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

Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area

Analizzare i comportamenti di mobilità urbana attraverso i dati GPS: un'applicazione all'area metropolitana fiorentina

Chiara Bocci, Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni Leonardo Piccini

Abstract Big Data, originating from the digital breadcrumbs of human activities, let us observe the individual and collective behaviour of people at an unprecedented detail. In this paper we investigate the informative potential of the digital tracking that GPS-enabled devices can offer to academic research and to policy makers, with a specific attention for urban and metropolitan settings. The unstructured nature of the dataset requires a careful consideration and correction of possible biases which could lead to unreliable results. We use the 2011 census commuting matrix as a validation tool for our proposed methodology. GPS data contain information that would not be otherwise available, i.e. non-systematic mobility patterns. The produced estimates are then used to analyse mobility patterns within the Florence Metropolitan Area in a more exhaustive and detailed form.

Abstract *L'evoluzione tecnologica ha portato, nel corso degli ultimi anni, ad un notevole incremento dei dispositivi in grado di produrre e memorizzare tracce digitali dei nostri comportamenti quotidiani. In questo lavoro vogliamo indagare il potenziale informativo contenuto nelle tracce prodotte da apparecchi dotati di sistemi GPS per scopi di ricerca o di pianificazione delle politiche, con particolare riferimento all'ambito urbano e metropolitano. La natura spontanea e non strutturata dei dati richiede un'attenzione particolare alle possibili fonti di distorsione. Utilizziamo la matrice di pendolarismo del censimento 2011 come strumento di validazione dei risultati. I dati GPS contengono informazioni altrimenti di difficile reperibilità, come i comportamenti di mobilità non-sistematica. Le stime ottenute sono utilizzate per un caso di studio incentrato sull'Area Metropolitana Fiorentina.*

Key words: big data, mobility, urban planning, O-D matrix validation

Chiara Bocci, Leonardo Piccini
IRPET, Firenze, Italy e-mail: name.surname@irpet.it

Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni
KDDLAB ISTI CNR, Pisa, Italy e-mail: name.surname@isti.cnr.it

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

Technological evolution brought along, in recent years, a remarkable increase in the diffusion of devices that can record digital footprints of our behaviour on a daily basis, tracking a vast degree of activities. Constant and basically unintentional production of such tracks generates huge datasets that contain a precious quantum of information about socio-economic behaviour that may be extracted and used for socio-economic research and for policy analysis [1].

Big data sources may support policy makers in the ex-ante phase of policy implementation, by providing a more sophisticated depiction of the socio-economic environment and may be used for ex-post evaluation purposes in quasi-experimental design and counterfactual settings.

Literature on the matter and practical experiences have highlighted pros and cons of this approach [5]. Some of the pros include timeliness, cost effectiveness, spatial and temporal disaggregation, emergence of unexpected and/or unobservable phenomena. On the other hand, since the relative novelty of the methodologies used to deal with these data, extra carefulness needs to be used to acknowledge possible shortcomings in terms of quality, accessibility, applicability, relevance, privacy policy and ownership of the data, all of which may affect the quality of policy evaluation and appraisal. Nonetheless, we believe that big data sources can be successfully used to foster the capabilities of the public institutions to deal with complex problems, to plan effective policies and to evaluate the outcomes of their actions. To this extent, we propose a methodology that allows us to use data collected from GPS-enabled devices, installed on private vehicles for insurance purposes, to analyse and understand mobility patterns within a urban setting.

2 Research statement, objectives and data sources

The aim of the paper is to find a viable method to use GPS data to produce a non-biased Origin-Destination matrix for the selected study area, i.e. the Florence Metropolitan Area. Since the GPS dataset is derived from private car mobility, our focus will be mainly on this type of flows. However, since we want to use our estimates to assess the intensity and characteristics of the relations between different geographic zones within the metropolitan area, we need to find a way to correct for the different propension on public transport usage which we expect to observe across the different Origin-Destination pairs.

Typically, this kind of data is collected systematically every 10 years, during the nationwide official census. However, census data, while very rich with information and details, has two major drawbacks: the temporal lag between census, during which we have no information on mobility, and the focus on what we call systematic mobility, i.e. the mobility which happens almost every day and is mainly related to home-to-school or home-to-work trips, leaving out an increasingly relevant segment of non-systematic mobility, which, by its own nature, is difficult to capture with tra-

ditional methods. If our methodology is correct, we can thus increase our analytical capability with an informative base that can be updated almost continuously and that includes all mobility and not only the systematic one.

