A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas

Angelo Furno, *Member, IEEE*, Marco Fiore, *Member, IEEE*, Razvan Stanica, Cezary Ziemlicki, and Zbigniew Smoreda

Abstract—Urban landscapes present a variety of socio-topological environments that are associated to diverse human activities. As the latter affect the way individuals connect with each other, a bound exists between the urban tissue and the mobile communication demand. In this paper, we investigate the heterogeneous patterns emerging in the mobile communication activity recorded within metropolitan regions. To that end, we introduce an original technique to identify classes of mobile traffic signatures that are distinctive of different urban fabrics. Our proposed technique outperforms previous approaches when confronted to ground-truth information, and allows characterizing the mobile demand in greater detail than that attained in the literature to date. We apply our technique to extensive real-world data collected by major mobile operators in 10 cities. Results unveil the diversity of baseline communication activities across countries, but also provide evidence of the existence of a number of mobile traffic signatures that are common to all studied areas and specific to particular land uses.

Index Terms—Mobile networks, mobile traffic data analysis, communication activity profiles, mobile traffic signatures, land use

1 Introduction

It is commonly accepted that most individuals exhibit mobility and activity patterns—driven by their family life, work obligations, hobbies, occupations and personal habits, as well as by the presence of infrastructures, services and amenities—that are highly repetitive and yet very distinctive. These considerations apply to mobile network subscribers and their communication habits as well, as highlighted in a recent, extensive survey of mobile traffic data analyses [2]. The regular but variegated behavior of mobile users leads to heterogeneity in subscribers' profiles [3], temporal periodicity of the aggregated demand [4], load fluctuations in presence of large-scale social events [5], or to geographic diversity of mobile communications [6].

In this paper, we focus on this latter aspect. Specifically, it has been shown that there exist strong relationships between the mobile communication activity and what we refer to as *urban fabrics*, i.e., the combination of infrastructure (e.g., roads, transportation systems, and sports, education, or healthcare facilities) and land use (e.g., residential, industrial, or commercial) that characterizes different zones within a same metropolitan area. Important correlations were found between the mobile demand and the underlying city cartography: notable examples include the spatial diversity of mobile activity within a conurbation [7], the similarity of temporal dynamics of traffic in residential areas [8], or the fact that load peaks undergo geographic shifts between precise urban areas throughout the day [9] and during weekday-to-weekend transitions [10]. Recently, mobile phone data was even leveraged to validate urban planning theories on conditions that promote life in a city [11].

Motivated by these results, we delve deeper in the characterization of the spatial heterogeneity of mobile communication activities. More precisely, we inherit the notion of *mobile traffic signature* to denote the typical activity pattern of the mobile demand at one specific geographic zone [12]. We find that such signatures can provide an evident association of prototypical mobile communication dynamics to precise urban fabrics. Moreover, many of these signatures appear to be general in nature, since they emerge in different cities and countries. The work leading to these conclusions yields a number of original contributions, summarized as follows.

i) We propose a novel methodology to construct mobile traffic signatures and classify them, which builds on the understanding and refinement of previous approaches. We demonstrate the superiority of our approach over state-of-the-art solutions, by showing that it creates mobile demand profiles that better agree with land use ground-truth data.

ii) We apply our proposed methodology to real-world mobile traffic data collected by major network operators in ten different cities during several months. This is a much larger dataset than those employed in previous analyses: it allows generalizing our results, and investigating similarities and diversity across a substantial set of different cities.

iii) The reference signatures we obtain characterize the geographic diversity of mobile communications with a significantly higher granularity than previously achieved in the literature. Related works only investigate five to ten major profiles of mobile traffic activity in the cities they studied; instead, we identify tens of signatures in each city. We discuss a relevant subset of signatures that characterize common dynamics

- *A. Furno is with IFSTTAR-ENTPE, Université de Lyon, Bron F-69675, France. E-mail: angelo.furno@ifsttar.fr.*
- *M. Fiore is with CNR-IEIIT, Torino 10129, Italy. E-mail: marco.fiore@ieiit.cnr.it.*
- *R. Stanica is with Univ Lyon, INSA Lyon, Inria, CITI, Villeurbanne F-69621, France. E-mail: razvan.stanica@insa-lyon.fr.*
- *C. Ziemlicki and Z. Smoreda are with SENSe, Orange Labs, Issy-les-Moulineaux F-92794, France. E-mail: {cezary.ziemlicki, zbigniew.smoreda}@orange.com.*

Manuscript received 31 May 2016; revised 26 Oct. 2016; accepted 25 Nov. 2016. Date of publication 9 Dec. 2016; date of current version 29 Aug. 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TMC.2016.2637901

typical of many city neighborhoods, as well as peculiar behaviors pinpointing specific city locations.

iv) We identify the baseline signatures of mobile traffic that can be associated to residential urban fabrics in two different countries. Our results highlight the significant dissimilarity emerging between countries in this regard.

v) We identify a number of distinctive signatures that are linked to particular urban fabrics, such as offices, universities, industrial areas, transportation hubs or leisure centers. Interestingly, we find many of such mobile demand patterns to be consistent across countries, proving that the inter-country diversity observed in residential behaviors tends to disappear when focusing on precise human activities.

2 Related Work

The dynamics of human presence in an urban area are inherent to the notion of urban fabric: a city is composed of funtionally diverse regions, and inhabitants move among them according to regular patterns, so as to carry out activities related to their destination region [13]. Using information on human presence and activity allows then telling apart the functional regions in a city, as shown by Yuan et al. [14], who exploited to that end geo-localized taxi trips.

Similarly, there exists an intertwining between urban fabrics and mobile network usage. This is an intuitive phenomenon that has been known for a long time, and which mobile operators are keen to take into account during access network planning [15], [16]. Yet, the phenomenon is far from being completely understood, since the exact influences of urban fabrics on the behavior of mobile network customers are not easily characterized. This has led to the emergence of a significant literature specifically addressing this problem.

In a seminal work, Girardin et al. [12] introduce the notion of mobile traffic signatures, i.e., condensed representations of the typical mobile demand dynamics observed in a given geographical region. They demonstrate that different urban fabrics can indeed generate diverse mobile traffic signatures: specifically, the work focuses on the mobile phone activity of roaming users, used to detect touristic areas in Rome, Italy. The approach is then elaborated by other studies. Calabrese et al. [21] use Wi-Fi associations of staff and students to map activity areas within the MIT campus. Becker et al. [17] study mobile traffic signatures in the city of Morristown, NJ, USA, and identify differences between the demand in downtown and that in high school areas. In a larger-scale study, Toole et al. [18] analyze the entire conurbation of Boston, MA, USA. The authors use information on five land use types (residential, commercial, industrial, parks, and others) as ground truth, and associate each cellular base station with one of these areas. They then build signatures for the five surfaces through a random forest classifier: this allows predicting land use from mobile traffic signatures with 57 percent accuracy. These studies all focus on specific scenarios, whereas we aim at associating accurate signatures to an exhaustive set of urban fabrics. Moreover, most of the works above make use of accurate ground truth information to train mobile traffic signatures for specific land use zones. Our proposed solution is instead unsupervised, and does not require a-priori ground truth.

Recently, an original spectral analysis approach is taken by Secchi et al. [19], who apply wavelet transforms and principal component analysis to mobile traffic signatures in the urban area of Milan, Italy. This allows drawing heatmaps of the most significant human activities in the city. This study, strongly focused on activity characterization, is complementary to ours, which provides instead a fine-grained representation of the prototypical demand observed over space. Another recent work that is orthogonal to ours is that by Lenormand et al. [20], who aim at understanding the scaling laws of (possibly mixed) land uses and at modelling them through theoretical characterization. They find interesting properties for four land uses in Spain; instead, our approach unveils a much higher variety (i.e., tens) of land uses without any inclination towards theoretical modelling.

Three works are the closest to ours. Soto et al. [22], Grauwin et al. [23] and Cici et al. [24] all propose traffic signature clustering techniques, and apply them to urban-scale scenarios. We detail their proposed methods in Sections 3.1, 3.2, and 3.3, respectively, and show that our approach outperforms them in Section 4. Moreover, our signature analysis in Section 5 builds on a much larger dataset than those considered in previous works. This allows for unprecedented detail and generality of the characterization.

3 Mobile Traffic Signatures

Let us consider a generic dataset $\mathcal{D}$, describing the communication activity of a mobile subscriber population during a set of days $\mathbf{d} = \{d\}$. For each day, the mobile demand is stored as the aggregate of the traffic generated by all users in a same area during a given time interval; the size of the area and duration of the interval determine the spatial and temporal granularity of the dataset, respectively. We name *unit area* the spatial aggregation level: the whole geographic region under consideration $\mathbf{a} = \{a\}$ is thus divided[1] into unit areas a. The time granularity is instead characterized by the duration of a *time slot*, i.e., the interval during which user activity is aggregated in each unit area. Each day $d \in \mathbf{d}$ is thus split into a set $\mathbf{t} = \{t\}$ of time slots t. Overall, $\mathcal{D} = \{v_a(d,t)\}$, where every element $v_a(d,t)$ describes the total mobile communication activity within each unit area a at time slot t of day d.

The techniques for the construction of a representative set of mobile traffic signatures process the dataset $\mathcal{D}$ through six phases. These phases aim at: (i) summarizing the mobile traffic activity in each unit areas into a meaningful profile, i.e., the unit area signature (first three phases); (ii) grouping similar unit area signatures into a limited set of classes, each exhibiting a unique behavior (last three phases). Next, we discuss each phase in detail.

1. The *signature metric* indicates the nature of subscriber activity to be represented. Examples of metrics are the number or duration of voice calls, the number of short text messages (SMS), the volume of Internet data traffic, or the kind of mobile services consumed by the users. The metric controls the actual information in each dataset entry $v_a(d,t)$.
2. The *signature support* is the time interval over which the signature is defined. Denoted as a set of days $\boldsymbol{\delta} = \{\delta\}$, the support entails the level of compression of the data into the signature. It can range from a couple of days (implying a high level of compression, since datasets typically span weeks or months) to the entire observation period, i.e., $\boldsymbol{\delta} = \mathbf{d}$ (no compression).

1. The definition of unit area is general, and can accommodate any tessellation of space. Unit areas can map to, e.g., cell sector boundaries, coverage zones of base stations, Voronoi cells, or elements of a grid.

3. The *data denoising* component extracts information deemed to be representative of the typical mobile traffic activity in a unit area, isolating it from the inherent noise in the data. In cases where the signature support is smaller than the observation period, implicit denoising is realized through compression, which increases data robustness by merging multiple $v_a(d,t)$ samples into a single value.
4. The *signature normalization* makes signatures independent from the absolute volume of mobile traffic recorded at a unit area. This allows comparing the mobile communication activity at different unit areas on the sole basis of the mobile demand variations.
5. The *signature pairwise distance measure* determines the degree of similarity of two signatures.
6. The *signature clustering algorithm* groups together signatures that are alike, leveraging the distance measure above. Ultimately, this last phase returns a set of classes of archetypal signatures, denoted as $\mathbf{c}$. Each class $c \in \mathbf{c}$ maps to a distinct type of human activity.

In Sections 3.1, 3.2, and 3.3, we will survey the current state-of-the-art definitions for mobile traffic signatures provided by Soto et al. [22], Grauwin et al. [23], and Cici et al. [24]. We will then introduce our own definition in Section 3.4.

3.1 Weekday-Weekend Signature (WWS)

In the definition by Soto et al. [22], mobile traffic signatures correspond to the average voice and text volume observed during (i) a typical working day, and (ii) a typical weekend day. Thus, in this case, the signature metric is the sum of voice and text volumes,[2] and the signature support is two days, i.e., $\boldsymbol{\delta} = \{\text{WD}, \text{WE}\}$. We will thus refer to this approach as Weekday-Weekend Signature.

