Computer Communications 000 (2016) 1–14

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Filtering and windowing mobile traffic time series for territorial land use classification

Gloria Re Calegari*, Emanuela Carlino, Diego Peroni, Irene Celino

CEFRIEL - Politecnico di Milano, via Fucini 2, 20133, Milano, Italy

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Mobile traffic time series
Time series smoothing
Time series decomposition
Time series filtering
Time series windowing
Land use classification
Urban planning
Smart cities

ABSTRACT

Analytics of mobile traffic information may take into account the time-series nature of the data itself. When employing mobile traffic data in a predictive setting to derive useful knowledge to characterize the city environment, the most suitable time series processing methods must be identified. In this paper, we propose an approach to process mobile traffic data using specific time series techniques – smoothing, decomposition, filtering, time-windowing – and to establish the best approach to exploit information extracted from those time series to classify land use, according to sensitivity/specificity metrics. We apply our methodology to a large-scale mobile traffic dataset, we assess its feasibility and we discuss the suitability of different methods for land use classification.

1. Introduction

Mobile traffic data is increasingly available for analytics; this is due to the fact that telecommunication operators are more and more aware of the potential value hidden within the data they manage everyday. Still, mobile traffic information comes in different formats and with various constraints (e.g., users' anonymity), depending on the country and operator. Furthermore, the rise of big data technologies led to an increasing demand for mobile traffic data which are processed jointly with data of different nature; the smart city vision of leveraging ICT solutions to improve urban development encompasses also the integration of heterogeneous datasets in order to form a holistic view of the city itself.

The specific interest in mobile traffic information can be explained because mobile phone usage often reflects human actions: people use their mobile phones on the move as part of their activities on the territory. Moreover, mobile traffic data is produced as a "collateral effect" of the mobile service and thus it is available almost for free. As such, mobile traffic data is very relevant for analytics and adaptable to various purposes.

In this paper, we present our original contribution to the analysis of large-scale mobile traffic data to classify land use. More specifically, we propose a mobile time series processing approach that includes different and complementary methods – smoothing, decomposition, filtering – to characterize the territory in terms of mobile activity and a predictive analytics approach to derive land use classification thereof. Our experiments test the proposed methodology on a specific mobile traffic dataset related to the area surrounding the city of Milano in Italy and are aimed to identify the best time series processing method as well as the most suitable time-windowing strategy. Indeed, mobile traffic data can be very large, since they are continuously generated, and smart approaches are required to get patterns and summarizing indicators out of a wide dataset.

The remainder of the paper is organized as follows: Section 2 reviews the state of the art of mobile traffic data analytics with special reference to land use classification, to better motivate and frame our work; Section 3 sets the problem space and formulate our main research questions; we propose an approach for mobile traffic data processing in Section 4; our experimental settings are illustrated in Section 5 and we discuss the experimental results in Section 6, comparing the different processing methods and time-windowing strategies together with a qualitative analysis of errors; finally, in Section 7 we draw some conclusions and foresee our future work.

* Corresponding author.
E-mail address: gloria.re@cefriel.com (G. Re Calegari).

http://dx.doi.org/10.1016/j.comcom.2016.04.016
0140-3664/

2. Related work

Data from mobile phones, originally intended as a communication tool, is increasingly used as a new source in geography and social sciences research for improving urban planning and for reducing inefficiencies in present-day urban systems. This space-time data represents a way to better understand the urban environment and its dynamics. Actually, it can reveal how the citizens relate to the urban contexts [1]. In this sense, mobile traffic data can be used to infer information about the urban environment and the land use coverage. The mobile traffic data is used in its aggregated form, which consists in taking the total calling time (or counting) in a certain temporal interval and within a given area. The literature review below concerns the objectives for which the phone activity data has been used and the different methods adopted for processing mobile traffic data.

The majority of the studies in literature aim at finding common behaviour in the phone activity signatures in order to automatically identify how citizens use the different geographic regions within a city. Several unsupervised learning techniques are adopted, as *k*-means [2] or hierarchical clustering [3,4]. Other studies follow a predictive approach, trying to derive land use information from mobile phone data. Pei et al., [5] proposes a semi-supervised learning method, the fuzzy c-means clustering (FCM), that uses unlabeled data in conjunction with a small amount of labeled data. Supervised learning methods are adopted by Toole et al., [6] which trained different classification models (Random Forest, Support Vector Machine and *k*-Nearest Neighbor) to infer the land use of the Boston region. In both these researches, the detection rates of the classes are considerably different: the best predicted class is always the one with the highest cardinality, whereas the classes with only few samples are often misclassified. In our research we aim to propose a more robust and more flexible classification algorithm able to predict in a comparable way all the land use classes. To this end we aim at overcoming the problem of unbalanced classes. We also want to investigate whether different kinds of land use require different methods to be better predicted.