For this study we use GPS data that are provided by a leader company in the Insurance Telematics that deals with about the 2% of the total vehicles circulating in Italy. Our dataset counts about 150k private vehicles crossing Tuscany in the month of June 2011, and represents a primary source of information for studying the mobility behaviours. Data on vehicle fleet in Italy provided by the Italian Automobile Club (ACI), Census data provided by ISTAT, and trip duration and distances with different transportation means computed using Google services are used to re-scale the vehicle sample to the real mobility flows. Once we validate our data and estimate a reliable O-D matrix, we can use the data to carry out an extensive descriptive analysis of mobility patterns in our selected geographic area. To demonstrate the informative potential of this kind of data, we choose the Florence Metropolitan Area as a case study.

3 Estimating a detailed O-D Matrix using GPS data

As we previously discussed, GPS data contain an inherent bias: they account only for private cars usage (specifically, for the fraction of vehicles that have a GPS device installed for insurance purposes and that are being monitored by our provider).

Since we want to use GPS data for socio-economic analysis and for policy planning and evaluation, we need to find a way to scale back the flows that we observe towards our real population, which means accounting for (at least) three missing dimensions:

1. We observe vehicles, but we want to estimate the number of people actually travelling, which means accounting for average car occupation;
2. We observe a fraction of vehicles that is geographically heterogeneous, so we want to account for different market penetration by our provider;
3. We observe only private cars, so we want to account for an heterogeneous share of public transport users.

In order to estimate a complete O-D Matrix, we use the 2011 Census Origin-Destination Matrix as a validation tool. Such matrices are usually released with a territorial detail that corresponds to the administrative units of municipalities. These matrices contain information on municipality of origin, municipality of destination, time of departure, duration of the trip, mean of transport, gender and purpose of the trip (work- or school-related). A geographically more detailed matrix is also released by ISTAT, with a disaggregation to the census zones, but with less information on the characteristics of the trip (only the purpose). Since we want to be able to estimate an O-D Matrix to analyse urban areas, we want a sub-municipality disaggregation for lager municipalities. We therefore use the more detailed matrix for our validation. Since we also need at least the mean of transportation for our validation

we split our flows using the share of public transportation registered between the corresponding municipalities.

Our starting dataset is comprised of systematic trips observed over the month of June 2011 and aggregated by the 2011 Census zone partitions. If we hypothesise our data to be a random sample extracted from the population of all car movements happening within Tuscany borders during our time frame, we can estimate our target values (a census zone O-D matrix of people using all available means of transportations) with the following formula

$$XFlow_{i,j} = flow_{i,j} * car.pen_i * avg.occ_i * public.t.ind_{i,j}$$

where $XFlow$ is our desired estimate from zone i to zone j, $flow$ is our observed flow from zone i to zone j, *car.pen* is the market penetration of our data provider for the municipality within which zone i falls, *avg.occ* is the average occupancy rate for systematic mobility departing in municipality within which zone i falls (derived from census data), *public.t.ind* is a public transport accessibility index calculated between zone i and zone j (calculated using google services).

4 Validating GPS data using the Census O-D Matrix

Once we estimate our O-D Matrix, we want to check how our estimates perform against our reference values, i.e. the 2011 census O-D Matrix. Literature on matrix comparison has produced different indicators that asses how similar two matrices are (see [2] to an detailed presentation and discussion of these indicators). Moreover, one recent thread of research has been trying to evaluate the similarity of O-D matrices by using image quality assessment techniques mutuated from image processing methodologies [6, 7]. We test the performance of our estimated matrix applying different measures: some classical statistical indicators (like the R^2 association measure, the Root Mean Square Error (*RMSE*) and the Pearson χ^2 test), the Geoffrey E. Havers (*GEH*) statistic which evaluate the level of closeness of each pair of cell of the two matrices, and the recently proposed (and still under study) Mean Structural Similarity Index (*MSSIM*) by Van vuren and Day-pollard [6], which compare two O-D matrices considering the means, variances and covariance of contiguous matrix cells evaluated within a moving block of cells in each matrix.

5 Using the estimated Matrix for socio-economic analysis

Once we have validated the estimation, we can use our matrix to produce a variety of indicators that can help policy makers understand the connection and mobility patterns which operate within their territories. The case study area that we selected is the Florence Metropolitan Area.

