This is a preprint. Please cite this work as:

Machine Learning and Process Intelligence

Research Group

A stigmergy-based analysis of city hotspots to discover trends and anomalies in urban transportation usage

Antonio L. Alfeo, Mario G. C. A. Cimino, Sara Egidi, Bruno Lepri, and Gigliola Vaglini

Abstract—A key aspect of a sustainable urban transportation system is the effectiveness of transportation policies. To be effective, a policy has to consider a broad range of elements, such as pollution emission, traffic flow, and human mobility. Due to the complexity and variability of these elements in the urban area, to produce effective policies remains a very challenging task. With the introduction of the smart city paradigm, a widely available amount of data can be generated in the urban spaces. Such data can be a fundamental source of knowledge to improve policies because they can reflect the sustainability issues underlying the city.

In this context, we propose an approach to exploit urban positioning data based on stigmergy, a bio-inspired mechanism providing scalar and temporal aggregation of samples. By employing stigmergy, samples in proximity with each other are aggregated into a functional structure called trail. The trail summarizes relevant dynamics in data and allows matching them, providing a measure of their similarity. Moreover, this mechanism can be specialized to unfold specific dynamics.

Specifically, we identify high-density urban areas (i.e. hotspots), analyze their activity over time, and unfold anomalies. Moreover, by matching activity patterns, a continuous measure of the dissimilarity with respect to the typical activity pattern is provided. This measure can be used by policy makers to evaluate the effect of policies and change them dynamically. As a case study, we analyze taxi trip data gathered in Manhattan from 2013 to 2015.

Index Terms—Green intelligent transportation systems (ITS), Global Positioning System, Stigmergy, Emergent Phenomena, Hotspot, Pattern Analysis, Taxi-GPS traces.

I. INTRODUCTION AND MOTIVATION

SMARTNESS and sustainability are two key aspects of the forthcoming transportation systems. Smartness provides transportation monitoring and control with qualities like real-time sensing and fast decision making. Sustainability aims to manage travel demand efficiently by means of environmentally friendly strategies, providing transportation systems with policies for long-term economic suitability [16]. Specifically, a sustainable urban development demands adequate policy instruments aimed to handle and mitigate the increasing volume of traffic congestion, carbon emission, and air pollution. One of the most frequently used policy tools for the measurement and evaluation of transportation sustainability performance are indicators. Indicators can be defined as quantitative measures aimed to explain and communicate complex phenomena simply, including trends and progress over time [19]. In order to provide effective measures of sustainability of transportation activities, it is essential to define indicators purpose and scope. During the last two decades, a number of international initiatives addressed the development of indicators aimed to achieve a more sustainable transportation on the local, regional, and global levels, by involving both scientific community and policy-makers [31]. However, as of today, there is no standard or common agreement about the set of indicators to be used to assess transportation sustainability. Many works in the field perform an impact-based classification by employing a three-dimensional framework based on economic, environmental, and social impacts [17]. The proposed indicators are in general calculated by exploiting commonly available data sources [31]. Thanks to the pervasive technology supporting the smart city strategy, some of these indicators may be calculated via big data fed by on-board or fixed sensors. As an example, data obtained by smart cards [25] can be used to describe the characteristics of public transit usage, such as the number of trips for different transit modes, and travel time distribution for all transit modes and user types. Another example could be the GPS-enabled vehicles, which can provide a more comprehensive view of the factors shaping transportation emissions and efficiency, by analyzing passenger occupancy and trip density by location and time [26]. Finally, air quality monitoring systems can be used to monitor local pollution emission([27],[29]) and noise emission [28]. These sources allow enhancing the precision of the investigation, providing insights about sustainability issues on specific urban locations and the moment in time. According to [30], [31], [32], [33], some of the widely accepted indicators assessing the sustainability of transportations are derivable from localized sensors data. Below we present a list of indicators arranged according to their addressed purpose.

1) Transportation system efficiency:

   a) Total vehicle-mileage traveled by motorized (private and public) and non-motorized (bikes and pedestrians) traffic participants;

   b) Portion of travels made by private car;

   c) Total vehicle-mileage traveled in urban-peak con-

Antonio L. Alfeo is with the Department of Information Engineering, both at the University of Pisa, (largo Lucio Lazzarino 1, Pisa, Italy) and the University of Florence (Via di Santa Marta, 3, Florence, Italy). Email: luca.alfeo@ing.unipi.it,

Mario G. C. A. Cimino, Sara Egidi, and Gigliola Vaglini are with the Department of Information Engineering, University of Pisa. Email: mario.cimino@unipi.it, s.egidi1@studenti.unipi.it, gigliola.vaglini@unipi.it.

Bruno Lepri is with Bruno Kessler Foundation, via S. Croce, 77, Trento, Italy. Email: lepri@fbk.eu.

ditions;

    d) Delay per trip;

    e) Vehicle occupancy per travel (as passenger-vehicle);

    f) Frequency and mean duration of traffic congestions;

2) Pollutions levels:

    a) Total emissions of greenhouse gases (e.g. CO_2);

    b) Total emissions of air pollutants (e.g. PM_{10});

    c) Total emissions of noise pollutants;

3) Land use:

    a) Percentage of population living in proximity to public transportation facilities (e.g. transit station);

    b) Total land area consumed by transportation infrastructure;

    c) Total land area consumed by cars;

According to the moment in time and the location observed, an indicator may show different patterns. The detection of significant patterns usually involves a complex system modeling, due to largeness and complexity of data underlying the indicator. Thus there is the need for a data mining algorithm aimed to provide pattern identification, detection of behavioral regularities, and comparison between different traffic phenomena [18].

In this paper, we propose a novel approach to analyze urban positioning data. Both complexity and largeness of data can be handled by employing computational techniques belonging to the *emergent paradigm*. Emergent paradigms allow avoiding the explicit modeling of dynamics, which works only under the assumption formulated by the designer. Emergent paradigms, instead, offer model-free computational approaches, characterized by adaptation, autonomy, and self-organization of data [8].

In particular, the emergent paradigm provides a biologically inspired aggregation mechanism known as *stigmergy*. By using stigmergy each sample position is transformed into a digital pheromone deposit. Deposits accumulate with each other according to their proximity while progressively evaporating. This simple mechanism intrinsically embodies the time domain, thus it is used to unfold significant spatiotemporal dynamics in the data [9].

