Spatial and Temporal Exploratory Factor Analysis of Urban Mobile Data Traffic

Angelo Furno[1*], André Felipe Zanella[2,3], Razvan Stanica[4], Marco Fiore[2]

[1]Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT-ECO7, Lyon, France.
[2]IMDEA Networks Institute, Madrid, Spain.
[3]Universidad Carlos III de Madrid, Madrid, Spain.
[4]Univ Lyon, INSA Lyon, Inria, CITI, Villeurbanne, France.

*Corresponding author(s). E-mail(s): angelo.furno@univ-eiffel.fr;
Contributing authors: andre.zanella@imdea.org; razvan.stanica@insa-lyon.fr;
marco.fiore@imdea.org;

Abstract

Mobile data traffic is characterized by complex spatiotemporal fluctuations that are linked in entangled ways to the mobility and diverse activities of the mobile network subscribers. Unraveling such dynamics and understanding their root causes are challenging tasks that call for dedicated, complex data analysis tools. In this paper, we propose to employ Exploratory Factor Analysis (EFA) as a unified approach to identify both spatial and temporal structures hidden in the mobile data traffic. We provide a brief introduction to the EFA methodology, discuss how it can be tailored to a networking context, and outline its advantages in terms of versatility, unsupervised nature and interpretability of results. Experiments with large-scale measurement data collected in two urban regions demonstrate the effectiveness of the approach, which allows recognizing and explaining a variety of fundamental structures that underpin real-world spatiotemporal traffic dynamics. A thorough discussion of the results provides interesting insights, including that a reasonably small number of latent factors can describe well the majority of temporal and spatial structures observed in mobile traffic demands, providing valuable insights into key spatiotemporal patterns of population and becoming a valuable asset in understanding the attractiveness factors in urban areas.

Keywords: Mobile data traffic; exploratory factor analysis; spatiotemporal dynamics; latent factors.

1 Introduction

Mobile data traffic has been steadily surging over the past two decades, and forecasts from major players in the telecommunication ecosystem anticipate that this trend has not exhausted yet. As a representative example, Ericsson indicates that global mobile network data traffic is rapidly approaching 100 Exabytes per month, with a year-on-year growth steady in the 40-50% range Ericsson (2021). The phenomenon has spurred increased interest towards the precise dynamics of the mobile data traffic consumption. Indeed, even at relatively small scales, such as within single urban areas, the demand for mobile data is not homogeneous; rather, it undergoes substantial fluctuations in time and space due to the varying mobility and digital activities

of the users. Characterizing these spatiotemporal changes and understanding their root causes has important practical applications, as it helps to establish new links between human endeavors, city fabrics and the utilization of mobile services, and can support more informed network infrastructure management.

Early efforts in the analysis of mobile data traffic have revealed important features of its dynamics. For instance, the traffic generated by mobile subscribers is strongly periodic Shafiq (2011) and geographically localized Willkomm (2008), which enables its effective prediction. Also, social events can determine significant variations Shafiq (2013), with the consequent need for dedicated resource management policies in mobile networks. Similarly, the bandwidth consumed by individual subscribers is highly heterogeneous Paul (2011), yet it is captured by a limited number of typical profiles Keralapura (2010); Mucelli (2015); Li (2015); Yang (2015), which enables, *e.g.*, the informed tuning of traffic plans. Moreover, specific mobile services can generate traffic patterns that are highly heterogeneous in time Marquez (2017), while differences over space depend on the considered geographical scale Shafiq (2012); Peltonen (2018); Singh (2019), hence calling for tailored resource management strategies in network slicing environments.

Analyses of mobile traffic demand in the literature can be divided into two broad categories Blondel (2015); Naboulsi (2016). On the one hand, there are 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. On the other hand, there are studies that take the viewpoint of a mobile network operator, 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. In this second category, the study of mobile phone traffic data has often supported transportation studies as a larger-scale and cost-efficient alternative to travel surveys for, *e.g.*, the understanding and modelling of travel demand at regional scale and its spatial classification Fekih (2022), the identification of the most-travelled routes on a road network Toole (2015), the analysis of travel demand between transportation hubs in urban areas Chen (2022), and the reconstruction of travel mode for inter-city trips Breyer (2022).

Our work falls in the second category above. Here, despite previous efforts and as detailed in Section 2, we still lack dependable tools to explore complex relationships in mobile data traffic. In particular, while partial solutions have been proposed to detect either temporal or spatial structures in traffic demands, little attention has been paid to the more challenging *concurrent* inspection of both space and time dimensions. Development of a tool capable, at the same time, of automatically segmenting an urban territory into homogeneous areas and providing a temporal description of each identified area, could provide valuable information for travel demand estimation, as well as macroscopic traffic modeling.

In this paper, we present an original methodology for the spatiotemporal classification of the mobile data traffic observed in operational networks, so as to fill the gap above. The proposed solution builds upon *Exploratory Factor Analysis* (EFA), a well established data analysis instrument in psychology research. As explained in Section 3, EFA aims at identifying, in a fully automated way, latent factors that cause the dynamics observed in the data. This is achieved by identifying the variables of interest in the data, and describing their covariance relationships in terms of the underlying and unobservable factors. We tailor EFA to the specific problem of identifying recurrent behaviors in the mobile data traffic, as discussed in Section 4. The approach yields significant advantages over previous proposals:

- *Versatility in spatiotemporal analyses.* The methodology represents a unified approach to recognize factors that are temporal or spatial in nature. Along the time dimension, EFA can detect temporal structures in the network-wide communication activity, revealing time periods that show a similar, stable spatial distribution of the mobile traffic demand. On the spatial dimension, EFA identifies hidden spatial structures, by automatically decomposing a target geographical area into zones where mobile data traffic follows homogeneous time dynamics.

- *Unsupervised and probabilistic nature of the approach.* EFA is a completely unsupervised tool that produces probabilistic structures. This allows overcoming the limitations of methods previously employed for mobile traffic analysis, such as clustering, which only produce deterministic temporal or spatial categories, or supervised techniques that require labeled data.
- *Interpretability of results.* Our proposed methodology eases the exploration of the root causes for the temporal and spatial structures above, by allowing an automated extrapolation of the structures hidden in the respective dual dimensions. In other words, EFA implicitly provides knowledge of the traffic geography that characterizes each temporal structure, and of the precise traffic time series that distinguish each spatial structure. This plays an important role in the interpretation of results.

We demonstrate these advantages with real-world mobile data traffic collected in production networks serving two major cities in France, in Sections 5 and 6. Our results highlight the vast range of unique temporal and spatial profiles that can be obtained from mobile traffic data through the use of EFA, as well as the ability of identifying short and long term spatial structures and mixed land usage in cities. Ultimately, as concluded in Section 7, our work opens interesting perspectives on the use of EFA as a dependable tool for the analysis of complex hidden structure in mobile data traffic.

2 Related work

The analysis and characterization of mobile data traffic has applications across research domains, including transportation research, urban planning, sociology, epidemiology, and, of course, telecommunications Blondel (2015); Naboulsi (2016). Studies in these fields have adopted a wide range of approaches to investigate the properties of mobile traffic demands. Yet, independently of the research domain, most works rely on basic statistical tools, and only consider the likes of mean values, deviations or distributions of temporal and spatial traffic samples. While such representation provide initial insights in the data, they are aggregated over the full set of samples, and cannot reveal the complex hidden structures that distinguish specific locations or time intervals.

Temporal analysis. Studies of temporal structures in mobile traffic can be as simple as straight comparisons of traffic volumes among whole seasons Cardona (2014), or observation of counting statistics over time for calls Bagrow (2011) or users Calabrese (2010). Although these approaches allow appreciating how planned and unplanned events induce significant variations in the typical temporal structure of the mobile traffic demand, investigations at a finer granularity require more elaborated methodologies. A more composite method for temporal analysis of mobile traffic consists in building an expectation from historical data, and labeling time slots deviating from the statistics as anomalous Goergen (2013). Still, the approach is specific to one purpose, *i.e.*, outlier detection, and cannot find the typical but hidden structures we aim at discovering.

Closer to our goal is the temporal profiling of network activity Furno (2016). In this case, snapshots of the mobile traffic demand in a target region collected at different time slots are clustered via a dedicated distance measure. This allows unveiling time periods with similar patterns in the spatial distribution of traffic. Recently, tools from spectral analysis have also been adopted to achieve a similar goal, on a per-service basis Marquez (2020). However, both approaches above rely on traditional clustering algorithms, hence forcing each time slot into a single specific category. This loses nuances in the data, *e.g.*, during time periods where the spatial traffic is in fact at the boundary of two or more behaviors. Moreover, the approaches do not provide information about the root causes that lead to the resulting clusters, whose interpretation is thus left to the expert knowledge of the system. EFA offers a sound mathematical framework that overcomes both limitations above.

Spatial analysis. A larger body of works addresses spatial structures in mobile traffic, focusing on urban settings that show especially rich and interesting traffic patterns. The majority of these studies rely on unsupervised learning approaches to reveal geographical areas with

equivalent traffic dynamics. Specifically, they collect mobile traffic time series in different geographical zones, compress such time series into normalized signatures that summarize the observed temporal patterns of traffic, and finally adopt legacy clustering algorithms to group the zones based on their signatures Soto (2011); Toole (2012); Lenormand (2016); Grauwin (2015); Furno et al. (2017). In a similar spirit, spectral methods have been explored to define signatures of routine activities and deviations from them, with the same objective of using them for spatial clustering Cici (2015). All these works have the same limitations highlighted for temporal analyses based on clustering, *i.e.*, they force an inflexible binary association of geographical zones to clusters, and they do not offer straightforward handles for interpretation of the obtained clusters.

