Space-time clustering for identifying population patterns from smartphone data

Clustering spazio-temporale per dati smartphone sulla distribuzione della popolazione

Francesco Finazzi and Lucia Paci

Abstract In this work we aim at studying spatio-temporal patterns of the population movement across a large city. We exploit the information on people position collected by the smartphone application of the Earthquake Network project and we adopt a dynamic model-based clustering approach to identify the patterns. The approach is applied to smartphone data collected in Santiago (Chile) over the period February-April 2016. Some preliminary results are presented and discussed.
Abstract *L'obiettivo di questo lavoro è studiare i pattern spazio-temporali di movimento della popolazione su una grande città. Sfruttiamo l'informazione sulla posizione delle persone raccolta dall'applicazione smartphone del progetto Earthquake Network ed applichiamo un approccio di clustering dinamico per identificare i gruppi. L'approccio è applicato ai dati smartphone raccolti per la città di Santiago (Cile) lungo il periodo febbraio-aprile 2016. Alcuni risultati preliminari sono presentati e discussi.*

Key words: Finite mixture models, Markov chain Monte Carlo, spatio-temporal modeling, state-space, crowd-sourcing data

1 Introduction

Detecting population dynamics over short periods (e.g. daily movements) may provide the public with useful information to improve traffic infrastructure associated with spatio-temporal commuting patterns, upgrade accessibility or attractiveness of

Francesco Finazzi
Department of Management, Information and Production Engineering, University of Bergamo, Dalmine, Italy, e-mail: francesco.finazzi@unibg.it

Lucia Paci
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy e-mail: lucia.paci@unicatt.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

areas interested by less people than others, enhance public transportation according to infrastructure/open space utilization. Indeed, population patterns are characterized by drastic changes during the day according to several activities such as eduction, working, recreation, visiting and shopping activities, among others.

Customary, population studies are based on census data that do not allow to capture population movements in short periods. Rather, mobile-based data collected over a given region at high temporal scale offers new opportunities to study population distribution and movement patterns over such region. For instance, Secchi et al (2015) proposed a non-parametric method for the analysis of spatially dependent functional mobile network data to identify subregions of the metropolitan area of Milan sharing a similar pattern along time, and possibly related to activities taking place in specific locations and/or times within the city.

Alternatively, we can identify potential partitions of the space and study their evolution over time to extract useful and concise information from smartphone-based data that is helpful to investigate population dynamics. Recently, Paci and Finazzi (2017) proposed a model-based approach to identify clusters in data collected at fixed spatial locations and time steps. Within finite mixture modeling, spatio-temporally varying mixing weights are introduced to allocate observations at nearby locations and consecutive time points with similar cluster's membership probabilities. As a result, a clustering varying over time and space is accomplished. Conditionally on the cluster's membership, a state-space model is deployed to describe the temporal evolution of the sites belonging to each group.

In this work we employ the dynamic space-time clustering approach to explore population dynamics and motion patterns over the city of Santiago (Chile) using data coming from the Earthquake Network project (`www.earthquakenetwork.it`). The project implements a crowdsourced earthquake early warning system based on smartphones networks (Finazzi and Fassò, 2016) and it requires to collect the precise location in space of smartphones at regular time steps. Here, it is assumed that the smartphone location is also the position in space of its owner.

2 Bayesian space-time mixture modeling

Let $y_t(\mathbf{s})$ be a response variable observed at time t ($t = 1, ..., T$) and location $\mathbf{s} \in \mathbb{R}^2$. We assume that observation $y_t(\mathbf{s})$ comes from a finite mixture model, that is

$$f(y_t(\mathbf{s}) \mid \pi, \Theta) = \sum_{k=1}^{K} \pi_{t,k}(\mathbf{s}) f(y_t(\mathbf{s}) \mid \Theta_k) \tag{1}$$

where K is the number of components. The distribution under the k-th component ($k = 1, \ldots, K$) is denoted by $f(\cdot \mid \Theta_k)$ where f is a density function of specified form and Θ_k denotes the set of parameters of each component distribution. The mixing probability $\pi_{t,k}(\mathbf{s})$ is the probability that the location $\mathbf{s}$ belongs to component k at time t and it satisfies $\pi_{t,k}(\mathbf{s}) > 0$ with $\sum_{k=1}^{K} \pi_{t,k}(\mathbf{s}) = 1$ for each $\mathbf{s}$ and t.

