Fusing GPS Probe and Mobile Phone Data for Enhanced Land-Use Detection

Angelo Furno, Nour-Eddin El Faouzi, Marco Fiore, Razvan Stanica

Fusing GPS Probe and Mobile Phone Data for Enhanced Land-use Detection

Angelo Furno*, Nour-Eddin El Faouzi*, Marco Fiore†, Razvan Stanica‡
* Univ Lyon, ENTPE, IFSTTAR, LICIT UMR_T 9401, F-69518, Lyon, France – name.surname@ifsttar.fr
† CNR – IEIIT, Corso Duca degli Abruzzi 24, 10129 Torino, Italy – marco.fiore@ieiit.cnr.it
‡ Univ Lyon, INSA-Lyon, INRIA, CITI, F-69621 Villeurbanne, France – razvan.stanica@inria.fr

***Abstract*—Profiling the diversity of land use in modern cities by mining data related to human mobility represents a challenging problem in urban planning, transportation and smart city management. Previous work on mobile phone data (i.e., Call Detail Records) has shown the existence of strong correlations between the urban tissue and the associated mobile communication demand. Similarly, GPS traces of vehicles convey information on transportation demand and human activities that can be related to the land use of the neighborhood where they take place.**

In this paper, we investigate the land use patterns that emerge when studying simultaneously GPS traces of probe vehicles and mobile phone data collected by network providers. To this end, we extend previous definitions of mobile phone traffic signatures for land use detection, so as to incorporate additional information on human presence and mobility conveyed by GPS traces of vehicles. Leveraging these extended signatures, we exploit an unsupervised learning technique to identify classes of signatures that are distinctive of different land use. We apply our technique to real-world data collected in French and Italian cities. Results unveil the existence of signatures that are common to all studied areas and specific to particular land uses. The combined use of mobile phone data and GPS traces outperforms previous approaches when confronted to ground-truth information, and allows characterizing land use in greater detail than in the literature to date.

I. Introduction and Related Work

Urban landscapes present a variety of socio-technical environments that are associated to diverse and complex human activities. Such undertakings have a significant impact on the way individuals connect with each other, move from place to place and, in general, behave. Therefore, a strong link exists between the urban geography and the consumption of mobile communications and transportation services [1], [2]. By characterizing such consumptions, it is possible to develop solutions for the automatic identification of land use. In turn, land use classification can drive core-business investments, sustainable planning and decision-making processes across domains like transportation, telecommunication, or urbanism.

Traditional urban land-use classification approaches leverage questionnaire-based travel surveys of properly selected population samples [1], or remote sensing imagery data [3]. Such approaches are expensive, time-consuming and subject to rapid obsolescence. Moreover, state-of-the-art solutions typically exploit single-source datasets, which bounds the spectrum of land use information to the inherent dynamics and spatio-temporal granularity of the data source. As an example, land use detection based on the analysis of the pick-up/set-down dynamics of taxis may reveal the social function of certain urban areas [4], but is unhelpful in regions with low taxi traffic: this results in a sparse land use characterization. Conversely, the pervasiveness of mobile phones allows discovering a large variety of land uses, with a high geographical coverage [2]: however, the spatial granularity is limited by the sparsity of cellular base stations [5]. Land use classification based on large-scale and multi-domain data related to human mobility, communication and social activities is therefore emerging as a very promising and cost-efficient alternative, but is still at an early-stage [6], covering single-city scenarios and few days of data.

In this paper, we aim at investigating land uses that emerge when simultaneously studying multiple sources of urban big data. To that end, we propose a novel approach, presented in Sec. II and based on a generalization of the methodology originally introduced in [7] for Call Detail Records (CDR). Our key contribution consists in extending our previous CDR-based approach by supporting fusion of additional data, e.g. Taxi and Floating Car Data (FCD), to improve land use detection. In Sec. III, we target the heterogeneous datasets considered in our analysis: mobile phone traffic data, i.e. CDR, collected by network providers, as well as GPS traces of floating vehicles provided by traffic operators in multiple urban areas. To the best of our knowledge, this is the first paper proposing a methodology and analysis in a multi-city scenario of land use that emerge when jointly classifying mobile phone and floating car data. Our results, discussed in Sec. IV and Sec.V, show that such an approach leads to finer-grained and more accurate detection of land use.

II. Multi-domain Signature-based Clustering for Land-use Detection

Let us consider the generic dataset $\mathcal{D}^k$, describing aggregate and geo-localized measures of a metric k of human activity (e.g., communication volumes, car stops, tweets, etc.) in a reference area during a set of days $\mathbf{d}^k = \{d\}$. We name *unit area* the spatial aggregation level for dataset $\mathcal{D}^k$: the whole geographic region under consideration $\mathbf{a}^k = \{a\}$ is 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

[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 geometric grid.

human activity is aggregated in each unit area. Each day $d \in \mathbf{d}^k$ is thus split into a set $\mathbf{t}^k = \{t\}$ of time slots t. Overall, $\mathcal{D}^k = \left\{v_a^k(d,t)\right\}$, where every element $v_a^k(d,t)$ describes the total observed amount of metric k within each unit area a at time slot t of day d.