Formally, the set of days $\mathbf{d}$ is split into two sets $\mathbf{d}^{\text{WD}}$ and $\mathbf{d}^{\text{WE}}$, which contain all Mondays-to-Fridays, and all Saturdays and Sundays, respectively. Then, the generic element in the signature of a unit area a is

$$s_a(\text{WD}, t) = \frac{1}{|\mathbf{d}^{\text{WD}}|} \sum_{d \in \mathbf{d}^{\text{WD}}} v_a(d,t), \quad \forall a \in \mathbf{a}, \tag{1}$$

for time slots t during working days, and

$$s_a(\text{WE}, t) = \frac{1}{|\mathbf{d}^{\text{WE}}|} \sum_{d \in \mathbf{d}^{\text{WE}}} v_a(d,t), \quad \forall a \in \mathbf{a}, \tag{2}$$

for time slots t during weekends.

We remark that this approach induces a significant level of compression, squeezing months of data into a two-day support. Thus, further data denoising is unnecessary. The signature of a is built from the elements in (1) and (2) as

$$\mathbf{s}_a = \Big\|_{\delta \in \boldsymbol{\delta}} \left(\Big\|_{t \in \mathbf{t}} s_a(\delta, t) \right), \quad \forall a \in \mathbf{a}. \tag{3}$$

In (3), $\|$ indicates the time-ordered concatenation of all elements in a set: $\mathbf{s}_a$ is thus the concatenation of all elements computed at every time slot during the average working day and the average weekend day.

Signatures then undergo a standard score normalization. To that end, each element obtained in (1) and (2) is normalized with respect to the mean and standard deviation of all elements referring to the same unit area. Formally, for a generic element of unit area a

$$\hat{s}_a(\delta, t) = \frac{s_a(\delta, t) - \mu(\mathbf{s}_a)}{\sigma(\mathbf{s}_a)}, \quad \forall \delta \in \boldsymbol{\delta}, t \in \mathbf{t}, a \in \mathbf{a}, \tag{4}$$

where $\mu(\mathbf{s}_a)$ and $\sigma(\mathbf{s}_a)$ denote the mean and standard deviation of the set of elements concatenated in the signature $\mathbf{s}_a$. Then, the normalized signature $\hat{\mathbf{s}}_a$ is simply obtained by concatenation of $\hat{s}_a(\delta, t)$ for all $\delta \in \boldsymbol{\delta}$ and $t \in \mathbf{t}$, as in (3).

As far as the similarity between signatures is concerned, WWS considers a simple Euclidean distance. Given the signatures of unit areas a and a', their distance is

$$\Delta_{a,a'} = \sqrt{\sum_{\delta \in \boldsymbol{\delta}} \sum_{t \in \mathbf{t}} \left(\hat{s}_a(\delta, t) - \hat{s}_{a'}(\delta, t)\right)^2}, \quad \forall a, a' \in \mathbf{a}. \tag{5}$$

Finally, the clustering of signatures is performed by running a k-means algorithm over the set of all signatures $\hat{\mathbf{s}}_a, \forall a \in \mathbf{a}$, using (5) as the k-means distance measure. The algorithm requires the parametrization of k, i.e., the desired number of signature classes: WWS selects k according to the validity index proposed in [25]. In all original case studies, the best results are always obtained with $k = 5$.

3.2 Typical Week Signature (TWS)

Grauwin et al. [23] propose a variation of WWS, named Typical Week Signature. Also in this case, the signature metric adds up voice and text volumes. However, the support is one week, from Monday to Sunday, i.e., $\boldsymbol{\delta} = \{\text{MON}, \text{TUE}, \text{WED}, \text{THU}, \text{FRI}, \text{SAT}, \text{SUN}\}$. Let us denote as $\mathbf{d}^{\delta} \subset \mathbf{d}$ the set of days in the dataset $\mathcal{D}$ that correspond to the day of the week δ, with $\bigcup_{\delta \in \boldsymbol{\delta}} \mathbf{d}^{\delta} = \mathbf{d}$. For instance, $\mathbf{d}^{\text{MON}}$ groups all Mondays in the dataset. Then, the generic element in the signature of unit area a is

$$s_a(\delta, t) = \mu\left(\left\{v_a(d,t) | d \in \mathbf{d}^{\delta}\right\}\right), \quad \forall a \in \mathbf{a}, \tag{6}$$

for time slots t during day δ. We recall that $\mu(\cdot)$ represents the mean of the set within parentheses. Also in this case, $\boldsymbol{\delta}$ is small with respect to the overall set of days $\mathbf{d}$, which implies high compression and makes denoising pointless.

Signatures are then obtained by concatenation of time-ordered elements, through (3). The normalization procedure is different from WWS, as each element is normalized with respect to a signature average, as

$$\hat{s}_a(\delta, t) = \frac{s_a(\delta, t)}{\mu(\{s_a(\delta, t) | \delta \in \boldsymbol{\delta}, t \in \mathbf{t}\})}, \forall \delta \in \boldsymbol{\delta}, t \in \mathbf{t}, a \in \mathbf{a}. \tag{7}$$

Normalized signatures are again clustered with a k-means algorithm using (5) as the distance measure. In this regard, the only difference from the WWS approach is that the choice of k is guided by the local maxima of the Silhouette Index [26]. This leads to a value $k = 6$ as the best choice in all considered scenarios.

3.3 Seasonal Communication Series (SCS)

The solution by Cici et al. [24], named Seasonal Communication Series, considers the whole timeserie in each unit

2. As shown throughout this section, all proposed definitions of mobile traffic signatures leverage voice and text activity, while they discard data traffic. The reason is that the former are an excellent proxy of human endeavors—and thus of the urban fabrics that affect them. Instead, data traffic is often generated autonomously by applications running or updating in background, and it is thus less representative of the actual occupations of the user.

area. In other words, $\boldsymbol{\delta} = \mathbf{d}$, and the signature of area a is

$$\mathbf{s}_a = \mathop{\Big\|}_{d \in \mathbf{d}} \left(\mathop{\Big\|}_{t \in \mathbf{t}} s_a(d,t) \right), \quad \forall a \in \mathbf{a}, \tag{8}$$

where $s_a(d,t) = v_a(d,t), \forall a \in \mathbf{a}, d \in \mathbf{d}, t \in \mathbf{t}$, and $v_a(d,t)$ corresponds to the volume of voice call and text messages.

In such a definition, the number of elements that compose a signature is not fixed, but depends on the timespan of the dataset $\mathcal{D}$. Also, (8) does not involve any compression, which calls for denoising: to that end, SCS applies a Fast Fourier Transform (FFT) to the signature, so as to clean it from irregular patterns. More precisely, once converted to the frequency domain with FFT, only the highest power frequencies are kept, and the time signal is reconstructed with an Inverse FFT (IFFT) from the retained frequencies. The filtering returns a so-called seasonal (i.e., typical) component of the original signature.

Normalization of whole-time series filtered signatures is then performed using the standard-score approach in (4), where, clearly, $\boldsymbol{\delta} = \mathbf{d}$.

The pairwise signature similarity is based on the Pearson correlation coefficient, which, for two unit areas a and a', is

$$C_{a,a'} = \frac{\sum_{\delta \in \boldsymbol{\delta}} \sum_{t \in \mathbf{t}} (\hat{s}_a(\delta,t) - \mu(\hat{\mathbf{s}}_a))(\hat{s}_{a'}(\delta,t) - \mu(\hat{\mathbf{s}}_{a'}))}{\sqrt{\sum_{\delta \in \boldsymbol{\delta}} \sum_{t \in \mathbf{t}} (\hat{s}_a(\delta,t) - \mu(\hat{\mathbf{s}}_a))^2} \cdot \sqrt{\sum_{\delta \in \boldsymbol{\delta}} \sum_{t \in \mathbf{t}} (\hat{s}_{a'}(\delta,t) - \mu(\hat{\mathbf{s}}_{a'}))^2}}. \tag{9}$$

We recall that $\boldsymbol{\delta} = \mathbf{d}$ in (9). The distance measure is then

$$\Delta_{a,a'} = 1 - C_{a,a'}, \quad \forall a, a' \in \mathbf{a}. \tag{10}$$

Concerning signature clustering, SCS adopts an agglomerative hierarchical clustering, namely, the linkage clustering algorithm with average distance criterion. This hierarchical clustering outputs a whole family of solutions that can be represented as a dendrogram: it thus returns a richer information than a single-cluster set solution, as in the case of, e.g., k-means. However, this also implies that some criterion must be adopted to select the best clustering among all those in the family. To that end, the skewness of the cluster sizes is evaluated at the different levels of the dendrogram built by the hierarchical clustering: selecting the level with minimum skewness allows grouping unit area signatures into classes of relatively comparable sizes. It is important to note that, by using the lowest-skewness criterion, the number of generated signature classes can be high, in the order of hundreds. Since this makes the analysis cumbersome, SCS limits the analysis to the 10 largest classes, which they consider to represent the most relevant urban fabrics in the considered region.

3.4 Median Week Signature (MWS)

The current approaches presented above are based on a variety of signature definitions, pairwise distance measures and clustering approaches. Here, we introduce a novel signature model that aims at combining the advantages of previous proposals, while overcoming their limitations. Specifically, our definition of a *Median Week Signature* is based on the following considerations.

- First, it has been repeatedly shown that there exists a strong weekly periodicity in human occupations [27], [28], which implies that most of the diversity in mobile traffic activity occurs within a one-week period. We thus speculate that a signature describing the typical weekly behavior of the mobile demand at one unit area contains the vast majority of the significant information about the nature of that area. This lets us consider a week-long signature, avoiding dimensionality problems in presence of long time series (which can instead affect SCS [29]), and not discarding any important knowledge (an issue in highly-compressed WWS [24]).
- Second, we deem the median to be a more reliable statistical measure than, e.g., the average or the absolute values, when it comes to assessing the typical activity in mobile traffic. As a matter of fact, the median is much more robust to outliers, which are frequent in mobile traffic due to special events of social, political, sports, or cultural nature [4], [7].
- Third, understanding (i) whether denoising is beneficial to the signature definition, (ii) which normalization works the best, and (iii) how signature distance measures affect the results is not trivial. A sensible choice needs substantial empirical tests on representative data.

The MWS is computed according to the guidelines above, as follows. The metric is the sum of voice and text activity volumes, as assumed by all techniques in the literature. The support is the same considered in TWS, i.e., $\boldsymbol{\delta} = \{\text{MON}, \text{TUE}, \text{WED}, \text{THU}, \text{FRI}, \text{SAT}, \text{SUN}\}$. By using the same notation, the element associated to time slot t of day $\delta \in \boldsymbol{\delta}$ in the signature of unit area a is

$$s_a(\delta,t) = \mu_{1/2}\big(\{v_a(d,t) | d \in \mathbf{d}^\delta\}\big), \quad \forall a \in \mathbf{a}, \tag{11}$$

where, $\mu_{1/2}(\cdot)$ represents the median of the set within parenthesis. The MWS is then defined as the concatenation of time-ordered samples according to (3). We remark that this defintion realizes our first two considerations above.

Taking as a pivot the MWS model just defined, we explore the design space of a complete solution for urban fabric detection. This implements the last consideration listed before, and results in the MWS variants below.

1. We assess the impact of denoising, considering both the case where the signature is filtered via the FFT/IFFT procedure proposed in [24] and described in Section 3.3, and the case where it is used as is.
2. We evaluate two different techniques to normalize MWS. One option is the standard score normalization introduced above; in this case, signatures are normalized according to (4), where $\boldsymbol{\delta} = \{\text{MON}, \text{TUE}, \text{WED}, \text{THU}, \text{FRI}, \text{SAT}, \text{SUN}\}$. The other option is daily normalization, where the signature element of unit area a at time slot t of day δ is

$$\hat{s}_a(\delta,t) = \frac{s_a(\delta,t)}{\sum_{t \in \mathbf{t}} s_a(\delta,t)}, \quad \forall a \in \mathbf{a}. \tag{12}$$

   The expression in (12) normalizes each element with respect to the total activity during the weekday the element belongs to.
3. We test the distance measures used in WWS, TWS and SCS, i.e., the Euclidean distance in (5) and the distance based on the Pearson correlation coefficient in (10); in both cases $\boldsymbol{\delta} = \{\text{MON}, \text{TUE}, \text{WED}, \text{THU}, \text{FRI}, \text{SAT}, \text{SUN}\}$.