As usual in Bayesian analysis, a hierarchical formulation of the mixture model is exploited to facilitate the computation. For each observation, we introduce a latent allocation variable, $w_t(\mathbf{s})$, that identifies the component membership of $y_t(\mathbf{s})$, that is $Pr(w_t(\mathbf{s}) = k) = \pi_{t,k}(\mathbf{s})$. In other words, we assume that the allocation variables $w_t(\mathbf{s})$ are conditionally independently distributed given $\pi_{t,k}(\mathbf{s})$ and they come from a multinomial distribution. Given the latent $w_t(\mathbf{s})$, the observations $y_t(\mathbf{s})$ are independent with $f\big(y_t(\mathbf{s}) \mid w_t(\mathbf{s}) = k, \Theta\big) = f\big(y_t(\mathbf{s}) \mid \Theta_k\big)$. As customary in model-based clustering, we interpret each mixture component as a cluster, such that observations are partitioned into mutually exclusive K groups.

The mixing probabilities, $\pi_{t,k}(\mathbf{s})$, are allowed to vary from observation to observation, i.e., across space and over time. Space-time dependence in the observations is introduced through the prior distribution of the weights such that observations corresponding to nearby locations and consecutive time points are more likely to have similar allocation probabilities than observations that are far apart in space and time. For each location $\mathbf{s}$ and time t, the weights take the form

$$\pi_{t,k}(\mathbf{s}) = \frac{\exp\big(\mathbf{x}_t'(\mathbf{s})\beta_k + \phi_{t,k}(\mathbf{s})\big)}{\sum_{l=1}^{K} \exp\big(\mathbf{x}_t'(\mathbf{s})\beta_l + \phi_{t,l}(\mathbf{s})\big)} \tag{2}$$

where $\mathbf{x}_{t,k}(\mathbf{s})$ is a $p \times 1$ vector of covariates, $\phi_{t,k}(\mathbf{s})$ are spatio-temporal random effects and $\beta_1 = 0$ and $\phi_{t,1}(\mathbf{s}) = 0$ $(t = 1, \ldots, T)$ to ensure identifiability. The logistic-type transformation in (2) guarantees that the two conditions mentioned in Section 2 are satisfied (Fernández and Green, 2002). When available, covariates may help in predicting group membership's probabilities while random effects provide adjustment in space and time to the explanation provided by covariates. Therefore, the response distribution is allowed to vary in flexible ways across time, space and covariate profiles.

To allow for dynamics over time and dependence over space we assume, for $k = 2, \ldots, K$,

$$\phi_{t,k}(\mathbf{s}) = \rho_k \phi_{t-1,k}(\mathbf{s}) + \zeta_{t,k}(\mathbf{s}) \tag{3}$$

where $\zeta_{t,k}(\mathbf{s})$ are independent-in-time spatially correlated errors coming from a zero-mean Gaussian Process (*GP*) equipped with an exponential spatial covariance function. Although the $K-1$ spatio-temporal random effects $\phi_{t,k}(\mathbf{s})$ are assumed to be independent, the corresponding weights are not independent given their definition in (2). The space-time structure of random effects $\phi_{t,k}(\mathbf{s})$ allows to borrow strength information from nearby sites and consecutive time steps. As a result, similar outcomes at near space and time points are assigned with similar cluster membership's probabilities.

Model (1) requires the specification of the sampling density $f\,(y_t(\mathbf{s}) \mid \Theta_k)$. The approach pursued in this work is based on dynamic linear modeling, often referred to as state-space models. In particular, we assume a dynamic linear model to describe the temporal dynamic evolution of all the sites within component k.

Let $\mathbf{y}_t = (y_t(\mathbf{s}_1), \ldots, y_t(\mathbf{s}_n))'$ be the $n \times 1$ observation vector at time t, where n is the number of locations. Conditionally on the allocation variables, the space-state model is provided by

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{z}_t + \varepsilon_t \tag{4}$$

$$\mathbf{z}_t = \mathbf{G}\mathbf{z}_{t-1} + \eta_t \tag{5}$$

where $\mathbf{z}_t = (z_{t,1}, \ldots, z_{t,K})'$ is the $K \times 1$ state vector, $\mathbf{H}_t$ is a $n \times K$ matrix defined below, and $\mathbf{G}$ is a $K \times K$ stable transition matrix. Finally, $\varepsilon_t \sim N(0, \sigma^2 I_n)$ is the $n \times 1$ measurement error vector and $\eta_t \sim N(\mathbf{0}, \Sigma_\eta)$ is the $K \times 1$ innovation vector.

