Smart Urban Planning Support through Web Data Science on Open and Enterprise Data

Gloria Re Calegari
CEFRIEL - Politecnico di Milano
Milano, Italy
gloria.re@cefriel.com

Irene Celino
CEFRIEL - Politecnico di Milano
Milano, Italy
irene.celino@cefriel.com

ABSTRACT

Urban information abound today in open Web sources as well as in enterprise datasets. However, the maintenance and update of this wealth of information about cities comes at different costs: some datasets are automatically produced, while other sources require expensive workflows including human intervention. Regression techniques can be employed to predict a costly dataset from a set of cheaper information sources.

In this paper we present our early experiments in predicting land use and demographics from heterogeneous open and enterprise datasets referring to the city of Milano. The results are encouraging, thus demonstrating that a data science approach leveraging diverse data can be actually worth for a smarter urban planning support.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: correlation and regression analysis; H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Open Web Data; Enterprise Internal Data; Smart Urban Planning

1 Introduction and Motivation

Digital information about cities abound today. The sources of such information are constantly increasing, due to the pervasiveness of information and communication technologies in the so-called Smart Cities domain.

With the advent of the *open data* movement, with its call for transparency and knowledge sharing, a very large number of data sources has been made available on the Web. Some examples of those datasets are: demographics and statistics from municipalities (e.g. distribution of population, family income, crime statistics), listing of local businesses from chambers of commerce, various levels of descriptions about the environment from an urban planning perspective (e.g. land use or land cover, cadastre information), and so on. Additionally, the so-called Internet of Things (IoT) has led to the availability of massive *real-time and streaming information*, like climate sensors from environmental agencies, smart meters and GPS traces from public utilities.

After the Web 2.0 boom, also *user generated information* about cities has become ubiquitous, through crowdsourcing initiatives like OpenStreetMap[1], which popularized the Volunteered Geographic Information paradigm of "citizens as sensors" [1], and location-based social networks like Foursquare, Twitter, Flickr with their stream of "check-ins" and geo-located information.

While a large part of the aforementioned sources can be considered open or at least openly accessible on the Web, there exist also *closed data sources* about cities, produced and maintained by enterprises, like public utilities information. For example, telco companies collect data about the phone activity over time and also over space (due to the positioning of transceiver towers), which is a strong indicator of people presence and movement in the urban environment.

The collection, cleansing, curation and maintenance of that wealth of specialized data sources can require a complex and expensive process. This is the case of datasets requiring a *manual intervention*: demographics data, for example, needs a human-based census activity; in other cases, there can be an error-prone *(semi)automatic processing*: land use maps are derived from aerial or satellite imaging and characterize the environment with reference to domain-specific classifications. The cost of data management is therefore highly variable with respect to the diverse data origins.

Our current investigation is aimed to answer to the following research question: would it be possible to (semi)automatically generate or revise an outdated dataset (which would require an expensive manual work), on the basis of the content of other up-to-date information sources (which come almost for free)? In other words, would it be possible to use cheap datasets as a "proxy" for more expensive data sources?

To answer the question above, we propose to adopt a *predictive analytics* [2] approach: using available (cheap) data sources as predictors, we select the best regression model that is able to predict, as outcome, an (expensive) dataset; more specifically, we fit or train several regression models with data about city POIs or human activities in the urban environment, to get the land use or the population census as response variables. Comparing different regression algorithms according to specific evaluation metrics, we select the best possible technique to solve the task.

Once a suitable model is available, this can be applied to new revised versions of the predictor datasets, in order to obtain an estimation of the update required in the outcome variables. While it can be hard to automatically update an information source on the basis of other "proxy" datasets, the regression model can help in identifying change, i.e. which areas of the urban space could have had a significant variation in terms of land use or demographics;

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742131.

[1] Cf. `http://www.openstreetmap.org/`.

this can be helpful for urban planners to focus their manual intervention to update the (expensive) data sources only where needed.

In this paper, we present our early results in applying the above approach to the city of Milano: we detail the experimental setting, the adopted regression techniques and the obtained results; we give some insights on the evaluation of those methods and the possible extensions in terms on model and variable selection.