Among all the available sources of movement data for smart city application, location aware vehicles provide us the opportunity to help urban planners and policy makers in mitigating traffic, planning for public services and resources, and properly manage infrequent events [23]. However, both public transportation and private vehicles provide quite predictable GPS traces, because they are due to predetermined routes or personal routines (i.e. to and from work). On the other hand, GPS-enabled taxis, represent both a transit-complementary door-to-door transportation mode and a source of real-time human mobility information [1]. Indeed, taxicabs play a prominent role as a transportation mode in metropolitan areas, e.g., in New York City, over 100 companies operate more than 13,000 taxicabs with a daily demand of 660,000 passengers [21]. Moreover, by continuously serving a wide diversity of passengers in the city, taxis GPS traces can provide a detailed

glimpse into motivation and characterization of populations urban mobility. However, regular taxicab services becomes inefficient during urban-peak conditions, e.g., extreme weather or special events ([22], [24]), producing unnecessary traffic, pollution, energy consumption, and causing the increase of passenger's waiting time [44]. Thus, further investigation aimed at analyzing taxi-based transportation system is needed. In this context, an exhaustive survey of taxis' trip data analysis is provided in [2]. According to *Castro et al*, existing works can be divided into 3 categories: (i) *Social Dynamics*, which aims to analyze the collective behavior of urban population, detect the most visited areas in the city (hotspots), characterize their functionalities, and study the mutual relationships among different urban areas; the purpose of this kind of analysis is to determine the effectiveness of transportation systems and to provide guidance for needed improvements; (ii) *Traffic Dynamics*, i.e. the analysis of the traffic flow through the city road network, providing results about congestion durations and levels in a given city area; congestions have a remarkable impact over the travel time and on the occurrence of adverse events (e.g., accidents), and can be used to estimate pollution levels [36]; and (iii) *Operational Dynamics*, i.e. the analysis of taxis trajectory aimed to investigate taxi drivers behavior, providing route planning insights for unfolding anomalies in urban mobility. According to the authors in [2], many works in the field deal with more than one of the proposed categories in their analysis, and this is the case of our work. Specifically, we exploit taxis' trip data provided by Taxi and Limousine Commission (TLC) of New York City. All taxis of NYC are equipped with FCD (floating car data) devices, which manage localization and card payments data, and enable taxicab drivers and passengers to receive information from the Taxi and Limousine Commission [14]. FCD records include pick-up and drop-off positions, timestamp, and number of passengers, which feed the Taxi Trip Origin-Destination (OD) dataset.We analyze it in order to uncover city hotspots, characterize human mobility patterns, and detect anomalous occurrences by using an approach based on stigmergy.

The paper is organized as follows. In section 2 we discuss the related works. In section 3 we present our approach. We detail data preprocessing and experimental setup in section 4. In section 5 results obtained by analyzing taxi traces gathered in NYC during the years 2013, 2014 and 2015 are shown. Finally, we conclude our work and discuss future avenues of research in section 6.

II. RELATED WORK

Recently, the wide availability of taxi trip data has produced a significant number of works aimed to mine urban dynamics by exploiting this kind of data.

In [38] the authors use non-negative matrix factorization (NMF) algorithm to decompose taxi activity levels and extract three basic patterns. Those patterns represent respectively: (i) commuting between home and workplace, (ii) business traveling between two workplaces, and (iii) leisure trips from or to other places. Furthermore, authors model the relative daily deviation of the traffic flow in each category. In [35]

authors analyze taxi traces in order to model the typical pattern of passenger flow in an urban area; by applying this model authors were able to compute the probability that an event happened, and measure the impact of the event by analyzing anomalous patterns in passenger flow via Discrete Fourier Transform. An Interactive Voting-based Map Matching Algorithm is used in [41] to map GPS trajectories. This mapping is aimed to characterize typical drivers behaviors and discover abnormal ones. Finally, the authors mine the cause of the anomaly by checking data gathered by social networks.

One of the main issues concerning the analysis of this kind of data is their dimensionality. Many approaches handle it by focusing on specific areas (hotspots) whose high concentration of events or samples can summarize the most relevant dynamics occurring in data [43], [37].

In the literature, urban hotspots are typically divided in two categories: (i) regular and (ii) occasional. Areas comprising many points of interest such as movie theaters, commercial buildings, hospitals, schools, colleges, etc. are prime examples of regular hotspots. Occasional hotspots are those areas where any incident has taken place. An incident is defined as an unexpected event that temporarily disrupts the mobility flow, e.g., car crash, marathon, VIP passing area, etc. However, the most of the studies firstly consider regular hotspots. Li. et al [39] proposed and developed an improved auto-regressive integrated moving average (ARIMA) for detecting urban mobility hotspots using taxi GPS traces; moreover, the patterns of pick-ups occurring in these city locations are extracted and analyzed. Other works, such as the one from Makrai [40], provide a statistical approach for the detection of hotspots in New York City by means of a distributed environment. Authors in [10] use OPTICS in order to exploit taxi drop-off positions, extracting hotspots from density-connected point clusters. Cluster results are then assigned as daily taxi drop-off hotspots. Recently, Lu et al. [42] developed a monitoring system performing spatiotemporal analysis on taxi trip data to find seasonal hotspots. This result is achieved by using DBSCAN clustering algorithm with pick up and drop-off locations every fixed amount of time.

III. APPROACH DESCRIPTION

Our idea is to represent taxi traces as deposits of pheromones, akin to the ones ants release when exploring and searching for food. Isolated pheromones progressively evaporate and disappear, whereas pheromones which are subsequently deposited in proximity with each other aggregate as a trail. The trail guides the ants while seeking for food, providing them an effective self-organization mechanism, known as stigmergy [6]. The emulation of this mechanism in the context of data processing enables the unfolding of spatial and temporal dynamics in data [9] by providing information self-organization [8]. Specifically, the principle of stigmergy is used (i) to discover a set of locations (hotspots) characterized by the highest spatiotemporal density in data; (ii) to unfold hotspot activity patterns, (iii) to match those patterns in order to detect anomalies in hotspot activity. The overall processing schema has been developed and integrated with the MATLAB framework.

A. Hotspot Discovery

With the aim of discovering locations characterized by high density of taxi passengers activity (activity, for short), we exploit the number of people being picked up or dropped off in a given time slot and in a given location. During each step of the analysis, samples corresponding to a given time slot are provided to the system and processed by a four-stages procedure. At the beginning, the Smoothing process removes irrelevant samples' values and highlights significant ones (Fig.1b), by treating them with a sigmoidal function. The marking process releases a mark in a tridimensional spatial environment in correspondence of each smoothed sample position (Fig.1c). Each mark is defined by a truncated cone with a given width and an intensity (height) equal to the sample value. The marks aggregate form the trail, whose intensity is subject to evaporation, i.e., the trail intensity is decreased by a constant value δ at each step of the analysis (Fig.1d). Eq. 1 describes the trail T at time instant i.

$$T_i = (T_{i-1} - \delta) + Mark_i \tag{1}$$

Due to aggregation and evaporation provided by the trailing process, sparse marks progressively disappear. On the contrary, subsequent deposits in proximity with each other counteract the evaporation producing a stable trail, which highlights spatiotemporal density in data.

The hotspots are determined as the overlap of the city areas corresponding to the most relevant trails, obtained by analyzing data corresponding to early morning (i.e., 3a.m.-8a.m.), morning (i.e., 9am-2pm), afternoon/evening (i.e., 3pm-8pm), and night (i.e., 9pm-2am) time slots. Each hotspot is represented by a set of coordinates that bound a city area inside a polygon. The hotspots identified in Manhattan (New York City) are shown in Fig. 1e. Their locations correspond to East Harlem - Upper East Side (A), Midtown East (B), Broadway (C), East Village - Gramercy - MurrayHill (D), Soho - Tribeca (E), Chelsea (F) and Time Square - Midtown West - Garment (G).