Different and more original approaches are rare in the literature. Information theory tools have been used to group large statistical zones across a whole country based on the local consumption of mobile services Singh (2019). Or, techniques from signal analysis relying on eigendecomposition have been adopted to categorize buildings in a university campus based on the observed Wi-Fi traffic Calabrese (2010). These methods have very different targets than ours, as they focus on very large (*i.e.*, nationwide) or localized (*i.e.*, a single campus) geographical regions; moreover, they still aim at generating rigid associations of zones to behaviors with little explainability, as in the legacy clustering strategies above.

Novelty of our work. The approach we propose in this paper builds upon EFA techniques that root into the work of Spearman, over a century ago Spearman (1904). Since those early studies, EFA has emerged as one of the dominant classes of factor analysis, and has been widely employed in statistical psychology research Fabrigar (1999). To the best of our knowledge, the earlier version of this manuscript was the first work to introduce factor analysis for the study of mobile traffic data and, more generally, in the field of networking Furno (2017). Only recently, a similar approach has been used to detect network anomalies problems in a campus Wi-Fi network Camacho (2020), which is a different task than our target, *i.e.*, the inference of spatiotemporal structures in mobile traffic.

Relying on EFA allows overcoming multiple limitations of previous methods. First, while all previous works explored the temporal and spatial structures of the mobile traffic separately, EFA provides a unified framework that can be cast to explore both dimensions. Second, it surpasses the strictness of clustering, and allows for a probabilistic association of archetypal mobile traffic patterns to time periods or geographical areas, allowing to appreciate composite dynamics that are overlooked by traditional methods. Third, it outputs implicit information on the mobile traffic behaviors in space (respectively, time) that tell apart and characterize diverse periods (respectively, areas), thus easing the explanation of results.

It is also worth noticing that the vast majority of previous works have analyzed Call Detail Records (CDR) that encompass voice calls and text messages, but do not capture the activity of mobile subscribers in terms of data traffic. While interesting from a sociological perspective, calling and texting play an increasingly diminishing role in network traffic, hence the results of many studies in the literature hardly apply to modern networks. This is also the case of the earlier version of this work Furno (2017), among many others. Instead, our analysis fully focuses on data traffic collected in an operational 3G/4G network, and thus provides an up-to-date view on hidden structures in today's mobile traffic that can point towards a dynamic description of human presence and, therefore, travel demand.

3 Exploratory Factor Analysis

We start the technical discussion by presenting the foundations of EFA and providing a primer to its operation. To this end, in Table 1 we first introduce the terminology used in the remainder of the paper, using the toy example in Figure 1 to illustrate their semantics.

3.1 Fundamental model

Given a set of observed variables of interest, factor analysis is formally defined as *"a model of hypothetical component variables that explain*

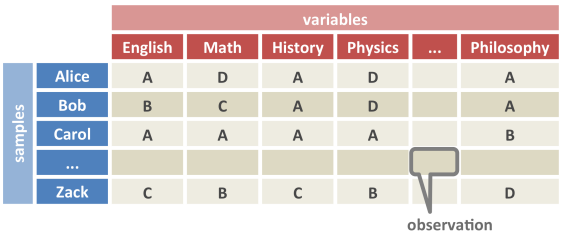


Fig. 1: EFA toy example: student grading across subjects. In this case, EFA can be used to identify a limited set of latent abilities of the students that may explain their grades.

the linear[1] *relationships existing between observed variables"* Mulaik (2009). 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 factor 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.*, $\mathrm{Cov}(\mathbf{F}, \mathbf{U}) = 0$. It follows that

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

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[2] are made, and it is thus generally assumed that all factors are orthogonal, *i.e.*, mutually uncorrelated and with unit variances. Therefore, $\mathbf{\Phi}$ can be replaced by the identity matrix in (2), and

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

Term	*Description*
Variables	The set of phenomena of interest, related to some population of individuals, *e.g.*, subjects taught to primary school students in Figure 1.
Samples	The set of individuals from the given population for which all phenomena of interest can be measured, *e.g.*, students from the same class.
Observations	The realizations of all variables for each sample, *e.g.*, the grades of examination tests in all subject obtained by each student.
Common factors	Complex interrelationship among the observed phenomena that can be reasonably assumed to exist, allowing for an interpretation of the phenomena, *e.g.*, inferring factors such as verbal and mathematical intelligence of the students, whose combination may explain the aptitude of the students to each subject.
Factor loadings	Numerical relationships that describe to what extend each common factor explains each variable, *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 students in mathematics-related disciplines.
Factor scores	Estimated values that relate samples to common factors, *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	Help explain the unique variance associated to each variable, pinpointing outlying behaviors in the data, as common factors alone cannot explain all variables, *e.g.*, unique factors could account for a rare talent of one student towards a specific discipline.

Table 1: EFA terminology and examples

[1] The linearity of relationships among variables in the specific context of mobile traffic will be discussed in Section 4.3.

[2] Next, we will use *common factor* and *factor* interchangeably.

whose i-th diagonal element can be written as

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

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.

3.2 Factor extraction

Several methods have been developed to estimate the unknown variables $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ in (3). We next present two popular approaches, which we will adopt in our analysis.

Maximum Likelihood Estimation (MLE) allows inferring the unknown variables $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ in (3) in a way that is efficient and robust Mulaik (2009). 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. The information provided by $\mathbf{S}$ may also be represented by a correlation matrix $\mathbf{R}$ and a set of standard deviations $s_1, s_2, \ldots, s_N$.

MLE maximizes the likelihood function

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

with Tr indicating the matrix trace operator. The $\mathbf{\Sigma}$ matrix maximizing (5) also minimizes the following fit function Lawley (1940)

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

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 Mulaik (2009).

[3]MLE is known to yield good estimations even when the actual distribution of $\mathbf{X}$ is not multivariate Gaussian Mulaik (2009). This is proven in the context of mobile data traffic analysis by the minor discrepancy of the results attained with MLE and with MINRES, which does not require the Gaussian distribution assumption.

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

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

where the diagonal matrix $[\gamma_i - 1]_K$ contains the K largest eigenvalues of $\mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$, 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

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

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

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

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

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.

Minimum Residuals (MINRES) is an alternative approach to factor extraction. Unlike maximum likelihood estimation, MINRES does not rely on any assumption about the distribution of observed variables and can produce valid solutions even when applied to singular matrices Joreskog (1978). The working principle of MINRES is to minimize the sum of off-diagonal squared residuals of the correlation matrices, *i.e.*, differences between the observed ($\mathbf{R}$) and reproduced ($\mathbf{\Lambda}$) correlations, without requiring any estimation of the communalities, *i.e.*, h_i in (4). In other terms, MINRES minimizes the fit function

$$F_K(\mathbf{\Lambda}) = \|[\mathbf{R} - \Im] - [\mathbf{\Lambda}\mathbf{\Lambda}^\intercal - \mathsf{Diag}(\mathbf{\Lambda}\mathbf{\Lambda}^\intercal)]\|, \quad (10)$$

where K refers to the number of common factors considered for the estimation of the $\mathbf{\Lambda}$ factor loadings matrix, and $\mathbf{H} = \mathsf{Diag}(\mathbf{\Lambda}\mathbf{\Lambda}^\intercal)$ is the diagonal matrix of reproduced communalities.

Equation (10) can be written in explicit form

$$F_K(\mathbf{\Lambda}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (r_{ij} - \sum_{k=1}^{K} \lambda_{ik} \cdot \lambda_{jk})^2, \quad (11)$$

which is a function of the $N(N-1)/2$ off-diagonal residual correlations. MINRES minimizes $F_K(\mathbf{\Lambda})$ by successive approximations of the values of the factor loadings λ_{ik}. The communalities $h_i = \sum_{j=1}^{k} \lambda_{ij}^2$ are then obtained as a by-product of such method.

The minimization approach traditionally adopted with MINRES is based on a Gauss-Seidel procedure, originally proposed in Harman (1966). The procedure starts with the choice of the number of factors K and the selection of an arbitrary loading matrix, for the given N variables and selected K factors. At each iteration, the arbitrary factor loading matrix is modified on a per-row (*i.e.*, per-variable) basis, by considering an increment x_{ik} of the loading value of the considered variable i on each factor k, and by selecting the displacements vector X_i that minimizes the objective function. This optimal displacements can be iteratively determined by zeroing the partial derivatives of Equation (11) with respect to the considered variable i, considering the loading matrix modified at previous iteration, *i.e.*,

$$\mathbf{X}_i = \mathbf{R}_i^0 \mathbf{\Lambda} (\mathbf{\Lambda}_{)i(}^{\intercal} \mathbf{\Lambda}_{)i(})^{-1}. \quad (12)$$

Here, $\mathbf{X}_i$ is the row vector of incremental changes of the factor loadings for variable i; $\mathbf{\Lambda}_{)i(}$ is the factor matrix obtained by replacing with zeros the elements in row i of the current factor matrix; and, $\mathbf{R}^0$ is the vector of (observed) residual correlations of variable i with all other variables, with zeros for self-residuals.

Successive adjustments of the factor loadings matrix are performed multiple times on all variables, until a numerical convergence criterion, related to the rate of change of F_K, is satisfied, and the final loadings matrix is obtained. Specific details on the factor estimation procedure and suggested convergence criteria and parameters can be found in Harman (1966), while later optimizations are discussed in Joreskog (1978); Briggs (2003); Comrey (2013) as well as in the `fa` function implementation of the MINRES method, from the R `psych` package Revelle (2021).

3.3 EFA, PCA and CFA

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 Jolliffe (2002).

