Joint Spatial and Temporal Classification of Mobile Traffic Demands

Angelo Furno, Marco Fiore, Razvan Stanica

Joint Spatial and Temporal Classification of Mobile Traffic Demands

Angelo Furno*†, Marco Fiore‡, Razvan Stanica*
* Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA, F-69621 Villeurbanne, France – name.surname@inria.fr
† Université de Lyon, IFSTTAR, ENTPE, F-69675 Lyon, France – angelo.furno@ifsttar.fr
‡ CNR – IEIIT, Corso Duca degli Abruzzi 24, 10129 Torino, Italy – marco.fiore@ieiit.cnr.it

***Abstract*—Mobile traffic data collected by network operators is a rich source of information about human habits, and its analysis provides insights relevant to many fields, including urbanism, transportation, sociology and networking. In this paper, we present an original approach to infer both spatial and temporal structures hidden in the mobile demand, via a first-time tailoring of Exploratory Factor Analysis (EFA) techniques to the context of mobile traffic datasets. Casting our approach to the time or space dimensions of such datasets allows solving different problems in mobile traffic analysis, i.e., network activity profiling and land use detection, respectively. Tests with real-world mobile traffic datasets show that, in both its variants above, the proposed approach *(i)* yields results whose quality matches or exceeds that of state-of-the-art solutions, and *(ii)* provides additional joint spatiotemporal knowledge that is critical to result interpretation.**

I. Introduction

The surge in mobile data traffic –estimated globally at 3.7 exabytes in 2015, with an overall 4,000-fold growth over the past ten years [1]– has fostered a widening interest in better understanding the dynamics of the mobile demand. Knowledge mining of real-world datasets has revealed important features that characterize mobile traffic: examples include a strong temporal periodicity [2] and geographic locality [3] that enable effective prediction of the demand; the appearance of significant fluctuations induced by social events [4] with the consequent need for dedicated resource management policies; or, a neat heterogeneity of the capacity consumed by subscribers [5] that is however captured by a limited number of typical profiles [6], which enables the informed tuning of traffic plans.

Analyses of mobile traffic demand in the literature can be divided into two broad categories [7]: *(i)* works that take a user perspective, and study the behavior of individual subscribers in terms of their mobility, the traffic they generate, and the mobile services they consume; *(ii)* works that take an operator perspective, and investigate properties of the demand aggregated over all users present in a given area, typically a cell sector or the coverage region of a base station.

Our work falls in the second category. Specifically, we focus on the problem of classification, i.e., finding hidden regular structures in the aggregate traffic generated by mobile users. As set forth in Sec. II, previous works have proposed solutions to detect either temporal or spatial structures in the data, but little attention has been paid to the more challenging *concurrent* inspection of both space and time dimensions.

This paper fills the gap above, by presenting an original methodology for the joint spatiotemporal classification of the aggregate demand supplied by a mobile network operator. The proposed methodology stems from Exploratory Factor Analysis (EFA), a well established instrument in psychology research. As detailed in Sec. III, EFA aims at identifying, in a fully automated way, latent factors that cause the dynamics observed in the data. When tailored to the specific use case of mobile traffic classification, as discussed in Sec. IV, EFA offers the possibility of exploring the space and time dimensions of the data at once. This yields two significant advantages.

First, the same methodology can be cast to recognize factors that are temporal or spatial in nature, solving the following two problems in mobile traffic analysis.

• **Network activity profiling** aims at detecting temporal structures in the network-wide communication activity, and allows classifying together time periods that show a similar, stable spatial distribution of the mobile traffic demand. Network activity profiles are expected to find applications in cognitive networking [8], where they can drive the establishment, modification, release and relocation of resources, in concert with the temporal variations in the mobile demand [9].

• **Land use detection** is the identification of spatial structures in the mobile traffic data, through the decomposition of a geographical area into zones where the mobile traffic dynamics over time are homogeneous. When considering metropolitan-scale regions, these zones correspond to land uses – i.e., the combination of urban infrastructures and predominant undertakings of people at those locations. This has applications in geoinformatics, as an effective way to automatically label the urban tissue, at lower cost and with higher accuracy than traditional survey methods [10]. It is also relevant to cognitive networking, since the discovery of city regions with analogous demand evolution may ease the dynamic allocation of spectrum at individual base stations, helping mitigating high fluctuations of resource needs in small network areas [9].

Second, our proposed methodology allows immediate extrapolation of the structures hidden in the secondary dimension of both problems above. In other words, it provides, at no additional cost, knowledge of the spatial patterns that characterize each network activity profile, and of the precise temporal dynamics that distinguish each land use. This plays an important role in the interpretation of classification results.

We demonstrate these advantages by performing a spatiotemporal classification of real-world mobile traffic data recorded by national operators in two major European cities, in Sec. V. When compared with current state-of-the-art techniques dedicated to the two data analysis problems above, our solution easily matches them in terms of quality of the classification; in addition, it simplifies the interpretation of classes through a combined spatiotemporal view of the same.

II. Related work

The study of traffic data collected by mobile operators has found applications across research domains such as urbanism, transportation, sociology, epidemiology, and telecommunication networking [7]. Related to our problem are classifications of the temporal and spatial distributions of the mobile demand.

Fine-grained temporal classifications have mainly focused on outlying situations [11]. Specifically, both planned [12] and unplanned [13] events were found to induce significant variations in the typical temporal structure of the mobile traffic demand. A more complete approach to network activity profiling has been recently presented in [14]: it is based on a dedicated, fine-tuned clustering of *snapshots* of the mobile traffic demand at different time periods. We will compare the results of our proposed solution to those obtained with this framework in the performance evaluation in Sec. V.

As far as high-detail spatial structures in the mobile traffic are concerned, a number of works have revealed the correlations between the geographic diversity of mobile traffic and the urban landscape [15], [16], and employed them for automated land use detection [10], [17], [18]. All previous land use detection algorithms represent the mobile demand in different geographical zones as time series, process them though compression, filtering and normalization, and finally classify them via clustering. Among the proposed solutions, that in [18] has been shown to provide the most accurate results, and we will this consider it as our benchmark in Sec. V.

It is important to note that all previous works explored the temporal and spatial structures of the mobile traffic separately, and we still lack a comprehensive methodology that can address the two dimensions at once. In this paper, we borrow from EFA techniques to achieve such a goal. The roots of factor analysis date back to the work of Spearman, over a century ago [19]. Following those early studies, EFA has emerged as one of the dominant classes of factor analysis, and has been widely employed in statistical psychology research [20]. To the best of our knowledge, this is the first time factor analysis is leveraged for the study of mobile traffic data and, more generally, in the field of wireless networking.

III. Exploratory Factor Analysis

Here, we provide an introduction to EFA fundamentals. We start with some terminology used in the remainder of the paper.