B. Stigmergic Receptive Field

The result of the hotspot identification is a set of urban areas in which the most relevant activity dynamics occur. For each of them, we extract the activity time series by gathering the amount of activity occurred in the hotspots during the day. The hotspot activity time series are analyzed by means of the Stigmergic Receptive Field (SRF), a stigmergy-based processing schema which provides a similarity measure of a given couple of time series ($a(k)$ and $\bar{a}(k)$ in Fig. 2).

The overall architecture of the SRF is composed by a five-stages processing pipeline, shown in Fig. 2.

Input samples are firstly treated by the Clumping process. It is aimed to reduce microfluctuation in data while highlighting the dynamics occurring among relevant information levels. In our application, 3 levels are used to specify the activity behaviors, namely Low, Medium and High. The Clumping process is implemented by treating each input sample with a double sigmoid function, which is parametrized by means of
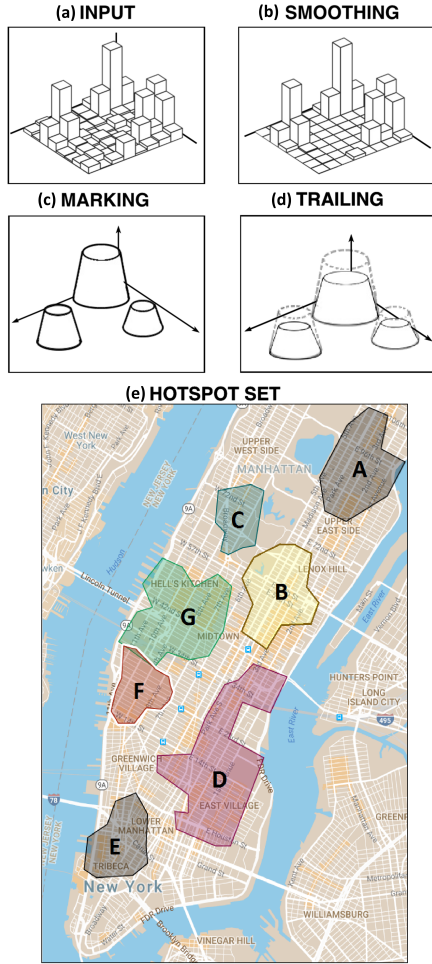
Fig. 1. The overall process of hotspots discovery in Manhattan.

$$f(x, \alpha_a, \beta_a) = \frac{1}{(1 + e^{-\alpha_a(x - \beta_a)})} \tag{3}$$

In order to have an effective sample processing the SRF should be properly parameterized. For example, low trail evaporation causes early activation, whereas high trail evaporation generates a trail consisting of the latest marks only, and preventing pattern reinforcement. Specifically, the SRF parameters are: (i) the clumping inflection points $\alpha_{c1}, \beta_{c1}, \alpha_{c2}, \beta_{c2}$; (ii) the mark width ϵ; (iii) the trail evaporation δ; and (iv) the activation inflection points α_a, β_a. The SRF parameters are adjusted by the Adaptation process. It uses the Differential Evolution (DE) algorithm [37], in order to minimize Mean Square Error (MSE, Eq. 4), which is computed as the difference between desired ($S'(h^*)$) and actual ($S(h^*)$) output value on a training set of N labeled couples of activity time series ($\{a(k^*), S'(h^*)\}$).

$$Fitness = \frac{\sum_{i=1}^{N}(|S(i^*) - S'(i^*)|^2)}{N} \tag{4}$$

C. Stigmergic Perceptron

Let us suppose to have a pure form time series which embodies a behavioral class (i.e., an archetype). The SRF can detect this specific behavior in the actual time series, by processing it together with the archetype. An example of behavioral class in our domain is Rush-Hour (Fig. 4a), which correspond to the behavior of the activity occurring in the hotspot when people movement is at its highest rate. Other classes provided are Asleep (Fig. 4g), i.e., the hotspot at its lowest activity level; Falling (Fig. 4f), i.e., the transition between regular activity and its calm down; Awakening (Fig. 4e), i.e., the waking up of urban activity following a calm phase; Flow (Fig. 4d), i.e., the hotspot at its operating capacity; Chill (Fig. 4c), i.e., the calm down of the hotspot activity after a rush hour; Rise (Fig. 4b), i.e., the transition to the most intense activity level.

A set of SRFs aimed to recognize these archetypes can be arranged into a connectionist topology, obtaining a Stigmergic Perceptron [5]. By forming a linear combination of the SRFs outcomes, the Stigmergic Perceptron provides an assessment of the current behavior of the input time series among all the classes provided. Specifically, the output of the SP is calculated as the average of the SRFs enumerations (represented as 1-to-7 in our application case) weighted by their output similarities (Eq. 5). The SP output is called activity level and is defined between 0 and N (i.e., the number of archetypes).

$$ActivityLevel = \frac{\sum_{i=1}^{N}(S(i) * i)}{\sum_{i=1}^{N}(S(i))} \tag{5}$$

In order to prevent multiple activations of SRFs in the same SP, their Adaptation process is a two phases procedure: (i) the Global Training phase is aimed to provide a suitable interval for each SRFs parameters according to their sensitivity; specifically, the interval for the evaporation rate, that is the most sensitive SRF parameter, is determined considering the narrowest interval including the fitness values above the 90th

its inflection points ($\alpha_1, \beta_1, \alpha_2, \beta_2$). In correspondence of each clumped sample, a mark is released in a bidimensional virtual environment by the Marking process. The mark is defined as a trapezoid with given intensity (i.e., height) and width (ϵ). The Trailing process handles the aggregation of the marks in the trail, and the trail temporal decay, i.e., the decrease of its intensity of a given amount δ at each step of time. The trail can be considered as a short term memory summarization of the spatiotemporal dynamism occurring in data.

Both time series ($a(k), \bar{a}(k)$) provided to the SRF undergo these processing stages independently, until the Similarity process compares their trails (T_1, T_2) using the Jaccard coefficient [3] as defined in Eq. 2.