The proposed technique generalizes that originally introduced in [7]. It is based on the construction of a representative set of multi-metric signatures through seven phases. These phases aim at: (i) summarizing the human activity in each unit areas into a meaningful profile, i.e., the unit area signature; (ii) grouping similar unit area signatures into a limited set of classes, each exhibiting a unique behavior.

1. The *signature metric* k defines the nature of the observed human-related activity. Examples of metrics include the number of taxi pickups/drop-offs, vehicle in/out flow, the number of voice calls, the number of short text messages (SMS), the volume of Internet traffic, etc. The metric controls the actual information in each dataset entry $v_a^k(d,t)$.
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). Our approach leverages the definition of a *Median Week Signature (MWS_k)* for each kind k of activity. As widely discussed in [7], the typical weekly behavior of the mobile demand at each unit area contains the vast majority of the significant information about the nature of that area. We assume that this statement applies to different kinds of activity k. The support is therefore defined as $\boldsymbol{\delta} = \{\text{MON,TUE,WED,THU,FRI,SAT,SUN}\}$.
3. The *data denoising* component extracts information deemed to be representative of the typical observed activity k 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 $v_a^k(d,t)$ samples into a single value. Our definition of MWS exploits the median aggregator to denoise and compress the available samples. Thus, the generic element associated to time slot t of day $\delta \in \boldsymbol{\delta}$ in the k-signature of unit area a is
$$s_a^k(\delta,t) = \mu_{1/2}\left(\left\{v_a^k(d,t) \mid d \in \mathbf{d}_\delta^k\right\}\right), \quad \forall a \in \mathbf{a}^k, \quad (1)$$
where, $\mu_{1/2}(\cdot)$ is the median of the set within parenthesis.
4. The *signature normalization* makes signatures independent from the absolute volume of activity k recorded at a unit area. This allows comparing activity k at different unit areas on the sole basis of its variations. Our approach adopts a standard score normalization. To that end, each element obtained in (1) 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
$$\hat{s}_a^k(\delta,t) = \frac{s_a^k(\delta,t) - \mu(\mathbf{s}_a^k)}{\sigma(\mathbf{s}_a^k)}, \quad \forall \delta \in \boldsymbol{\delta}, t \in \mathbf{t}^k, a \in \mathbf{a}^k, \quad (2)$$
where $\mu(\mathbf{s}_a^k)$ and $\sigma(\mathbf{s}_a^k)$ denote the mean and standard deviation of the set of elements concatenated in the signature $\mathbf{s}_a^k$. Then, the normalized signature $\hat{\mathbf{s}}_a^k$ is simply obtained by concatenation of $\hat{s}_a^k(\delta,t)$ for all $\delta \in \boldsymbol{\delta}$ and $t \in \mathbf{t}^k$.
5. The *signature pairwise distance measure* determines the degree of similarity of two signatures and is based on the Pearson correlation coefficient.
6. The *signature clustering algorithm* groups together signatures that are alike, leveraging the distance measure above. Our methodology 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. To select the best clustering among all those in the family, we evaluate the skewness of the cluster sizes 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. This phase returns a set of classes of archetypal signatures, denoted as $\mathbf{c}_k$.
7. The *class fusion* procedure aims at combining sets of land use classes $\mathbf{c} = \{\mathbf{c}_k\}$ related to different domains of observed activities into an organic and enhanced representation of land use. The procedure strongly depends on the type of activities considered: an example of class fusion procedure is proposed in Sec. V, to refine classes from the clustering of mobile phone datasets by means of land use information retrieved from GPS traces of floating vehicles.

III. Datasets and Preprocessing

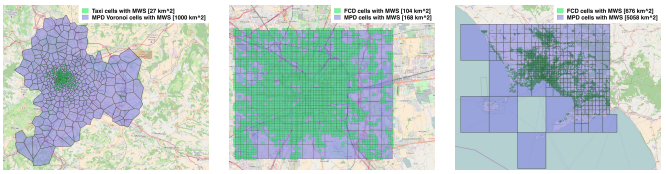
We leverage three different data domains for land use classification: mobile phone traffic logs, taxi trips and floating car data. Multiple sources of data have been accessed to generate the final datasets used in our analysis, as follows.

