SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board
A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Mining Mobile Phone Data to Detect Urban Areas

Analisi di dati di telefonia mobile per l'individuazione di aree urbane

Maarten Vanhoof, Stephanie Combes and Marie-Pierre de Bellefon

Abstract The production of Urban Areas zonings at national level is characterized by long delays between consecutive updates. As mobile phone data has recently shown promising results for automated land use classification, we investigate the possibility to reproduce the French Urban Area Zoning (ZAUER). We exploit a dataset of hourly mobile phone activity profiles at cell-tower level, discuss methodological challenges, and find Urban Centers to be most correctly classified. Our findings frame the possibilities and limits for using mobile phone data to automatically, and continously produce urban zonings

Abstract *In questo articolo esaminiamo labilt dei dati del telefono cellulare nel riprodurre la zonizzazione dellarea urbana francese. Partendo un dataset di profili di attivit di telefonica oraria registrati dalle antenne di uno dei maggiori operatori telefonici francesi, analizziamo le sfide metodologiche coinvolte, e identifichiamo i centri urbani pi facilmente predicibili. I risultati proposti mostrano alcune delle possibilit e dei limiti legati allutilizzo dei dati del telefono cellulare per produrre zonizzazioni automaticamente e continuamente.*

Key words: Supervised classification, Mobile phone data, Spatial autocorrelation, Urban areas, Map comparison

Maarten Vanhoof
Open Lab, Newcastle University, Newcastle-Upon-Tyne, UK and Orange Labs, Paris, FR
e-mail: M.Vanhoof1@newcastle.ac.uk

Stephanie Combes
INSEE, 18 boulevard Adolphe PINARD,Paris, France
e-mail: stephanie.combes@gmail.com

Marie-Pierre de Bellefon
INSEE, 18 boulevard Adolphe PINARD,Paris, France
e-mail: marie-pierre.de-bellefon@insee.fr

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

1 Introduction

The growth of cities, and with it the extension of urban agglomerations, has become characteristic for contemporary times [Galster et al., 2001]. In this context, the identification of economically integrated areas is crucial for the adequate implementation of policy measures [Duranton and Puga, 2014] and thus calls for the definition of other typologies than purely administrative regions. As a mean of defining integrated (urban) areas, [Berry et al., 1969] suggested to rely on commuting patterns toward a predefined urban core to delineate metropolitan areas.In the same spirit, the National Statistics Office of France (INSEE) nowadays produces a zoning (ZAUER: Urban Area and Rural Employment Areas Zoning) that identifies cities areas of influence at a national level. Producing urban area zonings is a complex task involving multiple actors and methods. As a consequence, long delays are observed between consecutive publications (between five and ten years in France) which contrast with the fast pace of change in territories. [Floch and Levy, 2011]. In this context, alternative sources of more timely, high-resolution data associated with simpler procedures could contribute in a meaningful way to the production of more recurrent releases of such typologies.

In this paper we investigate the capabilities of French mobile phone data to reproduce the ZAUER zoning and explore a procedure that could lead towards a data-driven and recurrent production of the typology between official releases. Our contribution is twofold. First we demonstrate how mobile phone data can be mobilized to develop a nationwide typology of urban areas. Secondly, we elaborate a case that demonstrates how supervised classification tools can be of interests to official statistics. In addition to their ease-of-use, supervised classification techniques provide for both classification outputs and a quality evaluation. The latter being key in official statistics, we deem that a wider investment in these techniques could be profitable.

2 Urban Areas in France

The official french Urban Areas classification (ZAUER), as produced by the French National Statistical Institute (INSEE), consists of 9 classes, being distinguished mainly by the size of the employment pole in the 'central' Urban Unit. Major, medium and small centers are Urban Units offering respectively more than 10,000, between 5,000 and 10,000 and between 1,500 and 5,000 jobs that are inhabited by at least 2,000 people and cover a continuous build-up area with no more than 200 meters between buildings. Surroundings of urban centers are municipalities for which more than 40 % of the working population commute daily to an adjacent urban center. Special cases are being recognized for municipalities that have several urban centers to commute to (multi-polarized municipalities) or municipalities that are not 'influenced' by any urban centers.