$$S = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \tag{2}$$

The provided similarity values are between 0 (completely different trails) and 1 (identical trails). Finally, the Activation process transforms the similarity values by treating them with a sigmoidal function (Eq. 3). Specifically, activation process lowers insignificant similarity values and enhances relevant ones depending on the values of the sigmoid inflection points (α_a, β_a).
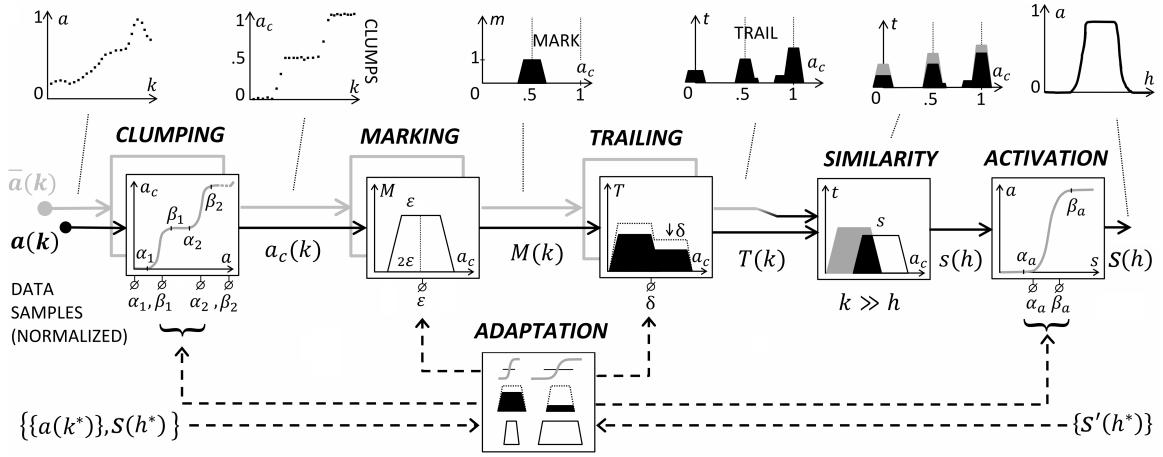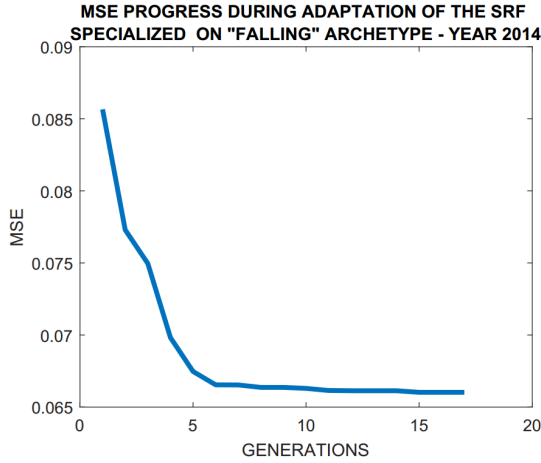
Fig. 2. The architecture of a stigmergic receptive field.

Fig. 3. The progressive decrease of the error in the assessment of the "Falling" archetypes (in year2014), during adaptation process.

Fig. 4. The overall processing of the activity samples.

percentile, while the intervals for the other parameters can be provided according to the application domain constraints; (ii) the Local Training phase is aimed to find the optimal values for every parameter and each SRF by exploiting their Adaptation process and the interval determined in the Global phase; the training set for each SRF is made by half signals belonging to its behavioral class, and half belonging to the behavioral classes of adjacent SRFs. As an example, the progressive decrease of the MSE provided by this adaptation phase is shown in Fig. 3.

A properly trained SP produces a time series of activity levels by transforming a given time series of activity samples. The activity levels time series can be considered as a higher level characterization of the hotspot activity during an entire day. The source code of our Stigmergic Perceptron has been publicly released on the MATLAB Central File Exchange [46].

D. Anomaly Degree Computation

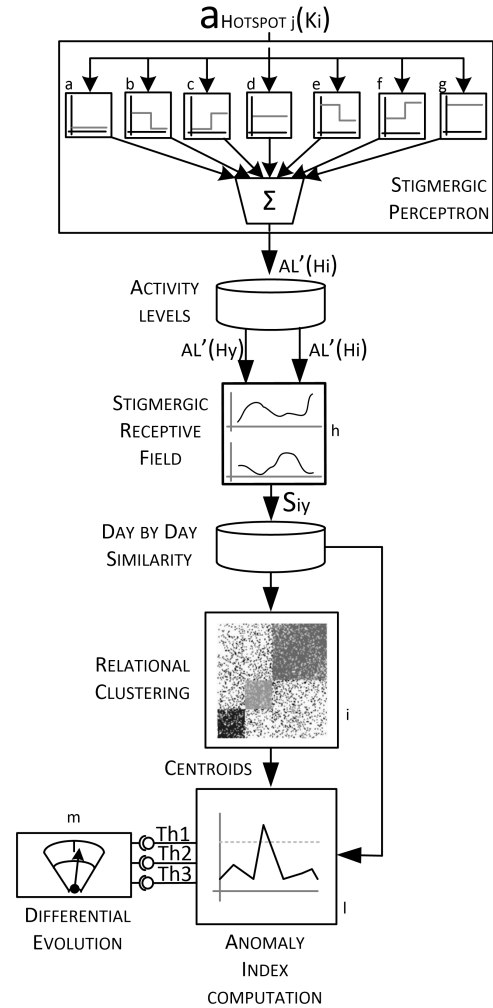In order to detect anomalous activity level patterns, we employ a further SRF aimed to measure of the similarity between two activity levels time series gathered in different days (Fig. 4h).

The Adaptation process of this SRF aims to minimize the difference (in terms of MSE) between computed and ideal
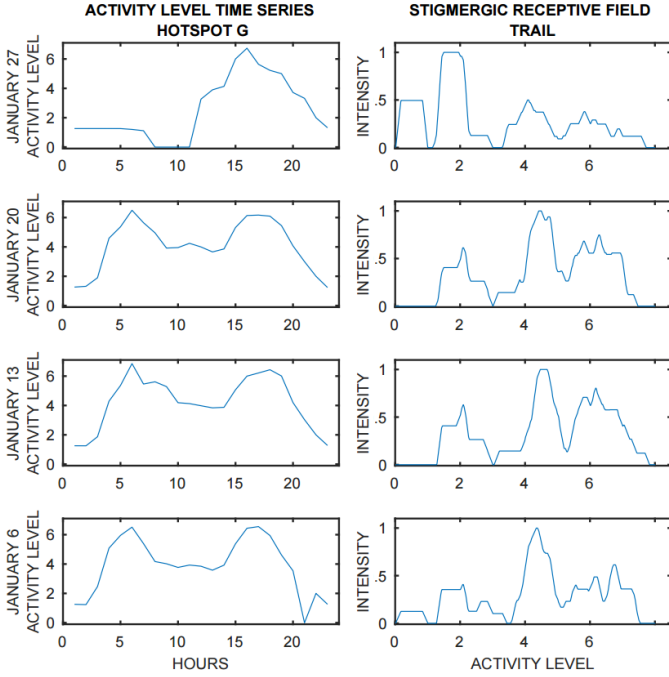
Fig. 5. Examples of trails generated by the activity level time series gathered in different working days. Jan-27 seems to be characterized by an anomalous pattern.

Fig. 6. Similarity matrix obtained by analyzing patterns gathered during year 2015.

similarity over a training set, by tuning mark width ϵ, trail evaporation δ, and activation thresholds α_a, β_a. The training set is composed by different (i.e., 900) couples of activity level time series whose similarity is supposed to be 1, if they belong to the same activity level behavioral class, 0 otherwise. As an example, an activity level behavioral class can be "Working-Day" which is a day whose activity is mainly affected by working routines. Other activity level behavioral classes provided in our analysis are "Entertainment-Days" (usually occurring in Fridays and Saturdays) and "Leisure-Days" (usually occurring in Sundays).

