelines in three projects, namely: (i) a Web Intelligence project related to the estimation of enterprises characteristics from enterprises' websites, (ii) the Social Mood on Economy Index, computed from Twitter data and (iii) the Land Cover project, producing statistics and maps from Sentinel-2 imagery.

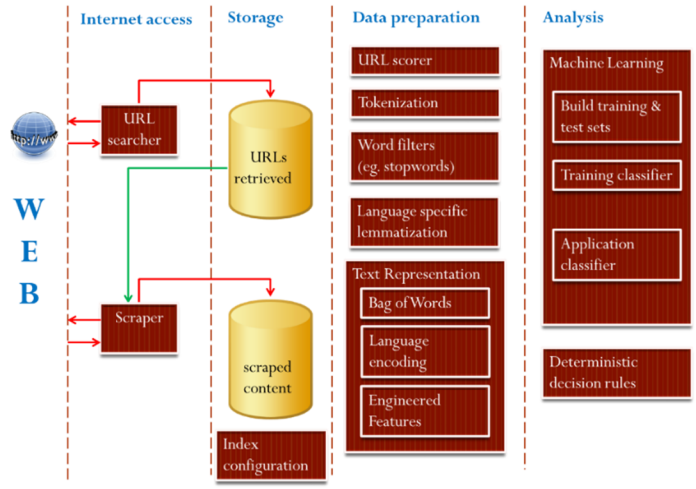
# 2. Some Istat Projects Using Textual Data

## 2.1 Estimating Enterprise Characteristics

In 2018, Istat started producing experimental statistics on the activities that enterprises carry out through their websites (web ordering, job vacancy advertisement, link to social media, etc.)[4]. They are a subset of the statistics currently produced by the "Survey on ICT usage and e-Commerce in Enterprises" and are computed starting from enterprise websites' contents, acquired by web scraping tools and processed with text mining techniques. A machine learning approach is adopted to estimate models in the subset of enterprises for which the survey and the web sources are both available, with survey data serving as training set for the machine learning task. The content scraped from successfully reached websites is used as input to predict the target values by applying the model fitted in the previous step. The experimental statistics are obtained using two different estimators: (i) a full model-based estimator; (ii) an estimator that combines model and survey based estimates. Considering the various domains for which they have been calculated, the three sets of estimates (survey, model and combined) in most cases are not significantly different (i.e. model and combined estimated values lay in the confidence intervals of survey estimates). Simulations have demonstrated that the Mean Square Errors of these new estimates are competitive as compared to those produced in the traditional way. The detailed description of the project can be found in (Barcaroli and Scannapieco, 2019).

Figure 2.1-1 shows a generic pipeline for dealing with enterprise websites. It includes two logical blocks, namely: URL searcher and Scraper, while the Storage phase includes a URLs Retrieved storage block and Scraped content one.

**Figure 2.1-1**
**Generic pipeline for processing textual data from enterprise websites**




[4] https://www.istat.it/en/archive/216641

In order to start a collection of data by scraping websites, first a list of URLs identifying the home pages of the sites to be reached is needed. If this is not available, there is the opportunity of setting up a dedicated URL retrieving activity (see (Barcaroli and Scannapieco, 2019) for details). The Data preparation phase includes several blocks related to text transformation that are to be performed in a preliminary way before the Text Representation can be actually done: this is an essential step when dealing with textual data that are unstructured or partially structured data so an explicit step is necessary to "represent" texts in a way that can be used for the subsequent analyses. The Text Representation logical block includes the traditional Bag of Words approach, a Language modelling block, e.g. recent Word Embeddings approaches, and an Engineered Features block, related to the specific features that can be selected for subsequent machine learning tasks. Finally, the Analysis phase for websites is based on the use of a Machine Learning approach, in which either Supervised or Unsupervised Machine Learning methods can be applied. In addition, for specific use cases, deterministic decision rules can also be put in place: this is the case for instance of a use case aiming at computing if an enterprise is present or not on a social media, which can be checked by looking at the presence on the enterprise website of a specific social media, i.e. Twitter, Facebook, etc.