The remainder of the paper is as follows: Section 2 details the information sources about Milano used in this research; Section 3 illustrates our regression experiments, explaining the adopted methodology(§ 3.1), the details on the employed data variables (§ 3.2) and the experimental results (§ 3.3); related works are presented in Section 4, and Section 5 concludes the paper with some perspectives on our future extensions to this work.

2 Overview of the Milano Datasets employed in our work

Our case study deals with the predictive analytics of diverse urban datasets related to the municipality of Milano in Italy. The datasets used in the analysis are illustrated in Table 1: the open data about population demographics from Milano municipality[2]; the land use classification elaborated within the CORINE European initiative[3] and made available as open data by Lombardy Region[4]; the Points of Interest (POIs) of the city provided by both Milano municipality and OpenStreetMap; two months of mobile call data records provided by the Telecom Italia mobile operator. Those datasets represent a meaningful mix of open data, volunteered geographic information and enterprise data.

Domain (content)	Data Source	Data Format	Spatial Resolution	Time Period	Volume (records)
Demographics (population)	Milano Open Data	Shape file	Census area	2011	10s
Urban Planning (land use)	Lombardy Region	Shape file	Building resolution	2012	10Ks
Mobile Phones (call records)	Telecom Italia	Tabular	Grid cells (250m)	2013	100Ms
Points of interest (POIs)	Milano Open Data	Shape file	Points (lat-long)	2013	1Ks
Points of interest (POIs)	Open Street Map	Shape file	Points (lat-long)	2014	1Ks

Table 1: Characteristics of the used datasets

As the table reveals, besides content heterogeneity, the datasets also differ in terms of spatial granularity and this is why a preprocessing phase was required in order to make them comparable. The population dataset was collected using the census area resolution level and the land use data has even a building level resolution, Telecom data is mapped into a grid of square cells and POIs datasets consist of points. Therefore we selected the more suitable spatial resolution and we interpolated the other sources to map all data into the Telecom data records resolution, a grid of 3538 square cells of 250 m (cf. Figure 1).

Telecom dataset records every ten minutes the activity occurred in Milano in Nov-Dec 2013. Five different phone activities are stored: incoming and outcoming calls, incoming and outcoming SMSs (text messages) and Internet connection. To reduce the dataset size and to take into account the spatial information, we compressed all the data of each cell into a "footprint", i.e. a summarizing data structure which records for each time slot of ten minutes the average activity of that cell, distinguishing between working days and holidays. The resulting data consists of one footprint for each cell and for each activity type. More specifically, each cell is described by a vector with 1440 elements, as we have, for each type of phone activity, 144 values (activity every 10 minutes). There are 5 phone activity blocks for the weekdays and 5 for the holidays. Picture 2 shows an example of the footprint visualization of the "Brera"district for the incoming calls.

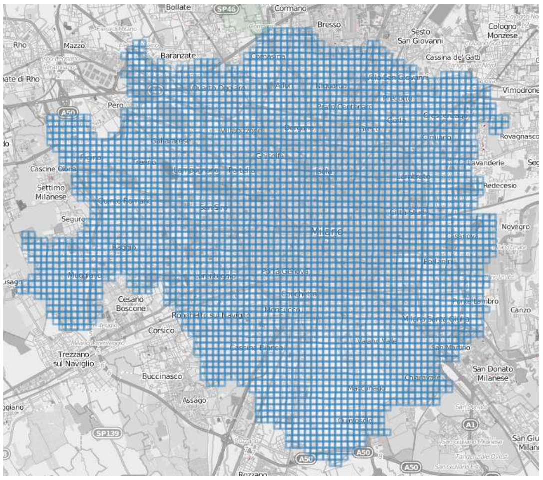

Figure 1: Spatial resolution level used in the analysis

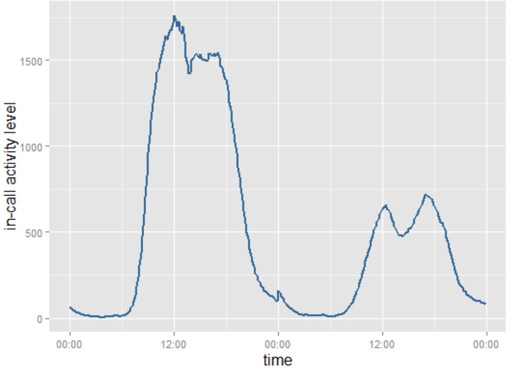

Figure 2: Telecom footprint of "Brera" district with concatenation of weekdays activity and holiday activity

The completeness and multifaceted nature of Telecom dataset allowed us to select the data that could better fill our specific needs, giving to the analyses flexibility and adaptability. On the one hand, we are interested in analysing the changing of phone activity during the day, thus we can extract only data at certain timestamps; on the other hand, we may require an overall picture of the phone activity, thus we can compute a global mean value. Similarly, we may want to focus our attention on a particular phone activity as incoming calls or out-coming text messages or we may want to assess the influence of all these components together.

