UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

This is a preprint. Please cite this work as:

A. L. Alfeo, M. G. C. A. Cimino, S. Egidi, B. Lepri, and G. Vaglini, "An emergent strategy for characterizing urban hotspot dynamics via GPS data", in proceedings of "The 5th Conference on scientific analysis of mobile phone datasets", organized by F. Calabrese, E. Moro, V. Blondlel, A. Pentland. (NetMob 2017) 5-7 April 2017, Milan, Italy.

Machine Learning and Process Intelligence

Research Group

An emergent strategy for characterizing urban hotspot dynamics via GPS data.

Antonio L. Alfeo[1], Mario G.C.A. Cimino[1], Sara Egidi[1], Bruno Lepri[2] and Gigliola Vaglini[1]
[1]*Department of Information Engineering, Università di Pisa, largo Lazzarino 1, Pisa, Italy*
[1]*Bruno Kessler Foundation, via S. Croce, 77, Trento, Italy*
luca.alfeo@ing.unipi.it, mario.cimino@unipi.it, s.egidi1@studenti.unipi.it, lepri@fbk.eu, gigliola.vaglini@unipi.it.

Keywords: Urban mobility, taxi-GPS traces, stigmergy, emergent paradigm, hotspot.

INTRODUCTION AND MOTIVATION

The increasing volume of urban human mobility data arises unprecedented opportunities to monitor and understand city dynamics. Identifying events which do not conform to the expected patterns can enhance the awareness of decision makers for a variety of purposes, such as the management of social events or extreme weather situations [1]. For this purpose GPS-equipped vehicles provide huge amount of reliable data about urban dynamics, exhibiting correlation with human activities, events and city structure [2]. For example, in [3] the impact of a social event is evaluated by analyzing taxi traces data. Here, the authors model typical passenger flow in an area, in order to compute the probability that an event happens. Then, the event impact is measured by analyzing abnormal traffic flows in the area via Discrete Fourier Transform. In [4] GPS trajectories are mapped through an Interactive Voting-based Map Matching Algorithm. This mapping is used for off-line characterization of normal drivers' behavior and real-time anomalies detection. Furthermore, the cause of the anomalies is found exploiting social network data. In [5] the authors employ a Multiscale Principal Component Analysis to analyze Taxi GPS data in order to detect traffic anomalies. The most of the methods in the literature can be grouped into four categories: distance-based, cluster-based, classification-based, and statistics-based [6]. Typically, due to the complexity of this kind of data, the modeling and comparison of their dynamics over time are hard to manage and parametrize [7]. In this paper, we present an innovative technique aimed to handle such complexity, providing a study of urban hotspot dynamics.

APPROACH DESCRIPTION

The developed approach is based on stigmergy, a mechanism belonging to the emergent paradigms. Emergent paradigms allow to avoid the explicit modeling of a system, which works only under the assumption formulated by the designer. Emergent paradigms offer model-free computational approach, characterized by adaptation, autonomy and self-organization of data [8]. In particular, with the emergent mechanism based on computational stigmergy, each sample position is associated to a digital pheromone deposit (mark). Marks are defined in a three-dimensional space, and characterized by *evaporation* over time, i.e. a progressive decay of mark intensity. Marks aggregate according to their spatio-temporal proximity, forming a stigmergic trail. As an effect, the evaporation is counteracted while new marks arrive in a given region. Thus, aggregation and evaporation can produce an emerging mechanism which acts as an agglomerative spatio-temporal clustering with historical memory. Employing the principle of stigmergy, we exploit positioning data to identify high-density areas (*hotspots*) within a city and to analyze their activity (presence of people) over time. Fig. 1 shows the positional stigmergy applied to hotspot identification. First, input data undergo the *smoothing* process, which focuses the analysis, highlighting relevant dynamics while removing insignificant activity levels. At a periodic time instant, a mark is released in a computer-simulated spatial environment (stigmergic space). The mark intensity is proportional to the number of people. Marks aggregation forms a trail, which progressively evaporates. For a given threshold of trail intensity, areas with intensity higher than the threshold correspond to hotspots. As an example, Fig. 2 shows the hotspots identified in Manhattan (New York). 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 - Garmet (G).