## 2.2 Social Mood on Economy Index

The Social Mood on Economy Index is an experimental statistic published by Istat (see https://www.istat.it/it/archivio/219585). It provides daily measures of the Italian sentiment on the economy, these measures derived from samples of public tweets in Italian language are captured in real time. The index production procedure collects and processes only tweets containing at least one word belonging to a specific set of filter keywords, which has been designed by subject-matter experts.

The tweets, collected through the Twitter's Streaming API, use a filter made by 60 keywords that allow filtering the data collected in order to address the gathering phase only on the relevant ones to produce the social mood on economy index.

The statistical production process handles all the tweets collected in a single day (about 47,000) as a single block. As shown in **Figure 2.2-1**.2-1, the messages collected, after a cleaning and normalization activity, undergo a sentiment analysis procedure through an approach based on an unsupervised lexicon. The lexicon consists of a set of lemmas associated to pre-computed sentiment scores. Through the comparison between the words of each message (tweet) and the lexicon, a sentiment score is computed and is used to cluster the messages into three mutually exclusive classes: Positive, Negative and Neutral tweets. Lastly, the daily index value is derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes.

To prevent off-topic tweets from passing the filter and undermining the robustness of the index has been put in place a surveillance system that searches for anomalous values in the time series of the daily index, continuously. Daily values detected as potential outliers generate a set of dedicated diagnostic reports that will be analysed by human reviewers with the aim of deciding whether the detected values are actually proper data or truly anomalous. All daily index values classified as truly anomalous are imputed via nearest-neighbour interpolation.

Data collection for the Social Mood on Economy Index started in February 2016 and has been active since then almost without interruptions. The first publication of the index as Experimental Statistic involved daily data covering the period 10th February 2016 – 30th September 2018. For organizational reasons, updates of the index are currently published on a quarterly basis. The daily time series of the Social Mood on Economy index has been released as Experimental Statistics "for users' consultation and evaluation" (see Figure 2.2-2). More details on the project can be found in (Zardetto et al., 2019).

In the next section, we will focus on how evaluating the quality of a filter of a Twitter data stream, like the one used for the Social Mood on Economy Index.

**Figure 2.2-1**
**Pipeline to produce the Social Mood on Economy Inde**

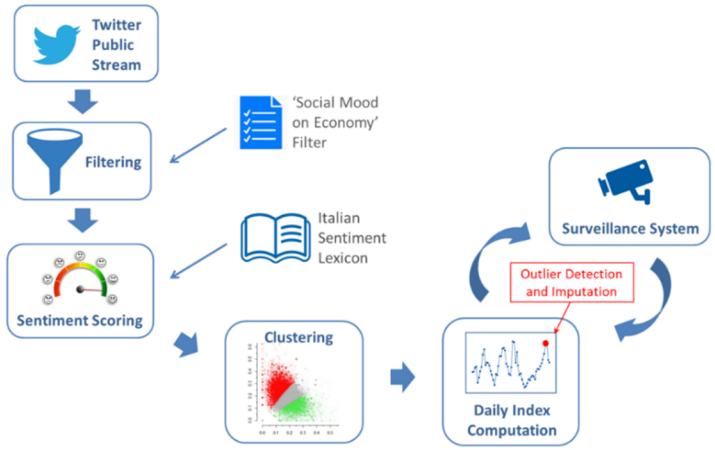
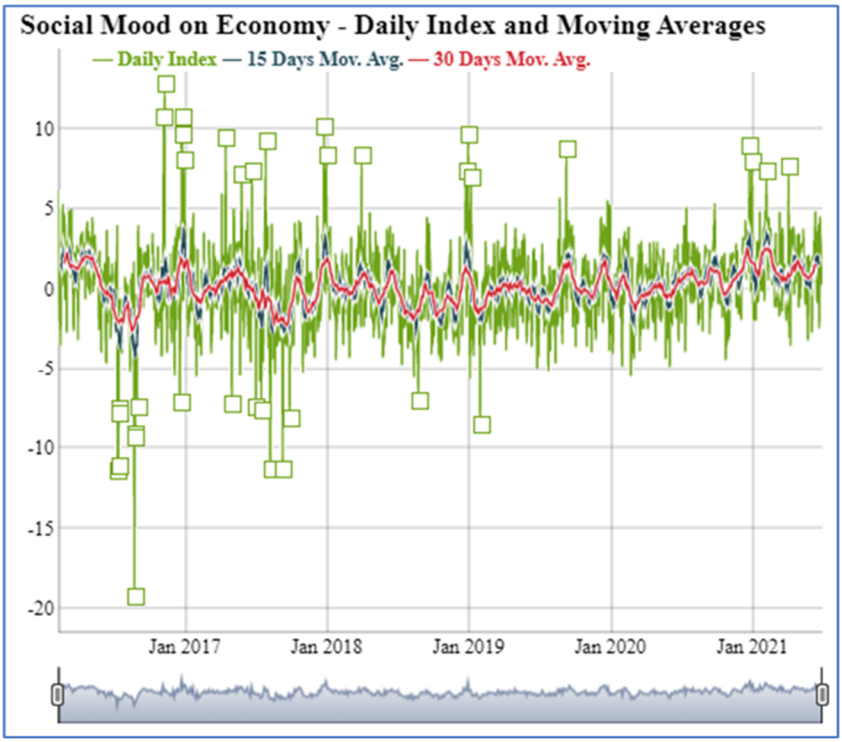