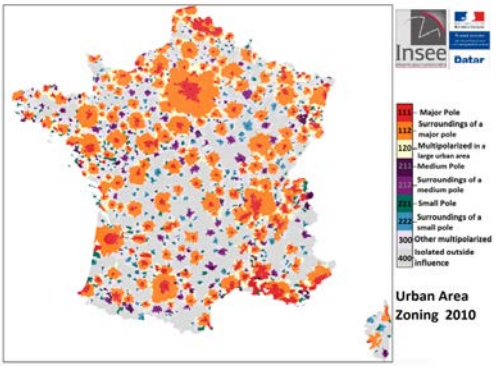

Fig. 1 Spatial distribution of ZAUER classes. The 2010 ZAUER classification is founded on data collected between 2008 to 2010 in the national census survey

3 Activity profiles from Mobile Phone Data

Call Detail Record (CDR) data are collected by mobile phone service providers for billing and network maintenance purposes. CDR data gather locational, temporal, and interactional information (who contacts whom) every time a phone call or text message is initiated by a user. The spatial resolution of observations is restricted to the locations of cell-towers, which are not uniform in space because of demand-driven placement.

In this study we use an aggregated CDR dataset from France provided by Orange and based on the activities of 18 million subscribers during a period of 154 consecutive days in 2007 (May 13 to October 14). Anonimysation at individual level was complemented with aggregation at the cell-tower level (see next paragraph) hereby ensuring full privacy of individual users as demanded by the French Data Protection Agency (CNIL) in light of the EU General Data Protection Regulation.

Literature validates the hypothesis that cell-tower profile activity (amounts of events measured at an antenna over time) can be informative for territory qualification [Soto and Martinez, 2011]. Therefore, we construct antenna activity profiles as a time series of the amount of activities registered each day at every hour and standardize the series for comparative purposes. Next, the obtained relative hourly profiles are averaged per hour of the day for the entire six-months window resulting in activity profiles for each antenna ('signatures'). In total we build four distinct signatures per antenna averaging each hour on i) weekdays in non-summer months, ii) weekends in non-summer months, iii) weekdays in summer months and iv) weekends in summer months. This results in 24×4 features per antenna (figure 2).

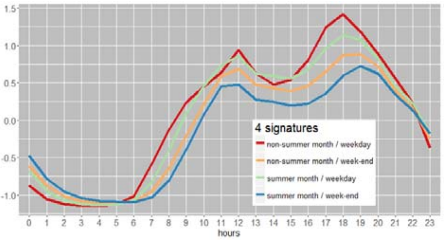

Fig. 2 Average relative activity profile for all antennas grouped by summer/non-summer months and week-/week-end days.

Because standardizing the activity profiles implies a loss of information about the absolute amount of activities, we add the circumference of the Vorono polygon for each antenna as a complementary feature. Lower circumferences indicate locally higher antenna densities and, given demand-driven placement, higher expected amounts of activities by the operator.

4 Methodology

4.1 Classification methods

For our classification task, we consider each antenna as an observation that needs classification in one ZAUER class. The output of the procedure will form a zoning that can be compared to the offical one. Multiple algorithms are available to carry out multiclass classification procedures. We implement the random forest approach described by [Breiman, 2001], Boosting Trees [Schapire, 2003] and the Elastic-Net penalized Logistic Regression [Zou and Hastie, 2005].

The Logistic Regression with Elastic Net penalty consists of maximizing the likelihood under a constraint expressed on the coefficients' amplitude [Zou and Hastie, 2005]. Specifically, this approach is better in accounting for multicollinearity between features (which is likely to happen here since our features are extracted from a temporal profile) than the initial LASSO [Tibshirani, 1996]. In contrast to the Logistic Regression, Random Forests and Boosting Trees do not assume linear interactions between variables. Random forests [Breiman, 2001] aggregate classification trees built on bootstrap samples, but introduce randomness by sampling a set of regressors from the initial set of variables at each separation step of each tree. Boosting Trees [Schapire, 2003] is rather different. It is an additive adaptive procedure which takes into account the biggest forecasting errors at a given iteration when calibrating the next iteration, by actualizing observations' weights.