We have considered two sources of mobile phone traffic logs: 1) per-base station CDR collected by Orange SA in the whole of France; 2) aggregate CDR released by Telecom Italia for several Italian cities, in the context of their 2015 Big Data Challenge (TBD) [8]. Our analysis is focused on three city-wide case studies, i.e., the metropolitan areas of Lyon, Milan, and Naples. Thus, we extracted three different CDR datasets (i.e., **Ly-CDR**, **Mi-CDR**, and **Na-CDR**) from our mobile traffic sources, as detailed in Table I. The CDR signature metric v_a^{CDR} is defined as the hourly volume of in-coming and out-going subscribers' calls and SMS. The communication activity in the data covers the period from March 2 to April 26, 2015 for Italian cities, and from September 1 to November 30, 2014 for Lyon. The volume is provided on a per-antenna basis, in the case of Lyon. Thus, unit areas from Ly-CDR map the coverage zones of the mobile network antennas, which are approximated as the cells of a Voronoi tessellation. For the Italian cities, volumes are provided according to a spatial grid-based tessellation of the city surface, with cell sizes that vary according to the actual coverage of the base stations.

A fourth dataset, namely **Ly-TAXI**, has been extracted from anonymized taxi trips, each described by means of

TABLE I
LABELS AND DESCRIPTION OF THE REFERENCE DATASETS.

Label	Source	City	Unit Areas	Activity	Period
Ly-CDR	Orange	Lyon	454 Voronoi cells	SMS+Calls	Sept-Nov'14
Mi-CDR	TBD15	Milan	429 variable-size cell grids	SMS+Calls	Mar-Apr'15
Na-CDR	TBD15	Naples	534 variable-size cell grids	SMS+Calls	Mar-Apr'15
Ly-TAXI	Taxi-radio	Lyon	18,718 235m×235m cell grids (487 reliable MWS)	Taxi drop-offs	Sept-Nov'12
Mi-FCD	TBD15	Milan	3,162 235m×235m cell grids (1,881 reliable MWS)	FCD stops	Mar-Apr'15
Na-FCD	TBD15	Naples	135,761 235m×235m cell grids (12,244 reliable MWS)	FCD stops	Mar-Apr'15



(a) Lyon (b) Milan (c) Naples

Fig. 1. Cells with reliable CDR, TAXI or FCD MWSs.

multiple timestamped GPS positions of the associated taxi. We considered data collected by Taxi-Radio over a fleet of about 400 taxis in the Grand Lyon agglomeration [9], from September to November 2012[2]. The GPS positions of the taxi in each trip are reported according to a variable sampling interval (between 10 and 60 seconds), with a global average of 800,000 measurements per day. We considered a regular-cell spatial tessellation of the surface of the city of Lyon, to an extent corresponding to the Voronoi tessellation from Ly-CDR. Each cell has a 235×235 m^2 area, i.e., the lowest spatial resolution observed for CDR unit areas. The TAXI signature metric v_a^{TAXI} is defined as the count of taxi drop-offs within each TAXI unit area, in 4-hour time slots (i.e., 00-04, 04-08, 08-12, etc.)[3].

The last two datasets (i.e., **Mi-FCD**, and **Na-FCD**) have been extracted from floating car data collected by InfoBlu in the period March-April 2015 for major Italian cities. This data was provided as part of the 2015 TBD challenge. Similarly to the TAXI data, it describes, by means of sampled GPS positions, trips of vehicles (i.e., cars, motorbikes, trucks, vans and other unspecified vehicles) that are equipped with mobile connected devices. We considered a regular 235×235 m^2-cell tessellation of the surface of Milan and Naples (same extent of the CDR tessellations) to identify FCD unit areas. Consistently with Ly-TAXI, the FCD signature metric v_a^{FCD} is defined as the number of trip stops (i.e, periods of non movement, with engine off for more than 30 minutes) for the generic FCD unit area a. Volumes have been aggregated over 4-hours time-slots.

From the 5 datasets above, we built MWSs for CDR, TAXI and FCD unit areas. To improve the reliability of the analyzed MWSs, we filtered out all the signatures with zero median activity for more than 80% time slots of the week. The resulting number of unit areas with reliable MWS is reported in Tab. I and graphically presented in Fig. 1. Reliable TAXI MWSs are mainly associated to the city center of Lyon and its airport area, while Ly-CDR MWSs refer to a much larger geographical extent (the whole Grand Lyon agglomeration is covered by Voronoi cells with reliable CDR MWSs).

In the Italian cities, FCD unit areas with reliable MWSs cover a larger surface, comparable in size to the one from CDR datasets as a result of a significantly larger number of observed vehicles in the InfoBlu raw data.