**Figure 2.2-2**
**Daily time series of the Social Mood on Economy index**




## 2.2.1 On the Quality of Twitter's filters

In this Section we will describe the use of Word Embeddings to evaluate the quality of Twitter's filters. Word Embeddings (WE) are data-driven models where the words are mapped as vectors in a vector space in a way for which words that are closer in the vector space are expected to be similar in meaning.

Word embedding models help devising domain-specific 'filters', namely sets of keywords that we use to filter out off-topic tweets with respect to the intended statistical production goal.

To improve the quality of the filter keywords used in the tweets collection, the use of methodology based on word embedding models could be very useful, but the exploration of these models is not easy. The high dimension of the models in the order of 200/300 vector coordinates, and the enormous quantity of words, therefore of vectors, in the embedding space, does not help their investigation.

To exploit the potential of the WE models, it would be very useful to be able to explore these models around a semantic area and represent the relationship between words that emerge from a specific context. With graphs is possible to achieve this objective and overcome the difficulties of WE models exploration. A graph is a structure amounting to a set of objects in which some pairs of the objects are "related". The objects correspond to mathematical abstractions called nodes, and each of the related pairs of nodes is called edge or link. In our case, the nodes correspond to the words and the edge to the relations between the words (syntactic or semantic relationships).

In Istat, we have developed a software tool, named *WordEmBox* that combine WE models and graphs structure. This software allow to visualize WE models in two dimensions only, and to investigate the relationship between words in a very user-friendly manner.

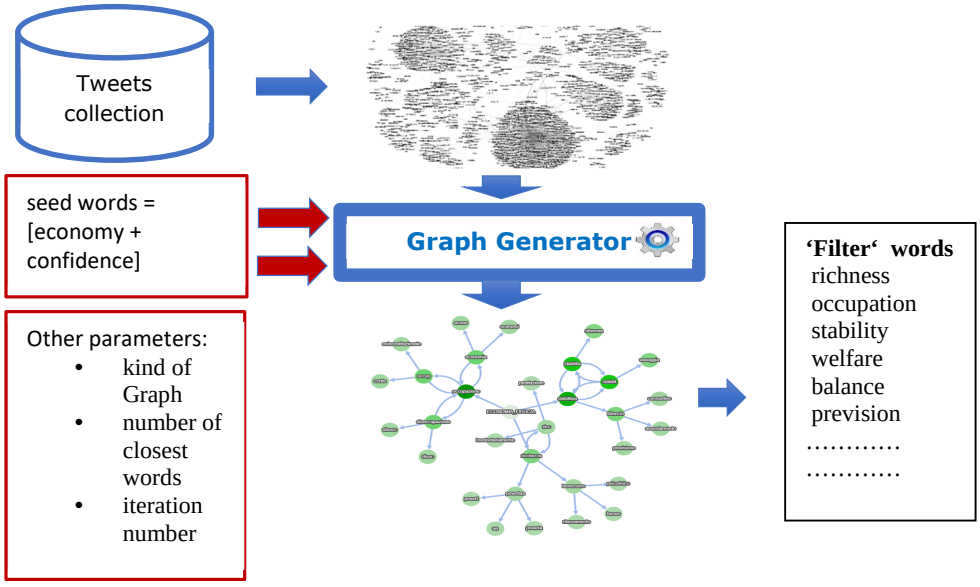
The WordEmBox includes three functions: graph representation, word-analogy and word-similarity.