Finally, signature clustering is performed as in SCS, using the agglomerative hierarchical algorithm described in

TABLE 1
Summary of the Considered Techniques for the Detection of Classes of Mobile Traffic Signatures

Name	Signature	Filtering	Normalization	Distance	Clustering
WWS	average weekday-weekend	–	standard score	Euclidean	k-means, $k = 5$
TWS	average week	–	average rescaling	Euclidean	k-means, $k = 6$
SCS	whole time series	FFT/IFFT	standard score	Pearson correlation	linkage, minimum skewness
MWS-stdscr-pearson	median week	–	standard score	Pearson correlation	linkage, minimum skewness
MWS-stdscr-euclidean	median week	–	standard score	Euclidean	linkage, minimum skewness
MWS-daily-pearson	median week	–	daily	Pearson correlation	linkage, minimum skewness
MWS-daily-euclidean	median week	–	daily	Euclidean	linkage, minimum skewness
MWS-fft-stdscr-pearson	median week	FFT/IFFT	standard score	Pearson correlation	linkage, minimum skewness
MWS-fft-daily-euclidean	median week	FFT/IFFT	daily	Euclidean	linkage, minimum skewness

Section 3.3, and considering minimum skewness as the stopping rule. We favor this approach over simpler ones based on k-means, since it is fully unsupervised. Also, unlike SCS, our MWS approach does not limit relevant classes to an arbitrary number, but considers that all signatures covey some unique behavior of mobile network subscribers and thus deserve to be studied and understood.[3]

4 Comparative Evaluation

In this section, we provide a complete comparative evaluation of the techniques for signature classification described in Section 3. A summary of the different solutions we test is provided in Table 1. In order to carry out our study, we gather mobile traffic as well as ground-truth data in two major cities in Italy, as detailed in Section 4.1, and we define a set of relevant metrics, presented in Section 4.2. We then discuss the results of our study in Section 4.3.

4.1 Comparative Evaluation Datasets

We consider two citywide case studies. The mobile traffic data is provided in both scenarios by Telecom Italia Mobile (TIM), as part of their Big Data Challenge initiative [30]. The ground-truth information consists instead in land use data retrieved from open databases of local authorities. As a matter of fact, we assess the quality of signature classes identified via each technique by verifying their congruence with the nature of the underlying urban fabrics. This methodology is consistent with those employed in the literature [24], and stems from the expectation that human activities, including mobile communications, are strongly related to the type of city facilities around them. A detailed description of the data follows.

4.1.1 Milan

The first urban scenario is that of Milan, Italy. The mobile traffic dataset is referred as *Mi-13*, and describes the communication activity of TIM subscribers in the conurbation of the city for a two-month period in November and December 2013. The dataset differentiates among incoming and outgoing calls, providing information about their number and duration. The dataset also contains the number of received and sent SMS and amount of Internet data traffic generated by TIM mobile devices.

Mobile traffic information is aggregated over 10-minute time intervals, according to a regular-cell spatial tessellation of the surface of the city of Milan. Each cell has a 235×235 m^2 size, i.e., an area of 0.055 Km^2, and maps to a unit area in our analysis. In our study, we consider a region of approximately 150 Km^2 containing 2,726 cells. The *Mi-13* dataset is the same used in [24] for the evaluation of the SCS approach with respect to WWS. For the sake of fairness in the comparison, here we focus on the same subset of cell-phone activity as in [24]: a 4-week period ranging from November 4, 2013 to December 1, 2013.

As far as ground-truth information is concerned, we leverage the same data used in [24], which was retrieved from publicly available databases [31]. The data conveys information on urban infrastructures and land use that can be associated to different kinds of human activities. Specifically, it reports, for each unit area in Milan, the number of local inhabitants, business activities, sport centers, universities, schools and bus stops, as well as the percentage of unit area that is covered by green spaces, such as parks or woods. Fig. 1a displays a map of Milan unit areas with markers pointing out business, universities, and green spaces.

4.1.2 Turin

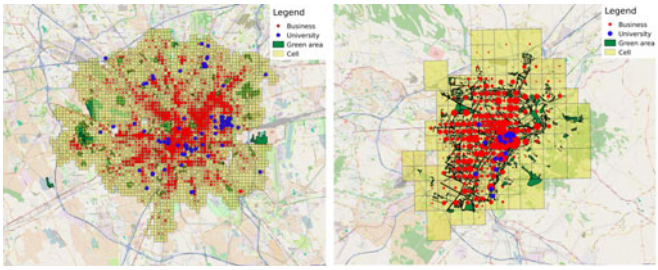
The second scenario we consider is that of Turin, Italy. Mobile traffic information in the region is provided by a dataset, referred to as *Tu-15*, which describes the mobile traffic activity of TIM customers in the region during March and April 2015. The spatial tessellation of the geographical surface provided by the network operator is different from that in *Mi-13*; cells, i.e., unit areas, are not regular, but feature heterogeneous sizes that mimic the non-uniform coverage provided by each base station in the region. Overall, the data refers to an area of approximately 150 Km^2 containing 261 unit areas whose size ranges from 255×325 m^2 to 2×2.5 Km^2. For the sake of consistency with the Milan case study, we limit our analysis to four weeks of data, from the March 1 to March 28, 2015.

We collected ground-truth data for each unit area in *Tu-15* by leveraging open data published by the local municipality [32]. We selected information related to the latitude-longitude coordinates of schools, universities and business activities, and we associated them to individual unit areas, exactly as done for Milan in [24]. We also leveraged open data on green zones and population distribution in order to determine their presence in each unit area. Fig. 1b shows a map of the resulting unit areas for the city of Turin, together with a representation of ground-truth data of universities, business activities and green zones.

4.2 Metrics

In order to evaluate the consistency of the signature classes with respect to the ground-truth data, we introduce a set of

3. Code and supporting material are available at the project website, http://mobile-traffic-analysis.project.citi-lab.fr/



(a) Milan (b) Turin

Fig. 1. Spatial tessellation into unit areas (*cells* in the legend), and partial ground-truth data for the (a) Milan and (b) Turin citywide scenarios. Figure best viewed in colors.

suitable metrics, presented next. We remark that, for the sake of comparability of our study with previous research, we include in our list the metrics proposed in [24].

4.2.1 Density

The *density* $D_{\mathcal{G}}(\mathbf{c}, c)$ is a measure of the frequency of ground-truth elements of type $\mathcal{G}$ within a signature class $c \in \mathbf{c}$. Let us define as $\mathbf{k}_{\mathcal{G}}$ the set of elements of type $\mathcal{G}$ (e.g., the set of universities) in the ground-truth data; also, $\mathbb{1}_c(k)$ is an indicator function, equal to one if a ground-truth element $k \in \mathbf{k}_{\mathcal{G}}$ is located in a unit area whose signature matches class c, and zero otherwise. The density is then

$$D_{\mathcal{G}}(\mathbf{c}, c) = \frac{1}{|c|} \sum_{k \in \mathbf{k}_{\mathcal{G}}} \mathbb{1}_c(k), \tag{13}$$

where $|c|$ denotes the cardinality of the class $c \in \mathbf{c}$, i.e., the number of signatures of individual unit areas that it includes. The density allows quantifying the prevalence of some land use type $\mathcal{G}$ within each signature class $c \in \mathbf{c}$.

4.2.2 Entropy

The *entropy* $H_{\mathcal{G}}(\mathbf{c})$ associated to ground-truth elements of type $\mathcal{G}$ for a given signature classification $\mathbf{c}$ estimates the dispersion of $\mathcal{G}$ across the classes in $\mathbf{c}$. It is defined as [33]

$$H_{\mathcal{G}}(\mathbf{c}) = -\sum_{c \in \mathbf{c}} P_{\mathcal{G}}(c) \log P_{\mathcal{G}}(c). \tag{14}$$

In (14), $P_{\mathcal{G}}(c)$ is the probability that a ground-truth element of type $\mathcal{G}$ falls into a unit area whose signature matches class c, i.e.,

$$P_{\mathcal{G}}(c) = \frac{1}{|\mathbf{k}_{\mathcal{G}}|} \sum_{k \in \mathbf{k}_{\mathcal{G}}} \mathbb{1}_c(k). \tag{15}$$

Lower entropy is thus an indicator of a less random, i.e., more precise, assignment of land use data of a given type to classes defined by the considered technique.

4.2.3 Coverage

The *coverage* $C_{\mathcal{G}}(\mathbf{c})$ of ground-truth elements of type $\mathcal{G}$ for a given signature classification $\mathbf{c}$ is the percentage of such elements included within those classes of $\mathbf{c}$ that are the most relevant to $\mathcal{G}$. Specifically, let us define a subset of signature classes $\mathbf{c}_{\mathcal{G}} \subseteq \mathbf{c}$ that have higher-than-average density for ground-truth data type $\mathcal{G}$, i.e., $\mathbf{c}_{\mathcal{G}} = \{c \in \mathbf{c}\ s.t.\ D_{\mathcal{G}}(\mathbf{c}, c) > |\mathbf{k}_{\mathcal{G}}| / \sum_{c \in \mathbf{c}} |c|\}$. Then, coverage is

$$C_{\mathcal{G}}(\mathbf{c}) = \sum_{c \in \mathbf{c}_{\mathcal{G}}} P_{\mathcal{G}}(c). \tag{16}$$

High coverage indicates that land use type $\mathcal{G}$ is well encompassed by unit areas that feature a limited set of highly-related mobile traffic signatures.

4.2.4 F-Score

The *F-score* provides a single value that summarizes the quality of the signature classes. It combines the entropy and coverage computed for each ground-truth type $\mathcal{G}$ into the harmonic mean

$$F_{\mathcal{G}}(\mathbf{c}) = 2 \cdot \frac{(1 - \hat{H}_{\mathcal{G}}(\mathbf{c})) \cdot C_{\mathcal{G}}(\mathbf{c})}{(1 - \hat{H}_{\mathcal{G}}(\mathbf{c})) + C_{\mathcal{G}}(\mathbf{c})}, \tag{17}$$

where $\hat{H}_{\mathcal{G}}(\mathbf{c}) = \frac{H_{\mathcal{G}}(\mathbf{c})}{\log(|\mathbf{c}|)}$ is the normalized entropy. The F-score index ranges in $[0, 1]$, with 1 indicating the best performance, i.e., minimum entropy and maximum coverage.

4.3 Results

We begin by comparing the first five techniques in Table 1. Figs. 2 and 3 show the entropy, coverage and F-score for the state-of-the-art approaches of *WWS*, *TWS* and *SCS* as well as for our proposed *MWS-stdscr-euclidean* and *MWS-stdscr-pearson* solutions. Histogram bars refer to different ground-truth types, and the two figures refer to the case studies of Milan and Turin, respectively.

As shown in Fig. 2a, regardless of the signature pairwise distance measure used (Euclidean or based on Pearson correlation coefficient), the solutions based on the MWS model attain a significantly lower entropy than that granted by previous approaches. As mentioned in Section 4.2, this indicates a reduced level of randomness, and thus a more precise taxonomy of unit areas with respect to the ground-truth data through signature classes. Also, the increased accuracy does not come at a cost in terms of coverage, as shown in Fig. 2b. In fact, the entropy gain granted by MWS is associated to an increase in coverage: this proves the higher effectiveness of a median-week representation of mobile traffic signatures.

These results are summarized in Fig. 2c, depicting the overall F-score. The plot outlines how our MWS solution improves current state-of-the-art techniques, i.e., it generates classes of mobile traffic signatures that better relate to the land use ground-truth data. The gain in terms of F-score ranges between 10 and 15 percent over *SCS*, and reaches typical values well above 100 percent when considering *WWS* or *TWS*.

The conclusions above are not specific to the Milan case study, as Fig. 3 shows that they hold also for the Turin scenario. The most notable difference emerging when comparing the two sets of plots is observed in Fig. 3a: there, *WWS* outperforms all other approaches in terms of lowest entropy. However, this is just an artifact of the lower number of signature classes returned by this solution, which entails higher heterogeneity in the probability distribution in (15). When entropy values are normalized with respect to the number of clusters, as done within the F-score in (17), MWS still emerges as a clear winner.