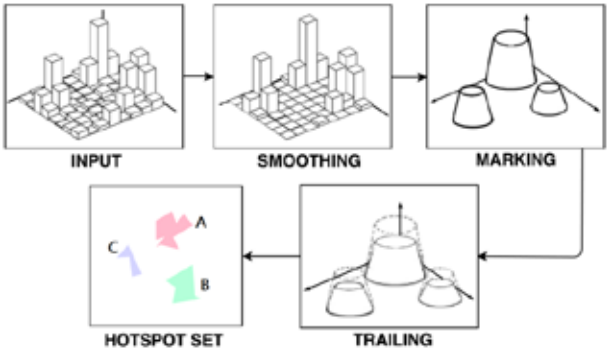

Figure 1. Positional stigmergy for hotspot identification.

Figure 2: Hotspots identified in Manhattan (2015).

In order to characterize the dynamics of the identified hotspots, the activity of each hotspot generates a mono-dimensional time series for every observation day. In a given time series, what is actually interesting is not the continuous variation of the activity over time, but the transition from one type of behavior to another. Each type of hotspot activity behavior should be defined by an expert in the field, in order to be general and reusable for many hotspots and many cities. More formally, each type is called an *archetype*, since it is an ideal time series segment defining a behavioral class of the hotspot activity. An example of archetype is *rising activity*, which means that the hotspot is moving from an ordinary activity level to its highest activity level. Fig. 3a and Fig. 3a' show such archetype and a real example of time series, respectively.

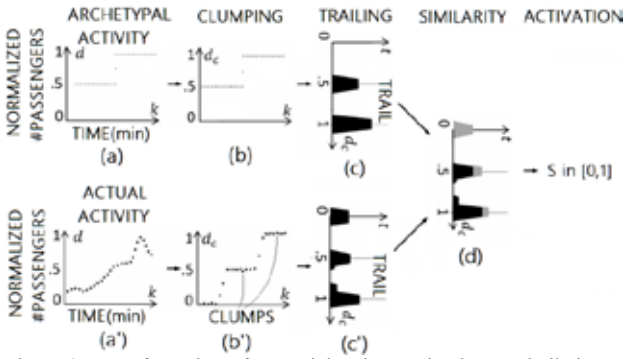

Figure 3. Transformation of an activity time series into a similarity time series with respect to an archetype.

In order to assess the match between an archetypal behavior and the current hotspot behavior, the time series are processed in sequential stages (Fig. 3). The initial samples first undergo the *clumping* process, which is a soft discretization of samples with respect to a set of levels of interest for any archetype (Fig. 3b and Fig. 3b'). Subsequently, in correspondence of each sample value, a trapezoidal *mark* is released (Fig. 3c and Fig. 3c'). Marks aggregate over time in a *trail*. Trail intensity is subject to evaporation. As an effect, a trail captures the spatiotemporal behavior of the time series in the short-term. A degree of similarity between the archetype trail and actual activity trail is then computed (Fig. 3d). Finally, to better sharpen the similarity value against the other archetypes, an activation function is applied to enhance only relevant similarity values, while removing insignificant values.

The overall processing schema is called *stigmergic receptive field* (receptive field for short), because it is receptive to a specific archetype, and it takes inspiration from the neurocomputing domain. A receptive field should be properly parameterized to effectively assess the archetypal behavior. For example, short-life marks evaporate too fast, preventing aggregation and pattern reinforcement, whereas long-life marks cause early activation. For this purpose, the receptive field is equipped with a parametric adaptation mechanism, based on the differential evolution algorithm [9]: parameters are adjusted by minimizing the mean square error over a set of annotated sample signals.