- **Variables** are the set of phenomena of interest, related to some population of individuals. E.g., subjects taught to primary school students.
- **Samples** represent the set of monitored individuals from the given population, for which all phenomena of interest can be measured. E.g., students from a same class.
- **Observations** are the realizations of all variables for each sample. E.g., the grades of examination tests in all subject obtained by each student.
- **Common factors** are complex interrelationship among the observed phenomena that the analyst can reasonably assume to exist. In practical cases, these latent features cannot be directly observed in the data, due to the very large number of variables. Then, the goal of EFA is the recognition of such factors[1], which are supposed to be small in number with respect to the variables, and thus allow for an easier interpretation of the phenomena. E.g., in our example, EFA may aim at inferring factors such as verbal and mathematical intelligence, whose combination could explain the average student's aptitude towards each subject.
- **Factor loadings** are numerical relationships that describe how much each common factor explains each variable. More precisely, the squared loading is the percent of variance in a variable explained by a common factor. Loadings close to the 1 or -1 extremes indicate that the factor strongly affects the variable, with a positive or negative correlation, respectively; instead, loadings close to zero indicate that the factor has a weak effect on the variable. As such, loadings are the main instrument to label the factors returned by EFA. E.g., a factor that has high loadings solely in algebra and geometry can reveal the existence of a common mathematical intelligence that explains the performance of the majority of students in mathematics-related disciplines.
- **Factor scores** are values that relate samples to common factors. For a given sample and factor pair, a high (low) score indicates that the sample has a ranking on the factor that is much above (below) the average. Interestingly, scores allow inspecting samples in the light of factors, and thus complement loadings when it comes to result interpretation. E.g., scores indicate if the good (poor) performance in scientific disciplines of any subset of students is especially well explained by their strong (weak) mathematical intelligence.
- **Unique factors** model situations where the data transcend common factors, by explaining the unique variance associated to each variable. Unique factors are thus useful to pinpoint outlying behaviors in the data. E.g., unique factors can account for a rare talent of one student towards a specific discipline. Due to its peculiarity, such a talent is not captured by any intelligence towards subjects that is commonly found in schoolchildren.

A. Fundamental model

Given a set of observed variables of interest, factor analysis is formally defined as *"a model of hypothetical component variables that explain the linear[2] relationships existing between observed variables"* [21]. Such a hypothetical set of component variables can be derived mathematically from the observed variables, as follows.

Let $\mathbf{X}$ be a $N \times 1$ vector of observed *variables*, distributed with expectation $\mathbb{E}(\mathbf{X}) = 0$ and covariance $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X})$. Let also $\mathbf{F}$ be a $K \times 1$ vector of unknown normalized *common factors*, having mean $\mathbb{E}(\mathbf{F}) = 0$, covariance $\mathbf{\Phi} = \mathrm{Cov}(\mathbf{F})$ and order $K < N$. Next, let $\mathbf{\Lambda}$ be an unknown $N \times K$ matrix of common factor pattern coefficients (i.e., *factor loadings*). Let also $\mathbf{U}$ be a $N \times 1$ vector of independently distributed error terms (i.e., *unique factors*), with mean $\mathbb{E}(\mathbf{U}) = 0$ and finite covariance $\mathbf{\Psi} = \mathrm{Cov}(\mathbf{U})$. Since each unique factors is specific to one variable, the error terms are independent, and $\mathbf{\Psi}$ is a diagonal matrix. Finally, we want common factors and unique factors to be uncorrelated, i.e., $\mathrm{Cov}(\mathbf{F}, \mathbf{U}) = 0$. Hence,

$$\mathbf{X} = \mathbf{\Lambda F} + \mathbf{U} \tag{1}$$

[1] We will use *common factor* and *factor* interchangeably in the following.

[2] The linearity of relationships among variables in the specific context of mobile traffic will be discussed in Sec. IV-B.

is the *fundamental equation of factor analysis*, stating that the observed variables in $\mathbf{X}$ are weighted combinations of the common factors in $\mathbf{F}$ and the unique factors in $\mathbf{U}$. From (1), the covariance of the observed variables $\mathbf{X}$ can be written as

$$\begin{aligned}\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{\Lambda F} + \mathbf{U}) = \\ \mathbf{\Lambda}\mathrm{Cov}(\mathbf{F})\mathbf{\Lambda}^\intercal + \mathrm{Cov}(\mathbf{U}) = \\ \mathbf{\Lambda\Phi\Lambda}^\intercal + \mathbf{\Psi},\end{aligned} \tag{2}$$

which represents the *fundamental theorem of factor analysis*. In the case of EFA, no hypotheses concerning the factors are made, and it is thus generally assumed all factors to be orthogonal, i.e., mutually uncorrelated and with unit variances. Thus, $\mathbf{\Phi}$ can be replaced by the identity matrix in (2), and

$$\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}^\intercal + \mathbf{\Psi}, \tag{3}$$

whose i-th diagonal element can be written as

$$\sigma_{ii} = \mathrm{Var}(x_i) = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii} = h_i + \psi_{ii}. \tag{4}$$

From (4), the variance of each observed variable, σ_{ii}, consists of two parts: the *communality* h_i, i.e., the portion of the variance shared with the other variables via the common factors, and the *unique variance* ψ_{ii}, i.e., the share specific to each variable, via the associated unique factor.

B. Factor extraction

Maximum Likelihood Estimation (MLE) allows inferring the unknown variables $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ in (3) in a way that is efficient and robust [21]. MLE assumes $\mathbf{X}$ in (1) to have a multivariate normal distribution[3] with mean $\bar{\mathbf{X}} = \frac{1}{M}\sum_{a=1}^{M}\mathbf{X}_a$ and covariance $\mathbf{S} = \frac{1}{M-1}(\sum_{a=1}^{M}\mathbf{X}_a\mathbf{X}_a^\intercal - M\bar{\mathbf{X}}\bar{\mathbf{X}}^\intercal)$ computed from the M observations. MLE maximizes the likelihood function

$$\ln L = -\frac{1}{2}(M-1)[\ln|\mathbf{\Sigma}| + \mathrm{Tr}(\mathbf{S\Sigma}^{-1})], \tag{5}$$

with Tr indicating the matrix trace operator. The $\mathbf{\Sigma}$ matrix maximizing (5) also minimizes the following fit function [22]

$$F_K(\mathbf{\Sigma}) = \ln|\mathbf{\Sigma}| + \mathrm{Tr}(\mathbf{S\Sigma}^{-1}) - \ln|\mathbf{S}| - N, \tag{6}$$

where K refers to the number of common factors considered. Using (3) in (6), the expression $F_K(\mathbf{\Sigma}) = F_K(\mathbf{\Lambda}, \mathbf{\Psi})$ can be used to compute the maximum likelihood estimates of the unknowns $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. We outline the main steps below, while full details are found in [21].

Firstly, F_K is minimized with respect to $\mathbf{\Lambda}$, where the minimizer $\tilde{\mathbf{\Lambda}}$ is computed by imposing $\frac{\partial F_K}{\partial \mathbf{\Lambda}} = 0$. Denoting as $\Im$ the identity matrix, the above condition leads to

$$\tilde{\mathbf{\Lambda}} = \mathbf{\Psi}^{1/2}\mathbf{\Omega}_K[\gamma_i - 1]_K^{1/2}, \tag{7}$$

[3]MLE yields good estimations even when the actual distribution of $\mathbf{X}$ is not multivariate Gaussian [21]. We ran tests with alternative methods like Minres and Principal Axis that do not rely on this assumption, and they provided results (omitted due to space limitations) consistent with those in Sec. V.

where the diagonal matrix $[\gamma_i - 1]_K$ contains the K largest eigenvalues of $\mathbf{\Psi}^{-1/2}\mathbf{S\Psi}^{-1/2}$, and $\mathbf{\Omega}_K$ contains the corresponding eigenvectors. Replacing (6) in (7) one can derive the expression of the conditional minimum for a given $\mathbf{\Psi}$, as

$$f_K(\mathbf{\Psi}) = -\sum_{j=K+1}^{N}\ln\gamma_j + \sum_{j=K+1}^{N}\gamma_j - (N-K), \tag{8}$$

where γ_j, with $j = K+1, \ldots, N$, are the residual eigenvalues of the matrix $\mathbf{\Psi}^{-1/2}\mathbf{S\Psi}^{-1/2}$.