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 experimentally shown, *e.g.*, in DeWinter (2016). 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 the identification of hidden regular structures in the target data. These structures are primarily driven by strong correlations that are difficult to observe in practice, as they are entangled within the mass of observations: thus, our 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.

Another rightful doubt on our choice of methodology may concern the higher suitability of EFA to solve the problem at hand, when compared against the other major class of factor analysis, *i.e.*, Confirmatory Factor Analysis (CFA). The

rationale for our selection is that, in EFA, the investigator has no expectations of the number or nature of the variables; this allows freely exploring the main dimensions to generate a novel theory or model from a relatively large set of latent constructs. Instead, CFA is a form of structural equation modelling, intended to test a proposed theory or model. In contrast to EFA, CFA has assumptions and expectations based on the a-priori theory, regarding the number of factors and which factor such theory best fits Mulaik (2009); Williams (2010). It is thus apparent that EFA is the correct tool for our problem, where no prior knowledge of the underlying temporal and spatial structure is granted. In the following, we will refer to EFA and *factor analysis* interchangeably.

4 Mobile data traffic analysis with EFA

We now introduce the mobile data traffic measurement dataset employed throughout our study, and discuss how EFA can be tailored to the problems of inferring temporal and spatial structures in such type of data. By doing so, we show how the EFA framework can be cast to solve the two tasks, which are in fact each other's dual.

4.1 Measurement dataset

We analyze mobile data traffic collected in the production infrastructure of a major network operator in Europe. The data was gathered during 12 consecutive weeks in the two largest urban areas of France, *i.e.*, Paris and Lyon, and covers the whole user base of the operator, which has a market penetration beyond 30% in in the target areas. The measurement data consists of total (*i.e.*, aggregated in the uplink and downlink directions) volume of data traffic generated by the whole subscriber base of the operator in the considered regions, consisting in several millions of unique (resident or temporary) users. The traffic data is geo-localized at the level of the radio access antenna serving each user, and is timestamped using a temporal granularity of 30 minutes.

To collect such data, the operator employed passive probes tapping at the interfaces of the 3G/4G network gateway. The probes monitor all individual uplink/downlink IP flows from/to over 1,000 NodeB and eNB antennas that provide

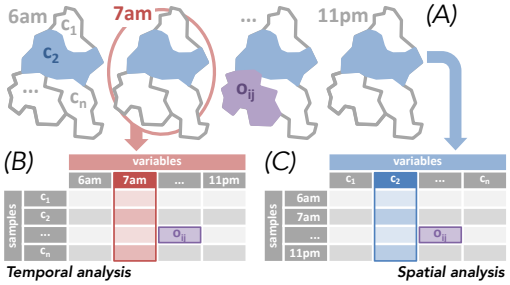


Fig. 2: Mobile data traffic analysis with EFA in a toy scenario. (A) The one-week demand in the target region is aggregated on a hourly basis with respect to a spatial tessellation of n cells, each representing the coverage of one antenna. The resulting demand in the i-th cell during the j-th time slot is the EFA observation o_{ij}. (B) Temporal analysis: the hourly time slots are the EFA variables, each characterized by a set of observations over the cell samples. (C) Spatial analysis: the geographical cells are the EFA variables, each characterized by a set of observations over the hourly samples. Figure best viewed in colors.

complete coverage to the two cities. The hourly antenna-level traffic volumes were computed by summing the contribution of such flows in secure servers within the operator's premises. The whole process took place under the supervision of the company Data Protection Officer (DPO) and of the national authority for privacy, in compliance with applicable national and European regulations. For our study, we only had access to the resulting de-personalized data, whose samples are aggregated over hundreds of data subjects at least, hence do not pose personal privacy risks.

4.2 Casting EFA for temporal and spatial analyses

As anticipated in Section 1, EFA is a versatile tool that can be cast to identify both temporal and spatial structures hidden in the dynamics of mobile data traffic. The input to either problem is an aggregate representation of the communication activity of the 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 (*e.g.*, voice calls, text messages, generic data usage, or consumption

of specific mobile services). In our specific case, the input format is aligned to the measurement data presented in Section 4.1 above: the activity maps to the volume of mobile data traffic, which is provided at a spatiotemporal resolution of the antenna location during every hour. We remark that the spatial mapping of traffic is then performed by adopting the legacy approach of approximating the coverage area of each antenna via a Voronoi tessellation, and assuming a uniform distribution of traffic within each Voronoi cell Paul (2011).

In order to explain how to cast EFA to solve the problems of temporal and spatial analysis of mobile traffic, let us consider the example in Figure 2. Here, the traffic demand in a given geographical region is aggregated on a hourly basis in n spatial cells, each corresponding to the coverage area of one antenna. One EFA observation o_{ij} matches the mobile data traffic volume recorded at the antenna cell i during the hourly time slot j. Then, the two problems are set apart by the mapping of variables $\mathbf{X}$ in (1), as follows.

Temporal analysis. We model *time slots* as the EFA variables. Each variable is described by the mobile traffic demand (*i.e.*, the EFA observations o_{ij}) recorded over all spatial cells $c_1 \dots c_n$ during a given hourly time interval (*e.g.*, 7:00 to 7:59 AM), as shown in plot (B) of Figure 2. In this EFA configuration, each variable (*i.e.*, hour) is represented by a snapshot of the spatial distribution of traffic across cells. The common factors sought by EFA are then temporal structures that explain at what time instants the geographical distribution of the mobile demand is comparable.

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 especially important for a given temporal profile. This allows interpreting the temporal analysis results from a spatial dimension.

Spatial analysis. EFA variables correspond to *geographical areas*. Each variable consists in the mobile traffic demand (*i.e.*, the EFA observations) recorded in a specific cell through the complete monitoring period, as in plot (C) of Figure 2. 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.

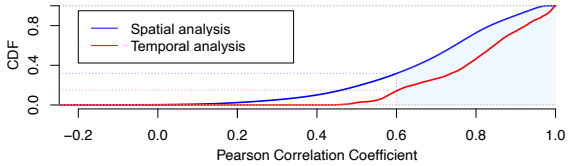


Fig. 3: Distributions of the Pearson correlation coefficient computed between all pairs of EFA variables in the temporal and spatial mobile traffic analysis problems.

Interestingly, time slots become the EFA samples in this configuration. Therefore, the EFA scores now explain what time periods are especially distinctive of the mobile traffic usage in different areas, offering again a unique spatiotemporal interpretation of the results.

4.3 Tuning EFA for mobile data traffic analysis

The two configurations of EFA outlined in Section 4.2 highlight how the temporal and spatial analyses of mobile data traffic are in fact dual problems. Beyond such a high-level mapping of variables and samples, several adjustments are needed in order to adapt the baseline EFA scheme to the specific problems at hand, including data verification and method parametrization. We discuss these aspects next.

Suitability of mobile traffic data for EFA. The definition of factor analysis in Section 3.1 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 analyzed. 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* Kaiser (1974) on our reference datasets presented in Section 4.1, which measures the proportion of the variance in the variables of the data that could be caused by the same common factors across all variables, and not just common correlations across pairs of variables. 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 Figure 3; (*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 appears 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. Many heuristics have been proposed to automatically determine such a number: most of them measure properties of the correlation matrix, search for the value that maximizes or minimizes the observed property, and suggest this value as the number of factors to retain Cattell (1966).

We rely on *parallel analysis* (PA) Horn (1965), which is recognized as one of the best methods for deciding how many factors to extract Thompson (2004). The PA method 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 Section 3, loadings are the main instrument to link factors and variables: thus, the presence of stronger (*i.e.*, closer to 1 or -1) loadings outlines more neatly structures in the data and eases result interpretation.

We use *VARIMAX rotation* Kaiser (1958) to identify an 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

$$\sum_{j=1}^{K} \frac{N \sum_{i=1}^{N} (a_{ij}^2/h_i^2)^2 - \left(\sum_{i=1}^{N} a_{ij}^2/h_i^2\right)^2}{N^2}, \quad (13)$$

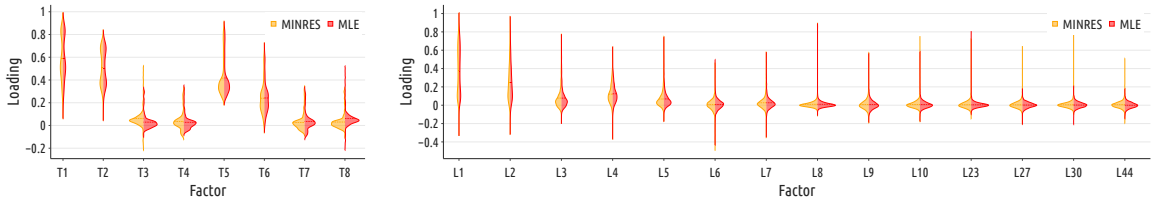
where a_{ij} are the elements of $\mathbf{\Lambda T}$ and h_i is the communality of the i-th variable defined as in (4) and computed from $\mathbf{\Lambda T}$.

Solving method. As detailed in Section 3.2, we consider two different methods to extract the common and unique factors via EFA, *i.e.*, MLE and MINRES. In order to measure the quality of the factors returned by each approach, we rely again on the loading values. As already mentioned, higher loadings on an equivalent number of factors indicate a sharper and more explainable decomposition of the data, since variables are linked to hidden factors in a tidier way.

In fact, solving the EFA problems with MLE and MINRES returns the exact same factors, which already proves the robustness of the whole framework. It also simplifies the direct comparison of the loadings on such factors: specifically, Figure 4 illustrates the distributions of loadings on the returned factors for all variables, as violin plots. The main takeaway is that the MLE and MINRES yield almost identical distributions of the loadings, as discrepancies are minimal, and only affect a few factors. We conclude that there is no operational difference in adopting one solving method or the other, and focus on results obtained by the computationally faster MINRES in the following.