Using mean values instead of the entire information not only help in managing lower-dimensional datasets, but also overcome possible local anomalies that are not specifically relevant when looking for a global "picture".

[2] Cf. `http://dati.comune.milano.it/`.

[3] CORINE multi-level taxonomy of land cover, cf. `http://www.eea.europa.eu/publications/COR0-landcover`.

[4] Cf. `https://www.dati.lombardia.it/`.

As regards *demographics information*, we mapped the number of inhabitants of each census area onto the grid (by overlapping the shapefile layers with a GIS) and we obtained a single value representing population density for each cell.

CORINE 2012 dataset provides data about the *land use* types of Milano territory and it classifies them by using CORINE multi-level taxonomy. Land use types can range from more general definition of residential/agricultural/industrial/wild areas to a more specific characterization as hospitals, roads, railways, construction sites and so on. After analysing the distribution of the different land uses in the Milano area, we selected the level of taxonomy more suitable for our case studies (by selecting specific categories or by grouping some classes together). The 5 types of land use that we assumed could better feature a metropolitan area as Milan are: residential areas, agricultural areas, commercial/industrial areas, parks and green areas, and sports centres. Therefore, each cell was described with a vector of 5 elements, representing the percentage shares of those land use categories over the cell area.

Lastly, as regards *points of interest* (both from OpenStreetMap and from Milano municipality), as they are data points described by latitude and longitude pairs, we computed their density in each cell. The POIs provided by the two sources are slightly different in terms of categories: they both have POIs about transports, schools and sport facilities, but in addition OpenStreetMap provides information about shops and amenity places of the city. On the other hand, data coming from Milano municipality are “official”, whereas the OSM dataset, being user generated data, may be less reliable and incomplete.

3 Regression Models with Milano Datasets

Our experiments are aimed to answer the question: would it be possible to use one or more “cheap” datasets as proxy for more “expensive” data sources? Therefore, our work is oriented to predict land use and demographic data (the “expensive-to-maintain” datasets) employing diverse “cheap-to-produce” datasets as predictors (telecommunication and points of interest data). In our first experiments we employed two different regression approaches, a statistical learning method and a machine learning one.

Hereafter, we explain in details the followed methodology, the input/output data used and the experiments performed. At the end of the section we discuss the obtained results.

3.1 Methodology

The aim of this study was to compare different regression approaches for predictive analytics.

As regards the statistical learning approach, we fit multiple linear regression models (MLR [3]) to have as dependent variables demographic and land use data, considering as possible independent predicting variables our multi-faceted information of phone activity and POIs.

With respect to the machine learning approach, we started with the Random Forest algorithm [4] which is an ensemble learning method that extends the concept of regression/classification trees (CART). Random Forest basically generates hundreds of regression trees starting from a random selection of a subset of data (bagging) with a randomized selection of predictors involved at each split. The various tree solutions are then averaged in order to predict the output variable with the smallest mean squared error (MSE). This approach has the advantage of reducing the variance of the model and also helps avoiding overfitting.

To build a model in both approaches, the standard methodology requires splitting the datasets in two groups, the training and the test sets.

The goal is first to learn the general rules that maps inputs to outputs with the training set data, and then evaluate the generated model using the unseen data of the test set. In both fields, a major emphasis is placed on avoiding overfitting, so as to achieve the best possible performance on an independent test set that follows the same probability distribution as the training set. Therefore in our experiments we divided the dataset into training and test sets, randomly selecting respectively the 90% of the data and the remaining 10%. We also applied k-fold cross validation (with 10 folds) to assess the generalization power of our models. Actually, the advantage of using cross validation is that all observations are used for both training and validation, and each sample is used for validation exactly once.