In addition to demonstrating the superiority of our MWS technique, Figs. 2 and 3 allow commenting on the relative performance of *MWS-stdscr-euclidean* and *MWS-stdscr-pearson*. Both solutions are based on MWS, but leverage different signature pairwise distance measures. The results show that the distance measure based on Pearson correlation

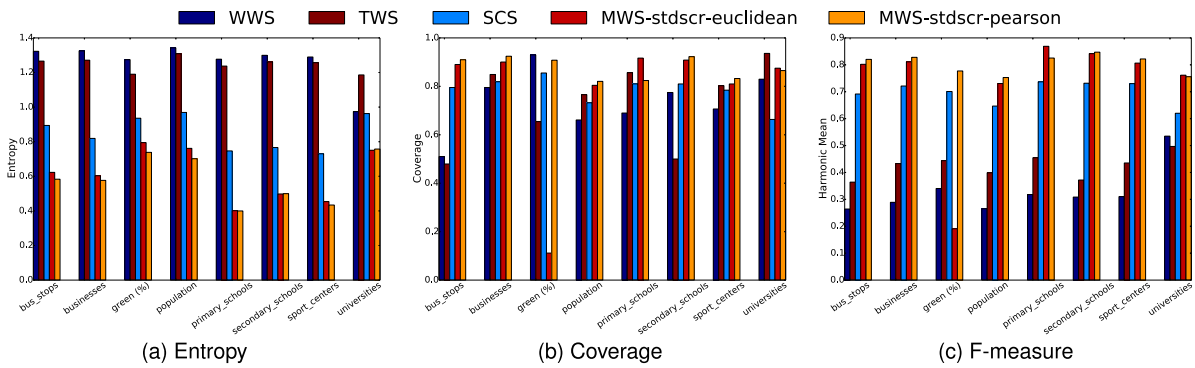


Fig. 2. Milan. Performance comparison among *WWS*, *TWS*, *SCS* (lowest skewness at 92 clusters), *MWS-stdscr-euclidean* (lowest skewness at 77 clusters), and *MWS-stdscr-pearson* (lowest skewness at 63 clusters).

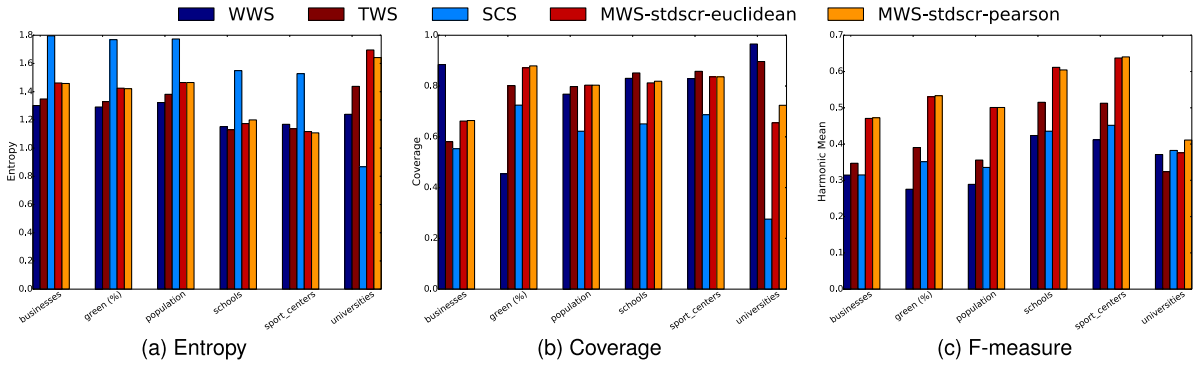


Fig. 3. Turin. Performance comparison among *WWS*, *TWS*, *SCS* (lowest skewness at 60 clusters), *MWS-stdscr-euclidean* (lowest skewness at 33 clusters), and *MWS-stdscr-pearson* (lowest skewness at 33 clusters).

coefficient achieves slightly better and more stable performance (see, e.g., green zones in the Milan scenario).

Wefurther explore the design space of MWS-based solutions by testing the impact of diverse normalization approach. Figs. 4a and 4b show the performance of *WWS*, *TWS*, and *SCS* against those attained by the MWS model combined with daily normalization instead of the standard score used before, when using both distance measures. The two figures refer to the Milan and Turin scenarios, respectively. For the sake of brevity, we limit results to the F-score, which is a more comprehensive metric according to our previous analysis. We remark that MWS-based solutions still tend to outperform techniques based on other signature models. In this case, however, slightly better performance is achieved by the *MWS-daily-euclidean* approach, and thus euclidean distance appears to work better in combination with a daily normalization. Again, results are consistent through different urban scenarios.

An ultimate comparison between the schemes providing the best performance in the previous tests is shown in Fig. 5. The plots portray the F-score attained by *MWS-daily-euclidean* and *MWS-stdscr-pearson*, as well as by *SCS* as the top-scoring representative of current state-of-the-art signature classification techniques. When confronting the three solutions above, *MWS-stdscr-pearson* emerges as the preferred approach, although the difference with respect to the other MWS-based technique is not dramatic.

Fig. 5 also includes two additional versions of MWS. These are based on *MWS-daily-euclidean* and *MWS-stdscr-pearson* but also include a further step, i.e., data denoising via FFT/IFFT as proposed in [24]. The results for these versions, denoted as *MWS-fft-daily-euclidean* and *MWS-fft-stdscr-pearson*, show that using such a filter on a median-week signature not only does not improve performance, but instead degrades it. This suggests that the median week compression already provides a sufficiently denoised representation of the typical mobile traffic observed at a unit area: further attempts at removing noise only risk to disrupt the embedded information.

In conclusion, our comparative performance evaluation highlights *MWS-stdscr-pearson* as the technique that produces the signature classes that best match the underlying urban fabrics. We believe that all components of our proposed technique are instrumental to its effectiveness. This thesis is supported by the fact that competitor solutions use portions of *MWS-stdscr-pearson* (i.e., *TWS*, consider week-long aggregation, and *SCS* uses agglomerative hierarchical clustering), yet *MWS-stdscr* outperforms them all. In all cases, the result is a proxy of the capability of *MWS-stdscr-pearson* to separate the mobile communication activities

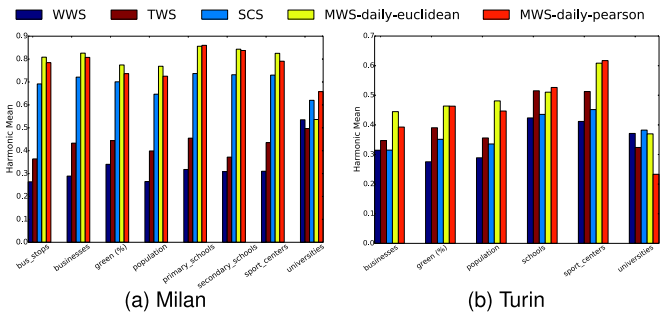


Fig. 4. F-score comparison among *WWS*, *TWS*, *SCS* (lowest skewness at 92 and 60 clusters), *MWS-daily-euclidean* (lowest skewness at 103 and 80 clusters), and *MWS-daily-pearson* (lowest skewness at 120 and 60 clusters) in the (a) Milan and (b) Turin scenarios.

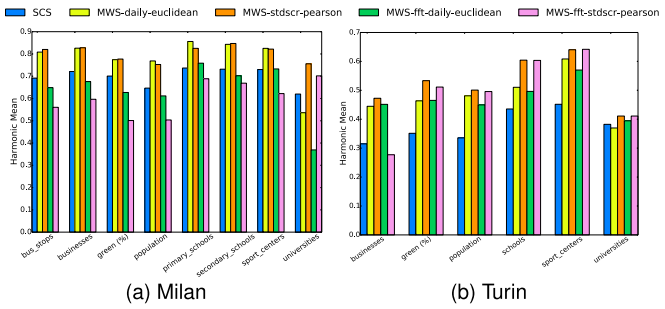


Fig. 5. F-score comparison among *SCS*, (lowest skewness at 92 and 60 clusters), *MWS-daily-euclidean* (lowest skewness at 103 and 80 clusters), *MWS-stdscr-pearson* (lowest skewness at 63 and 33 clusters), *MWS-fft-daily-euclidean* (lowest skewness at 120 and 100 clusters), and *MWS-fft-stdscr-pearson* (lowest skewness at 110 and 29 clusters) in the (a) Milan and (b) Turin scenarios.

recorded at each unit area in a more convenient manner than equivalent solutions proposed in the literature.

5 Signature Analysis

We leverage the *MWS-stdscr-pearson* technique to extract meaningful classes of mobile traffic signatures in a substantial set of urban scenarios in Italy and France. Such a study allows characterizing mobile communication dynamics and their intertwining with the urban landscape with high accuracy, across diverse cities and countries. To that end, we first introduce in Section 5.1 the mobile traffic datasets we employ in our study. An overview of the signature classes returned by *MWS-stdscr-pearson* is provided Section 5.2. Then, Sections 5.3, 5.4, 5.5, and 5.6 focus on a subset of such classes, especially interesting through their repeated occurrence or, on the contrary, their peculiarity.

5.1 Signature Analysis Datasets

Our datasets describe the mobile communication activity recorded in four major cities in Italy i.e., Milan, Turin, Rome and Trento, as well as in six major cities in France, i.e., Paris, Lyon, Marseille, Toulouse, Lille and Bordeaux. For the specific case of Milan, we consider two separate datasets, related to two different time periods. This sums up to ten urban scenarios and eleven datasets. Table 2 labels the datasets and summarizes their main features.

In all case studies, we consider a geographic region of 150 Km^2 around the city center. There, we collect time series

TABLE 2
Labels and Details for the Reference Mobile Traffic Datasets

Label	Source dataset	City	Unit Areas	Period
Mi-13	TIM 2014	Milan	2,763 cell grids	Nov./Dec. 2013
Tn-13	TIM 2014	Trento	152 cell grids	Nov./Dec. 2013
Mi-15	TIM 2015	Milan	434 cell grids	Mar./Apr. 2015
Rm-15	TIM 2015	Rome	341 cell grids	Mar./Apr. 2015
Tu-15	TIM 2015	Turin	257 cell grids	Mar./Apr. 2015
Pa-15	Orange	Paris	1,634 base stations	Aug.–Nov. 2014, Mar. 2015
Ly-15	Orange	Lyon	278 base stations	Aug.–Nov. 2014, Mar. 2015
Ma-15	Orange	Marseille	188 base stations	Aug.–Nov. 2014, Mar. 2015
To-15	Orange	Toulouse	220 base stations	Aug.–Nov. 2014, Mar. 2015
Li-15	Orange	Lille	156 base stations	Aug.–Nov. 2014, Mar. 2015
Bo-15	Orange	Bordeaux	158 base stations	Aug.–Nov. 2014, Mar. 2015

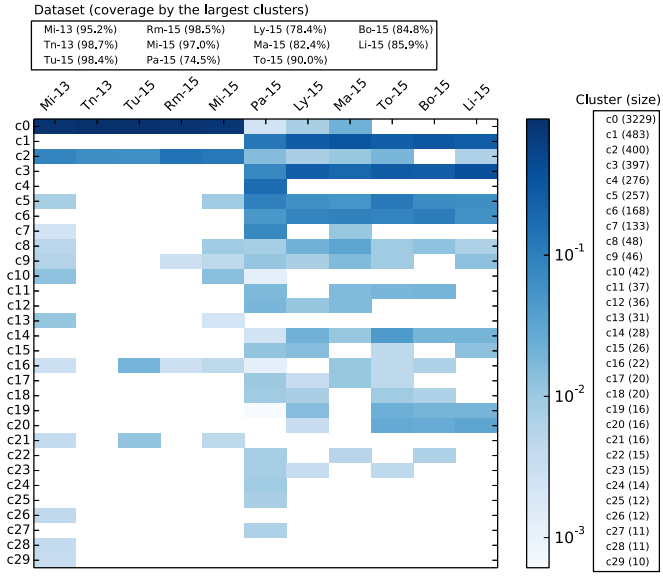


Fig. 6. Prevalence of signature classes in the reference urban regions.

of the mobile traffic demand generated by subscribers of major operators, i.e., Orange in France and TIM in Italy.