As an example, Fig. 5 shows the activity level obtained by analyzing activity in the hotspot G in every Tuesday of January 2015. The corresponding stigmergic trails are shown beside them. Here, an activity level time series differs from the other working days shown. Indeed, January 27 is characterized by an anomaly, due to winter storm 'Juno' that was hitting the city during that day.

By exploiting the SRF similarity measure we can match all the activity level time series of the SRF's training set and store their similarity values into a similarity matrix. As an example, in Fig. 6, we show the similarity matrix obtained by analyzing patterns gathered during the year 2015. Patterns are arranged by behavioral class, i.e., Working-Day (days 1-10), Entertainment-Day (days 11-20), and Leisure-Day (days 21-30). Here, the similarity value obtained by matching two patterns gathered in different days is represented by the color of the corresponding box. The whitest the box, the higher the similarity. As expected, the similarity values appear to be higher only with couples of days belonging to the same activity level behavioral class.

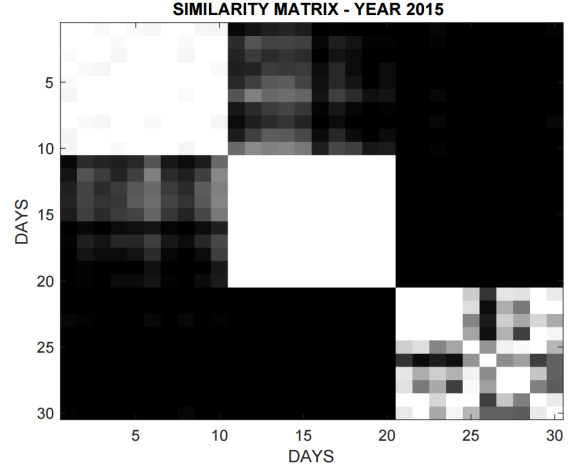This similarity matrix is processed by a relational clustering technique (Fig. 4i) in order to group similar daily activity levels. Specifically, we employ Fuzzy C-Means using as number of clusters the number of daily activity behaviors taken into account in the analysis (i.e. 3). The fuzzy clustering generates, for each activity level time series, a membership degree for each behavioral class. We can determine which are the most representative days for each behavioral class by selecting the activity level time series which are closest to the corresponding cluster centroid in terms of euclidean distance.

An activity level time series obtained by a typical day is expected to exhibit high similarity with respect to the activity level time series obtained by the most representative days of its behavioral class. Thus, we compute the Anomaly Index of current day d by exploiting the average of its similarity $S(d, i)$ with respect to its most representative N (i.e., 5) days, as detailed in Eq. 6. The Anomaly Index is defined between 0 (typical daily behavior) and 1 (very anomalous daily behavior).

$$AnomalyIndex(d) = |\frac{\sum_{i=1}^{N}(S(d,i))}{N} - 1| \qquad (6)$$

In order to discern typical days from anomalies, an Anomaly Index threshold for each activity level behavioral class must be defined. These thresholds have been determined by using DE (Fig. 4m) in order to minimize the classification error (i.e., the percentage of correctly classified days) over all the days of the year under analysis, given a set of known anomalies. As an example, Fig. 7 shows the classification of each day gathered during the year 2015. Known anomalies are characterized by a circle. The Anomaly Index thresholds are in dashed line.

IV. EXPERIMENTAL SETUP

We analyze data provided by the Taxi and Limousine Commission of New York City, containing details about all taxi trips occurred during 2013, 2014 and 2015 in Manhattan [12]. Each trip is reported with its taxi ID, number of passengers, together with latitude, longitude, and time-stamp of pick-up and drop-off. Data have been pre-processed in order to (i) remove missing values and (ii) discretize data in
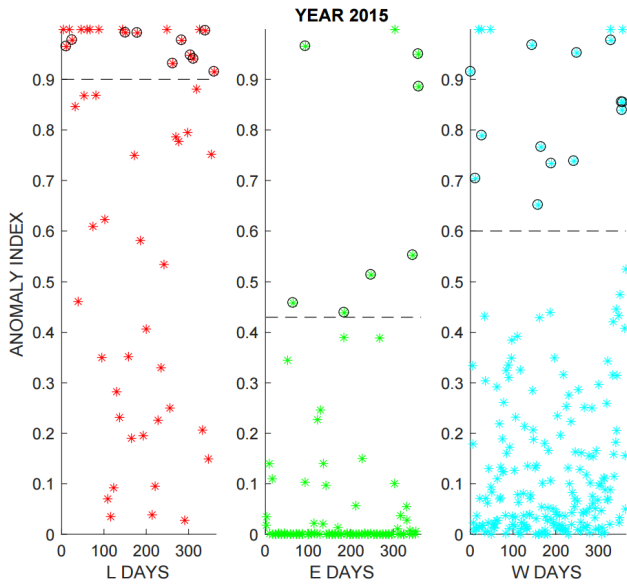
Fig. 7. Classification of anomalies (circle) and typical days (stars) obtained analyzing data gathered during 2015.The thresholds are reported in dashed line.

spatiotemporal buckets characterized by length and width of 10 foot, and duration of 5 minutes.

For the hotspots investigation, the period under analysis comprised both February and June. This period has been chosen since it can capture different seasonal behavior without being influenced by the presence of many holidays.

The activity time series extracted for each observed day has been normalized by using the min-max procedure.

Both global and local training phases of the SP are provided with a training set generated by applying random spatial noise and temporal shift to the pure archetype time series. The SP training set is composed of 70 time series (10 for each SRF), and the expected similarity is 1 if the current time series has been generated by the archetype on which current SRF must be specialized, 0 otherwise.

The SP outcome is processed by a further SRF, which is trained to measure activity level time series similarity, according to the daily behavioral classes provided, namely: (i) Working days which are expected to fall between Monday and Tuesday, when commuters and working routines deeply affects the crowd movements; (ii) Entertainment days, which are expected to fall on Friday and Saturday, and are characterized by high nocturnal activity due to the nightlife; (iii) Leisure days, which are expected to fall on Sunday, and are characterized by minor transportation usage. The training set is composed of 30 activity level time series, i.e., 10 time series representing the typical patterns of each behavioral class. The target similarity of each possible match is 1 if time series falls in the same behavioral class, 0 otherwise.

The typical pattern of each behavioral class has been determined by performing an exploratory analysis of the activity time series according to the following features: (i) A label (low, medium, high) describing the range of the values contained in the initial part of the time series; (ii) A label (low, medium, high) describing the range of the values contained in

the final part of the time series; (iii) The instant in which the mean value of the whole series is reached (initial and final part are not considered); (iv) The instant in which the maximum value of the whole series is reached (initial and final part are not considered); (v) The duration of the time series at its higher level. We select different activity series for each behavioral class (Working, Entertainment, and Leisure) in order to extract these features. A suitable range of values for each feature and each behavioral class is provided by choosing the minimum and the maximum among the values obtained. By means of this features range, we can describe the overall typical pattern of each behavioral class. Moreover, the time series whose features exceed the features range of its expected behavioral class, can be annotated as an anomaly. As an example, the features of an activity time series gathered during a Thursday, are expected to fit the feature ranges of Working-Day behavioral class. If it does not occur the time series is annotated as an anomaly.