5 Temporal structures

We start by discussing the results for the temporal analysis. We focus on a compressed version of the temporal data by adopting a median week representation Furno (2016): for each spatial cell, we compute the median traffic volume observed at each hour of the week (*e.g.*, for all Mondays, 7:00 am to 8:00 am), and use such values as EFA variables. The rationale for this choice is twofold. On the one hand, considering each hourly time slot in the 12-week period as an independent EFA variable would lead to a very large number of variables, higher than the number of cells, *i.e.*, EFA samples: this is a condition that shall be



(a) Temporal analysis

(b) Spatial analysis

Fig. 4: Violin plots of the loading values on (a) temporal and (b) spatial factors, when solving the EFA problem with MINRES and MLE. For the spatial analysis, the distributions are shown for a selection of the factors returned by EFA.

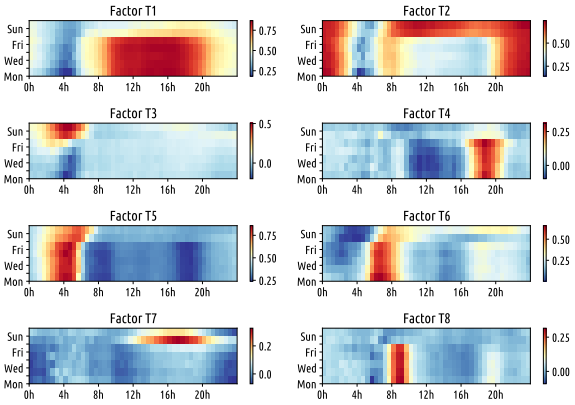


Fig. 5: Temporal factors obtained from EFA, for the median week mobile traffic demand. Each plot refers to one common factor returned by EFA. In every plot, the hourly time slots (EFA variables) are arranged along 24-hour daily cycles (on the abscissa) for 7 days (Monday to Sunday, on the ordinate), and colors illustrate the loading of the time slot on the considered factor. Figure best viewed in colors.

avoided, as it makes factor inference less dependable DeWinter (2009). On the other hand, hourly time variables recorded at the level of individual antennas are affected by substantial random noise; a median week compression is known to mitigate potential biases due to outlying behaviors, which, in our context, makes more representative traffic structures emerge.

5.1 Temporal structures across time

By solving the EFA problem in the temporal analysis configuration above, a total of 8 different common factors are obtained from our reference dataset, which we label T1 through T8. An early insight provided by our methodology is therefore that *the weekly dynamics of mobile data traffic can be summarized into a very small number of archetypal profiles*. Indeed, the dynamics observed in the 24×7 time slots can be explained as a linear combination of just 8 patterns. An interesting remark is that these results are obtained by merging the measurement data for Paris and Lyon, *i.e.*, using all cells in both cities as a single set of samples; this approach is preferred (rather than running for each city) as it allows the identification of more generalist temporal factors, which apply across cities and are not specific to a single urban area. Thus, both the above consideration and the remaining behaviors discussed to be discussed in this section will apply to both urban areas.

Factor	***Activity***	***Hours***
T1	Working	Weekdays, 8 am – 8 pm
T2	Relax	Weekday evenings, weekends
T3, T5	Nightlife	Midnight through 8 am
T4	After-work	From 6 pm until 9 pm
T6, T8	Early	From 6 am until 10 am
T7	Weekends	Saturday & Sunday noon

Table 2: Temporal factors in mobile data traffic identified by EFA.

The detailed loadings of all variables, *i.e.*, hourly time slots in a week for each factor, are shown in Figure 5. Such loadings are color-coded according to the scale on the right side of the plot, with red representing higher values, and blue mapping to lower values.

The representation in Figure 5 allows for a preliminary interpretation of the factors, recapitulated in Table 2. Factor **T1** shows higher loading values for time slots from Monday through Friday, starting at 8 am and decreasing after 8 pm; these are easily mapped to the standard working hours. Factor **T2** presents a complementary behavior to T1, as its loadings are lower during work times, and increase after 8 pm until 2 am; moreover, loadings on T2 are also high during the whole weekend, which makes us tag T2 as characterizing relax hours of the week. Profile **T7** yields some similarity to T2, but with high loadings limited to weekends, specifically Saturday afternoon.

Three factors can then be related to behaviors occurring at the start or end of the workdays. Two factors show high loadings in early hours: factor **T6** showing its peaks from 6 am to 8 am across all days of the week and interestingly a small 1h shift on the weekends (when peaks start appearing around 7 am), which can mean that this is a factor related to general early commuting; factor **T8** peaks are slightly shifted from the previous and go from 8 am until 10 am, but only for work days, which makes us believe this factor could relate to normal office commuting, as many workers in France start their work hours at the office between those times. Opposing the previous two, factor **T4** has its peaks at after work hours (5 pm to 8 pm) of weed days, which can relate to the general commuting related to leaving work and study spaces. It's interesting to note that there's no specific factor related to after work commuting on weekends, which can relate to the fact that this movement may be more dispersed throughout the weekend, with no clear pattern happening at our studied urban centers.

The two final profiles correspond to late night behaviors: factor **T3** focus in high activities exclusively during the weekends, with an initial growth in activity being seen Friday and peaking on Saturday and Sunday from midnight to 7am. Meanwhile, factor **T5** also has peaks of activity during late hours of the night, but with its peaks spread across all days of the week. Those differences may be due to factor **T3** being related to late night leisure while factor **T5** just as overall late night activity; further exploration of the geographical distribution of those factors during Section 5.2 can help confirm those hypothesis.

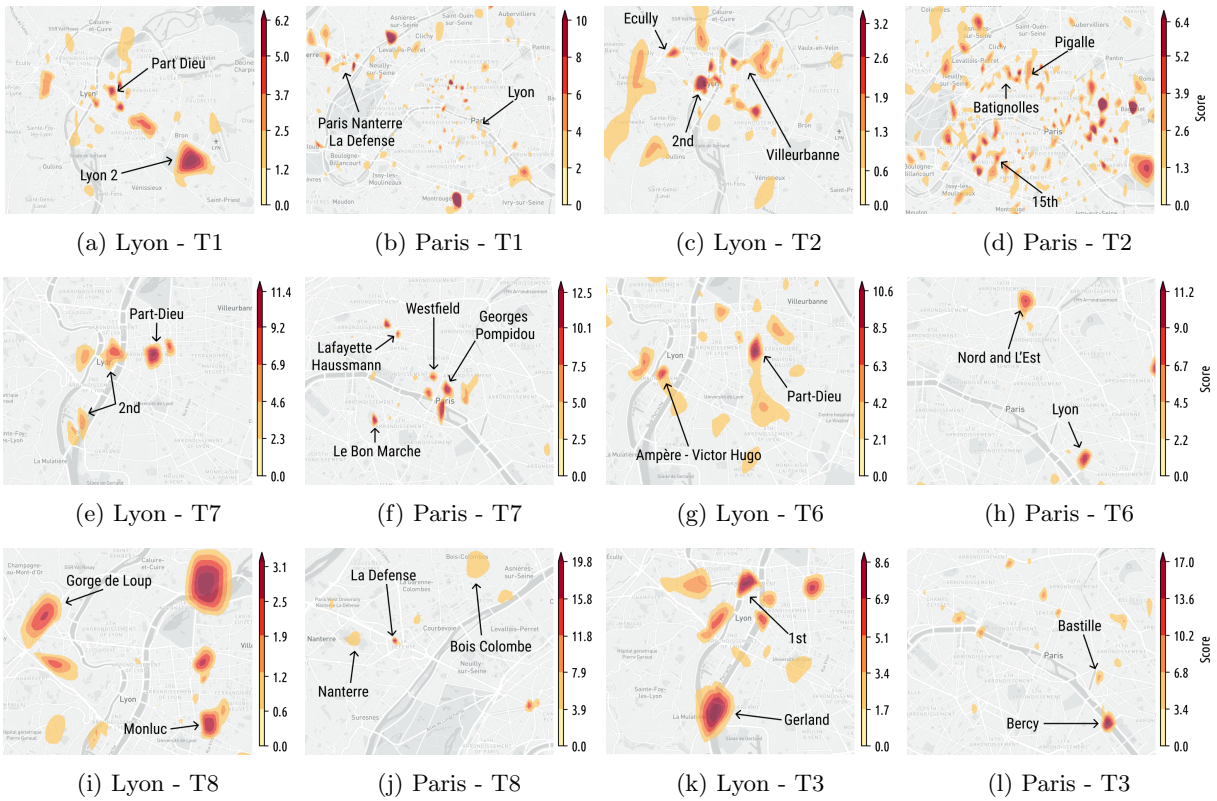
It is worth nothing that, unlike traditional clustering approaches Furno et al. (2017), EFA yields a non-deterministic association of time intervals and factors. Indeed, a specific same hour of the week may be affected by multiple concurrent profiles with changes of intensity; this is the case of Saturday afternoons (composed by T2 and T7, plus T1 to a lesser extent) or 8 am during working days (T8 and T6, plus T2 to a lesser extent). This representation highlights the composite nature of mobile traffic patterns observed during the week, represented as a combination of multiple factors.

5.2 Geographical distribution of temporal structures

Looking at the loadings in Figure 5 may not be enough to disambuiguate the phenomena that determine the temporal structures identified by EFA. Only obvious root causes, such as the work activities that underpin factor T1, can be pinpointed with some level of confidence.