Secondly, the function f_K is minimized with respect to $\mathbf{\Psi}$, by imposing $\frac{\partial f_K}{\partial \mathbf{\Psi}} = 0$, which leads to the expression

$$\mathrm{Diag}(\mathbf{\Psi}^{-1}(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^\intercal + \mathbf{\Psi} - \mathbf{S})\mathbf{\Psi}^{-1}) = 0. \tag{9}$$

At this point, the maximum likelihood estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ can be computed by means of an iterative procedure based on the Fletcher-Powell method and applied to the function f_K and its partial derivatives in (8) and (9), respectively.

C. EFA and Principal Component Analysis

The structure of the fundamental equation in (1) hints at the fact that factor analysis is a close relative of Principal Component Analysis (PCA), a popular tool for multivariate analysis. Therefore, a legitimate question is why EFA is more relevant than PCA to the problem we are trying to solve, i.e, the classification of mobile traffic demands.

To answer this question, we recall that PCA aims at finding orthogonal linear combinations of the variables that maximize the total variance in the data. In other words, PCA looks for the major sources of variation in data, or, equivalently, for the lowest number of components that explain the available observations. Such an objective lends itself to data dimensionality reduction, which is in fact the natural application of PCA.

EFA fundamentally differs from PCA in that it distinguishes between shared and unique variances in the data, modelled by common and unique factors, respectively. This isolates sampling noise (i.e., unique factors) during the process, and allows focusing more precisely on the actual latent variables that explain correlations in the observed data [23].

As a result, the decision whether to use PCA or EFA must be based on the purpose of the analysis, i.e., dimensionality reduction or identification of latent correlations, respectively. This is no minor difference, as shown, e.g., in a recent experimental evaluation [24]. By assessing the severity of errors due to PCA misuse, the study reveals that factor analysis consistently and significantly outperforms PCA in explaining correlation matrices. The conclusion is that one should never pretend that PCA components are common factors.

Reverting to our problem, we deal with classification, i.e., the identification of hidden regular structures in the data. These structures are primarily driven by strong correlations that are difficult to observe in practice, entangled as they are within the mass of observations: thus, our classification problem is in fact a correlation extraction problem. In the light of the considerations above, it is clear that EFA, and not PCA, is the appropriate tool for our purposes.

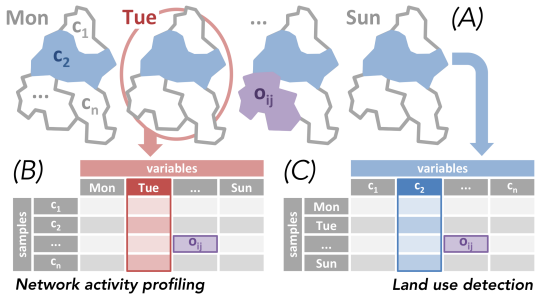


Fig. 1. Mobile demand classification with EFA in a toy scenario. (A) The one-week demand in the target region is aggregated daily with respect to a spatial tessellation of n cells. The resulting demand in the i-th cell during the j-th day is EFA observation o_{ij}. (B) Network activity profiling: days are the EFA variables, each characterized by a set of observations over the cell samples. (C) Land use detection: cells are the EFA variables, each characterized by a set of observations over the daily samples. Figure best viewed in colors.

IV. Mobile Demand Classification with EFA

We discuss how EFA can be tailored to the classification problems of network activity profiling and land use detection, as well as all system parametrizations required to that end.

A. Problem formulation

As anticipated in Sec. I, EFA can be cast to solve two dual problems in the context of mobile traffic demand classification. The input to both problems is an aggregate representation of the communication activity of mobile subscribers in the geographical region of interest. This definition of input is general and can accommodate any level of spatial and temporal aggregation, as well as any notion of mobile user activity (voice, text, data, specific services, etc.): a toy example is provided in Fig. 1. Then, the two problems are set apart depending on the mapping of variables $\mathbf{X}$ in (1), as follows.

Network activity profiling. We model *time intervals* as the EFA variables. Each variable is thus described by the mobile traffic demand (i.e., the EFA observations) recorded over all spatial cells during a given time interval, as shown in Fig. 1. In this EFA configuration, the common factors seeked by EFA are temporal structures that explain at what time instants the spatial distribution of the mobile demand is comparable: these structures are precisely network activity profiles.

An important remark is that, here, spatial cells map to EFA samples: hence, EFA scores relate cells to temporal profiles, revealing which geographical areas are important for a given temporal profile. This allows the joint inspection of the classification results in the space and time dimensions.

Land use detection. EFA variables correspond to *geographical locations*. Each variable consists in the mobile traffic demand (i.e., the EFA observations) recorded at a specific cell through the complete monitoring period, as in Fig. 1. In this EFA configuration, the EFA common factors represent structures in the geographical space that explain in what areas the mobile demand follows similar temporal dynamics: by definition, such areas correspond to land use classes.

Interestingly, time intervals become now the EFA samples. Therefore, EFA scores point up the time periods when the mobile demand is especially distinctive within each land use. This offers an unprecedented spatiotemporal perspective on land uses, and showcases again the potential of EFA for the concurrent spatiotemporal analysis of mobile traffic data.

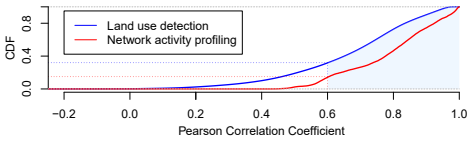


Fig. 2. Distributions of the Pearson correlation coefficient computed between all pairs of EFA variables in the two mobile demand classification problems.

B. Tuning EFA for mobile traffic data analysis

In Sec. IV-A, we shaped EFA so as to solve classification problems in mobile traffic analysis. However, several steps are needed in order to evolve such a fundamental scheme into an operational implementation. These include data verification and EFA parametrization choices, which are discussed next.

Suitability of mobile traffic data for EFA. The definition of factor analysis in Sec. III-A builds on two major hypotheses on the input data: (a) the existence of a non-zero correlation among the observed variables, and (b) the linearity of the functional relationships among the observed variables and the unknown hidden factors. In practical cases, it is important to verify if these assumptions hold for the data to be analysed. Thus, as a preliminary step in our study, we check the suitability of mobile traffic demand datasets for EFA.

Tests exist that are dedicated to this purpose. Specifically, we run the *Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy* [25] on our reference datasets (presented in detail in Sec. V-A). The test returns values in the range $[0, 1]$, where results close to 1 indicate a high suitability of the data to EFA. In both our classification problem formulations, and for all datasets, KMO returns values around 0.99.