We now turn to matrix $\mathbf{H}_t$. Suppose that site $\mathbf{s}$ belongs to component k at time t. Then, the i-th row of matrix $\mathbf{H}_t$ contains a single element equal to one at position k, while all the other elements are filled with zeros (Inoue et al, 2007; Finazzi et al, 2015). Note that, the one-zero structure of matrix $\mathbf{H}_t$ is allowed to vary over time according to mixing probabilities $\pi_{t,k}(\mathbf{s})$. Also, we benefit from the borrowing strength of information of all sites belonging to component k at time t, since they all contribute in estimating the common latent state $z_{t,k}$. Given the specification in (5), the desired temporal pattern of cluster k is represented by latent state $z_{t,k}$.

Fully inference is provided under a Bayesian framework. The hierarchy of the model is completed by independent noninformative prior distributions for all the hyperparameters and Monte Carlo Markov Chain (MCMC) algorithms are employed to approximate the joint posterior distribution; see Paci and Finazzi (2017) for all fitting details and posterior computation. Model fitting is carried out using the MATLAB code DYSC available online at the web page `https://github.com/graspa-group/DYSC`.

3 Analysis of smartphone data

Smartphones taking part in the Earthquake Network project send a heartbeat signal to a central server every around 30 minutes. Signals include the geographic location of the smartphones that is used to estimate the state of the network at any given time.

In this work, we exploit the information carried by the heartbeat signals to study the population movement across the city of Santiago. We consider 24′900 smartphones and we assume they are representative of the entire Santiago population. We partition the city of Santiago into a uniform lattice of $N = 354$ sites and for each site we consider the number of signals on a hourly basis. For each hour of the day, we aggregate signals observed over the period February-April 2016, assuming that the daily motion patterns of the population are stable over the 3 months. Moreover, we distinguish between working days and weekend in order to investigate possible differences. The aggregation leads to two $N \times T$ matrices for the working days and the weekend, respectively, with $T = 24$. Since we aim at studying the motion patterns independently from the number of signals, we standardize each time series with respect to its own mean and variance. This implies that sites are directly comparable. Hence, at each time step, the time series is interpreted in the following way: a negative value means that the number of signals coming from the site is below the site average, while a positive value means that the number of signals is above average.

Figure 1 shows the standardized number of signals received from each site during working days (left panel) and weekend (right panel) over the study period.

At each time step, thus, we apply model described in Section 2 to cluster sites which behave in a similar way with respect to their average and then we explore how the clusters evolve over the 24 hours of the day. We employ the diagnostic tool provided by Paci and Finazzi (2017) to select the number of clusters. The analysis suggests that only two clusters are needed. This is a consequence of the fact that time series are standardized and the number of signals from each sites can be either below or above average. Figure 2 shows the Posterior 95% credible interval of the temporal patterns $z_{t,k}$ for working day signals (left panel) and weekend signals (right panel), where each temporal pattern is related to a cluster. During working days, the separation between the temporal patterns is lower at 7 a.m. and 7 p.m., namely when people commute from home to work and vice-versa. During these hours, signals are more evenly distributed across city than in any other hour of the day. During the weekend, the same effect can be found at 10 a.m. and at midnight.

To provide the clustering, we assign each observations to their most likely group according to the maximum a posteriori probability (MAP) rule. In Figure 3 clustering result can be appreciated for 12 a.m., 8 a.m. and 8 p.m. and for both working days and the weekend. For any given hour of the day, blue and red points are sites with a number of signals below and above the average, respectively. During working days, the number of signals from the city center is below average at night and above average during the day. This pattern is disrupted during the weekend when people tend to move later in the morning and to return home later in the night.

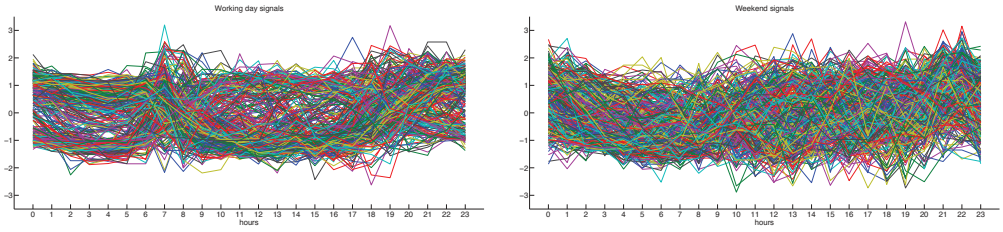

Fig. 1 Number of signals collected from each cell during working days (left panel) and weekend (right panel) over the period February-April, 2016.