IV. LAND USE CLASSIFICATION VIA MULTI-SOURCE DATA

Here, we analyze the MWSs extracted from CDR datasets, and provide a first land use classification based on those. Mobile phone traffic data is capable of capturing large-scale information on human presence and socio-economical activities [7]. When applied to our CDR datasets (i.e., Ly-CDR, Mi-CDR, Na-CDR), the approach presented in Sec. II produces eleven major land use classes[4]. Each class is associated with an archetypal signature that illustrates the typical weekly dynamics of the mobile phone traffic demand in such land use. Archetypal signatures, along with crowd-sourced information from the OpenStreetMap database[5] allow the interpretation and labeling of classes into land uses. We refer the reader to [7] for a thorough description of the labeling process.

Fig. 2 graphically shows the CDR unit areas belonging to each class with different colors. A brief description follows.

*1) French land with dominant residential fabric (**RE$_{fr}$**):* unit areas with negligible presence of noticeable infrastructures or particular activities of inhabitants. They encompass suburban and residential neighborhoods in Lyon, excluding city centers and popular points of interest (POIs), as in Fig. 2a. The associated signature shows a very high afternoon-to-morning peak ratio and higher-than-average weekend activity (Fig. 2f).

*2) French land with mixed residential-office fabric (**MR$_{fr}$**):* residential neighborhoods closer to Lyon city center (Fig. 2a). The archetypal signature (not shown due to space limitation) presents slightly higher morning communication activity than RE$_{\text{fr}}$, due to the presence of office buildings in the same areas.

*3) Cross-national land with dominant office fabric (**OF**):* business areas where the mobile traffic demand is concentrated over work hours, and drops on weekends (Fig. 2g). The pattern marks both French and Italian office-dense areas that host large companies headquarters, universities, industrial zones and hospitals as in Figs. 2b, 2d and 2e.

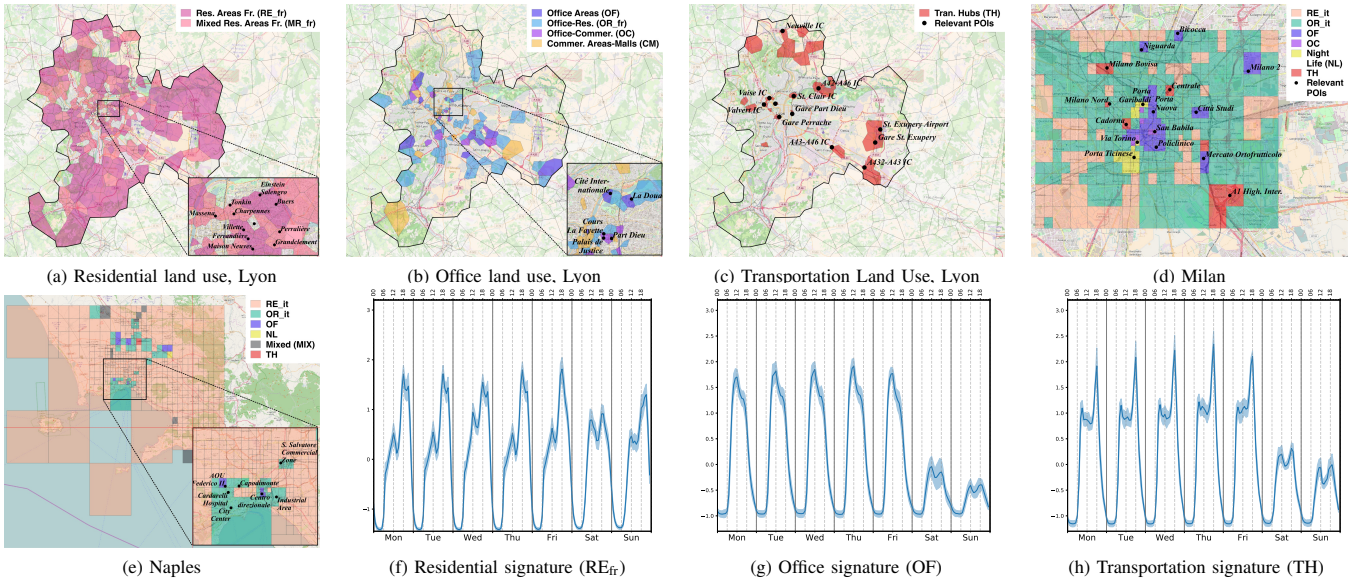
*4) Cross-national land with mixed office-commercial fabric (**OC**):* variations of the OF class that include areas in proximity of city centers, shopping malls and touristic POIs. The mobile phone traffic demand exhibits a minor afternoon peak and high weekend activity.

[2]Taxi and CDR data for Lyon are related to the same months but different years (i.e., Sept.-Nov. 2012 and 2014, respectively), being collected from different providers in the framework of previous collaborations. Thus, our analysis for Lyon is based on the hypothesis of yearly periodicity of taxi pick-up/drop-off dynamics.