As discussed in Section 3, EFA offers an automated way to better interpret the results via the factor scores. In the case of temporal profiles of mobile traffic, EFA scores allow quantifying the importance of each geographical location sample for every factor. In other words, we can draw maps of the EFA score, where higher values in one area indicate that the local traffic has a higher influence for the targeted profile. Such visualization highlights locations and points of interest, which can explain the social phenomenon underlying each temporal factor. We compile those results for selected factors in both cites on Figure 6.

Factor T1 has higher score around work areas: universities (Jean Moulin Lyon 3, Lyon 2 and INSEEC in Lyon; Paris Nanterre in Paris), working zones (3th arrondissement of Lyon; downtown Paris and La Defense, one of the biggest business centers of Europe) hospitals and big stations (Part Dieu, Lyon's busiest station; Gare de Lyon and Montparnasse in Paris). This matches the temporal structures seen on Figure 5, meaning those regions have their main peaks of mobile traffic consumption during work hours. In contrast, the relaxing hours Profile T2 presents a different side of both cities. Regions with higher scores are either located close to shops and restaurants (such as 2th arrondissement in Lyon and Passy and Batignolles neighborhoods in Paris), which expect most



(a) Lyon - T1 (b) Paris - T1 (c) Lyon - T2 (d) Paris - T2

(e) Lyon - T7 (f) Paris - T7 (g) Lyon - T6 (h) Paris - T6

(i) Lyon - T8 (j) Paris - T8 (k) Lyon - T3 (l) Paris - T3

Fig. 6: Geographical distribution of temporal factors across Lyon and Paris.

of their traffic activity after work, as well as places with an active night live (Pigalle neighborhood in Paris). Also, this profile emerges in more residential zones of both cities (Villeurbane, Ecully and Tassin in Lyon; 12th and 15th arrondissement in Paris). This opposition between working and residential places between both profiles is further seen when we compare Figures 6a and 6c for Lyon, as well as Figures 6b and 6b for Paris: bigger clusters of one factor are mainly in empty places of the other.

Factor T7, seen on Figures 6e and 6f, presents some of the favored locations during the weekends: department stores. In Lyon, a cluster is seen around 2th arrondissement (which known for its large commercial centers) and around the La Part Dieu Shopping Center. The same can be seen in Paris, where clusters are close to Le Bon Marche Department Store, Westfield Forum des Halles Mall and Galeries Lafayette, as well as the cultural center Georges Pompidou. In France, many stores are closed on Sundays, which could explain an extra influx of mobile traffic around those locations on Saturdays.

The geographical distribution of factors also helps further differentiate the three profiles that appear to be associated to commuting behaviors. Factor T6 shows higher loading from 6am to 8 am, and its scores on Figure 6g and 6h help relate it to general peak-hour commuting behavior. Clusters are seen around bigger stations, such as Ampère Victor Hugo and Part Dieu in Lyon, as well as Paris' Gare du Nord, de L'Est and Lyon. Those are highly connected stations that attract many for their daily commute. Factor T8 (seen on Figures 6i and 6j) relates to early day commuting, with regions located mostly in residential and suburban areas, such as Monluc and Gorge de Loup in Lyon, or Bois Colombe in Paris, as well as some business zones like Nanterre and La Defense, both located in the capital. Finally, T4 represents specifically after work commuting, from 5pm to 8pm. This reflects in regions with higher scores closer to the city's downtown. A few exceptions can be made,

like the La Defense region, which shows high values for both profiles (as well as T6). This is easily explained due to La Defense being one of the busiest work-only districts of Paris, attracting a strong influx and outflux of people who work in the area on a daily basis.

For the late night profiles, we note that the locations for Factor T5 can be considered a subset of the locations for Factor T3 (seen on Figures 6k and 6l), as most zones with high scores on the first profile are also present in the latter. This includes regions in Paris that are known for having an active nightlife, such as Bastille and Bercy in Paris or the 1st arrondissement area in Lyon. Some regions known for relatively higher crime incidence also emerge, such as Gerland in Lyon.

6 Spatial structures

The different ways spaces in a city are related can be better understood by setting EFA variables as the spatial cells related to where the mobile traffic is, and the EFA samples as the 30-minutes aggregated time slots during the complete 12 weeks span of the dataset. This results in factors that provide insights on similar uses of smartphones through different regions of a city. Also, exploring the scores of each EFA factor can shed more knowledge about temporal behaviors. We summarize patterns as *long-term behaviors* (such as commuting, working, relaxation) and *short-term events* (such as football matches, concerts).

A total of 47 factors resulted from the EFA analysis of spatial structures. This considerably big number is explained by the proportions of the dataset: a total of 2 287 cells from both Lyon and Paris were used as variables, with 4 080 30-minutes slots for mobile traffic used as samples, resulting in a long set with a significant variety of behaviors. Due to space limitations, a complete discussion on the complete 47 land use profiles is not possible. We note that not all profiles have the same importance; only a small portion capture large geographical areas with the majority being common to a small number of neighborhoods characterized by very specific human activity features. We highlight the number of cells for each profile that have *loading* ≥ 1 on Figure 7, where green marks the profiles of this analysis, including the 9 with over 100 relevant cells which represent long-term land use behaviors, as well as a few short-term land use behaviors that show interesting spots of the cities. We present a summary of all included profiles in Table 3, and discuss them in the following subsections.

Table 3: Trivial land use cases for EFA spatial structures analysis, for both long and short-term behaviors.

Factor	***Land Use***	***Examples***
L1	Residential, relaxation	Cities around Paris, Dense Residential such as 18th, 19th and 20th arrondissements.
L2	Working places	La Defense, Paris-Nanterre, La Part Dieu
L3	Long-range commuting (Train and RER)	Gare du Nord and L'Est, Rueil-Malmaison
L4	Short-range commuting (Metro)	Subway stations all around downtown Paris and Lyon.
L5, L7	Shopping places	La Confluence, Les Halles, Opera, Elysees.
L6	Cargo and maintenance sites	Gare du Nord, Saint Ouen, Noisy-le-Sec
L9	Clubs and night life	Bastille, Lyon 1st arrondissement.
L8, L30	Sport stadiums	Stade de France and Parc des Princes Stadium
L10, L23, L27	Expo centers	Expo Porte de Versailles, AccordHotels Arena, Euroexpo Center
L44	Catholic churches	Notre-Dame des Champs, Saint-Augustin, Sainte Trinité, Sainte Geneviève des Grandes Carrières

6.1 Long-term land use behaviors

By observing constant behaviors across weeks and months, a number of profiles can be tied together by their long-term usage pattern. We determine those by the types of land where loading values are higher as well as the scores obtained from the temporal samples. This results in a set of trivial land uses in cities, such as working places, residential neighborhoods and public transportation stations. The long-term land uses profiles present similar behaviors to the network profiling discussed

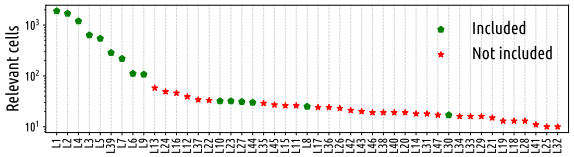
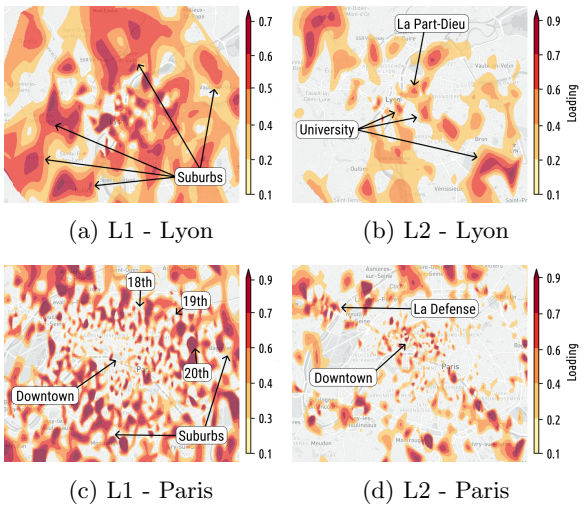


Fig. 7: For each spatial profile, the number of cells that are considered relevant for it($loading \geq 0.1$), where green marks the profiles included in this analysis and red marks the non-included.

in Section 5.1, but with more detailed spatial patterns than those in the network profiling maps.

Land use profiles **L1** and **L2** exemplify well the extra details obtained from the spatial structures. The first one presents spaces characterized by **relaxation activities** (dense residential and suburban regions), while the second relates to **work/study locations**. The duality of these factors can be seen in Figure 8: active locations during the work day, such as business districts (La Defense in Paris, La Part Dieu at the 3rd arrondissement of Lyon), universities and downtown areas are highlighted in L2, but disappear in L1. In the same way, suburbs and residential neighborhoods of Paris (such as the dense residential 18th, 19th and 20th arrondissements) are highlighted in L1 and not shown in L2. Scores seen on Figure 9 highlight the differences over time for the land uses: L1 has higher then average scores around night time, while L2 scores are higher around 8am till 8pm on working days and lower values around the weekend.