As additional checks, we verify: *(i)* the linearity of all pairwise relationships between EFA variables in the two mobile demand classification problems, finding strong correlation in 70–80% of cases, as shown in Fig. 2; *(ii)* the sample-to-variable ratio, finding that it is always much larger than one, which is typically considered as a good rule of thumb for a meaningful factor analysis. In the light of all these results, mobile traffic data appear as an excellent candidate for EFA.

Choice of the number of common factors. An important design choice concerns the number of common factors that EFA should target. We rely on *parallel analysis* (PA) [26], which uses the eigenvalues of the data correlation matrix as rough estimates of the actual common factors. Specifically, PA compares such eigenvalues against those of uncorrelated normal variables that mimic the data variables (i.e., come in the same quantity, with identical sample size). The presence of common factors shall induce large eigenvalues: the number of factors is set to the lowest rank above which all data eigenvalues are larger than those from the uncorrelated variables.

Factor rotation. The common factors that satisfy (1) are subject to *rotational indeterminacy*, i.e., they are not mathematically unique, and linear transformations allow moving across the full space of solutions. A sensible rotation of common factors can maximize high loadings and minimize low loadings. As explained in Sec. III, loadings are the main instrument to link factors and variables: thus, the presence of fewer strong (i.e., closer to 1 or -1) loadings outlines more neatly structures in the data and eases result interpretation.

We use *VARIMAX rotation* [27] to identify the most appropriate rotation of factors. Given the unrotated $N \times K$ loading

matrix $\mathbf{\Lambda}$, VARIMAX iteratively finds a $K \times K$ orthonormal transformation matrix $\mathbf{T}$ such that $\mathbf{\Lambda T}$ maximizes

$$\sum_{j=1}^{K} \frac{N \sum_{i=1}^{N} (a_{ij}^2/h_i^2)^2 - \left(\sum_{i=1}^{N} a_{ij}^2/h_i^2\right)^2}{N^2}, \qquad (10)$$

where a_{ij} are the elements of $\mathbf{\Lambda T}$ and h_i is the communality of the i-th variable defined as in (4) and computed from $\mathbf{\Lambda T}$.

V. Evaluation with measurement data

To assess the performance of EFA in the context of mobile demand classification, we leverage metropolitan-scale datasets of real-world mobile traffic, presented in Sec. V-A. We show a selection of results that cover network activity profiling, in Sec. V-B, and land use detection, in Sec. V-C, with diverse datasets. Full results obtained from all combinations of classification problems and datasets are consistent with those discussed below, and are omitted due to space limitations.

A. Datasets

We evaluate the performance of EFA in two heterogeneous scenarios, so as to avoid the risk that results are biased by the settings of one specific case study. The two scenarios refer to regions of comparable size (150 km^2 approximately) covering the conurbations of Milan, Italy, and Paris, France. The two cities are large enough for a study of mobile traffic to be statistically significant, but are located in different countries and have sensibly different population densities (around 7,000 and 21,000 inhabitants per km^2 for Milan and Paris, respectively).

Mobile traffic data in these scenarios was collected by major mobile network operators in each country, i.e., Telecom Italia Mobile (TIM) in Italy and Orange in France.

TIM-2013 dataset. The data was released by TIM as part of their Big Data Challenge. The dataset describes the mobile traffic generated by subscribers in the Milan conurbation over a two-month period spanning November and December 2013. The data includes voice, text, and Internet traffic of approximately 400,000 users, aggregated during time intervals of 10 minutes. Traffic volumes are georeferenced with respect to a regular-grid space tessellation of 235×235-m^2 cells.

Orange-2014 dataset. This dataset consists of Call Detail Records (CDR) collected for billing purposes by the operator. CDR describe hourly volumes of voice and text activity in the Paris metropolitan region, on a per-antenna basis. The data was collected from a sample of 100,000 users in September, October and November 2014. We employ a standard Voronoi tessellation to represent the spatial coverage of cells and the geographical distribution of traffic.

B. Network activity profiling

We investigate the performance of EFA for the profiling of network-wide mobile traffic activity over time. Let us first focus on a one-week period that is representative of the typical communication activity recorded in the TIM-2013 dataset. To that end, we condense the two months of data into one single *median week*, which has been shown to mitigate potential classification biases due to outlying behaviors [14]. For each cell in the Milan area: (i) we aggregate the demand of incoming/outgoing calls and texts on a hourly basis; (ii) we associate to every hour of the week the median value of all corresponding hourly demands (e.g., all Mondays at 8 am).

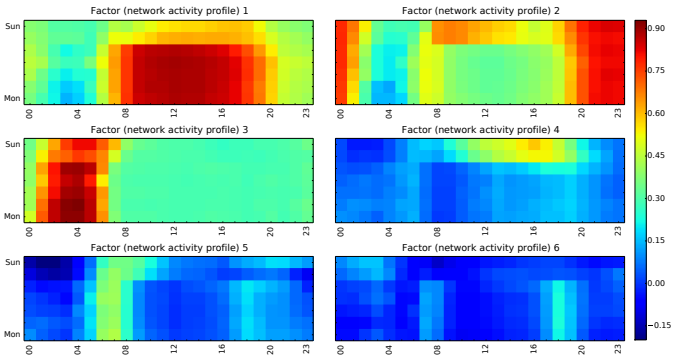


Fig. 3. Network activity profiling. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) over the median week in the TIM-2013 dataset. Loadings of the 24×7 hours of the week (i.e., EFA variables) on the six profiles (i.e., EFA factors). Figure best viewed in colors.

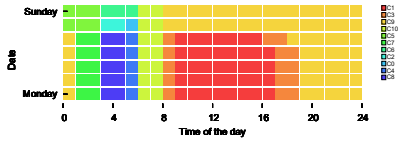


Fig. 4. Network activity profiling. Framework [14] on the total communication activity (sum of incoming/outgoing calls and SMS) over the median week in the TIM-2013 dataset. Classification of the 24×7 hours of the week into ten profiles. Figure best viewed in colors.

The network activity profiles (i.e., EFA factors) identified by EFA in the TIM-2013 median week are portrayed in Fig. 3. Each plot shows the loadings of every hour of the week (i.e., every EFA variable) on a specific profile, according to the color range on the right of the figure. E.g., hours from 8 am to 6 pm on Monday to Friday have very high loadings on factor 1, and low loadings on all other factors: hence, the first network activity profile characterizes the work hours.

Overall, EFA allows identifying the following profiles in Milan: (1) working hours; (2) relax hours in the evenings and weekend mornings; (3) overnight hours; (4) weekend afternoon hours; (5,6) morning and afternoon hours corresponding to the start and end of the work activity. One could reasonably map both profiles 5 and 6 to commuting behavior, but we will show in Sec. V-B2 how the additional spatial knowledge provided by EFA allows disambiguating them.

Takeaways. The profiles unveil that the overall mobile traffic activity in Milan yields a significant level of uniformity over time, being completely described by a very limited number of typical demand configurations. In other words, the temporal classification provided by EFA reveals major long-term dynamics (in the order of hours) in the aggregate demand.