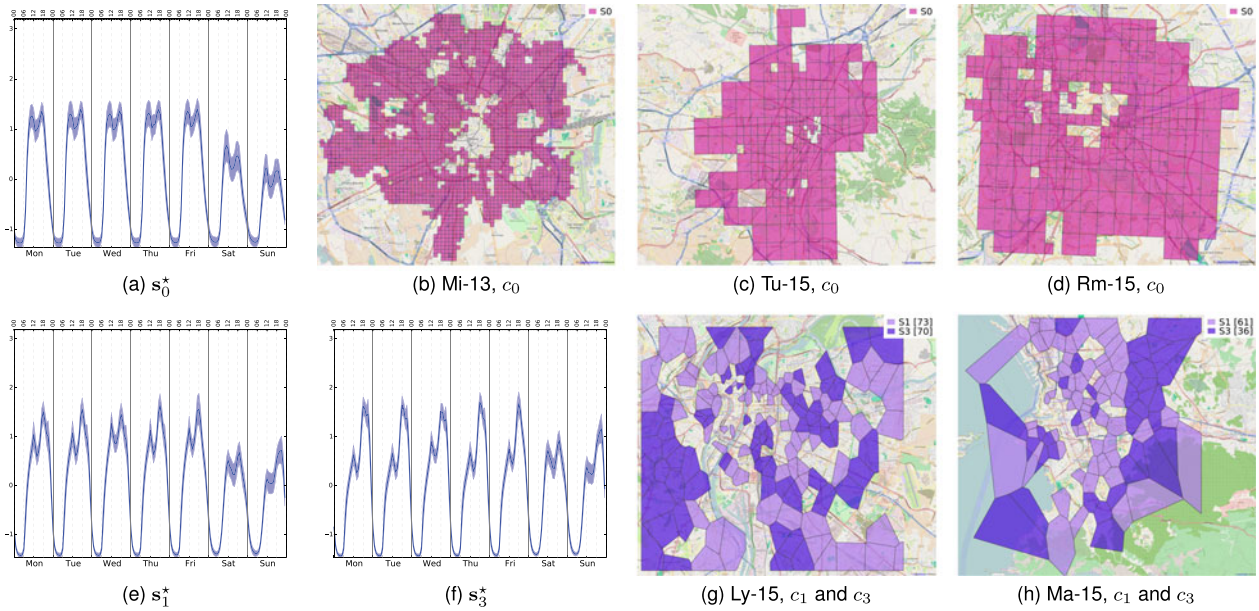
The source data provided by Orange consists of Call Detail Records (CDR) describing hourly volumes of voice and text message activity in the whole France, on a per-antenna basis. In French urban scenarios, our unit areas map the coverage zones of the mobile network antennas, which are approximated as the cells of a Voronoi tessellation. Time slots span one hour, as this is the maximum precision granted by the data. The communication activity in the data covers the period from August 12 to November 30, 2014, and from March 3 to March 25, 2015, for a total of 132 days.

The TIM datasets are the same we used for the comparative evaluation, described in Section 4.1. For the sake of consistency, we formatted the data so that it conforms to that provided by Orange in terms of temporal granularity, i.e., we aggregated traffic into 1-hour time slots.

5.2 Overview of Signature Classes

We use the *MWS-stdscr-pearson* methodology to determine mobile traffic signature classes for the 6,581 unit areas that cover the urban regions in the reference datasets of Table 2. Overall, a set of 514 classes is identified across all cities. These numbers underscore how the scale of our study, encompassing ten cities and hundreds of signature classes, is significantly larger than that of previous works, focusing on one to three cities and five to ten signatures.

Fig. 6 provides an overview of the signature classes returned by our methodology. The plot shows how classes (rows) are distributed across cities (columns). Colors map to the percentage of city unit areas belonging to a specific class: the darker the color, the more dominant the class within the urban region (see the color range on the right of the plot for the precise value). For the sake of clarity, we limit the plot to the largest 30 classes, i.e., those including at least 10 unit area signatures (see the cardinality of each class on the right listing). These classes account for 75 to 98 percent of the total surface in each city (see the percentage of city surface covered by the 30 largest classes in the top listing). However, in our discussion of the results, we will also present smaller classes that correspond to peculiar communication dynamics emerging in specific unit areas.



(a) $\mathbf{s}_0^\star$ (b) Mi-13, c_0 (c) Tu-15, c_0 (d) Rm-15, c_0

(e) $\mathbf{s}_1^\star$ (f) $\mathbf{s}_3^\star$ (g) Ly-15, c_1 and c_3 (h) Ma-15, c_1 and c_3

Fig. 7. Residential urban fabrics. Characteristic signatures (with standard deviation) and maps of related unit areas in representative city scenarios.

Some preliminary considerations on the signature classes are in order: i) class c_0 covers most of the Italian cities, while its presence in France is almost negligible; ii) the majority of the analyzed French cities are mainly covered by classes c_1 and c_3, which do not include instead any area of the Italian cities; iii) some classes, such as c_2, c_8, c_9 and c_{16}, appear in almost all reference cities, independently of the country; iv) other classes, e.g., c_4 and c_{10}, are very city-specific; v) Paris displays the highest heterogeneity of classes, and only 74.5 percent of its surface is covered by the 30 largest classes (the minimum percentage in all scenarios); vi) Trento and Rome show the least diversity, as the signatures of their unit areas almost exclusively end in classes c_0 and c_2, with a 98.7 percent and a 98.5 percent coverage by the 30 largest classes, respectively.

In the remainder of this section, we will explore the causes behind the classification features outlined above, and more. To that end, we will leverage the notion of a *characteristic signature* for each signature class. The characteristic signature provides an immediate intuition of the mobile communication dynamics in the urban areas whose signatures are classified together. Formally, for the ith class $c_i \in \mathbf{c}$, the characteristic signature $\mathbf{s}_i^\star$ is obtained as the time-ordered concatenation in (3) of all elements

$$s_i^\star(\delta, t) = \frac{1}{|\, c_i \,|} \sum_{a \in c_i} \hat{s}_a(\delta, t), \quad \forall \delta \in \boldsymbol{\delta}, t \in \mathbf{t}, \tag{18}$$

where $|\, c_i \,|$ represents the cardinality of class c_i, and $\hat{s}_a(\delta, t)$ is an element of the MWS signature in (11) normalized via the standard score in (4). It can thus be interpreted as an average of all *MWS-stdscr-pearson* signatures contributing to a specific class $c_i \in \mathbf{c}$.

Before proceeding, we would like to remark that unit areas in French cities appear to be more distributed across signature classes in Fig. 6. The consistency of this behavior lets us speculate that the preprocessing enforced on the TIM datasets may have induced important information loss, flattening the diversity of mobile traffic activity. Specifically, the source data for Italian cities has a spatial granularity that can be quite coarse in some cases: Milan grid cells (i.e., unit areas) yield the highest resolution, and the fact that the *Mi-13* and *Mi-15* datasets rank as the most heterogeneous among Italian ones corroborates our conjecture that the spatial discretization introduces some bias in the data. Still, as we will see, this does not prevent the identification of meaningful characteristic signatures in Italian cities as well.

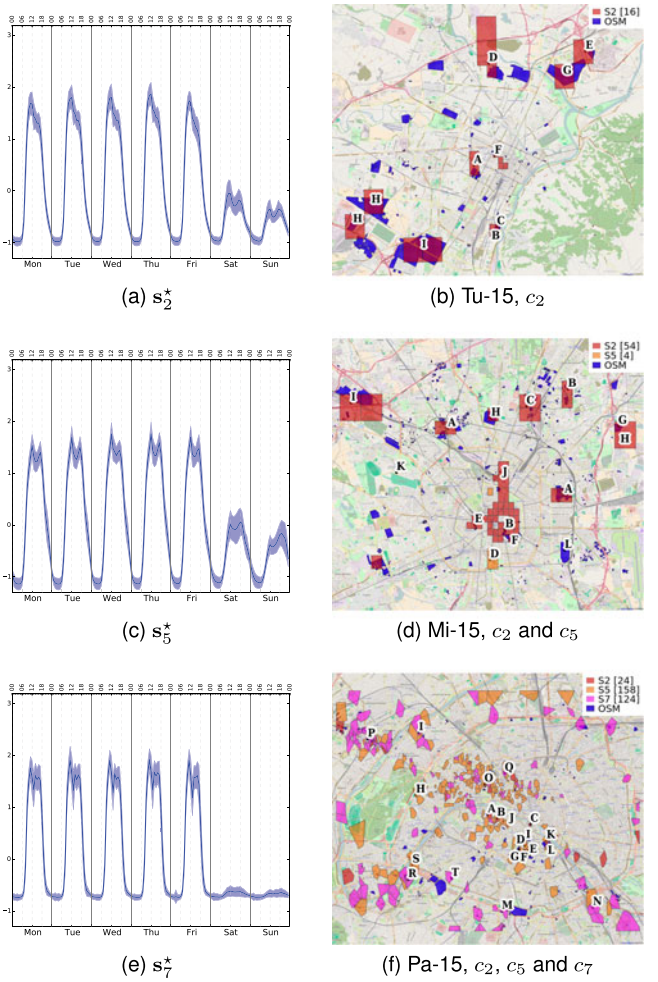
5.3 Residential Urban Fabrics

We start our analysis by studying the signature classes that appear the most frequently in the reference urban regions.

The characteristic signature $\mathbf{s}_0^\star$ of class c_0 is portrayed in Fig. 7a. This class characterizes all unit areas of the analyzed Italian cities that do not present any noticeable infrastructure and that do not draw any particular activity of inhabitants. This is outlined, e.g., in Figs. 7b, 7c, and 7d, which show the extent of unit areas in Milan, Turin and Rome whose signatures are in class c_0. The corresponding regions include suburban and mainly residential areas, and exclude city centers and popular points of interest (PoIs). Class c_0 can be thus associated to *residential urban fabrics in Italy*, which are denoted by a mixture of private housing and small business activity. It is thus representative of the baseline mobile traffic demand observed within urban regions in Italy.

By looking at the time series in Fig. 7a, we remark two comparable traffic peaks, at 11:00 and 17:00, repeating on all working days. The mobile activity in most urban areas in Italy is reduced during weekend, when the morning peak becomes dominant over the afternoon one, which is also shifted towards later hours.

A similar discussion holds in the case of signatures classes c_1 and c_3, this time for French cities. These classes designate *residential urban fabrics in France*, as exemplified by the geographic coverage of the associated unit areas in Lyon and Marseille, shown in Figs. 7g and 7h, respectively. The inspection of $\mathbf{s}_1^\star$ and $\mathbf{s}_3^\star$ shapes, in Figs. 7e and 7f, respectively, reveals significant similarities in the semantics of the two signatures. Both feature two traffic peaks, the afternoon one standing over the morning one; the activity during weekends is comparable in the two cases, just scaled up in $\mathbf{s}_3^\star$. The main difference between $\mathbf{s}_1^\star$ and $\mathbf{s}_3^\star$ appears thus to be the afternoon-to-morning peak ratio, higher in the latter. We conclude that both c_1



(a) $\mathbf{s}_2^\star$ (b) Tu-15, c_2

(c) $\mathbf{s}_5^\star$ (d) Mi-15, c_2 and c_5

(e) $\mathbf{s}_7^\star$ (f) Pa-15, c_2, c_5 and c_7

Fig. 8. Office fabric signatures $\mathbf{s}_2^\star$, $\mathbf{s}_5^\star$ and $\mathbf{s}_7^\star$ and maps of the related unit areas in Italian and French cities, with OpenStreetMap data.

and c_3 are representative of residential and small business areas in France, although c_3 is associated to a higher concentration of residential land use than c_1: indeed, the darker unit areas in Figs. 7g and 7h, mapping to c_3, are more present in the urban outskirts and less so in city centers.

It is also interesting to compare c_0, c_1 and c_3. The differences between the baseline profiles of the mobile traffic demand in Italy and France are striking. Activity peaks are uneven and shifted ahead of around one hour in France; their ratio is even reversed during weekends. This diversity is imputable to different routines in the two countries, and entails interesting sociological questions.