4.2 Challenges

Mobilizing mobile phone data for urban areas classification at a national level raises several challenges. First the official ZAUER classification consists of 9 imbalanced classes, meaning that both municipalities and antennas are heterogeneously distributed among the classes and with respect to the urban tissue. In anticipation of this problem, we regroup the existing 9 classes into 6 by merging medium urban centers, small urban centers, and their respective surroundings. More importantly however, we apply downsampling, which consists of removing instances from the majority class, to minimize the effect of imbalanced classes on our classifiers.

Secondly, the extended area of investigation implicates high degrees of spatial autocorrelation and high volatility of antenna activity profiles within the different classes. As such we pay special attention to spatial autocorrelation. To de-correlate testing and training sets, we first operate stratification sampling while segmenting the map of France in four comparable quadrants to produce an initial test set (guarantying some minimal representativeness among regions). Next, the nearest neighbours of each selected observation are added to the test set so that we can evaluate the algorithms on their ability to reproduce a zoning and not only a punctual classification of one antenna. Finally, once the test set is built, we consider every left antenna not located in a buffer region around the test observations as a training sample. This last step ensures spatial de-correlation. We use the same approach to build the data partitions mobilized for cross validation.

Thirdly, Urban Areas are characterized by various degrees of similarity and spatial entanglement (especially at the borders of areas where the validity of urban area typologies may be less reliable). We address this issue by recoursing to the use of the Fuzzy Kappa metric ([Hagen, 2003] and the improved Fuzzy Kappa [Hagen-Zanker, 2009]) allowing us to evaluate (and calibrate) our models while taking into account both location and category fuzziness.

5 Results

Following the procedures outlined before, we applied three classifiers in order to predict urban areas from signatures of mobile phone activity. Kappa and Fuzzy Kappa computed on the test sets are reported in table 1. Fuzzy Kappas values range from 0.66 to 0.70 whereas Kappas stand between 0.59 and 0.65. According to magnitude guidelines in literature (for example [Landis and Koch, 1977]), values between 0.61 and 0.80 are considered substantial (1 being the perfect agreement). Detection rates per class are represented in figure 3 for the different classifiers. We can see that synthetic measures like Kappa or Fuzzy Kappa mask an heterogeneity in the detection rates par class, highlighting the fact that some classes are more difficult to detect than others.

Table 1 Kappa and Fuzzy Kappa for the different classifier

Method	Kappa	Fuzzy Kappa
Random Forests (RF)	0.62	0.67
Boosting Trees (BT)	0.63	0.69
Elastic Net (ENet)	0.59	0.67

Fig. 3 Classification rate (in %) per class for each classifier

We observe small differences in the capabilities of the algorithms to detect classes. In general, major urban centers (class 1) present an excellent rate of detection (about 95 %) for every approach. Correct rates (50 to 70 %) can be achieved for classes 4,5, and 6 (medium and small urban centers and their surroundings, multipolarized municipalities, and isolated municipalities). Classes 2 and 3 are more difficult to discern. Especially Class 3 (multipolarized municipalities in a large urban area) whose detection rate varies from 30 % to 60 % in the best scenario. Class 2 (surroundings of major urban center) get properly detected for only 40 to 50 % of the cases. The results of our classification for Normandy, a region in the west of France that mixes all urban classes are displayed in figure 4.

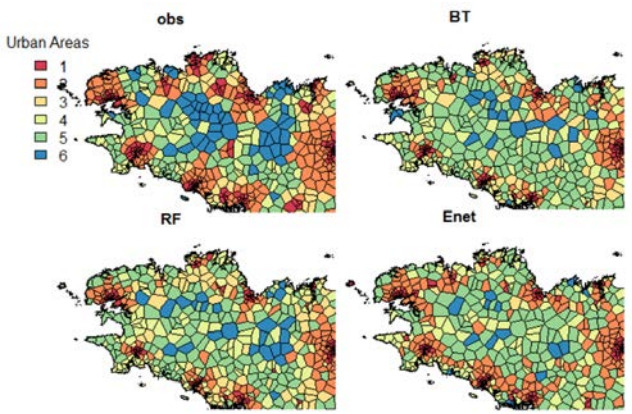

Fig. 4 Observed (obs) and predicted (based on the different classifiers) Urban areas for Normandy

6 Discussion

Our most remarkable findings are the difference between the accuracy of the prediction for major urban centers (class 1) and the heterogeneous performance in predicting the other classes (ranging between 30 and 70 %). Still, different classifiers show consistencies in results (with slight variations observed for one class or the other), which urges us to find explanations for these results based on the characteristics of our validation data (ZAUER) and mobilized data (mobile phone).