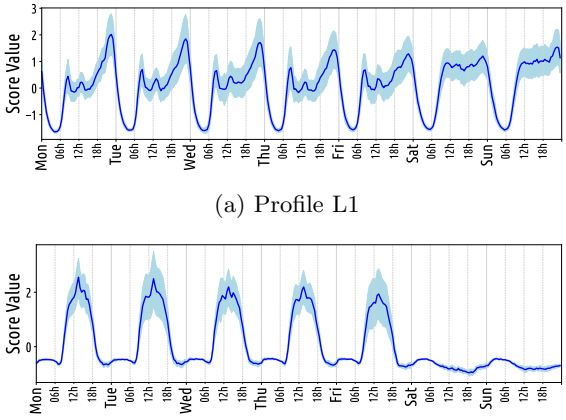
Profiles L3, L4 and L6 present land uses related to the different usages of **public transportation**. Profile **L3** is associated to stations used for medium/long-range commuting, which offer high speed trains (TGV), standard regional trains (TER) and metropolitan trains (RER, exclusive to Paris); most of the major train stations of the cities emerge, with details seen on Figures 10a and 10b. Profile **L4** links to short-range commuting (metro stations), which explains the higher density of relevant spatial cells in the downtown areas, as we can see on Figure 10c with a few key metro stations highlighted. Profile **L6** is a profile that shows operational spaces for metro/train, such as a maintenance center for the SNCF railway company in Saint Ouen and Noisy-le-Sec, both located in the metropolitan area of Paris and seen on Figure 10f; we also see the exit yard for trains of



(a) L1 - Lyon (b) L2 - Lyon

(c) L1 - Paris (d) L2 - Paris

Fig. 8: Land usage loading maps for long-term behavior profiles L1 and L2 in (a,b) Lyon and (c,d) Paris.



(a) Profile L1

(b) Profile L2

Fig. 9: Land usage scores across time for long-term behavior profiles (a) L1 and (b) L2, for both Lyon and Paris.

both Gare du Nord and de l'Est. Scores can assist the differentiation between all public transportation spaces. As seen on Figure 11a, L3 sees peaks on the beginning and end of week days, which relates to the use of those long-range stations for the commuting to work (a common behavior for both cities due to their business and economic importance), and less during weekends. L4 sees more uniform scores across the day, showing that

those short-range stations tend to have a more uniform usage across time, while the cargo and maintenance spaces of L6 see their activity rise late in the evening.

Profiles L5 and L7 relate to **leisure areas** of the cities, including entertainment spaces such as stores and restaurants. Profile **L5** presents locations known for their shops, bistros and galleries; this is seen on Figure 10e in Lyon for La Confluence, La Part Dieu and Cordeliers, as well as many shopping regions around Paris' 1st arrondissement (such as the Forum des Halles mall) and Opera regions (where many luxury galeries are present), seen on Figure 10d. Meanwhile, Profile **L7** presents commercial locations in Lyon's Ainay and Paris' Elysees and Madeleine, as seen on Figure 10g. We can better differentiate L5 and L7 when analyzing their scores over time. Profile L5 shows on Figure 11b a behavior previously seen in Section 5.1, with higher scores occurring on Saturdays (the preferred weekend shopping day in France); in the meantime L7 has more uniform scores from 8am to 8pm all days of the week. This might indicate that places around L5 are preferred by the population for shopping, while L7 shows a more mixed use that, although affected by a significant presence of commercial location, observes a more constant traffic due to the presence of offices.

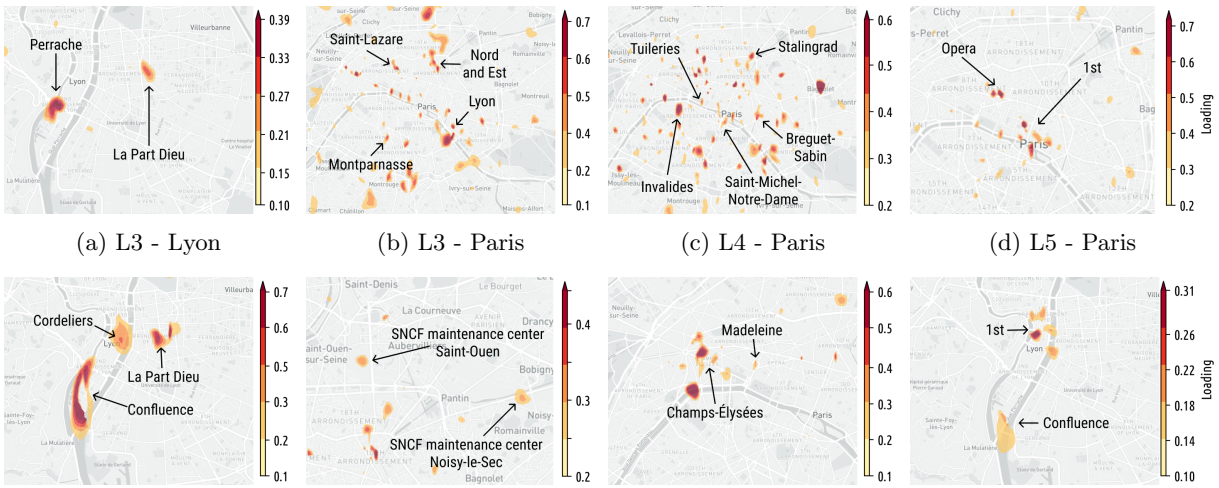
The final long-term land use profiles we explore is **L9**, seen for Lyon on Figure 10h highlighting locations known for their **late night attractions**. This is present on the 1st arrondissement of Lyon and the region of La Confluence, both having a good density of night clubs. Paris similarly has a good number of cells with high loading around Bastille, Pigalle and the 1st/2nd arrondissements, which are also known for clubs and restaurants opened until late night. When looking at the score for L9 on Figure 11c, the behavior in time of those locations match their usage, with values increasing on Fridays and Saturdays after 6pm, and being at the lowest from Sunday until Wednesday.

6.2 Short-term land use behaviors

The exploration of spatial structures also present regions of the cities used for short-term events that happen at punctual moments in time, *e.g.*, during a few hours of one or multiple days. Such short-term patterns can be found when analyzing EFA scores for the time samples, which exhibit above average values for short intervals only.

A good example of short-term behaviors is provided by a couple profiles that emerge for stadiums in Paris. Profile **L8**, seen on Figure 12a, is related to Stade de France (the biggest sports stadium in France), while Profile **L30** shows the Parc des Princes (used by the Paris Saint Germain football team). The scores for each profile indicate in which moments those spaces saw mobile traffic that differentiate them from their neighborhood. For State de France (L8), four dates during night emerged: Oct. 7th (football match between France and Bulgaria), Nov. 11th (football match between France and Sweden), Nov. 19th (rugby match between France and Australia) and Nov. 26th (rugby match between France and New Zealand). Those dates are seen on Figure 12c, which shows the scores for L8 across time. For Parc des Princes (L30), the scores show all of PSG home matches during the studied period: Sep. 9th (vs. Saint-Étienne), Sep. 13th (vs. Arsenal), Sep. 20th (vs. Dijon), Oct. 1st (vs. Girondins), Oct. 19th (vs. Basilea), Oct. 23th (vs. Basilea), Nov. 6th (vs. Stade Rennais) and Nov. 19th (vs. Nantes).

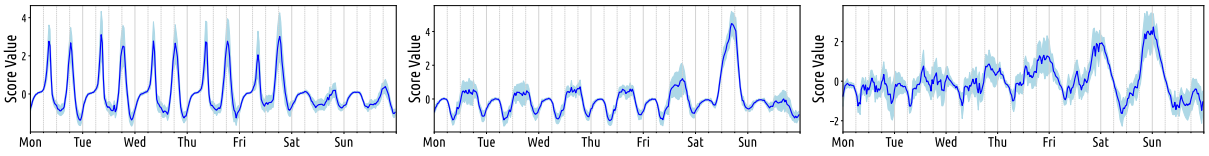
Equivalently, three factors relate to exposition centers. Profile **L10** shows both the Euroexpo center in Lyon, as well as the Expo Porte de Versailles in Paris; Profile **L23** shows AccorHotels Arena, an indoor arena and concert hall located in the neighborhood of Bercy in Paris and seen on Figure 12b; Profile **L27** shows exclusively the Expo Porte de Versailles in Paris. Like was seen with stadiums, the scores for the profiles represent specific events that made those regions unique for their land usage. L23 had three different events happening at the arena during the studied period: two shows happening on the nights of Sep. 20th and 21st, three music concerts happening on Oct. 15th, 16th and 18th and a tennis tournament hosted between Oct 29th and Nov. 6th. The scores of the events of Profile L23 are seen in Figure 12d: we note the concerts have higher values at the night time, while the tennis tournament starts around morning time and span throughout the entire week. It is also interesting to note the separation on scores for Oct. 17th, which present lower values since no event happened at the arena on this date.



(a) L3 - Lyon (b) L3 - Paris (c) L4 - Paris (d) L5 - Paris

(e) L5 - Lyon (f) L6 - Paris (g) L7 - Paris (h) L9 - Lyon

Fig. 10: Land usage loading maps for long-term behavior profiles: L3 in (a) Lyon and (b) Paris, (c) L4 in Paris, L5 in (d) Paris and (e) Lyon, (f) L6 in Paris, (g) L7 in Paris and (h) L9 in Lyon.



(a) Profile L3 (b) Profile L5 (c) Profile L9

Fig. 11: Land usage scores across time for long-term behavior profiles (a) L3, (b) L5 and (c) L9, for both Lyon and Paris.