Since any real signal is usually similar to more than one archetype, a collection of receptive fields specialized on different archetypes is arranged to make a *stigmergic perceptron*, i.e., a connectionist topology whose architecture is represented on the left of Fig. 4. More specifically, the archetypes are ordered for increasing activity of the hotspot: a) *asleep*, b) *falling*, c) *awakening*, d) *flow*, e) *chill*, f) *rising*, and g) *rush-hour* hotspot activity. The stigmergic perceptron combines linearly the similarity values provided by all receptive fields, providing a value between zero and *N*-1, where *N* is the number of archetypes, called *activity level.*

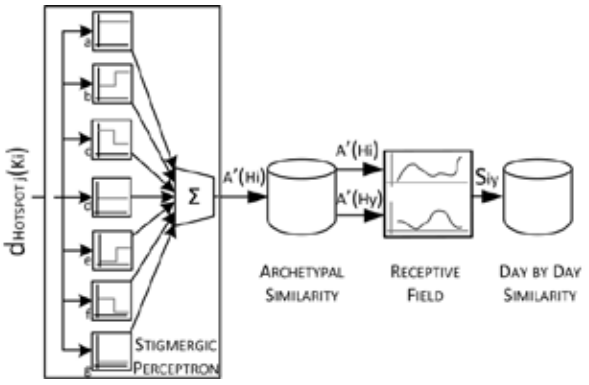

Figure 4: System architectural overview.

As a result, the stigmergic perceptron provides a new time series of activity levels for a given time series of activity samples. In order to compute the overall similarity between two days of hotspot activity, a further receptive field is used. Such receptive field compares the time series of two activity levels corresponding to two days. This measure of similarity is used to identify anomalies in the hotspot activity. For this purpose, we assume that during days characterized by an anomaly, the hotspot activity level is dissimilar with the activity level of equivalent days. Two days are considered *equivalent* if they fall on the same day of the week and in the same month. In order to detect the anomaly, we define the *normality index* of a given day as the average similarity of that day with its equivalent days. The normality index is 1 (or 0) if the day is perfectly similar (dissimilar) to its equivalent days.

The overall architecture allows the incremental detection of perturbations on the city routine that can involve the hotspots for a variety of reasons. As an example, the next section shows some experimental results on anomalies caused by adverse weather conditions.

EXPERIMENTAL RESULTS AND DISCUSSION

The analysis is based on taxis OD (Origin-Destination) traces provided by Taxi and Limousine Commission of New York City [10], which contains information about all medallion taxi trips from 2009 to 2016. Data are spatio-temporally discretized in bins characterized by 10 ft. wide and 5 minutes duration. In order to focus on urban hotspot dynamics the number of passengers in each bin, in terms of both pick-up and drop-off is first considered. Subsequently, the min-max normalization is applied. As a case study we focus on anomalies caused by an adverse weather condition. Specifically, we analyze Manhattan in January 2015. In January 26-27, 2015 many mobility issues, caused by a blizzard, were reported [11]. We refer this knowledge as the ground truth. In Fig. 5 the activity level and the corresponding stigmergic trail have been generated analyzing the activity in hotspot G in every Tuesday of January 2015. Here, it is apparent that the January 27 is characterized by an anomaly, for the notable differences with the other days both in terms of activity level (left side) and stigmergic trail (right side). A measure of these differences is based on the similarity values supplied by the receptive field, used to calculate the normality index of each hotspot. In Fig. 6 and Fig.7, a radar chart shows the normality indexes for all hotspots in each Tuesday and Monday of January, respectively.

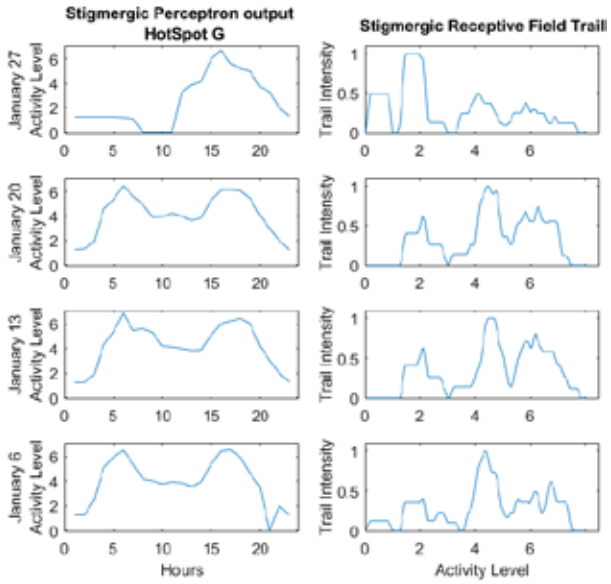

Figure 5. The activity levels on Hotspot G (on the left) during each Tuesday of January 2015 and the corresponding trails (on the right) which feed by the second level receptive field.