While word-analogy and word-similarity are the standard test functions to assess the quality of the learned word embedding, the graph is the real add-value of this application. It is an original graph-based methodology to explore, analyse and visualize the structure of the learned embedding spaces. In the graph function, the user have to fill the following parameters: one or more seed words from which to start the model exploration, the number of iteration desired and the width, that is the number of closest words to search at each iteration.

In the WordEmBox it is possible to generate and visualize three different types of graphs that we have named: *geometric*, *linear* and *geometric oriented*. In the Geometric graph the exploration range expands quickly, losing the initial semantic focus provided by the seed words; in the Linear one, the exploration and their representation stays much more focused, but explores just a “narrow” sub-model; the Geometric-Oriented graph is a compromise between the previous two.

To improve the quality of the filter keywords used to collect the tweets that feed the Social mood on economy index experimental survey, we have exploit the potentials of the WordEmBox.

**Figure 2.2.2-1**
**Pipeline to build a list of “filter words” starting from a tweet collection**



Starting from the original keyword filters, for each of them we carried out an exploration on the WE model generated using 24 million tweets filtered with the original keyword list.

Thanks to the WordEmBox, we have seen in a very simple manner, which filter keywords worked well (for our purpose) and which did not. This activity has allowed us to redefine the filter keywords more accurately in order to download only tweets that are more relevant to the topics of Social Mood on Economy.

Figure **2.2.2-1**2.2.2-1 represents the pipeline put in place to implement the process described above. More details can be found in (De Fausti et al., 2018).

# 3. Dealing with Imagery: the Land Cover Project

This section deals with a data source type that is very different from the textual data sources described in Section 2. In particular, we will describe a project dealing with images, i.e. the Land Cover project, and we will focus on the related pipelined designed and realized in order to produce statistics based on images.

## 3.1 Motivation for LC and Machine Learning

We launched a project to design and develop an automatic Land Cover estimation system both for the production of statistics and of maps concerning an area and a period of interest. The production of land cover maps and statistics at European level is nowadays performed mainly in two major projects: LUCAS (Eurostat, 2003) and CORINE (Bossard et al., 2000). Each of these projects disseminates its output on a multi-year basis according to its own taxonomy, moreover for their production a strong human intervention is required as far as the execution of field measurements or the photo interpretation of satellite images are concerned. Given these assumptions, a fully automated system providing land cover estimates and maps fed by satellite imagery available for the community today would be both useful and challenging. In fact, the good spatial and temporal resolution of European sentinel satellite constellations offers the possibility of providing landcover outputs with a temporal frequency hitherto unthinkable.

Machine learning-based classifiers such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree (DT), and Random forest (RF) are widely used in the classification of remote sensing imagery, RF and SVM classifiers in particular have shown better performance in terms of overall accuracy (Thanh Noi and Kappas, 2018).