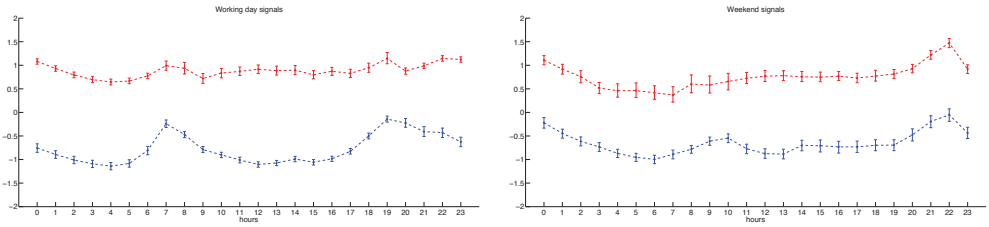

Fig. 2 Posterior 95% credible interval of the temporal patterns $z_{t,k}$ for working day signals (left panel) and weekend signals (right panel).

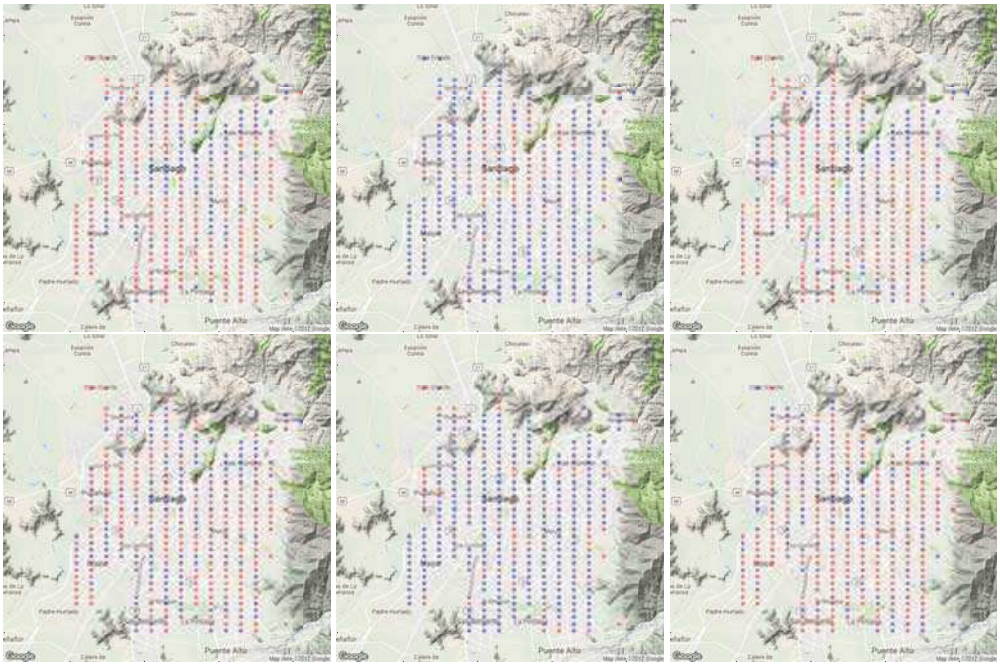

Fig. 3 Clustering result for working day (top row) and the weekend (bottom row) at 12 a.m. (left column), 8 a.m. (middle column) and 8 p.m. (right column). Blue and red dots refer to the blue and red temporal patterns in Figure 2, i.e., below and above the average, respectively.

References

Fernández C, Green PJ (2002) Modelling spatially correlated data via mixtures: A Bayesian approach. J R Stat Soc Ser B 64:805–826

Finazzi F, Fassò A (2016) A statistical approach to crowdsourced smartphone-based earthquake early warning systems. Stoch Environ Res Risk Assess Doi:10.1007/s00477-016-1240-8

Finazzi F, Haggarty R, Miller C, Scott M, Fassò A (2015) A comparison of clustering approaches for the study of the temporal coherence of multiple time series. Stoch Environ Res Risk Assess 29:463–475

Inoue LYT, Neira M, Nelson C, Gleave M, Etzioni R (2007) Cluster-based network model for time-course gene expression data. Biostatistics 8:507–525

Paci L, Finazzi F (2017) Dynamic model-based clustering for spatio-temporal data. Stat Comput DOI 10.1007/s11222-017-9735-9

Secchi P, Vantini S, Vitelli V (2015) Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Stat Methods Appl 24:279–300