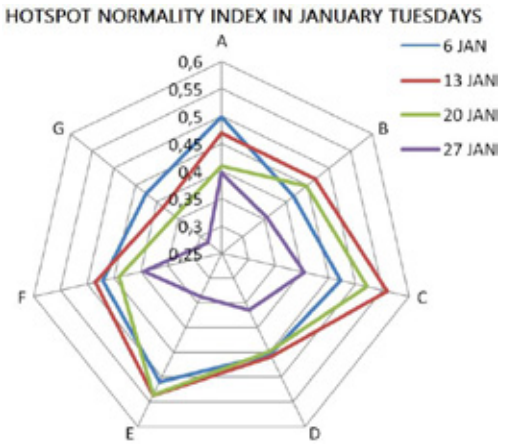

Figure 6. Normality index of January Tuesday for each Hotspot.

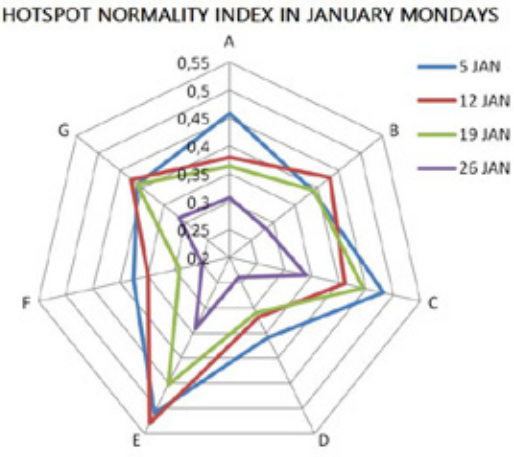

Figure 7. Normality index of January Mondays for each Hotspot.

Since a blizzard is a city-wide event which affects urban activity in the whole city, any hotspot during 26 and 27 January (purple line) results in the lowest normality index values. Table 1 shows, for equivalent days, the average normality index calculated between all hotspots. Again, it is apparent that the lowest values are related to the blizzard.

Table 1. The average of the normality index, over all city hotspots.

Mondays	05 Jan	12 Jan	19 Jan	26 Jan
Average Normality	0,45	0,42	0,38	**0,28**

Tuesdays	06 Jan	13 Jan	20 Jan	27 Jan
Average Normality	0,46	0,48	0,45	**0,35**

Although the study focuses on a blizzard, the analysis can be carried out on other events. Moreover, our approach provides a similarity measure that may be used with clustering techniques. Clustering can discover a structure in data and may generate data-driven prototypes of normal days, which are more effective than calendar-driven days. For this reason, to adopt a clustering technique is considered a key investigation task for future work.

REFERENCES

[1] Sagl, G., Loidl, M., & Beinat, E. (2012). A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. ISPRS International Journal of Geo-Information, 1(3), 256-271.

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

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

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

[5] Kuang, W., An, S., & Jiang, H. (2015). Detecting traffic anomalies in urban areas using taxi GPS data. Mathematical Problems in Engineering, 2015.

[6] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[7] Castro, P. S., Zhang, D., Chen, C., Li, S., & Pan, G. (2013). From taxi GPS traces to social and community dynamics: A survey. ACM Computing Surveys (CSUR), 46(2), 17.

[8] Vernon, D., Metta, G., & Sandini, G. (2007), 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.

[9] Cimino, M. G., Lazzeri, A., & Vaglini, G. (2015, June). Improving the analysis of context-aware information via marker-based stigmergy and differential evolution. In International Conference on Artificial Intelligence and Soft Computing (pp. 341-352). Springer International Publishing.

[10] *NYC.gov*, the official website of the City of New York: Taxi and Limousine Commission (TLC) Trip Record Data, http://www.nyc.gov/html/ tlc/html/about/trip_record_data.shtml

[11] Weather NYC: Thousands of transatlantic travellers face serious disruption caused by New York winter storm 'Juno'. The Independent. January 26, 2015.