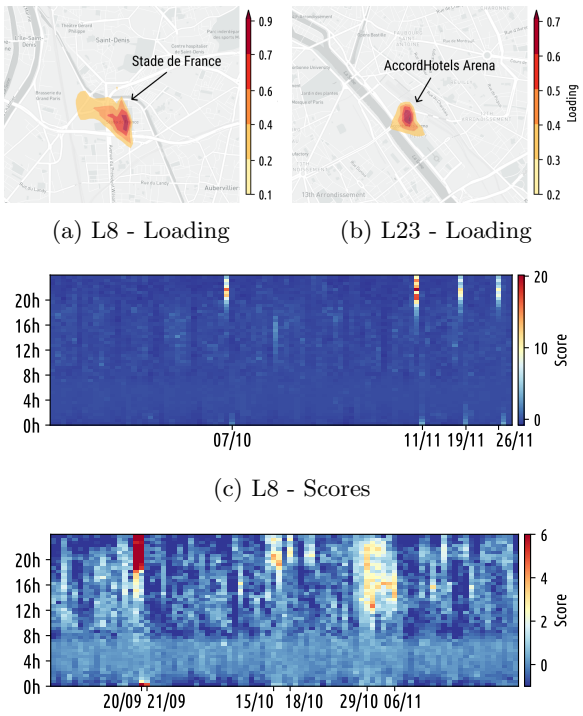
6.3 Mixed land use regions

We note that a significant number of cells present high loading values for multiple spatial profiles, which reveals the presence of mixed land usage in the cities. To find occurrences of cells with mixed land use, we utilize the Revealed Comparative Advantage (RCA) metric, a popular index in economics for measuring the relative advantage or disadvantage of traded commodities or services in the economy of a given region Balassa (1965). The RCA is calculated as:

$$RCA_{xf} = \frac{\lambda_{xf} / \sum_{x'=1}^{N} \lambda_{x'f}}{\sum_{f'=1}^{K} \Lambda_{xf'} / \sum_{x'=1}^{N} \sum_{f'=1}^{K} \lambda_{x'f'}} \quad (14)$$

where for our case $x \in \mathbf{X}$ is a cell that belongs to the $N \times 1$ vector of observed variables $\mathbf{X}$, $f \in \mathbf{F}$ is a profile that belongs to the $K \times 1$ vector of common factors $\mathbf{F}$, and λ_{xf} is the loading value of a given profile f for cell x that belongs to the $N \times K$ matrix of common factor pattern coefficients $\mathbf{\Lambda}$. This results in an index that presents the importance of one profile loading for that cell, pondered by all the loading values in that cell and all the values that profile have across the region. If $RCA_{xf} > 1$, a competitive advantage can be considered for factor f in cell x. Therefore, we consider that a cell presents mixed land usage in case it has an $RCA_{xf} > 1$ for more than one land use profile.

Analyzing the histogram for the number of land use profiles per cell for Paris and Lyon in Figure 13a, we note that a fairly small number of cells have a single land use; the majority of cells have instead between 3 and 4 significant profiles. It is important to highlight that regions with a



(a) L8 - Loading (b) L23 - Loading

(c) L8 - Scores

(d) L23 - Scores

Fig. 12: Land usage loading maps in Paris for profiles (a) L8 and (b) L23, as well as the scores across time for (c) L8 and (d) L23.

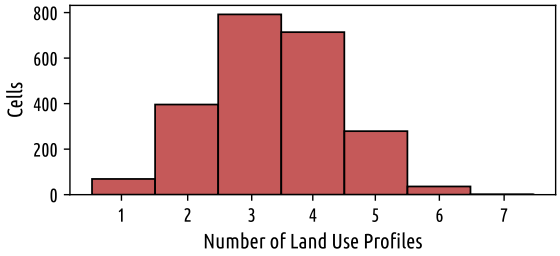
higher density of uses are mostly seen in downtown regions, like the Paris' regions north of the Seine river, as well as 1st and 2nd arrondissements in Lyon. Similarly, areas on the suburban parts tend to have less mixed usage, such as the southern zones of Paris and suburban areas.

To exemplify the identification of regions with mixed land use through EFA, we choose to explore a few examples of cells that had $RCA_{xf} > 1$ for two land use profiles. We present maps that showcase cells with mixed usage use of two profiles $[f_1, f_2]$ across cities by calculating for each significant cell $x \in \mathbf{X}$ that has $[RCA_{xf_1}, RCA_{xf_2}] > 1$ the value $(RCA_{xf_1} - 1) * (RCA_{xf_2} - 1)$, which will represent the *mixed usage RCA coefficient* for each set $[f_1, f_2]$. With this, we can better observe the incidence and spatial pattern of mixed use cells over the cities. Our first example encompasses cells with mix use of relaxing/residential (L1) and working (L2) hours profiles, a mixed land use that occurs on 209 of the total 1752 cells (11.93%) that are significant for either L1 or L2. Their distribution over space can be seen on Figure 13b, from which we can highlight the presence in Paris of many cells around the 1st and 7th arrondissement, which are highly touristic regions in the city, with many shopping stores, hotels and residential apartments. We also see the 10th arrondissemnt, which although not a typical tourist spot of the city, it's a densely residential area with the presence of the two main train stations of Paris (Gare du Nord and de l'Est). Those are regions that are active the entire day due to their extreme diverse zoning.

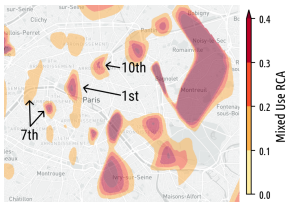
Another combination of land uses that is worth exploring is that between relax/residential (L1) and leisure (L5) profiles, where 332 of 1579 cells (21.03%) significant for either L1 or L5 have this mixed usage. Those locations prove to be active both in after-work hours and during weekends, and, as we see in Figure 13d, encompass zones featuring good mixture of parks, houses, restaurants and cultural spots, notably Pentes de la Croix-Rousse, Fourvière and Les Cordeliers in Lyon, as well as the 9th arrondissement and Marais in Paris. Finally, one the biggest intersections we identified is that between the leisure (L5) and late night (L9) land use profiles, where 460 of 1457 cells (31.57%) can be associated to both behaviors. As observed in Figure 13f, this happens at regions where people tend to spend their free time, both during the night and on the weekends, with popular neighborhoods in Paris such as Bastille, 3rd and 4th arrondissements, as well as the 1st and 2nd arrondissements of Lyon.

7 Conclusions and perspectives

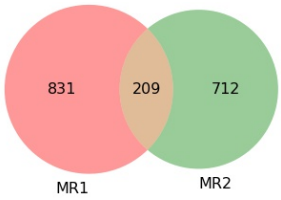
We proposed an original approach to the spatiotemporal classification of mobile traffic data, which relies on Exploratory Factor Analysis (EFA). Extensive tests with heterogeneous real-world datasets demonstrate the versatility of EFA, which provides a unifying framework to solve problems that have been 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.*,



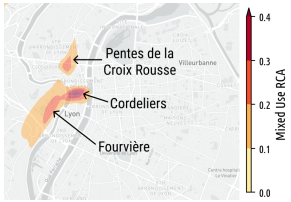
(a) Distribution of cells with multiple land use profiles



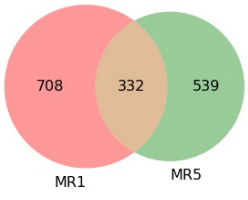
(b) L1 + L2 - Paris



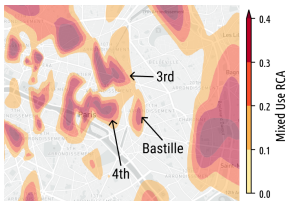
(c) L1 + L2 - Venn



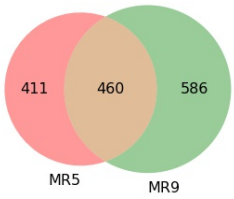
(d) L1 + L5 - Lyon



(e) L1 + L5 - Venn



(f) L5 + L9 - Paris



(g) L5 + L9 - Venn

Fig. 13: (a) Distribution of cells by the number of significant land use profiles; mixed use RCA coefficient and Venn diagrams for the mixed land use cases of (b-c) L1 + L2, (d-e) L1 + L5 and (f-g) L5 + L9.

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 multiple applications in data-driven studies, applicable across various levels and contexts. Firstly, our methodology offers fertile ground for travel-demand-related mobility studies. The temporal structures can provide precise insights into the key patterns of population presence across different days of the week. These patterns are linked to specific activities that people engage in within designated areas of the city. Factor loadings from the temporal description indicate times when individuals either currently require or will soon need access to transportation. For example, in the context of commuting factors, the associated loadings clearly delineate the moments people are arriving at or departing from major transportation hubs, train stations and transit stops. Concerning work or residential factors, loading variations can precisely pinpoint population transition to different activities potentially located in other parts of the city, thus indicating potential peaks in the associated travel demand. On the other hand, the spatial analysis underscores the potential of our methodology for studying the attractiveness of specific urban areas over time. This is true for both recurring and episodic scenarios, such as special events, as observed through mobile phone data. This information is invaluable for traffic simulation and transportation planning, serving again as a precious proxy for travel demand data that can be otherwise challenging to acquire in a real-time like fashion. Specifically, the identification of areas occasionally or regularly experiencing high levels of activity can enable the easier identification of accessibility bottlenecks and support the optimization of transportation resources and transit schedules.

Moreover, the temporal structures identified by EFA expose non-trivial long-term dynamics in the mobile traffic demand that are relevant to the allocation of mobile communication resources in, *e.g.*, Cloud Radio Access Networks (C-RAN) CogNet (2015). 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 transport resilience and anticipatory networking studies. 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 CogNet (2015). Overall, EFA-based classification is a potential brick for future big data-driven 5G systems Zheng (2016).

Future works using EFA together with mobile phone data can help in elucidating and modeling the specific links between the detected hidden structures and travel demand variations, finding ways to build models able to combine more traditional transportation data with information obtained from mobile network measurements to build even more robust models.

Declarations

Acknowledgments

The authors thank friends and colleagues in Milan and Paris who helped interpreting the classification results.

Ethical approval

Our work builds on mobile network traffic generated by users of a nationwide cellular infrastructure. The traffic measurements used to derive the data set were collected by the operator for network management and research purposes, and temporarily stored within a secure platform at their own premises. The aggregation at the level of the radio access antennas was also carried out in the same platform by personnel of the network operator, in full compliance with Article 89 of the General Data Protection Regulation (GDPR) of the European Commission. The data collection and processing was approved by the Data Protection Officer (DPO) of the operator, and authorized by the French National Commission on Informatics and Liberty (CNIL), within the context of a collaborative research project.