1) Comparison with the state-of-the-art: The current state-of-the-art solution for the inference of network activity profiles from mobile traffic data is the dedicated framework presented in [14]. The approach relies on a hierarchical clustering of network activity *snapshots*, i.e., representations of the load generated by mobile users on the access network during fixed-size time intervals. The clustering is based on a custom measure of the similarity between snapshot pairs, and profiles are obtained by applying standard stopping rules to the dendrogram generated by the clustering algorithm.

Fig. 4 shows the network activity profiles detected in the TIM-2013 median week by the framework above. There is a clear match between the result of the two approaches: a limited set of profiles emerge that distinguish working, relax, overnight and commuting hours. The main improvements of EFA are that: *(i)* the benchmark framework separates overnight hours into several low-significance profiles, which is an undesirable side effect of the low communication activity in those periods; *(ii)* EFA identifies a unique behavior during weekend afternoons that is not recognized by the reference framework.

More generally, the information provided by the benchmark is less rich than that returned by EFA. Indeed, the former assigns each hour of the week to exactly one profile, whereas the latter ascribes to each hour precise loadings on all profiles.

Takeaways. The quality of EFA network activity profiling is superior to that granted by a dedicated state-of-the-art solution.

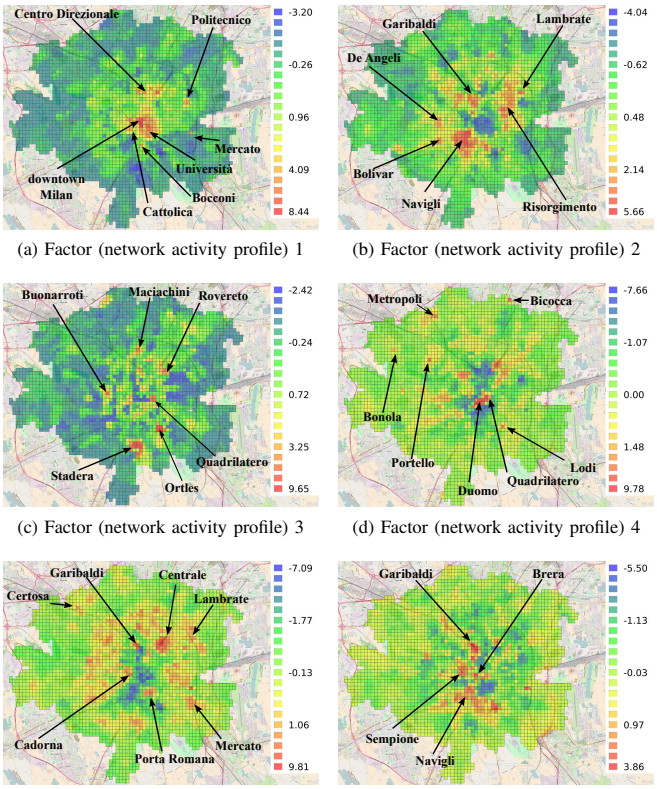
2) Advantages of EFA: As discussed in Sec. IV-A, EFA allows a joint spatiotemporal analysis of the mobile traffic demand. In the case of network activity profiling, EFA scores let us understand which geographical cells (i.e., EFA samples) are the most relevant to a specific profile (i.e., EFA factor). In other words, scores tell us *where* the mobile communication activity that characterizes a profile takes place. This is not possible with previous approaches such as [14].

Fig. 5 shows the scores, estimated via Thurstone's regression [28], of all geographical cells in Milan on the six network activity profiles[4]. As an example, we can remark how the cells interested by the first profile clearly highlight downtown Milan, where the business district and most offices are located. Additional geographical areas that have high scores on this profile are university campuses (Politecnico di Milano, Università di Milano, Bocconi, Cattolica) and commercial zones (public entrance to Mercato Ortofrutticolo). Clearly, these are the locations where mobile communication activities surge during the work hours, i.e., those hours that have high loadings on the first profile.

Equivalent analyses are possible for all of the other profiles. During evening and weekend mornings (profile 2), the network activity is much reduced in the city center, as shown in Fig. 5b. Prevalent areas are within the inner city beltway and characterized by a dense presence of bars, restaurants, and clubs (Navigli, Lambrate, Porta Garibaldi, Piazza Bolivar) or by a strictly residential nature (Risorgimento, De Angelis).

The most relevant areas to overnight hours (from 2 am to 5 am, profile 3) are depicted in Fig. 5c. Many are in fact ill-famed neighborhoods in Milan (Stadera, Maciachini, Rovereto), and not-so-legal undertakings may explain their importance late at night. Other areas are considered safe (Quadrilatero, Ortles, Buonarroti): the reason why they emerge at night remains an interesting open question. Explaining weekend afternoons (profile 4), in Fig. 5d, is easier. Mobile traffic is characterized by activities at touristic, shopping, and entertainment areas in the city center (Duomo, Quadrilatero), as well as large shopping centers in the suburbs (Bicocca Village, Piazza Lodi, Portello, Metropoli, Bonola).

[4]We stress that maps of cell scores are *not* maps of the aggregate mobile traffic volume. Scores are computed independently for each cell, and highlight if a cell is especially affected by a profile. They thus provide a geographical view of the profile that is suitably scaled on a per-cell basis. As a result, scores are neater and more insightful than plain traffic volumes.



(a) Factor (network activity profile) 1 (b) Factor (network activity profile) 2

(c) Factor (network activity profile) 3 (d) Factor (network activity profile) 4

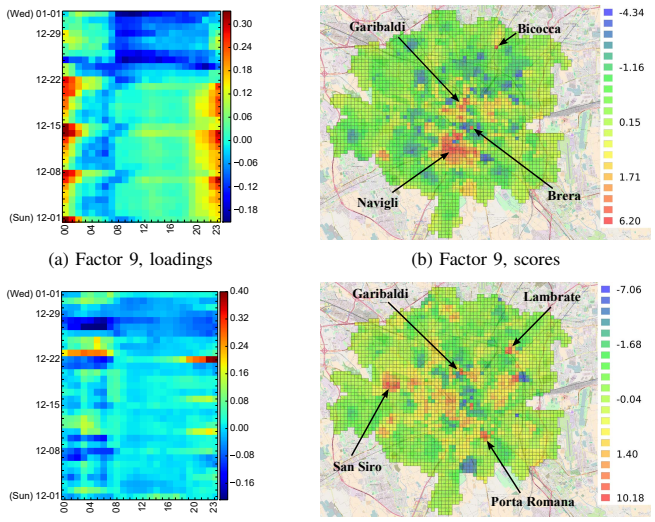
(e) Factor (network activity profile) 5 (f) Factor (network activity profile) 6

Fig. 5. Network activity profiling. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) over the median week in the TIM-2013 dataset. Thurstone's scores of the 434 cells (i.e., EFA samples) on the six profiles (i.e., EFA factors). Figure best viewed in colors.