5.4 Office Urban Fabrics

Unit areas with signatures matching class c_2 are extensively present in both Italy and France, as shown in Fig. 6. In order to understand the kind of urban fabrics they pinpoint, we extract layered information on all reference cities from the OpenStreetMap (OSM) database [34] and use it as a proxy for land use.[4] When superposing the urban surface covered by unit areas associated with c_2 to OSM data, we remark a good match with locations essentially related to office-hour work activities. Maps of exemplar case studies are provided in Figs. 8b, 8d, and 8f for Turin, Milan, and Paris, respectively. A full record of matchings between unit areas in c_2 and office fabrics is instead detailed in Table 3. Where applicable, capital letters link table entries to maps in Fig. 8.

The list is fairly extensive, and we omit a complete discussion for the sake of brevity. As the reader will remark by browsing Table 3, the signature class appears to highlight office-dense areas, universities, hospitals (especially those linked to research centers or universities), large companies headquarters, administrative centers, and commercial-only areas. We therefore consider c_2 to be representative of *office urban fabrics*, i.e., urban areas interested by socio-economical activities related to development, commercialization and fruition of services and goods, with a typical European working time during week days, 9:00-18:00. A confirmation comes from the analysis of the signature $\mathbf{s}_2^\star$ associated to c_2, in Fig. 8a. The signature is characterized by a fairly constant activity during office hours; more importantly, mobile activities tend to disappear during the weekend, when a very small fraction of offices is open.

It is also interesting to investigate situations where a mismatch is observed between the mobile traffic signature and the OSM data. As an example, an important commercial area is located in Southeastern Milan according to OSM, see L in Fig. 8d; however the mobile traffic profile in the area is not that of c_2. Indeed, this is the *Mercato Ortofrutticolo*, a wholesale market where most of the commercial activity is carried out very early in the morning, and that is nearly deserted in the afternoon. Unlike OSM data, mobile traffic signatures neatly detect the quite unique nature of this zone, which is moved apart from standard office areas and into a category per se. Similarly, the Expo 2015 zone in Milan is covered by an office fabric signature in the 2015 dataset, see I in Fig. 8d; however it is not in the 2013 dataset, when the area was still under construction.

Other popular signatures, mostly located in France, resemble $\mathbf{s}_2^\star$. In particular, the $\mathbf{s}_5^\star$ signature, in Fig. 8c, also presents a quite homogeneous activity with a peak late in the morning, and a high weekday-to-weekend traffic ratio. However, these peculiar features of mobile activity in office areas are blended with those observed for residential fabrics. This suggests that unit areas characterized by c_5 still contain mostly offices, but have a minor presence of residential fabrics. Support comes from Fig. 8f, Fig. 8f, and Table 3, as c_5 is mainly located in city centers and other mixed-use areas.

The opposite happens for $\mathbf{s}_7^\star$, in Fig. 8e, which shows the usual late morning peak, but also a significant reduction of activity at noon during working days and no traffic on Saturday and Sunday. This profile thus denotes pure office fabrics, not contaminated by other land usages. Proofs come from Fig. 8f and Table 3: e.g., in Paris the signature is associated to the Issy-les-Moulineaux area, where headquarters of important companies are located, and La Defense, the major business district of the metropolitan area.

4. OSM allows tagging geographic areas or even individual buildings so as to denote their primary purpose. The crowd-sourced nature of the information makes it often inaccurate, with tags that can be very generic, associated to multiple activities, or just missing. Thus, we do not treat OSM data as ground truth; rather, it provides hints towards a correct interpretation of the mobile traffic signatures.

TABLE 3
Office PoIs in Unit Areas of Classes c_2, c_5, and c_7

Class	Dataset	PoIs
c_2	**Tu-15**	Politecnico di Torino (**A**); University of Turin and St. Anne's Hospital (**B**); Le Molinette Hospital (**C**); Telecom Italia Labs (**D**); New Holland Constructions (**E**); Turin Police School, city-center office area (**F**); Iveco Trucks Plant (**G**), Turin Industrial Zone (**H**), Fiat Mirafiori (**I**) plants.
	Mi-15	Politecnico di Milano (**A**); University of Milan (**B**); University of Milano-Bicocca (**C**); Catholic University of the Sacred Heart (**E**);
	Mi-13	Policlinico Hospital (**F**); San Raffaele Hospital (**G**); Mediaset Milano 2 Television Studios and Industrial Pole (**H**); Expo 2015 area (**I**) (Mi-15 only); commercial city-center area (**J**); San Siro Hippodrome Betting pool (Mi-13 only) (**K**).
	Pa-15	Ministry of Defense (**A**); Ministry of Ecology and Development (**B**); Palace of Justice (**C**).
	Rm-15	Ministry of Defense and Intern; Montecitorio Palace, Chamber of Deputies; Policlinico Umberto I; La Sapienza University; Foro Italico University of Rome; Pontifical Universities; Bambino Gesù Children's Hospital; RAI Television.
	Tn-13	Interporto Industrial Zone; Trento Northern Commercial Zone; Spini di Gardolo and Lavis Industrial Zones.
	Ma-15	Aix-Marseille University; Marseille European Hospital.
	Ly-15	University of Lyon I; Parc Technologique Portes des Alpes.
	To-15	Airbus North-West industrial pole; Médipôle Garonne Hospital; Rangueil Hospital; Veolia Peche David plant.
	Li-15	Lille 1 University of Science and Technology; Synergie research-industrial park; Lesquin commercial-industrial park.
c_5	**Mi-15**	Milano Bicocca Village (Mi-13 only); Milano Bocconi University Campus (**D**).
	Mi-13	
	Pa-15	Paris I University (**D**); École Normale Superieure (**E**); Curie Institute (**F**); Paris Tech (**G**); Paris Dauphine University (**H**); Paris Sorbonne University (**I**); Paris Descartes University (**J**); Pierre and Marie Curie University (**K**); Polytech (**L**); Ministry of Interior (**O**); Georges Pompidou Hospital (**R**); France Television (**S**).
	Ly-15	Lumière Lyon II and Jean Moulin Lyon III universities; Saint Joseph Saint Luc Hospital center; Le Vinatier Hospital center; Hospital center and University of Lyon Sud; Edouard Herriot Hospital; New Palace of Justice; Cité administrative; Gare de Vaise commercial area; Parilly industrial area; Port Edouard Herriot.
	Ma-15	Timone University Hospital; École centrale de Marseille; Palace of Justice and Courthouse area.
	To-15	Toulouse National Center for Space Studies; Airbus Defence and Space area; Toulouse airport commercial area; Cité de l'espace; Purpan Hospital; Toulouse Sud industrial area; Fondeyre industrial zone; Thales Alenia Space Plant; ON Semiconductors Plant; Freescale Semiconductor Plant; Meteofrance and Aviation Civile centers.
	Li-15	Lille 2 University; Lille 3 Charles-de-Gaulle University; Lille Institute of Political Studies; Pasteur Institute Research Center; Saint-Vincent De Paul Hospital; Regional Hospital University Center; Oscar Lambret, Fontan, Albert Calmette, Roger Salengro and Saint Philbert Hospitals; Les Prés commercial business pole; Marcq-en-Baroeul commercial area.
	Bo-15	Palace of Justice and Courthouse area; French National School for the Judiciary; University School of Management (IAE); Bordeaux University (Campuses Carreire and Talence-Pessac); Pellegrin Hospital; Pessac Commercial area; Bruges Bordeaux Fret and René Ledoux industrial areas;
c_7	**Mi-15**	Minor industrial area (Mi-13 only).
	Mi-13	
	Pa-15	Pharmaceutical research laboratories (Sanofi, Pfizer) (**M**); Southern industrial commercial pole (**N**); La Defense business area (**P**); Google France (**Q**); Microsoft France (**R**); Orange France Headquarters (**T**).
	Ma-15	Port Logistic Platform Solaris; Administrative area of Marseille around Place Felix Berret (e.g., Administrative court, Banks and Consulates, etc.).

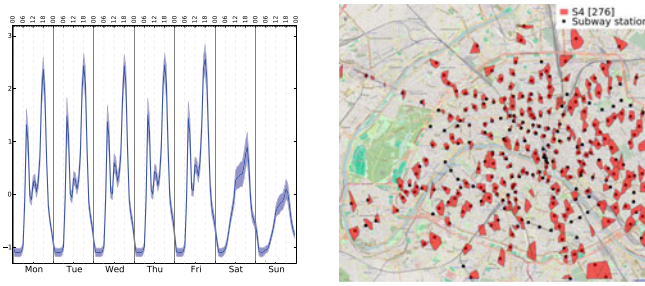
5.5 Transportation Urban Fabrics

A very distinctive class emerging from Fig. 6 is c_4, which only appears in Paris, and includes more than 10 percent of the unit areas in the city. In fact, we found c_4 to match almost perfectly the network of subway stations in Paris: as shown in Fig. 9b, more than 90 percent of the stations are precisely covered by a unit area characterized by c_4, with near-zero false positives. Indeed, signature $\mathbf{s}_4^\star$, in Fig. 9a is a clear portray of commuting behavior. During weekdays, it shows three clear activity peaks early in the morning, at lunch time (sensibly smaller than the other two), and late in the afternoon. During the weekend, all peaks basically vanish. We thus unequivocally associate c_4 to precise *subway station fabrics*. It remains the question of why we do not observe this signature class in Milan or Lyon, which also have subway networks. We believe that the answer lies in the much larger size of Paris, leading to a mass of subway commuters that is not comparable with those of other cities in our reference set. Mobile traffic dynamics such as those in c_4 most probably occur in Milan and Lyon as well, but not at a scale sufficient to emerge from the common activity.

Subways are not the only mean of mass transport in metropolitan regions, and we find other popular signatures in Fig. 6 to be related to different types of transportation hubs. This is the case of signatures $\mathbf{s}_9^\star$ and $\mathbf{s}_{10}^\star$, whose classes capture mobile traffic dynamics that are found in both Italian and French cities. The superposition of unit areas in c_9 and c_{10} with OSM layered data shows a match to train stations, as exemplified in Figs. 10b, 10d, and 10f for Lille, Milan and Paris, respectively. Letters in the figures pinpoint major railway infrastructures, as detailed in Table 4. The same table provides a complete listing of PoIs in unit areas associated to these two signature classes: again, most are relevant public transport facilities. Interestingly, in addition to train stations, unit areas in c_9 also cover important highway interchanges in, e.g., Milan, Marseille, Paris, or Toulouse: this underscores the common communication patterns characterizing all long-range commuters, independently of their transportation mode choice. We conclude that c_9 and c_{10} can be related to generic *transportation fabrics*.

Looking at the signatures themselves, in Figs. 10a and 10c, we can observe similarities with $\mathbf{s}_4^\star$, in Fig. 9a. Specifically, the three traffic peaks are still there, but the first two are dramatically reduced, whereas the late afternoon one remains comparable to that in $\mathbf{s}_4^\star$. We surmise that mobile communication dynamics are similar but not identical for commuters using subways, or railways and roadways.

In some cities, we noted that major train stations are not characterized by signatures c_9 or c_{10}, see Table 4. In these cases, very specific signatures are associated to the few (possibly one) unit areas covering the railway station. Their shape is not dissimilar from that of c_9 or c_{10}, but they have



(a) $\mathbf{s}_4^\star$ (b) Pa-15, c_4

Fig. 9. Characteristic signature and map of unit areas in c_4. The map includes the locations of subway stations in Paris.

peculiar behaviors that let our clustering approach classify them apart. As an example, Fig. 10e depicts $\mathbf{s}_{36}^\star$, a signature denoting the most important train stations in Lyon and Paris, as shown in Fig. 10f. We can remark a general pattern comparable to that of $\mathbf{s}_9^\star$, but for the very high peak on Friday afternoon: these stations are not only used by commuters, but also as arrival and departure points for local inhabitants leaving for the weekend, as corroborated by the higher returning peak on Sunday afternoon.

More examples are provided in Table 4, which shows how major highway interchanges, seaports, and trafficked tunnels are also areas characterized by signatures such as those discussed above.