Once a model has been trained, it may be important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. The R squared (R^2), calculated using the test set, is one of the commonly used indexes of goodness of fit. However, since the R^2 index automatically and spuriously increases when extra explanatory variables are added to the model, in our analysis we decided to use the adjusted R squared (R^2_{adj}) that has been proposed to fix this behaviour. R^2_{adj} adjusts for the number of explanatory terms in a model relative to the number of data points; so R^2_{adj} increases when a new explanatory is included only if the new predictor improves the R^2 more than would be expected by chance.

As regards the predictors to be used to fit the model, we decided to perform some tests to evaluate how the number of input variables impacts the goodness of the model. So we planned a first experiment involving only a subset of 15 predictors selected by hand and a second set with all the 49 variables available. As regards MLR, we also employed, on both experiments, an automated model selection approach, the so called Akaike information criterion (AIC [5]). It automatically selects the most relevant predictors with the aim of simplifying the model and avoiding overfitting.

Unlike linear regression, interactions between different predictors are automatically incorporated into the Random Forest algorithm, making complex, non-linear interactions between variables easier to handle than in linear regression modeling. Variable selection could be unnecessary in Random Forest because irrelevant predictors should be automatically excluded from the model. However, in our experiments, we decided to investigate the variable importance as well, by analyzing the information provided by the Random Forest algorithm about how important each predictor is in the model. Actually Random Forest computes the importance score of each variable in terms of the mean decrease in accuracy caused by removing that variable; the mean decrease in accuracy values are calculated during the out of bag error calculation phase. The higher decrease in accuracy due to the removal of a single variable, the more important that variable in the Random Forest model. Therefore variables with a larger mean decrease in accuracy are more important for data regression.

3.2 Input and Output Data

According to our long-term goal of employing “cheap-to-produce” datasets to predict or update “expensive-to-maintain” datasets, in the following analyses we chose demographics and land use data as outcome variables of both the linear model and the Random Forest algorithm. Regarding demographics, the dependent variable is the population density in the spatial unit, while regarding land use, we extracted from the CORINE dataset different 1-dimensional vectors representing the portion of each spatial unit covered by a specific land use (residential, commercial/industrial, agricultural, green and sports areas as explained in section 2).

As predictors, we divided the experiments in two steps: the first considers only a subset of predictors while the second takes into account all the variables available, as detailed below.

We first focused our attention on Telecom data. From the call data records we decided to take in consideration different “facets” of that dataset. To this end we computed a set of ten 1-dimensional vectors representing the average activity for each communication type (incoming/outcoming calls, incoming/outcoming text messages, internet connection) and distinguishing working week days (Monday to Friday) from week-end days (Saturday, Sunday and other holidays). We added also information about OpenStreetMap POIs of five categories: transportation, leisure, food, shops and schools. So, as a first step, we employed 15 variables, 10 coming from phone data records and 5 regarding open data points of interest.

Since we have a larger number of urban datasets available, as illustrated in Section 2, we can use a much higher number of predictors in our analysis. Therefore, as second step, we employed a set of 49 predictors that better represent the multi-faceted nature of the city. Besides the predictors listed above, we added 8 predictors representing categories of POIs as listed by the municipality of Milano in their open data, and we enriched the set of predictors based on Telecom data, by introducing the average activity at different times of the day (without distinguishing between activity type and working day/holiday); regarding the latter, we introduced one predictor for each hour interval.

The following Table 2 lists and explains all input/output variables included in the analyses. We remind that variables are vectors with 3538 components, one for each spatial unit.