Interestingly, the geographical information allows disambiguating profiles that look similar in Fig. 3. Commuting hours (profiles 5 and 6) can be now told apart by looking at Fig. 5e and Fig. 5f. Indeed, the areas interested by profile 5 highlight train stations (Centrale, Garibaldi, Lambrate, Porta Romana, Certosa), major railway lines leading to these stations, and residential areas beyond the inner city beltway. These zones are interested by the commuting activity during working days and by the presence of tourist (whose schedule is understandably shifted forward with respect to that of commuters, in Fig. 3) during weekends. There is also a notable widespread zone of interest in proximity of the city wholesale market for farm produce (Mercato Ortofrutticolo), when goods are stocked.

The geographical cells related to profile 6 are sensibly different: they correspond to popular areas for an after-work aperitif (Brera, Garibaldi, Navigli, Sempione), which is a common habit in Milan. Although these leisure occupations temporally overlap with the afternoon commuting hustle, EFA successfully classifies the two behaviors into separate profiles, and allows explaining them geographically.

A second major advantage of EFA is that it can be run on hourly data directly, instead of using a lumped representation like the median week. Using hourly data allows pinpointing additional profiles that map to more specific (and possibly outlying) behaviors in the mobile traffic. Instead, traditional classification approaches yield poor results in this case, as they tend to be sensible to noise in the data [14].



(a) Factor 9, loadings

(b) Factor 9, scores

(c) Factor 15, loadings

(d) Factor 15, scores

Fig. 6. Network activity profiling. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) over December 2013 in the TIM-2013 dataset. (a,c) Loadings of the 24×30 hours (i.e., EFA variables) on two profiles (i.e., EFA factors). (b,d) Thurstone's scores of the 434 cells (i.e., EFA samples) on the same two profiles. Figure best viewed in colors.

Fig. 6 shows two representative examples of the more varied profiles that are obtained when analysing one month of data on a per-hour basis. A new dedicated profile emerges for regular weekend nightlife that then disappears during the Christmas week, as shown by the loadings in Fig. 6a. The relevant areas are clearly indicated by the scores in Fig. 6b, and single out popular evening and late night meeting areas, especially for university students (Navigli, Garibaldi, Milano Bicocca, Brera). The additional profile in Fig. 6c concerns a limited set of nights (December 12, 16 and 22) with heterogeneous intensity. Scores in Fig. 6d evidence that the main geographical areas relevant to this profile are those around the soccer stadium in Milan (San Siro). Also concerned are key areas for transportation to the stadium (Lambrate, Porta Garibaldi and Porta Romana train stations). The combination of loadings and scores lets us conclude that this profile corresponds to major events at the stadium: indeed, soccer games were played on all the dates showing higher loadings (Ajax-AC Milan on December 12, AC Milan-Rome on December 16, and Inter Milan-AC Milan on December 22). The last game is the city derby: the associated activity spans throughout the following night as fans celebrate the winning team until early morning.

Takeaways. EFA does not only identify substantial temporal profiles. It also yields spatial information that is precious for the correct interpretation of the dynamics of the mobile traffic demand, in both routine and outlying situations. Moreover, EFA is robust to data sparsity, which enables its use at finer time granularity than previous clustering-based solutions.

C. Land use detection

The second application of EFA is the detection of land use based on the dynamics of the mobile traffic demand. We recall that such a classification problem is solved by mapping EFA variables to the spatial cells where the communication activity is aggregated, and EFA samples to time intervals (i.e., hours in the following analysis). EFA factors are then land uses.

TABLE I
LAND USE DETECTION. FULL LIST OF LABELED LAND USES IN PARIS.

Factor	Labeled land use (Code)	Examples
1	Dense Residential (DR)	Large areas of the populous 15th, 18th, 19th, and 20th arrondissements
2	Office-Industrial (OI)	La Defense, Issy Les Moulineaux
3	Short-range Commuting (SC)	Paris Subway Stations
4	Malls and Shopping Centers (Ma)	Galerie La Fayette Haussmann, Forum Les Halles, Le Millenaire, So Ouest
6	Expo Area (Ex)	Porte de Versailles: Pavillons 1, 3 and 4
7	Nightlife (Ni)	Rue Montmartre, Pigalle, Place St. Michel, Place St. Jacques, Rue de Lappe, Rue Guisard, Rue Saint-Jacques
8	Highway Interchanges (HI)	Porte Maillot ↔ Porte de Clignancourt, Porte de Charenton ↔ Porte de Bercy
10	Education (Ed)	Colleges (St. Sulpice, Notre Dame de Sion), Schools (Chartreux, Blanche de Castille)
13	Long-range Commuting (LC)	Gare de Lyon, du Nord, de Montparnasse, d'Austerlitz, de l'Est, Saint Lazare
14	Leisure (Le)	Bercy AccorHotels Arena, Théatre du Palais Royal, de la Michodière, du Gimnase
5, 9, 11, 12	Partially inactive base stations	These are groups of base stations that are only active during part of the three months, due to infrastructure upgrades or continuing roll-out of the data collection probes

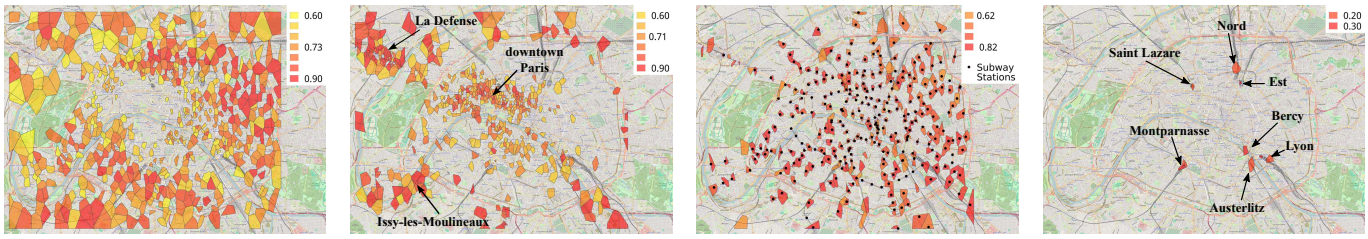
Fig. 7a–d show a selection of four out of fourteen land use classes detected by EFA in the Orange-2014 dataset. For the sake of clarity, in each plot we only show cells (i.e., EFA variables) that have a high loading on the class (i.e., EFA factor) the map refers to. The first two factors, in Fig. 7a and Fig. 7b, correspond to residential regions (mainly suburbs around the city center) and business areas (e.g., downtown Paris, La Defense or Issy-les-Moulineaux). Residential and business are often the two most distinctive land uses in urban areas, according to both traditional cartographies and previous studies [10], [16], [17], [18]: their detection as prevalent EFA factors is thus a promising first result.

However, EFA also identifies non-trivial land uses, such as that associated with factor 3, in Fig. 7c. This land use characterizes a large number of uniformly distributed cells in the Paris conurbation. In fact, cells with high (above 0.7) factor 3 loadings map very well (93% precision with 96% recall) to the network of subway stations in the city (black dots in Fig. 7c). Similarly, factor 13, in Fig. 7d neatly pinpoints all major train stations in Paris (i.e., Gare de Lyon, Austerlitz, Saint Lazare, Montparnasse, Bercy, de l'Est and du Nord).