5.6 Touristic and Leisure Urban Fabrics

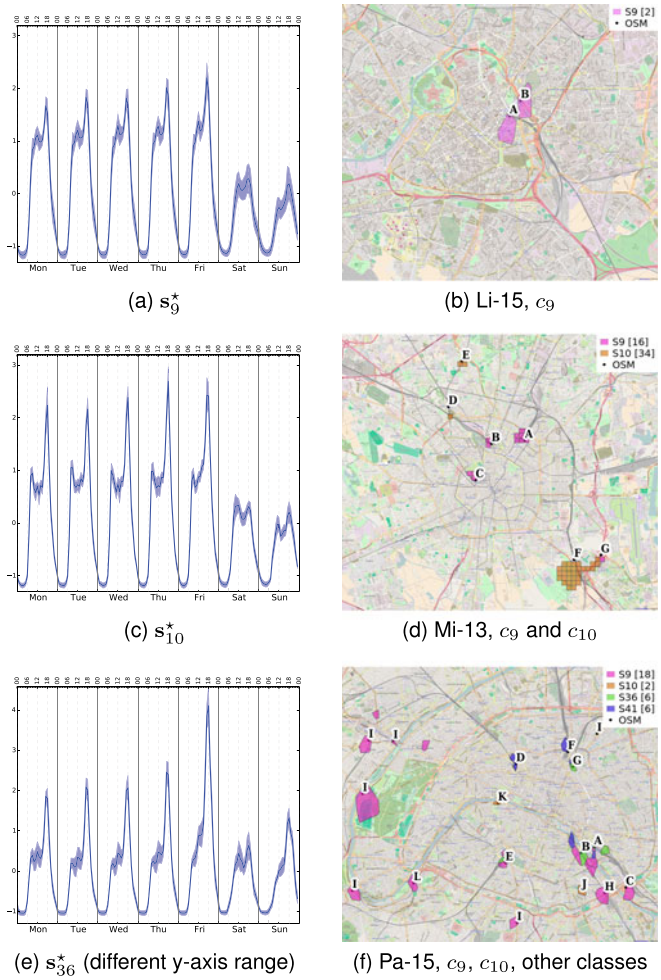
A signature class that appears to characterize areas in both Italian and French cities in Fig. 6 is c_{16}. The corresponding signature, $\mathbf{s}_{16}^\star$, portrayed in Fig. 11a, is representative of *touristic and leisure urban fabrics* typically found in city centers. Figs. 11b and 11c show examples for the unit areas in Milan and Turin, respectively: large parts of the cities historical centers, where famous monuments, museums and squares are located, have mobile traffic signatures that belong to this class. Details of the labeled locations in the maps are provided in Table 5.

We highlight how $\mathbf{s}_{16}^\star$ is in fact not very far from signatures associated to common and office fabrics. Often, city centers are not exclusively touristic places, but host administrative buildings, offices or universities, as well as residential neighborhoods. All these facets contribute to the mobile traffic activity, somehow blurring the tourist presence. Still, the clear indicator of visitor activity, which allows separating areas in c_{16} from office-only zones is the persistent mobile communication load during weekends.

Although found in both Italian and French cities, c_{16} is more frequent in the former than in the latter. City centers in France are characterized by slightly different mobile demand dynamics, represented by class c_{14}. Signature $\mathbf{s}_{14}^\star$, portrayed in Fig. 11d, shows a remarkable peak of activity on Saturdays. We find this peculiar pattern to characterize fashion and commercial streets in city centers, and peripheral areas where very large malls are located. Representative examples are shown in Figs. 11e and 11f for Lyon and Toulouse, respectively, while more matches are listed in Table 5. The sensible reduction of traffic on Sundays is due to the fact that most commercial and leisure areas, even those located in city centers, tend to be closed in France on that day.

5.7 Unique Urban Fabrics

All signature classes presented above describe mobile traffic dynamics that are common to a significant number of unit



(a) $\mathbf{s}_9^\star$ (b) Li-15, c_9

(c) $\mathbf{s}_{10}^\star$ (d) Mi-13, c_9 and c_{10}

(e) $\mathbf{s}_{36}^\star$ (different y-axis range) (f) Pa-15, c_9, c_{10}, other classes

Fig. 10. Transportation fabric signatures $\mathbf{s}_9^\star$, $\mathbf{s}_{10}^\star$, and $\mathbf{s}_{36}^\star$, and their maps for French and Italian cities, with OpenStreetMap data.

areas across cities of countries. However, very enthralling behaviors emerge when investigating peculiar signatures that pinpoint individual unit areas, i.e., *unique urban fabrics* in our dataset. Many of those are signatures that display dramatic surges in the communication activity that deviate from a regular activity pattern. One example is that of c_{67}, whose signature $\mathbf{s}_{67}^\star$ is shown in Fig. 12a: it yields a regular activity pattern similar to that of office and commercial areas, but for surging traffic on Wednesday and Sunday at noon. When looking at the corresponding unit areas, this class locates in a very precise manner St. Peter's church and square in Rome, in Fig. 12d. The traffic peaks match exactly the weekly blessing ceremonies of the Pope in that place, which regularly gather a large, diverse audience.

In fact, outlying mobile traffic signatures are often associated to large-scale social events. A representative case are sports events that attract supporters to stadiums. For instance, Fig. 12b shows the signature of c_{152} which maps to the Chaban-Delmas stadium in Bordeaux, in Fig. 12e. The structure is home to major games in both football and rugby, which occur during the weekend at different times during afternoons and evenings. It is interesting to remark that no single signature class is identified for all stadiums: depending on the diverse match schedule and nature of stadium events, their signatures present important differences, and, therefore, are associated to different clusters. For instance, c_{150}, whose signature and unit area are in Figs. 12b and 12f,

TABLE 4
Transportation PoIs in Unit Areas of c_4, c_9, c_{10}, and Similar Classes

Class	Dataset	PoIs
c_4	**Pa-15**	90% of the metro stations in Paris.
c_9	**Mi-13**	Milano Centrale Train Station (**A**); Milano Porta
	Mi-15	Garibaldi Train Station (**B**); Milano Cadorna Train Station (**C**).
	Rm-15	Termini Station.
	Pa-15	Gare de Lyon Train Station (**A**); Gare d'Austerlitz Train Station (**B**); Porte de Bercy Highway Interchange (**C**); Quai d'Ivry Highway Interchange (**H**); Gare de Montparnasse Train Station (**E**); Issy Val de Seine Train station (**L**); other relevant highway interchanges (**I**).
	Li-15	Gare de Lille Flandres Train Station (**A**); Gare de Lille Europe Train Station (**B**).
	Ma-15	Gare maritime de la Major Ferry Station; Vieux-Port, Tunnel du Vieux Port and Tunnel Prado-Carénage access areas
	Ly-15	Gare Part-Dieu area.
	To-15	Balma Gramont transportation hub.
c_{10}	**Mi-15**	Milano Nord Bovisa Train/Metro Stations (**D**);
	Mi-13	Affori Train/Metro Stations (**E**); Milano Rogoredo Train Station (**F**); starting point of A-51 Milan Highway (**G**).
	Pa-15	Gare des Invalides Metro Station (**K**); Gare des Olympiades Metro Station (**J**).
Other	**Pa-15**	Gare de l'Est (**G**, signature c_{36}); Gare Saint Lazare Train Station (**D**, signature c_{41}); Gare du Nord Train Station (**F**, signature c_{41}).
	Ly-15	Gare de Lyon Part Dieu Train Station (signature c_{36}); Perrache Train Station (signature c_{279});
	Ma-15	Marseille Saint Charles Train Station (signature c_{94}).
	To-15	Gare de Toulouse-Matabiau Train Station (signature c_{93}).
	Bo-15	Gare de Bordeaux-Saint-Jean Train Station (signature c_{278}).

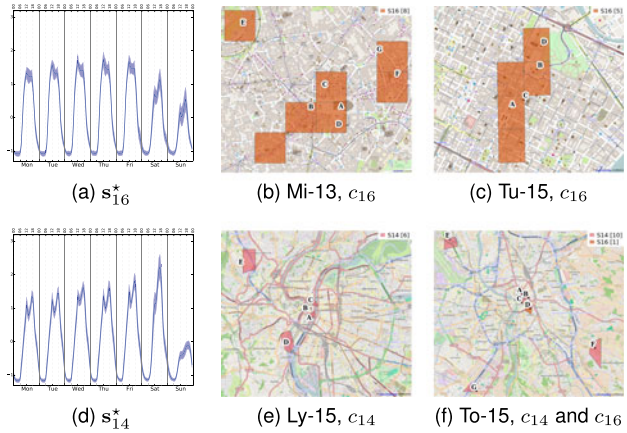
refers to the Vélodrome stadium in Marseille: the local football team plays both national (during weekends) and international (on Tuesday) matches, which reflects in unique peaks.

6 Discussion

Our signature analysis provides a number of interesting cues that stimulate discussion on the interplay between urban fabrics and mobile traffic dynamics. Below, we summarize the main takeaway messages, separating observations that we find intuitive from insights we deem surprising.

Intuitive Findings. Among the expected results, our analysis highlights a clear dichotomy between two prevalent classes of urban fabrics, i.e., residential and office. The former are characterized by a more uniform human presence in time, while the latter—including universities, business centers and company headquarters—are characterized by mobile traffic signatures with high weekday-to-weekend traffic ratios and higher load during typical working hours. Also, one can anticipate that residential areas occupy most of the urban surface, and, thus, that residential mobile traffic signatures are the most common temporal profiles that operators must assume their networks to accommodate: the results in Section 5 confirm all these speculations.

Another expected result is that touristic and leisure areas are characterized by a relatively high mobile traffic activity



(a) $s^{\star}_{16}$ (b) Mi-13, c_{16} (c) Tu-15, c_{16}

(d) $s^{\star}_{14}$ (e) Ly-15, c_{14} (f) To-15, c_{14} and c_{16}

Fig. 11. Touristic and leisure fabric mobile traffic signatures and maps.

during weekends. Equivalent considerations can be made for areas that host periodical events attracting a large number of people, which induce dramatic surges in the communication activity: the likes of arenas, theaters, stadiums or religious places are easily spotted via mobile traffic signatures. Similarly, some nation-specific behaviors could be easily envisaged with minimal knowledge of the local population habits: e.g., activity peak on Saturdays in large French commercial areas and the little or no human presence on Sundays are easily explained by the fact that commercial centers are typically closed on Sundays in France.

Finally, all the above demonstrates that mobile traffic signatures can detect urban fabrics with a higher level of accuracy than crowd-sourced databases of land use, and possibly pinpoint, at very fine grains, urban zones that yield fairly unique human undertakings. Moreover, mobile traffic analysis allows automatic updating of the urban tissue information, which is not possible with traditional survey-

TABLE 5
Touristic and Leisure PoIs in Unit Areas of c_{14} and c_{16}

Class	Dataset	PoIs
c_{16}	**Mi-13/ Mi-15**	Milan Cathedral (**A**); Milan Cathedral Square (**B**); Galleria Vittorio Emanuele II (**C**); Royal Palace of Milan (**D**); Sforza Castle (**E**); San Babila Church (**F**); Quadrilatero della moda (partially) (**G**).
	Tu-15	San Carlo's Square (**A**); Madama Palace (**B**); Egyptian Museum (**C**); Royal Palace and Chapel of the Holy Shroud (**D**).
	Rm-15	Campo Marzio area (including parts of Piazza di Spagna, Madama Palace and Piazza del Popolo).
	Pa-15	Champs Elysees (George V area); Place Edouard VII (Opera area).
	Ma-15	Marseille Cathedral.
	To-15	Grand Rond Park.
c_{14}	**To-15**	Place du Capitole (**A**); Saint Georges Mall (**B**); Place Esquirol (**C**); Blagnac Mall(**E**); Saint Orens Mall (**F**).
	Ly-15	Place Bellecour (**A**); Place de la Republique (**B**); Rue de La Bourse shopping area (**C**); La Confluence shopping area (**D**); Ecully Grand Ouest Mall (**E**).
	Pa-15	Rue de Rivoli shopping are); Bazar de l'Hôtel de Ville Mall; Rue de Commerce shopping area; Passy Plaza Mall.
	Ma-15	Cours Saint Louis shopping area; La Valentine Mall area.
	Li-15	Euralille Mall; Place du Général-de-Gaulle; Englos Mall.
	Bo-15	Alienor Mall; Merignac Mall; Rive d'Arcins Mall.