Output variable name	Variable description
population	population density
corine.resid	dense residential land use density
corine.agric	agricultural land use density
corine.comm	industrial/commercial land use density
corine.green	park/green area land use density
corine.sport	sport facility land use density
Predictor variable name	Variable description
poi.mun.school	school presence (low/high) from municipality open data
poi.mun.transport	transportation presence (low/high) from municipality open data
poi.mun.bike.car.share	bike/car sharing presence (low/high) from municipality open data
poi.mun.sport	sport facility presence (low/high) from municipality open data
poi.mun.pharmacy	pharmacy presence (low/high) from municipality open data
poi.mun.newsstand	newsstand presence (low/high) from municipality open data
poi.mun.conv.centre	convention center presence (low/high) from municipality open data
poi.mun.culture	cultural center presence (low/high) from municipality open data
poi.osm.transportation	transportation POI density from OpenStreetMap
poi.osm.leisure	leisure place presence (low/high) from OpenStreetMap
poi.osm.leisure	leisure place presence (low/high) from OpenStreetMap
poi.osm.food	restaurant/bar presence (low/high) from OpenStreetMap
poi.osm.shop	shop presence (low/high) from OpenStreetMap
poi.osm.school	school presence (low/high) from OpenStreetMap
sms.in.wd	average incoming text messages during working days
sms.in.hd	average incoming text messages during holidays
sms.out.wd	average outcoming text messages during working days
sms.out.hd	average outcoming text messages during holidays
call.in.wd	average incoming calls during working days
call.in.hd	average incoming calls during holidays
call.out.wd	average outcoming calls during working days
call.out.hd	average outcoming calls during holidays
internet.wd	average Internet usage during working days
internet.hd	average Internet usage during holidays
hour.hi.hj	average telecommunication activity in the i-j hour interval

Table 2: List of data variables included in regression analysis.

Finally, it is worth noting that, in order to apply linear regression, some conditions must hold for the data. To this end, we also applied some transformations to the considered variables; for example Telecom data, which showed a strongly left-skewed distribution, went through a logarithmic transformation, while some POI categories with limited number of distinct values were turned into categorical variables with two values indicating the presence/absence of that specific type of POI within the spatial unit. Although no assumption are made about the distribution of the data in Random Forest algorithm, we used this transformed dataset also in the machine learning phase in order to have comparable experiments and results.

3.3 Experimental Results

We performed five different tests: firstly we applied MLR to both the 15 and 49 predictors described in section 3.2; secondly we made an experiment to study how an AIC-based variable selection influences the linear regression results; lastly we trained our Random Forest model again with both the 15 and 49 variables.

The results are summarized in Table 3, in terms of R^2_{adj} on both training and test sets. The test set results are also plotted on the bar chart in Figure 3.

At first glance, it is evident that both the statistical and the machine learning approaches reach the best results in predicting population, residential and agricultural areas. Actually, the R^2_{adj} values of these output variables range from 0.4 to 0.66, in contrast with the values of commercial, green and sports areas that reach at most 0.25.

As regards the variance of the 10-fold cross validation process, we verified that on training set it is always lower than on test set; in both cases, variance is always sufficiently low, ranging respectively from 10^{-6} to 10^{-5} and from 10^{-3} to 10^{-4} orders of magnitude.

Another general tendency shown in Figure 3 is that Random Forest (blue and yellow bars) always equals or outperforms multiple linear regression (green, red and grey bars). This indicates that the data does not follow a linear distribution and so a non-linear, more complex model is needed. Actually, Random Forest, as an extension of tree-based algorithm, generates a model that implies the partition of the space into blocks according to different split variables.

If we compare the results of Random Forest with 49 and 15 predictor, besides higher R^2_{adj} values for the 15-predictors model, we can see that there may be overfitting using a larger number of variables, since the difference in R^2_{adj} between training and test sets is much higher in the 49-predictors case. This tendency to overfit is also evident comparing the MLR-49 predictors models with and without AIC: the R^2_{adj} of the full model is always lower than the one with variable selection.

Even though we selected by hand the subset of 15 predictors, the use of a restricted number of inputs improved significantly our prediction in all the Random Forest experiments except for the population case. We tried to investigate this result by looking into the variable importance ranking produced by each Random Forest model. As the right side of Figure 4 shows, the top-10 variables in the population model, ranked according to the mean decrease in accuracy (as explained in § 3.1), include only two of the 15 predictors used in the 15-RF analysis (poi.osm.transport and tlc.avg.internet.wd). On the contrary, if we look at the variable ranking of the residential case, in which less predictors led to a higher R^2_{adj}, we can find 7 predictors out of the manually-selected 15 in the top-10 variables (cf. left side of Figure 4). To sum up, we verified that variable selection is an essential step in optimizing a predictive model. Since further improvements can be achieved by smartly selecting predictors, the optimization of the variable selection process will be our next challenge in future analyses.

4 Related Works

The increasing availability of data related to urban environment has fostered the growing of various research studies that aim to explore spatio-temporal patterns of big cities. These datasets, coming from different and heterogeneous sources (as open data from social media or private data records), describe various aspect of the urban environment.