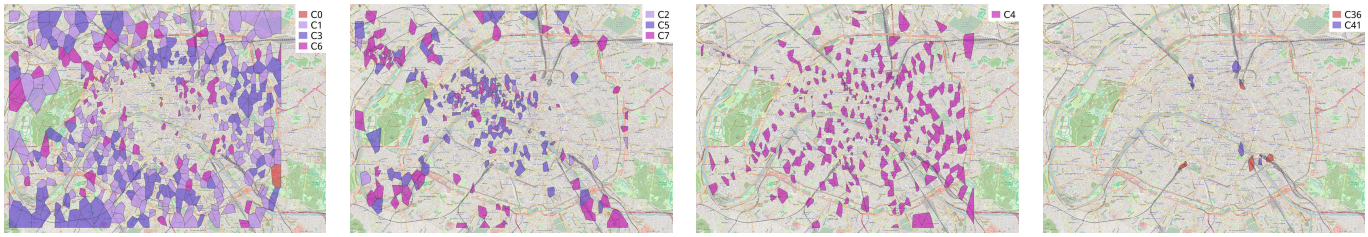
Due to space limitations, an extensive discussion of all land use classes is not possible: yet, all have meaningful interpretations that are summarized in Tab. I.

Takeaways. The classes extracted by EFA correspond well to expected macroscopic land uses (e.g., residential and business areas), yet they also reveal many microscopic land uses (e.g., public and private transportation hubs, shopping or leisure areas) where the mobile traffic demand follows distinctive dynamics. This showcases how EFA can be successfully employed to refine land use cartography, which is often complex and expensive to draw and update.

1) Comparison with the state-of-the-art: Previous solutions dedicated to land use detection from mobile traffic data rely on the notion of *signature*, i.e., a representation of the typical activity recorded over time at one geographical cell. Approaches in the literature group signatures using different pair-

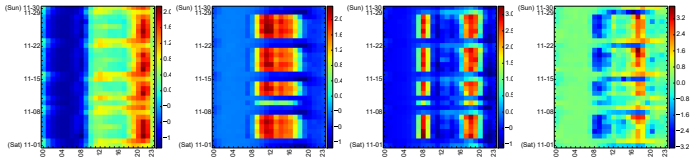


(a) Factor (land use class) 1 (b) Factor (land use class) 2 (c) Factor (land use class) 3 (d) Factor (land use class) 13



(e) Clusters 0, 1, 3, 6 (f) Clusters 2, 5, 7 (g) Cluster 4 (h) Clusters 36 and 41

Fig. 7. Land use detection. (a)-(d) EFA of the total communication activity (sum of incoming/outgoing calls and SMS) in the Orange-2014 dataset. Loadings of the 1596 Voronoi cells (i.e., EFA variables) on four (out of fourteen) representative classes (i.e., EFA factors). (e)-(f) Signature clustering [18] on the total communication activity (sum of incoming/outgoing calls and SMS) in the Orange-2014 dataset: Voronoi cell clusters that match our choice of EFA factors. Figure best viewed in colors.



(a) Factor 1 (b) Factor 2 (c) Factor 3 (d) Factor 13

Fig. 8. Land use detection. EFA of the total communication activity (sum of incoming/outgoing calls and SMS) in the Orange-2014 dataset in November. Thurstone's scores of the 91×24 hours (i.e., EFA samples) on a selection of the 16 classes (i.e., EFA factors). Figure best viewed in colors.

wise signature similarity measures and clustering algorithms. The resulting clusters of signatures are the land uses. For our comparative evaluation, we consider median weeks (see Sec. V-B) as signatures, and measure their similarity via Pearson correlation. Clustering is performed through linkage with an average distance criterion, using skewness minimization as the stopping rule. This configuration is the most suitable for land use detection, according to comparative evaluations [18].

Fig. 7e-h portray some classes (i.e., signature clusters) generated by the benchmark technique above. The match with EFA results is striking. E.g., Fig. 7e shows that signature clusters 0, 1, 3 and 6 identify the same cells as EFA factor 1, i.e., residential suburban areas in the Paris region.

Interestingly, different clusters in the plot coincide with cells that have various loadings on the EFA factor (e.g., see the similar geographical distribution of light and dark colors in Fig. 7a and Fig. 7e): this means that the diverse signature clusters in Fig. 7e just capture different intensities of a same phenomenon – in this case the actual preponderance of residential users in the cell over other user types. Similar considerations hold for the remaining plots, in Fig. 7f-h. In fact, we found each EFA factor to correspond to 2-4 signature clusters in most cases. Our conclusion is that EFA provides a more compact set of classes, grouping clusters that are akin; then, it allows in-depth intra-class analysis via loading values.

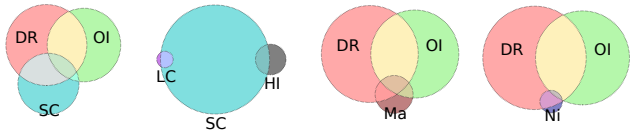


Fig. 9. Land use detection. Venn diagrams of the distribution of cells into partially overlapping land uses (coded as in Tab. I). Mixed land uses are represented by the set intersections.

Takeaways. The quality of the proposed EFA classification is equivalent to that granted by a dedicated state-of-the-art solution. In fact, the benchmark framework returns a much larger set of clusters that are often strongly related to each other in terms of their underlying demand dynamics: however, such inter-cluster relationships are unknown, and cumbersome to identify a posteriori. EFA eases significantly the interpretation of results, as it returns a more concise set of classes that can be further explored through loading analysis.

2) Advantages of EFA: Also under this problem formulation, scores can be leveraged to confirm and improve our understanding of factors. More precisely, EFA returns scores that indicate the relevance of time (i.e., EFA samples) to each class (i.e., EFA factor). Thus, scores tell us *when* each land use class shows an especially remarkable mobile communication activity[5]. This is not possible with any previous technique.

Fig. 8 shows the hourly scores during one months of Orange-2014 data on a representative subset of land use classes detected by EFA. We can observe that the suburban residential areas of factor 1 have distinctive traffic patterns during the evening hours of the working days, in Fig. 8a. These are the only areas of Paris where traffic surges between 8 pm and 10 pm on Mondays to Fridays. As expected, the working hours are most relevant to business areas in factor 2, shown in Fig. 8b: interestingly, morning hours seem more concerned than afternoon ones. One can also easily detect that the typical

[5]The same considerations in footnote 4 apply here as well, and time series of cell scores are *not* simple time series of the aggregate mobile traffic volume.

working day spans from 8 am to 6-7 pm: these are the morning and afternoon commuting times when the subway stations of factor 3 experience exceptionally high loads, as displayed in Fig. 8c. Train stations show a different temporal pattern, as the most characteristic activity occurs during the afternoon commuting hours only. In all plots it is easy to spot an irregularity, due to a public holiday (Armistice, November 11).

A second major advantage of EFA is that loadings explain how the activity in each geographical cell is related to all land uses at the same time. Thus, EFA naturally provides information on mixed land uses, i.e., city regions where different urban infrastructures merge. An example is provided in Fig. 9: set intersections in the Venn diagrams illustrate to what extent land uses overlap in Paris. Significant areas of mixed residential-business (DR-OI) nature exist, and they are fairly well served by short-commuting services (SC). Also, the same subway network (SC) links together all train stations (LC) as well as several major highway entry/exit nodes (HI). Malls and shopping centers (Ma) are uniformly distributed across residential and office areas, whereas nightlife (Ni) only thrives in densely populated regions of the city.