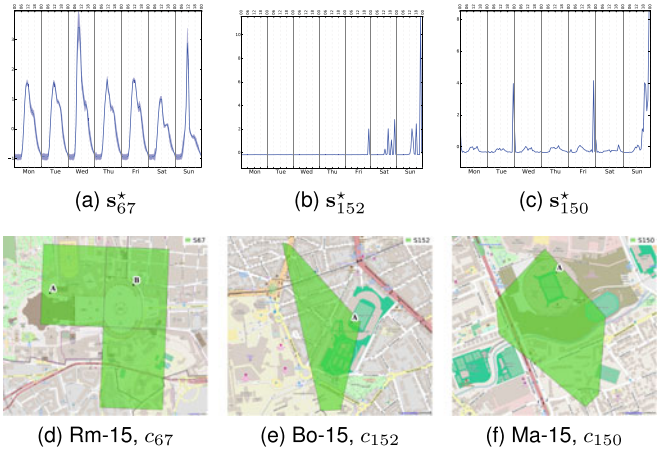


Fig. 12. Examples of unique urban fabrics signatures and maps.

based methods. Both observations were made in previous studies on land use detection based on mobile traffic data, and our analysis reinforces them.

Surprising Insights. Less obvious considerations also stem from the results in Section 5. First, residential mobile traffic in Italy and France shows striking differences, which one would hardly expect from countries that are in geographical and cultural proximity. This diversity lets us surmise that circadian rhythms are intrinsically different in the two countries. We hypothesize that the phenomenon could extend to many major countries in Europe, with important implications in, e.g., the cross-border competition among operators fostered by the new EU regulations on roaming.

Second, such a diversity disappears when moving away from residential city areas. Metropolitan regions that are driven by office, touristic or leisure activities, or that host mass transit infrastructures or main sports facilities have mobile traffic signatures whose traits are consistent in both counties, and across all our reference cities.

Third, ours is the first study to identify mobile traffic activity patterns that clearly denote major transportation hubs. The most distinctive feature is the very high traffic peak in the late afternoon: workers are thus more prone to use mobile services when commuting back home than when travelling to workplaces in the morning. Our analysis also unveils how the mobile demand of medium- and short-range commuters, using subways or urban railways, is semantically different from that of long-range commuters using trains or cars to reach their workplace: the former seem more inclined to mobile communications at all times during their commutes, whereas the latter tend to call and text during return trips mainly.

Fourth, considering ten different cities at once allows us to comment on the diversity observed across them. In the light of the results, our opinion is that the three aspects that drive most of the inter-city differences are (i) the country (ii) the size of the metropolitan area, and (iii) the spatial granularity of the mobile traffic data. We already discussed the dissimilarities in residential mobile traffic between France and Italy. In addition to this, we remark that Paris, three to ten times larger than all other cities, has unique signatures that tell it apart from the rest. Also, the per-antenna traffic recorded in French cities leads to a higher accuracy (and thus increased heterogeneity) of signatures than that observed in Italian cities. Instead, we do not note major differences between cities of comparable size in a same country.

Fifth, we recall that all our results relate to the mobile traffic activity, which does not necessarily maps to human presence directly. Although the general trends of the two dynamics are probably comparable, we spot noticeable differences in the vast majority of cases. For instance, it is well known that road traffic and railway usage follow a very typical daily pattern, with a high concentrated peak in the morning and a smoother lower peak in the afternoon: our transportation hub signatures are fairly different from this model. Another representative example is that of arenas and stadiums, where major events are characterized by a significant increase of presence, and an even more dramatic surge of mobile traffic demand: individuals attending live shows are keen to digitally share the experience with friends, which exacerbates the network activity peaks recorded in such occasions, with respect to the actual increment of population. Overall, we conclude that mobile traffic signatures are an effective way to detect urban fabrics, but not necessarily human presence.

7 Conclusion

Today, mobile communications permeate our social life. An interesting side effect of mobile device pervasiveness is the possibility of analyzing datasets collected by network operators for fine-grained analyses of subscribers' endeavors. In this paper, we unveiled the strong intertwining between the mobile traffic activity and the urban fabrics that characterize the areas when such activity takes place. We did so by (i) devising a novel signature classification technique that outperforms current state-of-the-art solutions, and (ii) using our technique on a dataset of unprecedented scale and heterogeneity. Our results demonstrate the concurrent presence of mobile communication behaviors that are common to all urban areas, and of habits that are unique to countries or cities. The proposed methodology has applications in automatic land use detection and network management.

Acknowledgments

This work was supported by the French National Research Agency under grant ANR-13-INFR-0005 ABCD and by the EU FP7 ERA-NET program under grant CHIST-ERA-2012 MACACO. This work is an extended version of a IEEE/ACM ASONAM 2015 paper [1].

References

[1] A. Furno, R. Stanica, and M. Fiore, "A comparative evaluation of urban fabric detection techniques based on mobile traffic data," *Proc. IEEE/ACM Int. Conf. Advances Social Netw. Anal. Mining*, Aug. 2015, pp. 689–696.

[2] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 124–161, Jan. 2016.

[3] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," presented at the IEEE INFOCOM, Shanghai, P.R. China, Apr. 2011.

[4] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," presented at the IEEE INFOCOM, Toronto, ON, Canada, Apr. 2014.

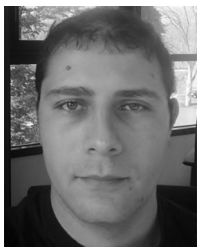
[5] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A first look at cellular network performance during crowded events," presented at the ACM SIGMETRICS/Int. Conf. Meas. Model. Comput. Syst., Pittsburgh, PA, USA, Jun. 2013.

[6] X.-R. Ahas, et al., "Everyday space–time geographies: Using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn," *Int. J. Geographical Inf. Sci.*, vol. 29, no. 11, pp. 2017–2039, Nov. 2015.

[7] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-scale measurement and characterization of cellular machine-to-machine traffic," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1960–1973, Dec. 2013.
[8] S. Almeida, J. Queijo, and L. M. Correia, "Spatial and temporal traffic distribution models for GSM," presented at the *IEEE VTS 50th Veh. Technol. Conf.*, Amsterdam, Netherlands, Sep. 1999.
[9] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: Connecting people, locations and interests in a mobile 3G network," presented at the 9th ACM SIGCOMM Conf. Internet Meas. Conf., Chicago, IL, USA, Nov. 2009.
[10] M. R. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Martinez, "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics," presented at the IEEE 2nd Int. Conf. Social Comput., Minneapolis, MN, USA, Aug. 2010.
[11] M. de Nadai, J. Staiano, R. Larcher, N. Sebe, D. Quercia, and B. Lepri, "The death and life of great italian cities: A mobile phone data perspective," presented at the 25th Int. Conf. World Wide Web, Montréal, QC, Canada, Apr. 2016.
[12] F. Girardin, J. Blat, F. Calabrese, F. Dal Fiore, and C. Ratti, "Digital footprinting: Uncovering tourists with user-generated content," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 36–43, Oct. 2008.
[13] M. Lee and P. Holme, "Relating land use and human intra-city mobility," *PLoS One*, vol. 10, no. 10, Oct. 2015, Art. no. e0140142.
[14] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," presented at the 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Beijing, P. R. China, Aug. 2012.
[15] K. Tutschku, "Demand-based radio network planning of cellular mobile communication systems," presented at the IEEE INFOCOM, San Francisco, CA, USA, Apr. 1998.
[16] H. Khedher, F. Valois, and S. Tabbane, "Real mechanisms for mobile networks modeling and engineering," presented at the IEEE Global Telecommun. Conf., Dallas, TX, USA, Dec. 2004.
[17] R. A. Becker, et al., "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 18–26, Jun. 2011.
[18] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," presented at the ACM SIGKDD Int. Workshop Urban Comput., Beijing, P.R. China, Aug. 2012.
[19] P. Secchi, S. Vantini, and V. Vitelli, "Analysis of spatio-temporal mobile phone data: A case study in the metropolitan area of milan," *Statist. Methods Appl.*, vol. 24, no. 2, pp. 279–300, Jul. 2015.
[20] M. Lenormand, et al., "Comparing and modelling land use organization in cities," *Roy. Soc. Open Sci.*, vol. 2, Apr. 2016, Art. no. 150449.
[21] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: Segmenting space through digital signatures," *IEEE Pervasive Comput.*, vol. 9, no. 1, pp. 78–84, Jan. 2010.
[22] V. Soto and E. Frias-Martinez, "Automated land use identification using cell-phone records," presented at the ACM 3rd ACM Int. Workshop MobiArch, Washington, DC, USA, Jun. 2011.
[23] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a comparative science of cities: Using mobile traffic records in New York, London, and Hong Kong," *Geotechnologies Environment*, vol. 13, pp. 363–387, Nov. 2015.
[24] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology," presented at the 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput., Hangzhou, China, Jun. 2015.
[25] S. Ray and R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," presented at the 4th Int. Conf. Advances Pattern Recognit Digit. Techn., Calcutta, India, Dec. 1999.
[26] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[27] H. Zang and J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," presented at the 13th Annu. ACM Int. Conf. Mobile Comput. Netw., Montreal, QC, Canada, Sep. 2007.
[28] F. Calabrese, G. di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 36–44, Oct. 2011.
[29] A. Hinneburg and D. A. Keim, "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering," presented at the 25th Int. Conf. Very Large Data Bases, Edinburgh, Scotland, U.K., Sep. 1999.
[30] Telecom italia big data challenge. (2016). [Online]. Available: http://www.telecomitalia.com/bigdatachallenge
[31] Milan open data. (2016). [Online]. Available: http://dati.comune.milano.it
[32] Turin open data. (2016). [Online]. Available: http://aperto.comune.torino.it/
[33] C. D. Manning, P. Raghavan, and H.Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press. 2008.
[34] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, Oct. 2008.

Angelo Furno received the PhD degree in computer science from the University of Sannio, Italy, in 2014. He is a researcher with IFSTTAR-ENTPE, Université de Lyon, France. He worked with INRIA, France, as postdoctoral researcher, from 2014 to 2016. His research interests include data analysis, mobile networking, distributed computing, and intelligent transportation systems, with special focus on applied machine learning for modelling human mobility from multi-source data. He is a member of the IEEE.

Marco Fiore (S'05-M'09) received the HDR degree from the Univeristé de Lyon, France, and the PhD degree from Politecnico di Torino, Italy. He is a researcher with CNR-IEIIT, Italy. He was an associate professor with INSA Lyon, France, associate researcher with Inria, France, and visiting researcher with Rice University, Texas, and the Universitat Politecnica de Catalunya, Spain. His research interests include the fields of mobile traffic data analysis and vehicular networking. He is a member of the IEEE.



Razvan Stanica received the MEng and PhD degrees in computer science from INPT, France, in 2008 and 2011, respectively. He also received the MEng degree from the Polytechnic University of Bucharest, Romania. He is an associate professor with INSA Lyon and a research scientist with the INRIA UrbaNet Team, CITI Laboratory. His research interests include wireless mobile networks, with a special focus on communication networks in urban environments.

Cezary Ziemlicki received the graduate degree in automation of industrial processes from the Warsaw University of Technology. He is a R&D engineer in the Laboratory Sociology & Economics of Networks and Services, Orange Labs, Chatillon, France. He is a research engineer and joined Orange Labs, in 2000. His work with Orange Labs is to develop methodologies of analysis of telco operator data for use in human sciences.

Zbigniew Smoreda received a PhD degree from Paris Est University. He is a researcher with Orange Labs. He was an assistant professor with Warsaw University, a researcher and lecturer with the Université de Paris 8, and a researcher with France Télécom and with Observatoire Mondial des Systèmes de Communication. His work at CNET/France Télécom R&D/Orange Labs is on the sociology of communication, with focuses on social uses of ICTs, social network forms, and transformations associated with technologies.