Quercia and Saez [6] aimed to determine whether social media could offer an alternative data source to study the relationship between the presence of specific physical venues in a London

R-squared	*MLR - 15 pred. w/ AIC*		*MLR - 49 pred. w/o AIC*		*MLR - 49 pred. w/ AIC*		*RF - 49 pred.*		*RF - 15 pred.*	
adjusted	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
Population	0.517	0.495	0.648	0.585	0.649	0.600	0.668	0.623	0.604	0.591
Residential	0.551	0.532	0.600	0.528	0.601	0.547	0.633	0.588	0.623	0.614
Agricultural	0.410	0.396	0.493	0.409	0.494	0.444	0.631	0.580	0.628	0.614
Commercial	0.107	0.075	0.172	0.021	0.174	0.068	0.255	0.159	0.234	0.209
Park/green	0.041	0.020	0.075	0.013	0.077	0.019	0.172	0.041	0.147	0.120
Sport	0.044	0.02	0.163	0.014	0.166	0.076	0.175	0.169	0.126	0.095

Table 3: R^2_{adj} on both training and test set obtained in the five different experiments

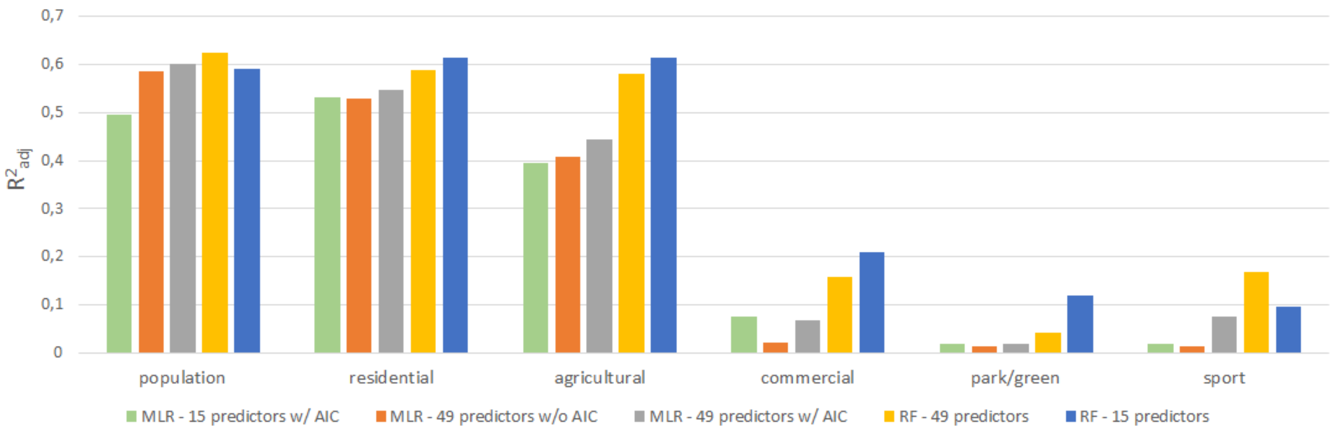

Figure 3: Graphical comparison of the five experiments in terms of R^2_{adj} on test set