We remark that the original network measurements contained personal identifiers (*e.g.*, the International Mobile Subscriber Identifier, or IMSI) and sensitive data (*e.g.*, locations of visited antennas, or mobile services consumed) about individual users, and were deleted upon aggregation. Instead, the aggregated data consist of time series of total traffic at the antenna level with a temporal granularity of 30 minutes, and do not contain personal identifiers or sensitive information, such as the device type, preference in terms of application consumption, or trajectories. In addition, the level of spatiotemporal aggregation ensures that no data subject can be re-identified, and that the statistics do not configure as personal data in the GDPR acceptation.

The researchers involved in the work presented in this paper only had access to such aggregated and privacy-preserving statistics for the purpose of carrying out the study. Ultimately, our dataset and research do not involve risks for the mobile subscribers.

Competing interests

The authors declared that there is no conflict of interest.

Author's contributions

A.F: Conceptualization, data curation, methodology, original draft preparation, review and editing; A.F.Z: Formal analysis and investigation, original draft preparation, review and editing; R.S: Conceptualization, data curation, methodology, original draft preparation, review and editing; M.F: Conceptualization, data curation, methodology, original draft preparation, review and editing.

Funding

The work of A.F. was supported by the French ANR research grant ANR-18-CE22-0008 PROMENADE. The work of A.F.Z. was supported by BANYAN project, which received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 860239. The work of M.F. was supported by NetSense, grant no. 2019-T1/TIC-16037 funded by Comunidad de Madrid, and by the research project CoCo5G (Traffic Collection, Contextual Analysis, Data-driven Optimization for 5G), grant no. ANR-22-CE25-0016, funded by the French National Research Agency (ANR).

Availability of data and materials

The data utilized on this work is proprietary and cannot be made publicly available.

References

P. Cerwall *et al.*, "Ericsson Mobility Report – June 2021", 2021.

M.Z. Shafiq, L. Ji, A. X. Liu, J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," *ACM SIGMETRICS*, 2011.

D. Willkomm, S. Machiraju, J. Bolot, A. Wolisz, "Primary Users in Cellular Networks: A Large-Scale Measurement Study," IEEE DySPAN, 2008.

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.

U. Paul, A.P. Subramanian, M.M. Buddhikot, S.R. Das, "Understanding Traffic Dynamics in Cellular Data Networks," *IEEE INFOCOM*, 2011.

R. Keralapura, A. Nucci, Z.-L. Zhang, L. Gao, "Profiling Users in a 3G Network Using Hourglass Co-Clustering," *ACM MobiCom*, 2010.

E. Mucelli Rezende Oliveira, A.C. Viana, K.P. Naveen, C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," *IEEE PerCom*, 2015.

H. Li, X. Lu, X. Liu, T. Xie, K. Bian, F. X. Lin, Q. Mei, F. Feng, "Characterizing Smartphone Usage Patterns from Millions of Android Users," *ACM IMC*, 2015.

J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," *IEEE Transactions on Emerging Topics in Computing*, 3(1), 2015.

C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, Z. Smoreda, "Not All Apps Are Created Equal: Analysis of Spatiotemporal Heterogeneity in Nationwide Mobile Service Usage," *ACM CoNEXT*, 2017.

M.Z. Shafiq, L. Ji, A.X. Liu, J. Pang, J. Wang, "Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network," *IEEE INFOCOM* 2012.

E. Peltonen, E. Lagerspetz, J. Hamberg, A. Mehrotra, M. Musolesi, P. Nurmi, S. Tarkoma, "The Hidden Image of Mobile Apps: Geographic, Demographic, and Cultural Factors in Mobile Usage," *ACM MobileHCI*, 2018.

R. Singh, M. Fiore, M. Marina, A. Tarable, A. Nordio, "Urban Vibes and Rural Charms: Analysis of Geographic Diversity in Mobile Service Usage at National Scale," *WWW*, 2019.

V.D. Blondel, A. Decuyper, G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Sci.*, 4(10), 2015.

D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, "Large-scale Mobile Traffic Analysis: A Survey," *IEEE Communications Surveys & Tutorials*, 18(1), 2016.

M. Fekih, L. Bonnetain, A. Furno, P. Bonnel, Z. Smoreda, S. Galland, T. Bellemans, "Potential of cellular signaling data for time-of-day estimation and spatial classification of travel demand: a large-scale comparative study with travel survey and land use data," *Transportation Letters*, 14(7), 787-805, 2022.

J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, M. C. González. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies, 58, 162-177, 2015.*

Y. Chen, Z. Wang, H. Sun, Y. Zhang, Z. He, "Analysis of travel demand between transportation hubs in urban agglomeration based on mobile phone call detail record data," Journal of Transportation Engineering, Part A: Systems*, 148(7), 04022041, 2022.*

N. Breyer, C. Rydergren, D. Gundlegård, "Semi-supervised mode classification of inter-city trips from cellular network data," Journal of Big Data Analytics in Transportation*, 4(1), 23-39, 2022.*

H. Assem, T. Sandra Buda, L. Xu, "Initial use cases, scenarios and requirements," H2020 5G-PPP CogNet*, Deliverable D2.1, 2015.*

J.C. Cardona, R. Stanojevic, N. Laoutaris, "Collaborative Consumption for Mobile Broadband: A Quantitative Study," ACM CoNEXT*, 2014.*

J.P. Bagrow, D. Wang, A.-L. Barabasi, "Collective Response of Human Populations to Large-Scale Emergencies," PLoS ONE*, 6(3), 2011.*

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.*

D. Goergen, V. Mendiratta, R. State, T. Engel, "Identifying Abnormal Patterns in Cellular Communication Flows," ACM IPTComm*, 2013.*

A. Furno, D. Naboulsi, R. Stanica, M. Fiore, "Mobile Demand Profiling for Cellular Cognitive Networking," IEEE Transactions on Mobile Computing*, 16(3), 2017.*

C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, Z. Smoreda, "Identifying Common Periodicities in Mobile Service Demands with Spectral Analysis," MedComNet*, 2020.*

V. Soto, E. Frias-Martinez, "Automated Land Use Identification using Cell-Phone Records," ACM HotPlanet*, Washington, DC, USA, 2011.*

J.L. Toole, M. Ulm, M.C. Gonzalez, D. Bauer, "Inferring Land Use from Mobile Phone Activity," ACM UrbComp*, 2012.*

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.*

S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, C. Ratti, "Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong," Geotechnologies and the Environment*, 13, 2015.*

B. Balassa, "Trade liberalisation and "revealed" comparative advantage," The manchester school *v. 33, n. 2, p. 99-123, 1965.*

A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, Z. Smoreda, "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas," IEEE Transactions on Mobile Computing*, 16(10), 2017.*

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.*

F. Calabrese, J. Reades, C. Ratti, "Eigenplaces: Segmenting Space through Digital Signatures," IEEE Pervasive Computing*, 9(1), 2010.*

C. Spearman, "General Intelligence Objectively Determined and Measured," The American Journal of Psychology*, 15(2), 1904.*

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.*

A. Furno, M. Fiore, R. Stanica, "Joint spatial and temporal classification of mobile traffic demands," IEEE INFOCOM*, 2017.*

J. Camacho, R. Bro, D. Kotz, "Automatic Learning coupled with Interpretability: MBDA in Action," IFIP TMA*, 2020.*

S.A. Mulaik, Foundations of Factor Analysis*, CRC Press, 2009.*

K. G. Jöreskog, Psychometrika*, 43(5):443–477, 1978.*

H. H. Harman, W. H. Jones. "Factor analysis by minimizing residuals (minres)," Psychometrika *31 (3), 1966: 351-368.*

A. K. Comrey, H. B. Lee, "A first course in factor analysis,". Psychology press*, 2013.*

N.E. Briggs, R.C. MacCallum, "Recovery of weak common factors by maximum likelihood and ordinary least squares estimation," Multivariate Behavioral Research*, 38(1), 2003.*

W. Revelle, "psych: Procedures for Psychological, Psychometric, and Personality Research," R package version 2.1.3*, [online, https:// CRAN.*

R-project.org/package=psych] (last checked, June 2021).

B. Williams, A. Onsman, T. Brown, "Exploratory factor analysis: A five-step guide for novices," Australasian Journal of Paramedicine, 8(3), 2010.

D.N. Lawley, "The Estimation of Factor Loadings by the Method of Maximum Likelihood," Proc. of the Royal Society of Edinburgh, 60(1), 1940.

I.T. Jolliffe, "Principal Component Analysis and Factor Analysis," Principal Component Analysis, Springer Series in Statistics, 2002.

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.

H.F. Kaiser, J. Rice, "Little Jiffy, Mark IV," Educational and Psychological Measurement, 34, 1974.

R.B. Cattell, "The Scree Test for the Number of Factors," Multivariate Behavioral Research, 1(2), 1966.

J.L. Horn, "A Rationale and Test for the Number of Factors in Factor Analysis," Psychometrika, 30(2), 1965.

B. Thompson, "Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications," American Psychological Association, 2004.

H.F. Kaiser, "The VARIMAX Criterion for Analytic Rotation in Factor Analysis," Psychometrika, 23(3), 1958.

J.C.F. de Winter, D. Dodou, P.A. Wieringa, "Exploratory factor analysis with small sample sizes," Multivariate behavioral research, 44(2), 2009.

L.L. Thurstone, "The vectors of mind", University of Chicago Press, 1935.

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.