[3]The choice of using 4-hours time slots with taxi and FCD data is motivated by the necessity to handle the high spatio-temporal sparsity of the data, and is the result of several empirical tests performed with different sizes (e.g., 1h, 2h, etc.). By using 4h-slots, the number of Ly-TAXI unit areas with reliable MWS increases from 123 to 487 with respect to considering 1-hour time slots. Similar increases were observed for the other cities.

[4]For the sake of brevity, similar classes are aggregated in this paper, under a common representative land use label.

[5]http://openstreetmap.org.



(a) Residential land use, Lyon (b) Office land use, Lyon (c) Transportation Land Use, Lyon (d) Milan

(e) Naples (f) Residential signature (RE_{fr}) (g) Office signature (OF) (h) Transportation signature (TH)

Fig. 2. Land use classification via mobile phone data. Maps of CDR unit areas, with representative characteristic signatures (with standard deviation).

*5) French land with dominant commercial-mall fabric (**CM**):* another variation of the OF class, with a remarkable activity peak on Saturday afternoons. Representative of fashion and commercial streets in downtown Lyon, and peripheral areas with large malls (Fig. 2b).

*6) French land with mixed office-residential fabric (**OR_{fr}**):* areas distant from the city center with dominant activity during work hours due to the presence of office buildings or industrial infrastructures. A minor peak during work day evenings and on Sundays highlights a combined office-residential use.

*7) Cross-national land with transportation fabric (**TH**):* areas with transportation infrastructure that are frequented by long-range commuters, independently of their travel mode and city. Train stations and major highway entry/exit nodes of almost all the analyzed cities belong to this class. The associated signature (Fig. 2h) portrays a typical commuting pattern, with activity peaks early in the morning and late in the afternoon.

*8) Italian land with nightlife-leisure fabric (**NL**):* areas characterized by a dominant presence of sport centers, parks, shopping, cinemas or weekend-life activity, often located nearby residential areas; they indicate the presence of a leisure fabric.

*9) Italian land with dominant residential fabric (**RE_{it}**):* areas in the outskirts of Naples and Milan. During work days, the signature is similar to that of MR_{fr}, suggesting the presence of minor business/commercial activity. However, on Sunday mornings the activity is much higher, probably due to the fact that commercial areas are open in Italy and closed in France.

*10) Italian land with mixed residential-office fabric (**OR_{it}**):* office-residential areas in Italy, including the city center of Naples and major industrialized areas around both Milan and Naples (Figs. 2d and 2e). These are the Italian equivalent of OR_{fr}; the major differences are found in the weekend activity, for reasons similar to those discussed for RE_{it}.

*11) Italian land with mixed fabric (**MIX**):* Mixed land use characterizing industrialized cities and towns in the Naples region, whose boundaries fall within single large CDR unit areas. These zones exhibit a common pattern of mobile traffic demand with two comparable peaks from Monday to Sunday.

Takeaways. Mobile phone traffic dynamics allow for effective large-scale classification of land use: both country-specific (e.g., malls, residential behaviors, etc.) and cross-national urban fabrics (e.g., offices, transportation hubs, etc.) are revealed. The major drawback of the CDR-based approach is the very coarse classification of areas without a dense geographical distribution of base stations, e.g., Naples city center or most peri-urban areas in all reference cities.

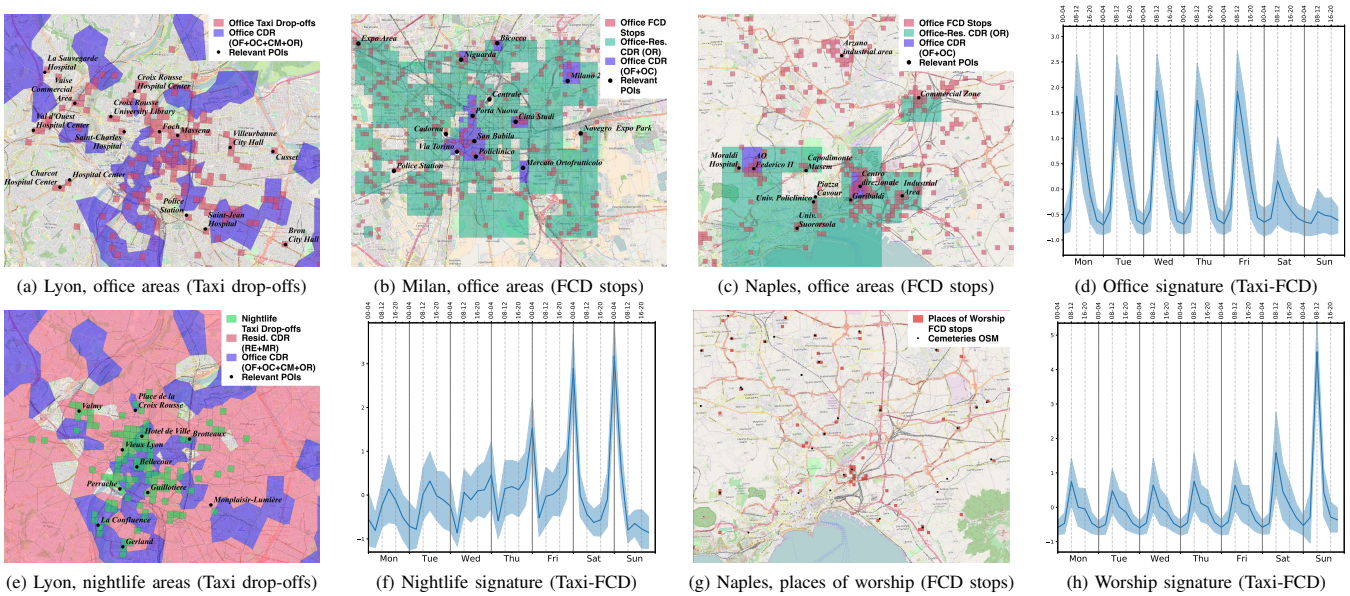
V. Enhancing CDR-based land use classification via taxi drop-offs and floating car trip stops