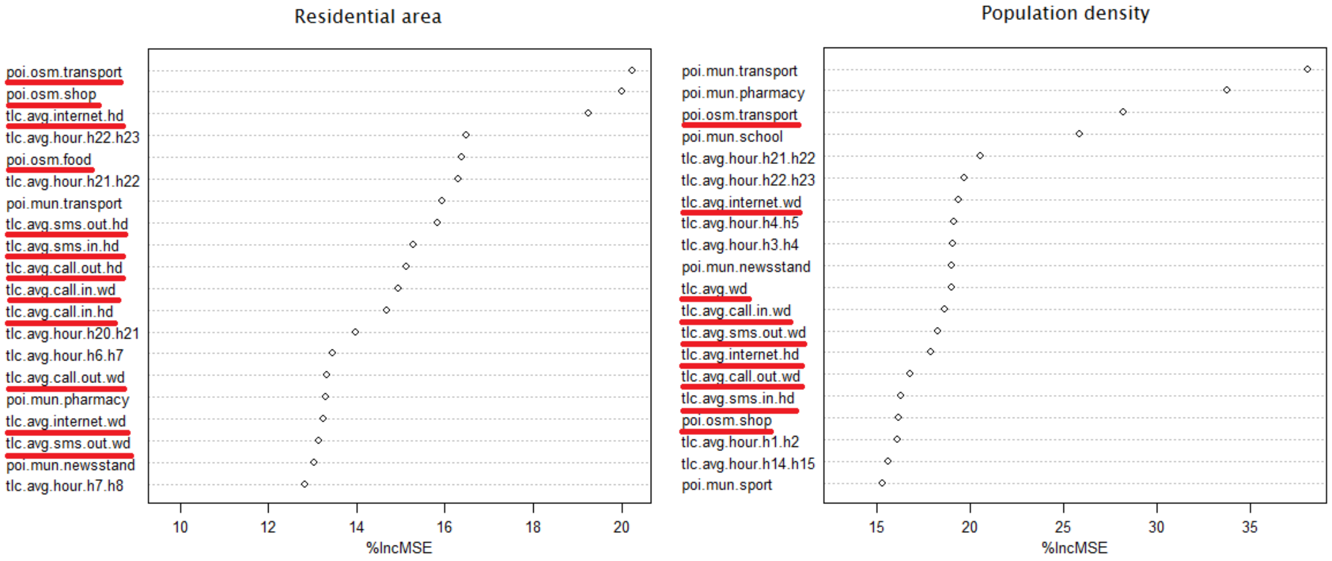

Figure 4: Ranking of variable importance for residential area and population

neighborhood (extracted from Foursquare) and the neighborhood's socio-economic deprivation. Other analysed social media data are geolocated Tweets, which were used as a complementary source of information for urban planning applications; Frias-Martinez et al. [7] aimed to automatically determine land uses in a specific urban area based on tweeting patterns and to identify urban points of interest as places with high Twitter activity.

The phone activity is one of the most used enterprise data record in research studies. For example, the aim of Reades et al. [8] was to extract recurring structures from overall mobile usage levels in order to find a correlation between phone activity and land use (the density of businesses categories in the urban environment was the method to label the land use of an area).

In an urban environment, mobile phone data was also used as a sensor to obtain information from users for discovering the Points of Interest (POIs) of the city, by looking at the presence of mobile signals [9]. When heterogeneous data sources are analysed together, the problem of having datasets at different spatial resolution levels is very common and the definition of a method to integrate these data is required. Ye et al. [10] faced this problem in their study on the global mapping of croplands: spatial cropland information of the entire world were available at different resolution levels, ranging from finer granularity of 30 m to coarser of 10 km. Another example related to the process of homogenizing different land cover datasets is the research of Tchuenté et al. [11] in which four different Africa's land cover classifications were available. Each of them has been produced by using different mapping initiatives and standards and by adopting diverse spatial resolutions.

As described in our case study, the predictive analysis between the different datasets can be performed with diverse methods accordingly to data complexity. In literature there are some examples of correlation and prediction analysis. When dealing with multiple heterogeneous datasets the most used methodology is the multiple linear regression, as in Maniquiz et al. [12]: their aim was to develop equations for estimating pollutant loads and event mean concentration as a function of rainfall variables. See et al. [13] applied the regression principles also to geographical data, by adopting the so called geographically weighted regression (GWR) to build an hybrid land cover map using crowdsourced validation data.

Machine learning techniques are employed for predictive analytics also at large scale. Chen et al. [14] use big data technologies and neural networks to predict $PM_{2.5}$ concentration in China from air quality and weather records as well as from open Web datasets. Spatio-temporal anomaly detection is proposed by Difallah et al. [15] with a scalable real-time stream processing approach employing clustering techniques. The integration of diverse datasets requires also to reconcile their different "semantics". In this regard, Kotoulas et al. [16] describe SPUD, a semantic environment for urban data processing that leverages Semantic Web technologies in information sense-making to address several challenges like city traffic diagnosis.

5 Conclusions and Future Work

Regression models are used to compute an outcome variable from a set of predictors. We applied regression analysis to evaluate if suitable models can be built to predict demographics or land use from other urban datasets. The aim is to support urban planned in the maintenance and update of relevant datasets that usually require an expensive human intervention. In this paper, we described our early experiments that gave encouraging results: indeed it is worth to employ diverse open and enterprise datasets, easily available on the Web, in regression models.