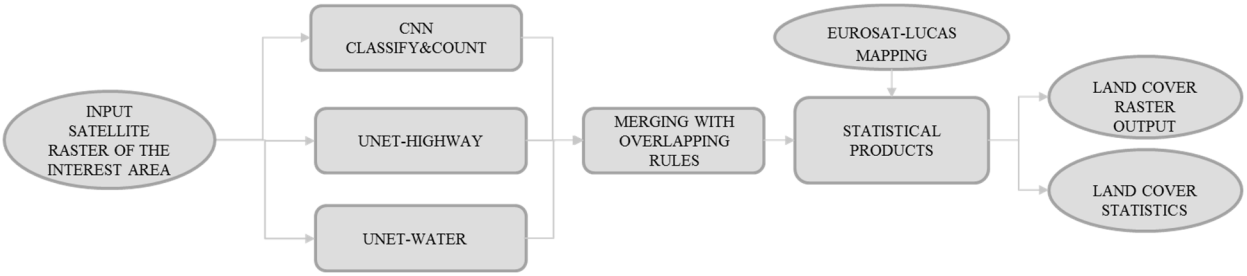
However, these approaches are primarily pixel based, and exploit the statistical relationship between land cover class and multispectral response as measured by the satellite's on-board instrument.

For the production of statistics and maps of land cover developed at Istat, we use a different approach that we consider more complete. In fact, for the classification of land cover in addition to the spectral response of pixel-based approaches we use the spatial patterns that characterize the land cover classes. Making an analogy it is as if for the recognition of objects we used the colors but also the shapes.
Deep learning algorithms today are the state of the art for spatial pattern recognition and are the ones we have implemented in our pipeline.

## 3.2 Pipeline for Land Cover Maps and Statistics

In this paragraph. we describe the inference process for the production of land cover statistics and maps. We will give just an overview due to the complexity of each individual step, a more detailed explanation of the process can be found in (Zardetto et al., 2021).

**Figure 3.2-1**
**Final pipeline for the production of map and statistics of land cover**




Looking at

Figure **3.2-1**, the input raster is composed merging many Sentinel 2 MSI multispectral raster downloaded from Copernicus Open Access Hub relatively to period and area of interest. For each pixel we process mainly the spectral response relative to red green blue and near infrared bands. For this bands the pixel resolution is 10 meters. Our pipeline can produce land cover statistics also of area wide like a region.

The input multispectral satellite image is processed by three different deep neural networks trained on different land cover classes, to produce three different output specific of land cover classes.

**CNN CLASSIFY&COUNT** is a architecture based by a Convolutional Neural Network multi-classes classifier INCEPTION-V3. We trained this classifier using an external EUROSAT dataset composed by a collection of Sentinel2 tiles of 640 X 640 meter, labeled with 7 classes: Annual-Crop, Forest, Herbaceous-Vegetation, Industrial, Pasture, Permanent-Crop, Residential.To feed the CNN we decompose the input image through several tiles of dimension 640x640 m and compose each prediction to produce the output classification matrix using an classify and count.

**UNET-HIGHWAY** is an architecture based by a U-net Neural Network producing a binary segmentation of raster in input. The output is a classification matrix of the highway land cover class. This architecture shows best performance for identify 1-dimentional land-cover classes such as highway. We trained this UNET network with a dataset that we built for this specific highway segmentation task using as input information Sentinel-2 tiles and
highway layer from the OpenStreetMap project.

**UNET-WATER** like UNET-HIGHWAY we used a U-net Neural Network to produce a binary segmentation of Water land cover class. We trained this UNET network with a dataset that we built for this specific water segmentation task using as input information sentinel2 tiles and water layer from High Resolution Level provided from Copernicus project[5].

In the merging operation the three classification matrixes are overlapped through the rules of priority between the classes and only one classification matrix is returned. We impose the class with higher priority highway, followed by the water class, and then by all the others. Finally, it is possible to produce outputs based on another land use classification e.g. LUCAS classification, by applying trans-coding tables.