Retrieving similar information from the output of traditional clustering-based solutions, which rigidly assign each cell to one land use, requires complex processing. This has only been attempted recently, and with a small set of land uses [10].

Takeaways. EFA does not only identifies relevant land uses in the urban landscape. It also provides temporal knowledge that helps interpreting them, and that immediately highlights how special events affect one or more land uses. In addition, EFA loadings implicitly bear mixed land use information, which are not easily inferred with traditional approaches.

VI. Conclusions and perspectives

We proposed an original approach to the spatiotemporal classification of mobile traffic data, which relies on Exploratory Factor Analysis (EFA). Extensive tests with heterogeneous real-world datasets demonstrate the versatility of EFA, which provides a unifying framework to solve problems that have been studied in isolation in the literature, i.e., mobile traffic profiling and land use detection. In both cases, EFA attains results that improve those of state-of-the-art solutions (e.g., the richer information of network activity profiles), or match them while yielding greater consistency (e.g., the better abstracted land use classes, where loadings can be leveraged for intra-class analysis). In addition, EFA provides supplementary knowledge (i.e., the geographical perspective of profiles and the temporal view of land uses) that proves paramount to the interpretation of the results, and eases tasks that are otherwise complex to perform (e.g., the analysis of per-hour temporal data, or the detection of mixed land uses).

EFA-based classification can find applications in data-driven network operations, at multiple levels. The temporal structures identified by EFA expose non-trivial long-term dynamics in the mobile traffic demand that are relevant to the allocation of resources in, e.g., Cloud Radio Access Networks (C-RAN) [9]. In addition, typical temporal profiles may serve as a basis for the detection of anomalous network usages, and for predicting the future demand in the context of anticipatory networking. In the spatial dimension, EFA classes neatly characterize the strong geographical locality of mobile demand. They can thus pave the way for cognitive network functions that aim at migrating network resources geographically, or at dynamically configuring the network topology; such functions are especially relevant to, e.g., Mobile Edge Computing (MEC) infrastructures [9]. Overall, EFA-based classification is a potential brick for future big data-driven 5G systems [29].

VII. Acknowledgments

The authors thank friends and colleagues in Milan and Paris who helped interpreting the classification results. This work was supported by the French National Research Agency grant ANR-13-INFR-0005 ABCD, and by the EU FP7 ERA-NET program grant CHIST-ERA-2012 MACACO.

References

[1] Cisco VNI Forecast, "Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020," 2016.
[2] M.Z. Shafiq, L. Ji, A. X. Liu, J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," *ACM SIGMETRICS*, 2011.
[3] D. Willkomm, S. Machiraju, J. Bolot, A. Wolisz, "Primary Users in Cellular Networks: A Large-Scale Measurement Study," IEEE DySPAN, 2008.
[4] M.Z. Shafiq, L. Ji, A.X. Liu, J. Pang, S. Venkataraman, J. Wang, "A First Look at Cellular Network Performance during Crowded Events," *ACM SIGMETRICS*, 2013.
[5] U. Paul, A.P. Subramanian, M.M. Buddhikot, S.R. Das, "Understanding Traffic Dynamics in Cellular Data Networks," *IEEE INFOCOM*, 2011.
[6] R. Keralapura, A. Nucci, Z.-L. Zhang, L. Gao, "Profiling Users in a 3G Network Using Hourglass Co-Clustering," *ACM MobiCom*, 2010.
[7] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, "Large-scale Mobile Traffic Analysis: A Survey," *IEEE Communications Surveys & Tutorials*, 18(1), 2016.
[8] R.W. Thomas, L.A. Da Silva, A.B. MacKenzie, "Cognitive networks," *IEEE DySPAN*, 2005.
[9] H. Assem, T. Sandra Buda, L. Xu, "Initial use cases, scenarios and requirements," *H2020 5G-PPP CogNet*, Deliverable D2.1, 2015.
[10] M. Lenormand, M. Picornell, O.G. Cantú-Ros, T. Louail, R. Herranz, M. Barthelemy, E. Frías-Martínez, M. San Miguel, J.J. Ramasco, "Comparing and modelling land use organization in cities," *Royal Society Open Science*, 2, 2016.
[11] D. Goergen, V. Mendiratta, R. State, T. Engel, "Identifying Abnormal Patterns in Cellular Communication Flows," *ACM IPTComm*, 2013.
[12] F. Calabrese, F.C. Pereira, G. Di Lorenzo, L. Liu, C. Ratti, "The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events," *Pervasive Computing*, 2010.
[13] J.P. Bagrow, D. Wang, A.-L. Barabasi, "Collective Response of Human Populations to Large-Scale Emergencies," *PLoS ONE*, 6(3), 2011.
[14] D. Naboulsi, R. Stanica, M. Fiore, "Classifying Call Profiles in Large-scale Mobile Traffic Datasets," *IEEE INFOCOM*, 2014.
[15] I. Trestian, S. Ranjan, A. Kuzmanovic, A. Nucci, "Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network," *ACM IMC*, 2009.
[16] J.L. Toole, M. Ulm, M.C. Gonzalez, D. Bauer, "Inferring Land Use from Mobile Phone Activity," *ACM UrbComp*, 2012.
[17] B. Cici, M. Gjoka, A. Markopoulou, C.T. Butts, "On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology," *ACM MobiHoc*, 2015.
[18] A. Furno, R. Stanica, M. Fiore, "A Comparative Evaluation of Urban Fabric Detection Techniques Based on Mobile Traffic Data," *IEEE/ACM ASONAM*, 2015.
[19] C. Spearman, "General Intelligence Objectively Determined and Measured," *The American Journal of Psychology*, 15(2), 1904.
[20] L.R. Fabrigar, D.T. Wegener, R.C. MacCallum, E.J. Strahan, "Evaluating the Use of Exploratory Factor Analysis in Psychological Research," *Psychological Methods*, 4(3), 1999.
[21] S.A. Mulaik, *Foundations of Factor Analysis*, CRC Press, 2009.
[22] D.N. Lawley, "The Estimation of Factor Loadings by the Method of Maximum Likelihood," *Proc. of the Royal Society of Edinburgh*, 60(1), 1940.
[23] I.T. Jolliffe, "Principal Component Analysis and Factor Analysis," *Principal Component Analysis*, Springer Series in Statistics, 2002.
[24] J.C.F. de Winter, D. Dodou, "Common Factor Analysis versus Principal Component Analysis: A Comparison of Loadings by Means of Simulations," *Communications in Statistics - Simulation and Computation*, 45(1), 2016.
[25] H.F. Kaiser, J. Rice, "Little Jiffy, Mark IV," *Educational and Psychological Measurement*, 34, 1974.
[26] J.L. Horn, "A Rationale and Test for the Number of Factors in Factor Analysis," *Psychometrika*, 30(2), 1965.
[27] H.F. Kaiser, "The VARIMAX Criterion for Analytic Rotation in Factor Analysis," *Psychometrika*, 23(3), 1958.
[28] L.L. Thurstone, "The vectors of mind", *University of Chicago Press*, 1935.
[29] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Network*, 30(1), 2016.