Comparing the results of our experiments with different techniques, we observe that population density, dense residential area and agricultural areas were adequately forecast by the predictive models, with explained variability reaching 62%. This means that, even if we considered quite diverse and heterogeneous datasets as predictors (call data records features and points of interest), there is a relation with land use and demographics.

Our next steps include extending the set of predictors with other open and enterprise datasets and employing further regression modeling techniques for predictive analytics [2] to choose the most accurate approach. We also intend to have our approach qualitatively validated with urban planning experts. Finally, since land use data is described with categorization systems and taxonomies like CORINE, we also intend to replicate the same approach with classification techniques [17].

Acknowledgments

This work was supported by the PROACTIVE project (id 40723101) co-funded by Regione Lombardia (POR-FESR 2007-2013).

6 References

[1] Goodchild, M.: Citizens as sensors: the world of volunteered geography. GeoJournal **69** (2007) 211–221

[2] Shmueli, G., Koppius, O.R.: Predictive analytics in information systems research. Mis Quarterly **35**(3) (2011) 553–572

[3] Freedman, D.: Statistical models: theory and practice. Cambridge University Press (2009)

[4] Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32

[5] Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19** (1974) 716–723

[6] Quercia, D., Saez, D.: Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. IEEE Pervasive Computing **13**(2) (2014) 30–36

[7] Frias-Martinez, V., Soto, V., Hohwald, H., Frias-Martinez, E.: Characterizing urban landscapes using geolocated tweets. In: SocialCom/PASSAT, IEEE (2012) 239–248

[8] Reades, J., Calabrese, F., Ratti, C.: Eigenplaces: analysing cities using the space-time structure of the mobile phone network. Environment and Planning B: Planning and Design **36**(5) (2009) 824–836

[9] Montoliu, R., Gatica-Perez, D.: Discovering human places of interest from multimodal mobile phone data. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia, ACM (2010) 12:1–12:10

[10] Yu, L., Wang, J., Clinton, N., Xin, Q., Zhong, L., Chen, Y., Gong, P.: From-gc: 30 m global cropland extent derived through multisource data integration. Int. J. Digital Earth **6**(6) (2013) 521–533

[11] Tchuenté, A.T.K., Roujean, J., Jong, S.M.D.: Comparison and relative quality assessment of the glc2000, globcover, MODIS and ECOCLIMAP land cover data sets at the african continental scale. Int. J. Applied Earth Observation and Geoinformation **13**(2) (2011) 207–219

[12] Maniquiz, M.C., Lee, S., Kim, L.H.: Multiple linear regression models of urban runoff pollutant load and event mean concentration considering rainfall variables. Journal of Environmental Sciences-china **22** (2010) 946–952

[13] See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A., Perger, C., Schill, C., Zhao, Y., Maus, V., Siraj, M.A., Albrecht, F., Cipriani, A., Vakolyuk, M., Garcia, A., Rabia, A.H., Singha, K., Marcarini, A.A., Kattenborn, T., Hazarika, R., Schepaschenko, M., van der Velde, M., Kraxner, F., Obersteiner, M.: Building a hybrid land cover map with crowdsourcing and geographically weighted regression. ISPRS Journal of Photogrammetry and Remote Sensing (2014)

[14] Chen, J., Chen, H., Pan, J.Z., Wu, M., Zhang, N., Zheng, G.: When big data meets big smog: a big spatio-temporal data framework for china severe smog analysis. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Nov 4th, 2013, Orlando, FL, USA. (2013) 13–22

[15] Difallah, D.E., Cudré-Mauroux, P., McKenna, S.A.: Scalable Anomaly Detection for Smart City Infrastructure Networks. IEEE Internet Computing **17**(6) (2013) 39–47

[16] Kotoulas, S., Lopez, V., Lloyd, R., Sbodio, M.L., Lécué, F., Stephenson, M., Daly, E.M., Bicer, V., Gkoulalas-Divanis, A., Lorenzo, G.D., Schumann, A., Aonghusa, P.M.: SPUD - Semantic Processing of Urban Data. J. Web Sem. **24** (2014) 11–17

[17] Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques (2007)