Besides the selection of the most suitable predictive model, the choice of the time series pre-processing technique needs to be carefully investigated. Several approaches for analysing the phone activity data aim at finding a smoothed and simplified version of the original signal. The simpler approaches consist in using information about the total call volume, which is an overall characteristic of how many people actually use mobile phones [5], or in describing the time series by a set of indicator points, such as the time at night with the lowest activity, or the time when call activity is increasing the fastest as well as the duration of the active day [7]. The time series have been also summarized using the simple concepts of mean and median: Toole et al., [6] and Soto et al., [2] summarized the whole time series into an hourly average footprint, obtaining a 24-hour time series of an average weekday and one for an average weekend-day, whereas Furno et al., [4] summarized the signal into the concatenation of the medians of the phone activity of each day (median for all Mondays, Tuesdays, Wednesdays, etc.). Reades et al., [8] proposed the eigen-decomposition method, a process similar to factoring and derived from signal analysis, to identify and extract the recurring patterns of mobile phone usage. Cici et al., [3] proposed a smoothing method based on the Fourier transform: the original signal is converted into the frequency domain, where the noise is filtered and only the principal frequencies are kept; the smoothed signal is transformed again in the time domain where the subsequent clustering operations are performed. We think that a deeper analysis of the signal in the frequency domain could bring great benefits, in particular by using the frequencies obtained by the Fourier Transform as predictors of the classification model. We believe that further study could be done in analysing the data as a time series, considering its cyclic, seasonal and trend components. These are the two aspects we investigate in this paper. In addition we would like to perform a comparative analysis of several smoothing methods to understand which is the best pre-processing approach in terms of land use classification accuracy.

3. Problem statement and objectives

Classifying and mapping the land use is far from being a trivial task. Indeed, it usually requires an expensive and partially manual process to collect, integrate and make sense of urban and territorial information. In Europe, the CORINE initiative[1] provides a shared categorization and procedure to support local and regional undertakings that periodically produce or update land use maps, usually starting from remote sensing images.

In this paper, we present our solution to extract urban land use and support smart cities' planning activities; more specifically, we focus on the problem of classifying land use by using mobile traffic data. Our first research question is therefore **RQ1**: *is it possible to detect land use from the analysis of mobile traffic data with an acceptable prediction accuracy?* This question, which has been partially addressed by other researchers, aims at providing further insights on the general scope of mobile traffic analytics when applied to regional planning. The original contribution of our work is twofold: on the one hand, we propose the idea to combine and compare various time series processing approaches for our classification objective (illustrated in Section 4); on the other hand, we illustrate and discuss the positive and interesting findings of our experimental results, obtained notwithstanding the aggregated nature of the employed mobile traffic datasets (explained in Section 5 and 6).

Traffic information is continuously produced by telecommunication operators as part of their billing and network engineering processes [9]. Therefore, this mobile (big) data is continuously available for analysis and elaboration and its size grows over time because of the regular recording activity. However, analytics methods on this kind of data at territorial level could imply very large-scale processing. While we live in the age of smart city big data [10], methods and tools for timely decision making may require reducing the amount of data.

Those considerations lead to our second research question **RQ2**: *which is the mobile traffic time series "simplification" method that produces the best result in terms of land use classification?* In other words, answering this question means understanding to which extent the mobile traffic signal contains "hints" about territorial dynamics and whether different methods are able to detect different land uses (agricultural, residential, industrial areas, etc.). We explore different time series processing techniques – smoothing, decomposition and filtering – and compare them to identify the best fit-for-use in terms of minimization of misclassification error.

Together with simplification in terms of signal complexity, another way to reduce mobile traffic data is to choose a suitable time-frame for analysis. Hence, we formulate our third and final research question **RQ3**: *what is the best windowing strategy on mobile traffic time series that minimizes the misclassification error in land use classification?* This question aims at understanding the influence of time, as well as of special temporal events, on mobile phone activity and, as a consequence, on its employment for land use classification. We consider different time windows and compare the classification results to provide some recommendations on mobile traffic data processing for exploration and decision-making.

[1] Cf. http://www.eea.europa.eu/publications/COR0-landcover.

Please cite this article as: G. Re Calegari et al., Filtering and windowing mobile traffic time series for territorial land use classification, Computer Communications (2016), http://dx.doi.org/10.1016/j.comcom.2016.04.016