## 3.3 Results

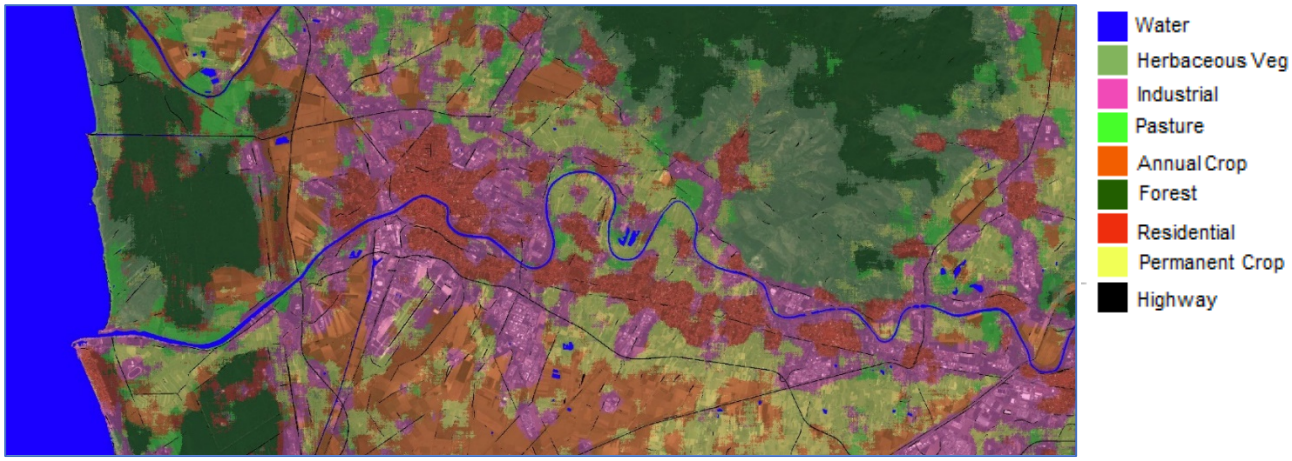
We have tested our system for the production of automatic land cover maps over large areas such as the Italian region of Tuscany. Such a processing requires about three days of computation on a server with NVIDIA-V100 GPU.

In
we show a clipping of an area (30120 x 14720 meters) relative to the Arno valley and the city of Pisa. In this detail we show the good agreement between the satellite image and the land cover map generated by the classification of the three neural networks. We show in particular the good segmentation of the Water class with the Arno River and the well-defined edges. It is also interesting to note the good separation within the urban area between the Residential and Industrial classes.

We believe that given the speed of creation of these maps and the good agreement of our system is a useful and promising tool for the analysis and control of the territory and its changes.

**Figure 3.3-1**
**Land Cover Automatic Map of Pisa Area**

[5] https://land.copernicus.eu/pan-european/high-resolution-layers/water-wetness




## 4. Concluding Remarks

Big Data processing pipelines are a reality in Istat already. The maturity degree of these pipelines and of the related projects is various: some of them are less mature and are related to pilot activities, some others are available as *experimental statistics* on Istat's official website having a higher maturity level, some others have already a full maturity and are in production.

Istat is planning to build a full production system based on the use of Big Data sources, with investments planned for the next years on continuing to build cross-cutting capabilities as well as subject-matter ones for a full-fledged Big Data based production.

## References

Barcaroli, G. and M. Scannapieco (2019), "Integration of ICT Survey data and Internet data from enterprises websites at the Italian National Institute of Statistics", *Statistical Journal of the IAOS - Journal of the International Association for Official Statistics* vol. 35, no. 4, pp. 643-656, 2019.

Bossard M, Feranec J, Otahel (2000), "CORINE land cover technical guide: Addendum 2000", Technical report No 40/2000, 2000.

De Fausti, F., M. De Cubellis, D. Zardetto (2018), "Word Embeddings: a Powerful Tool for Innovative Statistics at Istat", *International Conference on Statistical Analysis of Textual Data – JADT*, 2018.

Eurostat (2003), "The Lucas survey - European statisticians monitor territory", Office for Official Publications of the European Communities, 2003.

Zardetto, D., C. Fabbri, P. Testa, L. Valentino (2019), "New Experimental Statistics at Istat: the Social Mood on Economy Index", *Proceedings of the New Techniques and Technologies for Statistics Conference-NTTS*, 2019.

Zardetto, D., F. De Fausti et al. (2021), "Deep Learning Segmentation for Improved Land Cover Maps and Estimates", *Proceedings of the New Techniques and Technologies for Statistics Conference-NTTS*, 2021.