V. RESULTS

Depending on the land usage of the city area underlying each hotspot, some daily activity behaviors may not emerge. As an example, the Entertainment-Day behavior is mainly caused by the presence of clubs or other entertainment-oriented business that may attract the nightlife. Thus, the hotspots underlying a mixed usage zones are the most promising ones for the analysis, since our aim is to characterize all the aspects of the city life. According to official land use (publicy available at [13]), each city block can be classified into the following categories: commercial, residential, industrial, transportation space, institutional, open/recreational space, parking or vacant. By considering the distribution of these categories in each hotspot it can be evaluated how diversified the usage of that area is, and therefore the related amount of the mobility dynamics. Specifically (i) Hotspot A is primarily residential and secondly institutional; (ii) Hotspot B is mainly residential and commercial; (iii) Hotspot C is principally open space and residential; (iv) Hotspot D and E are characterized by an equal distribution of almost all usage classes; (v) Hotspot F is mainly commercial and residential, with some institutional blocks; (vi) finally, Hotspot G presents all usage categories, with a prevalence of the commercial category. The higher the variety of the usage of a hotspot, the better a candidate this hotspot is for our analysis: then hotspots D and E are chosen. For these hotspots, and for 2015, Table I shows the percentage of correctly classified patterns (among the normal and anomalous classes) obtained with 5 different trials in the form mean $\pm$ 95% confidence. The classification performance is also calculated by using two well-known time series distance measures: the Dynamic Time Warping [11], and the Fréchet distance [45]. Clearly, the SRF measure outperforms both the DTW and the Fréchet distances. In order to further test the assessment of anomalous patterns, the activity time series annotated as anomaly have been annotated by a triple according to their affinity with the typical pattern of each behavioral class. As an example, the triple W|E|L means that current time series is mostly similar to Working-Day typical pattern and secondly to Entertainment-Days one, whereas it shows only minor similarity with respect

Similarity Measure	Hotspot D	Hotspot E
SRF	95.61 ± 0.003	94.24 ± 0.24
DTW	90.57 ± 0.134	91.80 ± 1.387
FRECHET	90.14 ± 0.537	90.52 ± 2.34

Year	SRF	DTW
2013	92.71 ± 0.321	90.57 ± 0.134
2014	96.65 ± 0.109	92.27 ± 0.106
2015	95.61 ± 0.003	91.28 ± 0.106

to Leisure-Days. With the aim for measuring the capability of our similarity measure to generate a corresponding affinity assessment, we compute the average of the similarity of each time series annotated as an anomaly with respect to the most representative days of each behavioral class. Sorting them by similarity, we obtain the triple. The Mean Assessment Error is computed as the number of non-matching sort constraints for each pair of triples, averaged over all the set of the anomalies. As an example, the triples W|L|E and W|E|L have just one non-matching sort constraint, which is L<E, whereas both triples state that W<E and W<L.

A comparison is provided by repeating this procedure using the Dynamic Time Warping [11] distance. Resulting Mean Assessment Error are equal to 1.135 (SRF-based similarity measure) and 1.115 (DTW distance). According to these results both methods are suitable for pattern analysis, thus we provide the comparison of their performance in anomalous pattern detection.

Specifically, in order to compare the classification performances of our approach with respect to DTW, we collect the percentage of correctly classified days among 5 trials. During each trial, the DE generates a new set of Anomaly Index thresholds. If the Anomaly Index of an activity level time series exceeds the threshold, the corresponding day is considered anomalous. Obtained results are presented in the form mean $\pm$ 95% confidence interval in Table II.

Based on obtained results, our approach provides an effective detection of major anomalies. But, handling minor or potential anomalies could be more difficult. In order to evaluate the effectiveness of our measure while handling this kind of anomalies, we select a set of events including official holidays and days affected by special events with documented effect on the road in (or in close proximity of the) hotspot D. Such events could be days characterized by adverse weather condition (e.g., Juno storm), street closure (e.g., due to the Gay Pride parade) and so on. This set is provided for each year under analysis.

A set of ordinary days is also included. An effective anomaly measure is supposed to exhibit high correlation between its value and the set (events or ordinary days) which
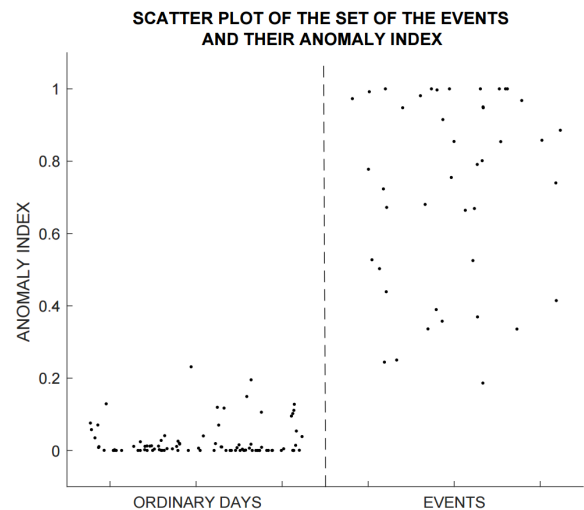

Fig. 8. Scatter-plot generated by considering Events, Typical Days and their corresponding Anomaly Index.

Year	SRF	DTW
2013	0.8963	0.7177
2014	0.9289	0.7236
2015	0.9210	0.6828

current day belongs to. In Fig. 8 the correlation obtained by using our SRF-based measure is shown.

In order to compare obtained results in terms of correlation between events and computed Anomaly Index, we provide it by using SRF-based approach and DTW. In Table III we present obtained correlation coefficient for each year under analysis.

VI. CONCLUSION

A novel approach for anomaly discovery in the context of taxi trip data have been proposed in this paper.

Our approach was able to identify city hotspots, characterize the daily patterns of their activity over time and detect days characterized by an anomaly. Our approach has been tested on real world dataset containing all taxi trips occurred in Manhattan during 2013, 2014 and 2015, for a total amount of 74GB of data.

On the basis of the results of this work, several conclusion have to be drawn.

First of all, we prove the suitability of our approach for the analysis of "big data". This is due to the information abstraction provided by our multilayer topology.

Moreover, it is worth recalling that our approach exploits the information self-organization obtained by using the principle of stigmergy and does not require the in-depth modeling of the dynamics under investigation. This quality together with the adaptation provided by the SP training process, allows to specialize the detection on any occurred dynamics

despite minor differences. Indeed, the performances of our approach have not been affected by differences among the typical activity patterns of each year. The source code of our Stigmergic Perceptron has been publicly released on the MATLAB Central File Exchange [46].