We now consider taxi and floating car datasets, and investigate how they can improve the CDR-based classification. In Subsec. V-A, we first apply the MWS-clustering approach directly to the datasets Ly-TAXI, Mi-FCD and Na-FCD, and compare the resulting classification with that produced using the CDR datasets: this lets us assess to what extent taxi and floating car data allow identifying land use. Then, in Subsec. V-B, we propose and evaluate a class fusion approach: we use the classes produced by the CDR-based classification to split the TAXI/FCD unit areas, and apply the MWS-clustering to each split separately.

A. Land use classification from taxi and FCD data

The classification of Ly-TAXI, Mi-FCD and Na-FCD produces 49 clusters. Due to space limitation, we briefly discuss a small subset of representative classes in the following.

1) Cross-national land with office fabric: unit areas associated with relatively high volumes of taxi drop-offs and trip stops during morning work hours, as confirmed by the signature in Fig. 3d. Figs. 3a, 3b and 3c draw these unit areas for Lyon, Milan and Naples, respectively, together with CDR-unit areas classified as office-related (i.e., OF, OR, CM and OC) in Sec. IV. Unsurprisingly, a high number of taxi and FCD unit areas overlap with the office-related CDR ones, in all



(a) Lyon, office areas (Taxi drop-offs) (b) Milan, office areas (FCD stops) (c) Naples, office areas (FCD stops) (d) Office signature (Taxi-FCD)

(e) Lyon, nightlife areas (Taxi drop-offs) (f) Nightlife signature (Taxi-FCD) (g) Naples, places of worship (FCD stops) (h) Worship signature (Taxi-FCD)

Fig. 3. Land use classification with taxi drop-offs and trip stops. Maps of related unit areas in representative city scenarios and characteristic signatures.

cities. Interestingly, this class also includes cells in densely populated areas or outskirts where important hospitals, police station buildings or commercial facilities are located. We remark that such areas were not classified as office-related via CDR datasets. Our speculation is that FCD and taxi datasets allow retrieving office-related land use at a higher resolution, especially in peripheral areas with a low base station density.

2) Cross-national land with nightlife fabric: unit areas with high volumes of taxi drop-offs and trip stops late in the evening or overnight, especially during week-end nights (Fig. 3f). A visual inspection of Fig. 3e for Lyon unequivocally pinpoints spaces with high concentration of restaurants, pubs, touristic and nightlife places, as well as peripheral hotels or residential areas, where taxis are used for homecoming. Similar patterns emerge in Milan and Naples. Unit areas in this class are spread across office and residential CDR-related classes (i.e., RE, MR, OR, etc.) for Lyon and Naples, while being mostly concentrated in the NL class for Milan. We argue that taxi/FCD dynamics can complement CDR in presence of nightlife.

3) Land with places-of-worship fabric (Naples): areas in the agglomeration of Naples, with very limited occurrence in either Milan or Lyon. The archetypal signature shows two neat morning peaks on Saturdays and Sundays (Fig. 3h). By comparing the corresponding unit-areas with OpenStreetMap crowd-sourced information, we discovered a very high match with cemeteries all over the city of Naples and neighboring towns (Fig. 3g), which tend to be frequented mainly during week-ends. Indeed, a few other unit areas from this class confirm this pattern, being located in the proximity of churches, parks or diverse sport facilities. This is a highly specific kind of land use, too fine-grained to be discovered via CDR data.