A first remark can be made by reflecting on municipalites in border areas between different zauer classes. In this perspective, antenna signatures of two distinct urban areas may sometimes be very similar and thus hard to distinguish (a case rather common in border areas). The difference of about 0.1 between Fuzzy Kappa and Kappa, however, denotes that our algorithms are able to, at least, partly cope with this difficulty by predicting sometimes wrongful but close classes (in terms of similarity of the classes or of spatial location).

A second remark urges us to consider the limitations of mobile phone as a data source for urban areas recognition. CDR data is, by design, subordinate to users' usage and the extracted activity profiles are subordinate to user's movement patterns, both of which might differ between regions and urban areas. In addition, local market shares of single operators are unknown making it impossible to correctly control for representativiness. Other uncertainties stem from uing spatial resolution of the cellular tower network. This resolution does, of course, not collide with administrative borders which renders a discrepancy between information gathering and the proposed classification task. Ultimately, antenna positioning may hinder the antenna signature to be characterized by the presence of (local) populations. Some antennas might capture only specific behavior of local subscribers when, for instance, being positioned along transport axes.

The choice of the methods seems less at stake. One alternative would have been to recourse to unsupervised techniques, which is often done in land use literature [Aguilra et al., 2014, Soto and Martinez, 2011]. Yet differences in antennas signatures can be interpreted in multiple ways. Exploring supervised classification therefore appears as a useful preliminary step for relevant features extraction. In this context, classification algorithms with feature selection designs like penalized logistic regression and random forests are extremely useful as they allow to identify the features contributing most to the discrimination of the classes of interest (amplitude of the coefficients in the penalized logistic regression, importance measure of the variables in random forests), hence leveraging interesting insights.

7 Conclusion

Concluding, we would like to consider improvements and applications. In terms of applications, our results encourage us to promote the use of mobile phone data as an alternative source for producing recurrent urban area zoning between official but less frequent releases. Specifically, We are quite optimistic on using supervised classification to, for example, show patterns on the emergence of urban centers or the progression of urban areas. We reckon, however, that assessments regarding the urbanization of rural and isolated areas should remain cautious as our classification tasks underperformed there. Thinking improvement, we hope that more recent sources of CDR data, or other sources like DDR (data detail record data) could provide for more dense and frequent observations in remote areas, improving our automated classification. Ultimately, we hope that the results of our classification case can encourage a more widespread use of machine learning techniques in official statistics.In our work we have shown that the application of such techniques is rather straightforward and can be instructive for both future and past work.

References

[Aguilra et al., 2014] Aguilra, V., Milion, C., and Allio, S. (2014). Territory analysis using cellphone data. *Transport Research Arena 2014, Paris*.

[Berry et al., 1969] Berry, B. J. L., Goheen, P. G., and Goldstein, H. (1969). *Metropolitan area definition: A re-evaluation of concept and statistical practice*, volume 28. [Washington]: US Bureau of the Census.

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

[Duranton and Puga, 2014] Duranton, G. and Puga, D. (2014). Urban land use.

[Floch and Levy, 2011] Floch, J. and Levy, D. (2011). Poursuite de la priurbanisation et croissance des grandes aires urbaines. *INSEE Premire*, 1375.

[Galster et al., 2001] Galster, G., Hanson, R., Ratcliffe, M. R., Wolman, H., Coleman, S., and Freihage, J. (2001). Wrestling sprawl to the ground: defining and measuring an elusive concept. *Housing policy debate*, 12(4):681–717.

[Hagen, 2003] Hagen, A. (2003). Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, 17(3):235–249.

[Hagen-Zanker, 2009] Hagen-Zanker, A. (2009). An improved fuzzy kappa statistic that accounts for spatial autocorrelation. *International Journal of Geographical Information Science*, 23(1):61–73.

[Landis and Koch, 1977] Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

[Schapire, 2003] Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer.

[Soto and Martinez, 2011] Soto, V. and Martinez, E. F. (2011). Automated land use identification using cell-phone records. *In Proceedings of the 3rd ACM international workshop on MobiArch, HotPlanet, New York, NY, USA*, 11:17–22.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, page 301320.