The performances of our approach is measured in terms of percentage of correctly classified daily patterns among typical and anomalous ones. Moreover, the effectiveness in handling minor anomalies is measured by computing the correlation between the Anomaly Index and the occurrence of urban events (e.g., local parade or official holiday).

Obtained results have been compared with respect to the one achieved by analyzing the activity time series with DTW. In both cases, and in every year under analysis, our approach outperforms DTW, achieving up to 96.65 percentage of correctly classified days and 0.9289 correlation coefficient.

Although mobility in Manhattan can be characterized by many dynamics for each day, both typical and exceptional, the stigmergy-based approach has proven to be a withstanding method to enhance significant patterns in taxi trip data, letting them emerge autonomously.

Finally yet importantly, our approach provides a continuous measure (the Anomaly Index) of the divergence with respect to typical activity patterns, which can be used by policy makers to evaluate the effect and the resilience of proposed policies or change them dynamically according to provided measures. Our approach can be potentially applied to any sustainable transportation indicator, to distinguish different spatiotemporal patterns occurring over time. For each indicator and for each basic pattern, a subset of labelled data is required to enable a proper calibration and interpretation of what can be discovered. This study focuses on a complete description, experimentation and discussion of our approach via a specific indicator extracted from a real world taxi trips data set. A more extensive experimentation on multiple indicators is clearly possible but it is out of the scope of this work. Thus, one of the most promising improvements for this study consists of cross-checking results obtained by exploiting different data sources and indicators. As an example, by incorporating pollution data or car crash data into the analysis, policy makers could gain insight about the mutual interaction of this phenomena, which is a fundamental knowledge for producing effective policies.

References

[1] M. Veloso, S. Phithakkitnukoon, C. Bento, Urban mobility study using taxi traces. ACM, In Proceedings of the 2011 international workshop on Trajectory data mining and analysis, 23–30 (2011).

[2] P.S. Castro, D. Zhang, C. Chen, S. Li, G. Pan, From taxi GPS traces to social and community dynamics: A survey. ACM Computing Surveys (CSUR), vol 46(2): Article No. 17, (2013).

[3] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, Using of Jaccard coefficient for keywords similarity. In Proceedings of the International MultiConference of Engineers and Computer Scientists, 1, 6, (2013).

[4] A.L. Alfeo, M.G.C.A. Cimino, A. Lazzeri, G. Vaglini, "Detecting urban road congestion via parametric adaptation of position-based stigmergy", Intelligent Decision Technologies, IOS Press, Pages 1-28, 2017, (ISSN 1872-4981)

[5] M. Avvenuti, C. Bernardeschi, M.G.C.A. Cimino, G. Cola, A. Domenici, G. Vaglini, "Detecting elderly behavior shift via smart devices and stigmergic receptive fields", in Proc. EAI The 6th EAI International Conference on Wireless Mobile Communication and Healthcare (MO-BIHEALTH 2016), pp. 1-8, Milan, Italy, 2016.

[6] G. Theraulaz, E. Bonabeau, A brief history of stigmergy. Artificial life, vol 5(2), pp. 97-116, (1999).

[7] A.L. Alfeo, F.P. Appio, M.G.C.A. Cimino, A. Lazzeri, A. Martini, G. Vaglini, "An adaptive stigmergy-based system for evaluating technological indicator dynamics in the context of smart specialization", in Proc. INSTICC The 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016).

[8] D. Vernon, G. Metta, G. Sandini, A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents. IEEE Transactions on Evolutionary Computation, 11(2), 151–180, (2007).

[9] A.L. Alfeo, P. Barsocchi, M.G.C.A. Cimino, D. La Rosa, F. Palumbo, G. Vaglini, "Sleep behavior assessment via smartwatch and stigmergic receptive fields", Personal and Ubiquitous Computing, Springer, Pages 1-17, 2017.

[10] A. Keler, M. Jukka Krisp.: Is there a relationship between complicated crossings and frequently visited locations? A case study with boro taxis and OSM in NYC. 13th International Conference on Location-Based Services, (2016).

[11] J. Taylor, "Method for investigating intradriver heterogeneity using vehicle trajectory data: a dynamic time warping approach." Transportation Research Part B: Methodological 73 (2015): 59-80.

[12] NYC.gov, Taxi and Limousine Commission (TLC) Trip Record Data, http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

[13] Zola, interactive map for zone and land use of NYC. http://maps.nyc.gov/doitt/nycitymap/template?applicationName=ZOLA

[14] Taxicab Passenger Enhancements Project (TPEP) Archive. http://www.nyc.gov/html/tlc/html/industry/taxicab_serv_enh_archive.shtml

[15] B. Schaller, (2006). The New York city taxicab fact book. Schaller Consulting, mars. At: http://www.schallerconsult.com/taxi/taxifb.pdf

[16] M.M. Haque, H.C. Chin, A.K. Debnath, Sustainable, safe, smartthree key elements of Singapores evolving transport policies. Transport Policy, Vol. 27, pp. 20-31, (2013).

[17] A. Dobranskyte-Niskota, A. Perujo, M. Pregl, (2007) Indicators to Assess Sustainability of Transport ActivitiesPart 1: Review of the Existing Transport Sustainability Indicator Initiatives and Development of an Indicator Set to Assess Transport Sustainability Performance. At:http://publications.jrc.ec.europa.eu/repository/handle/111111111/10416.

[18] R. Prabha, M. G. Kabadi, Overview of Data Collection Methods for Intelligent Transportation Systems. The International Journal Of Engineering And Science (IJES), vol 5(3), pp 16-20, (2016).

[19] EEA (2005). EEA core set of indicators: Guide. EEA Technical Report, Copenhagen. At: https://www.eea.europa.eu/publications/technical_report_2005_1.

[20] New York Taxi and Limousine Commission Taxi of tomorrow survey. 2011.

[21] D. Zhang, T. He, Y. Liu, J.A. Stankovic, CallCab: A unified recommendation system for carpooling and regular taxicab services. In Big Data, 2013 IEEE International Conference on (pp. 439-447). IEEE. (2013, October).

[22] Xianyuan Zhan, Xinwu Qian, and Satish V. Ukkusuri. Measuring the Efficiency of Urban Taxi Service System, August 2014.

[23] Y. Liu, F. Wang, Y. Xiao, S. Gao, Urban land uses and traffic source-sink areas: Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, vol 106(1), pp 73-87, (2012).

[24] Neuwirth, A., Chebchoub, F., Punjwani, J., Toure, M. Fare Share: Flow and Efficiency in NYCs Taxi System.

[25] J. Park, D.J. Kim, Y. Lim, Use of smart card data to define public transit use in Seoul, South Korea. Transportation Research Record: Journal of the Transportation Research Board, (2063), pp 3-9, (2008).

[26] S. An, X. Hu, J. Wang, Urban taxis and air pollution: a case study in Harbin, China. Journal of Transport Geography, vol 19(4), pp 960-967, (2011).

[27] T. D. Matte, Z. Ross, I. Kheirbek, H. Eisl, S. Johnson, J.E. Gorczynski, J. E. Clougherty, Monitoring intraurban spatial patterns of multiple combustion air pollutants in New York City: design and implementation. Journal of Exposure Science and Environmental Epidemiology, vol 23(3), pp 223-231, (2013).