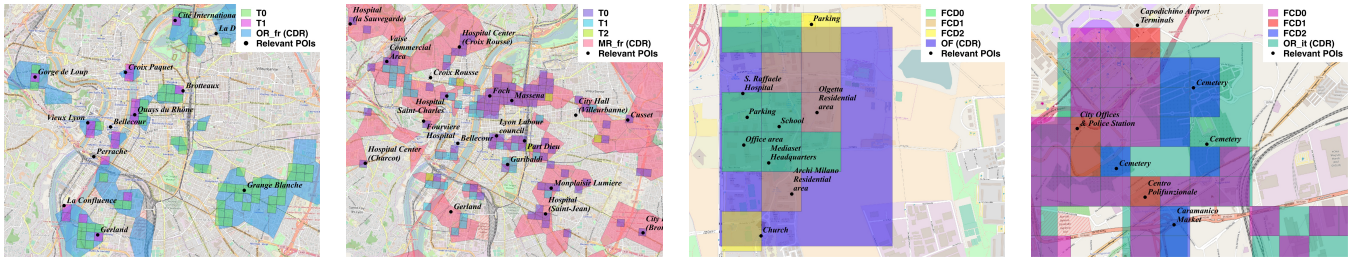
Takeaways. Taxi drop-offs and vehicles' trip stops can be used to retrieve fine-grained land use information. Their dynamics, even when largely aggregated over time, are capable of grasping office, nightlife and residential land use at a much finer spatial resolution than CDR data. Moreover, special fabrics (e.g., places of worship), often found in outskirts, may only be detected via this kind of data. However, taxi and floating vehicles have lower penetration rates than mobile phones, thus covering smaller samples of the observed population and only providing a partial view of land use. It thus appears natural to use them to refine CDR-based land use classes.

B. Land use detection with combined CDR and taxi/FCD data

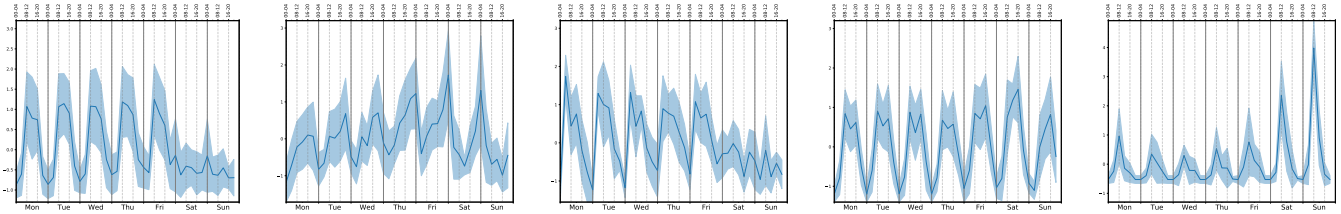
We use the classification results from the clustering of the CDR MWSs (i.e., the set of classes $\mathbf{c}^{\text{CDR}}$ of Sec. IV) to segment unit areas associated to Ly-TAXI, Mi-FCD and Na-FCD. For each class $c \in \mathbf{c}^{\text{CDR}}$ and dataset $\{\mathcal{D}^k\}$ with $k \neq$ CDR, we label the generic unit area a of datasets $\mathcal{D}^k$ as a^c if it intersects any of the CDR unit areas in class c. Then, for each class $c \in \mathbf{c}^{\text{CDR}}$, we apply the MWS clustering to the $\{a^c\}$ set of $\mathcal{D}^k$-unit areas. By using this approach, we preserve the original classification provided by CDR-based land use detection, and identify multiple subregions associated to different taxi or FCD dynamics within the same CDR-class.

1) Splitting the CDR-OR$_{fr}$ class: The clustering of taxi-drop-off MWSs that belong to the OR$_{\text{fr}}$ class produces two main clusters, i.e., *T0* and *T1* (Fig. 4a). The graphical inspection of these two clusters clearly distinguishes two facets of the OR$_{\text{fr}}$ land use: 1) T0 features drop-off dynamics concentrated in the 8-20 time range of work days, and denotes areas with office-residential use (Fig. 4e); 2) T1 has dominant drop-off activities during afternoons and evenings, especially on weekends (Fig. 4f). Indeed, important leisure/nightlife spots are mixed with both office and residential buildings in these unit areas, and therefore differently classified with respect to the T0 unit areas.