5.1 Filling the gaps and assessing mobility patterns

We can use our methodology to estimate an O-D Matrix for subsequent years and compare the results as a time series. Moreover, since we can unpack our matrix in a spatially detailed manner, we can assess mobility patterns within the municipality boundaries. As an example, in Figure 1 we can determine the average speed of the observed trajectories as it varies hourly within the day and for different partition of the city (in this case, the 5 administrative neighbourhoods of the city of Florence).

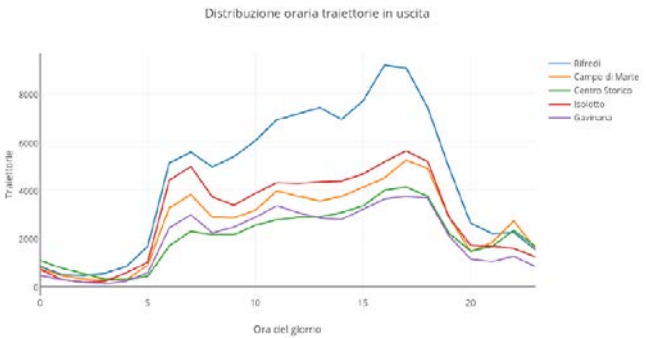

Fig. 1 Average speed by hour and zone of departure

5.2 The boundaries of the city

Generally the border of the city are measured looking at just census data i.e. population density in absolute terms, or the variation over time [4, 3]. We propose a clustering approach aimed at partitioning territories on the basis of human movements inferred using Big Data.

The aim of our work is to contribute to this debate, by providing a tool for policy makers to build a novel definition of regions, seen as functional areas. Focused on the Metropolitan Area of Florence, we aggregate territories that maximise internal traffic and minimise external one.

Given two generic nodes a and b, we define internal traffic, the sum of the flows from a to b and vice versa. For each pair a, b we calculate the distance matrix as the percentage of internal flow respect to the total flows. The clustering methods seeking the best partitioning minimises the distances contained in the matrix provided as input, so each pair of the distance matrix is calculated as $d = 1 - \%internal\,flows$.

We provide the distance matrix as input to DBSCAN and we evaluate the possible values of epsilon and we extract the territorial partition shown in Figure 2.

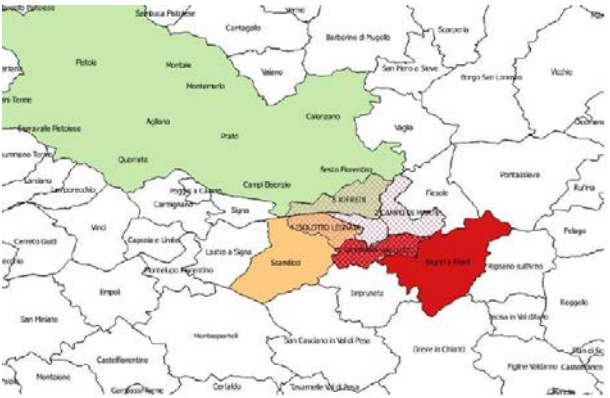

Fig. 2 Boundaries of Florence Metropolitan Area using GPS data

6 Conclusions and future research

The proposed methodology allows us to reliably use GPS data for urban mobility behaviour analysis, without relying on the snapshot provided every ten years by the national census. The informative potential of this source is very high and flexible. Future lines of research include expanding the methodology to validate non-systematic data and further validation using GSM data (call records from mobile phones usage).

References

1. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J **20**(5), 695–719 (2011)
2. Hollander, Y., Liu, R.: The principles of calibrating traffic microsimulation models. Transportation **35**(3), 347–362 (2008)
3. ISTAT: La nuova geografia dei sistemi locali. ISTAT (2015)
4. OECD: Redefining Urban: a new way to measure metropolitan areas. OECD (2012)
5. Scannapieco, M., Virgillito, A., Zardetto, D.: Placing Big Data in official statistics: a big challenge. In: NTTS 2013 Proceedings (2013)
6. Van Vuren, T., Day-Pollard, T.: 256 shades of grey - comparing OD matrices using image quality assessment techniques. In: 2015 Scottish Transport Applications Research Conference Proceedings (2015). URL http://www.starconference.org.uk/star2015.html
7. Zhou, W., Bovik, A., Sheikh, H.: Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process **13**(4), 600–612 (2004)