[28] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, E. Chang, Diagnosing New York city's noises with ubiquitous data. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 715-725). ACM.(2014, September).

[29] Y. Zheng, F. Liu, H.P. Hsieh, U-air: When urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD international

conference on Knowledge discovery and data mining (pp. 1436-1444). ACM. (2013, August).

[30] Xianyuan Zhan, Xinwu Qian, and Satish V. Ukkusuri. Measuring the Efficiency of Urban Taxi Service System, August 2014.

[31] T. Litman, T. (2011). Well measured: Developing indicators for sustainable and livable transport planning. Victoria Transport Policy Institute. At: http://www.vtpi.org/wellmeas.pdf

[32] York Region (2009), Transportation Sustainability Principles Transportation Master Plan (www.york.ca/Services/Regional+Planning/ Infrastructure/TMP_Reports.htm); summarized in Jim Helmer and Jim Gough (2010), Best Practices in Sustainable Transportation, ITE Journal ($www.ite.org$), Vol. 80. No. 3, March 2010, pp. 26-35.

[33] Christy Jeon, Adjo Amekudzi and Randall Guensler (2008), Sustainability Assessment at the Transportation Planning Level: Performance Measures and Indexes, TRB ($www.trb.org$) Annual Meeting.

[34] J. Miller (2008), Potential Performance Measures to Assess Transportation and Land Use Coordination, Transportation Research Board 87th Annual Meeting (www.trb.org).

[35] W. Zhang, G. Qi, G. Pan, H. Lu, S. Li, Z. Wu, City-scale social event detection and evaluation with taxi traces. ACM, Transactions on Intelligent Systems and Technology, 6(3), 40 (2015).

[36] A. Guhnemann, R. P. Schafer, K.U. Thiessenhusen, and P.Wagner, Monitoring traffic and emissions by floating car data. Working paper ITS-WP-04-07. Institute of Transportation Studies, University of Sydney, Australia. 2004.

[37] A.L. Alfeo, M.G.C.A. Cimino, S. Egidi, B. Lepri, A. Pentland, G. Vaglini, "Stigmergy-Based Modeling to Discover Urban Activity Patterns from Positioning Data" in Proc. SBP-BRiMS The International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2017), LNCS 10354, pp. 292-301, Washington, DC, USA, 5-8 July, 2017

[38] C. Peng, X. Jin, K.C. Wong, M. Shi, P. Li, Collective human mobility pattern from taxi trips in urban area. PloS one, 7(4), e34487. (2012).

[39] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, Z. Wang, Prediction of urban human mobility using large-scale taxi traces and its applications. Frontiers of Computer Science, 6(1), 111-121. (2012).

[40] G. Makrai, Efficient method for large-scale spatio-temporal hotspot analysis (GIS Cup).

[41] B. Pan, Y. Zheng, D. Wilkie, C. Shahabi, Crowd sensing of traffic anomalies based on human mobility and social media. ACM, In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 344-353. (2013).

[42] Y. Lu, G.G. Chua, H. Wu, C.S.Q. Ong, An Intelligent System for Taxi Service Monitoring, Analytics and Visualization. In IJCAI pp. 4256-4257, (2016, July).

[43] Hu, Yujie, Harvey J. Miller, Xiang Li. Detecting and analyzing mobility hotspots using surface networks. Transactions in GIS, 18.6, 911935, (2014).

[44] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 5(3), 38. (2014).

[45] A. Driemel, A. Krivosija, C. Sohler, Clustering time series under the Frchet distance. In Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 766-785). Society for Industrial and Applied Mathematics. (2016, January).

[46] Source Code of the Stigmergic Perceptron, Available online: http://www.mathworks.com/matlabcentral/fileexchange/66526

Mario G.C.A. Cimino is with the Department of Information Engineering (University of Pisa) as an Associate Professor. He is also a research associate at the Institute for Informatics and Telematics (IIT) of the Italian National Research Agency (CNR). In 2006, he was a visiting Ph.D. student at the Electrical and Computer Engineering Research Facility of the University of Alberta, Canada. In 2007, he received the Ph.D. degree in Information Engineering from the University of Pisa. His research focus lies in the areas of Swarm Intelligence and Business/Social Process Analysis, with particular emphasis on Stigmergic Computing, Workflow Mining and Simulation. He is (co-) author of more than 50 publications.

Sara Egidi was born in Ascoli Piceno, Italy, on 18 November 1990. She received the B.S. degree in industrial computer science from the University of Camerino, Ascoli Piceno, Italy, in 2014 and the M.S. degree in computer engineering from University of Pisa, Pisa, Italy, in 2017. Her research has been concerned with machine learning techniques applied in the field of city science.

Bruno Lepri Bruno Lepri leads the Mobile and Social Computing Lab (MobS) and is vice-responsible of the Complex Data Analytics research line at Bruno Kessler Foundation (Trento, Italy). Bruno is also research affiliate at the MIT Connection Science initiative and he has recently launched an alliance between MIT Connection Science and Bruno Kessler Foundation on Human Dynamics Observatories. He is also a senior research affiliate of Data-Pop Alliance, the first think-tank on Big Data and Development co-created by the Harvard Humanitarian Initiative, MIT Media Lab, Overseas Development Institute, and Flowminder to promote a people-centered big data revolution. In 2010 he won a Marie Curie Cofund post-doc fellowship and he has held post-doc positions at FBK and at MIT Media Lab. He holds a Ph.D. in Computer Science from the University of Trento. His research interests include computational social science, personality computing, urban computing, network science, machine learning, and new models for personal data management and monetization. His research has received attention from several international press outlets and obtained the James Chen Annual Award for best UMUAI paper and the best paper award at ACM Ubicomp 2014. His work on personal data management was one of the case studies discussed at the World Economic Forum.

Antonio Luca Alfeo was born in Taranto, Italy on September 4, 1987. He received the B.S. and the M.S. degree in Computer Engineering from the University of Pisa, in 2012 and 2015 respectively. He is currently a Ph.D. student of the International Ph.D. Program in Smart Computing and a visiting Ph.D. student at the Media Lab of the Massachusetts Institute of Technology. His research interests include the design and development of emergent and bio-inspired approaches for machine learning and data analysis, with applications in the context of cyber-physical systems including collaborative UAVs swarms, e-health, and smart cities.

Gigliola Vaglini was born in Pisa, Italy, on June 11, 1952. She received the M.S. degree in Computer Science from the University of Pisa. She was a research assistant at the University of Pisa, Department of Computer Science, an associate Professor at the University of Naples, Federico II, Department of Mathematics, and then at the University of Pisa, Department of Information Engineering; from 2002 she is a Full Professor at the same Department. Her research has been concerned mainly with formal methods for specification and verification of concurrent and distributed systems, in particular the model checking of concurrent systems. More recently her research activity was focused on the use of the stigmergic paradigma to aggregate data supplied by different sources and to obtain a distributed control on autonomous systems.