2) Splitting the CDR-MR$_{fr}$ class: Taxi drop-off dynamics allow to distinguish three main variations of the MR$_{\text{fr}}$ class (Fig. 4b): 1) the T0 area with taxi drop-offs mostly occurring during work hours in wealthy and commercial neighborhoods; 2) the T1 class characterized by taxi drop-offs mainly during



(a) OR_{fr} and inner taxi clusters, Lyon (b) MR_{fr} and inner taxi clusters, Lyon (c) OF and inner FCD clusters, Milano 2 (d) OR_{it} and inner FCD clusters, Poggioreale



(e) OR_{fr} (T0 signature) (f) OR_{fr} (T1 signature) (g) MR_{fr} (T0 signature) (h) OF (FCD1 signature) (i) OR_{it} (FCD2 signature)

Fig. 4. CDR-based classification refined via taxi drop-offs and FCD trip stops. Maps of unit areas in representative scenarios and characteristic signatures.

evenings in wealthy residential and hotel areas that people might tend to reach by taxis; 3) the T2 class, including popular transport hubs of Lyon that are characterized by the presence of long-range transportation infrastructures and a signature with peaks in the early morning of work days (Fig. 4g).

3) Splitting the CDR-OF class: The clustering of the OF class by means of the FCD MWSs produces 8 classes. To simplify their analysis, we focus on a special neighborhood of the Milan outskirts, namely Milano 2. This area is covered by one single large cell in the Mi-CDR dataset. The neighborhood hosts two large residential areas that are located around the headquarters of a major national television broadcaster, and in proximity of an important hospital. As shown in Fig. 4c, the neighborhood is split into 3 classes by FCD stops. Class FCD1 matches the most important residential neighborhoods. The associated signature presents a typical residential pattern (Fig. 4h). Class FCD0 covers the main office areas of the neighborhood, with a typical office signature. Finally, the FCD2 class covers two spots located at the north and south borders of the neighborhood, where large parking areas are located. The signature shows a main peak in the early morning, due to commuters reaching the parking lot by car.

4) Splitting the CDR-OR_{it} class: Similarly to Milano 2, we analyze the neighborhood of Poggioreale, Naples, classified as OR with CDR data. By clustering the FCD MWSs, three main clusters emerge in this area (Fig. 4d): 1) FCD0 covers the largest part of the neighborhood, with a significant presence of offices mixed with a minor presence of residential buildings; 2) FCD1 groups small zones with a pure office behavior; 3) FCD3 covers the cemetery areas in the neighborhood and a popular local market, mostly frequented during weekends (Fig. 4i).

Takeaways. The chained clustering of CDR and taxi/floating-car MWSs allows joining the large-scale and high-penetration features of mobile phone data to the accuracy and high spatial-resolution power of GPS data. This approach enables the discovery of special dynamics associated to city-center punctual areas, as well as higher levels of accuracy in the classification of peripheral neighborhoods, where the granularity of CDR data could be excessively coarse.

VI. Conclusion

Land use classification represents a fundamental aspect in transportation, telecommunication and urbanism. In this paper, we have described an approach for automatic land use detection based on fusion of mobile phone traffic logs, GPS taxi drop-offs and floating car stops. The evaluation of our solution in a multi-city scenario with real-world datasets shows its effectiveness in unveiling diverse land use classes with finer-grained spatial accuracy than literature to date.

The growing adoption of intelligent transportation systems and autonomous vehicles will increase data availability on real-time human mobility. We are thus planning to exploit such relevant sources of data and devise novel class fusion techniques in order to further increase the resolution power of our approach for land use classification.

References

[1] K. Maat, B. van Wee, and D. Stead, "Land Use and Travel Behaviour: Expected Effects from the Perspective of Utility Theory and Activity-based Theories", *Environment and Planning B: Planning and Design*, 32(1):3346, 2005.

[2] J.L. Toole, et al. "Inferring Land Use from Mobile Phone Activity." *ACM SIGKDD International Workshop on Urban Computing*, 2012.

[3] B. Chen, B. Huang, B. Xu, "Multi-source Remotely Sensed Data Fusion for Improving Land Cover Classification", *ISPRS Journal of Photogrammetry and Remote Sensing*, 124: 27-39, 2017.

[4] G. Pan, et al., "Land-use Classification Using Taxi GPS Traces", *IEEE Transactions on Intelligent Transportation Systems*, 14(1): 113-123, 2013.

[5] A. Furno, R. Stanica, M. Fiore, "A Comparative Evaluation of Urban Fabric Detection Techniques Based on Mobile Traffic Data," *ACM/IEEE ASONAM*, Paris, France, Aug. 2015.

[6] Y. Long, J.C. Thill., "Combining Smart-card Data and Household Travel Survey to Analyze Jobs-housing Relationships in Beijing." *Computers, Environment and Urban Systems*, 53:19-35, 2015.

[7] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, Z. Smoreda, "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas", *IEEE Transactions on Mobile Computing*, 1-14, Dec. 2016.

[8] Telecom Italia Big Data Challenge. [Online]. `http:/www.telecomitalia.com/bigdatachallenge`.

[9] C. Heinze, et al. "Transferring Urban Traveling Speed Model Fits Across Cities," *European Transport Research Review*, 8(3):1-12, 2016.