Multi-Scale Extraction of Regular Activity Patterns in Spatio-Temporal Events Databases: A Study Using Geolocated Tweets from Central Mexico

Pablo Lopez-Ramirez[1] and Oscar Sanchez-Siordia[1]

[1]Centro de Investigación en Geografía y Geomática Ing. Jorge L. Tamayo, CentroGeo

Abstract

This paper proposes a new technique for the extraction of regular activity patterns at different scales, mined from the micro-blogging platform Twitter. The approach is based on the recursive application of the DBSCAN clustering algorithm to the geolocated Twitter feed. This technique includes a novel way to obtain averaged regular activity zones based on the rasterization and aggregation of the Concave Hull of the clusters identified at each resolution level. Since the proposed technique uses only the spatio-temporal characteristics of the geolocated Twitter feed and it does not depend on the messages, it can be extended to work with different spatio-temporal event sources such as mobile telephone records. An experiment was carried out to demonstrate the effectiveness of the technique in the extraction of known activity patterns in the Mexico Central Region.

1 Introduction

The digital breadcrumbs left by social media users have proven to be a valuable source of geographic insights. In the GIS field, they have been used for the detection of events (Atefeh and Khreich) or the characterization of zones through social media activity (Frias-Martinez and Frias-Martinez; Lee et al.), among other things.

In most cases, extraction and characterization of regular activity patterns is of great importance. In this paper we propose a technique for the extraction of such patterns that relies only on the spatio-temporal properties of the Twitter feed. The purpose of this is, on the one hand, to improve on the current available techniques (Lee et al.; Frias-Martinez and Frias-Martinez) and, on the other hand, to be as independent from the nature of the Twitter feed as possible.

The proposed technique is based on the observation that human activity patterns exhibit a wide range of scales (Arcaute et al.), and that the current methods for determining this activity from Twitter messages do not consider this. The proposed approach is based on the recursive application of a clustering algorithm. This approach demands the development of a novel way for averaging the spatial patterns extracted from the data.

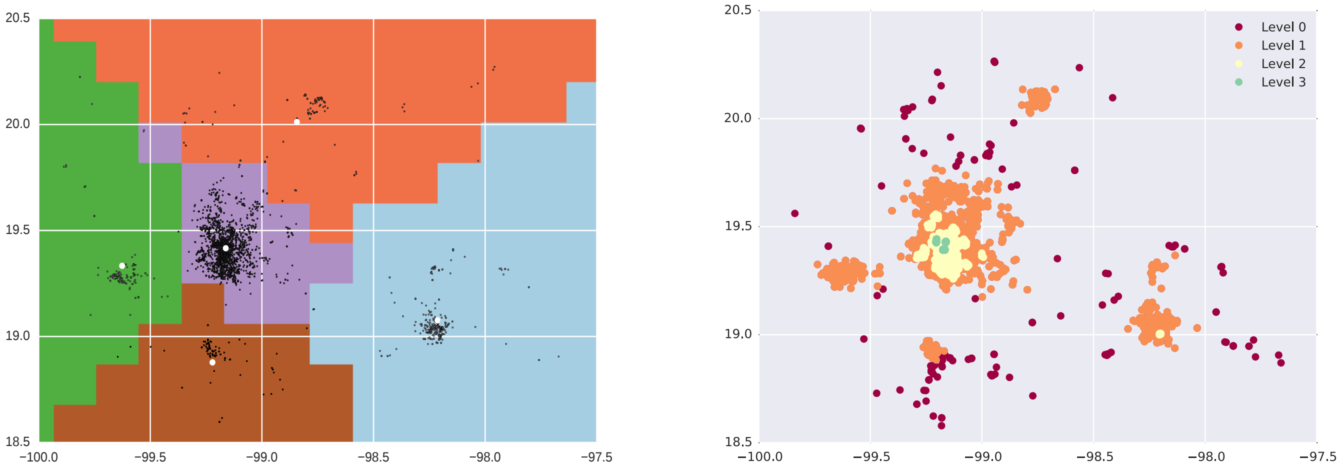
2 Methodology

The idea behind the technique proposed in this paper, is that activity patterns may exhibit a range of scales, and that this scales cannot be represented by a flat tessellation. To overcome this limitation, we propose the use of a recursive clustering strategy that is able to extract the structures present at different scales in the database.

2.1 Recursive Clustering

Several techniques produce a hierarchical representation of point samples, the focus being often on extracting the most significant clusters across all scales. In the case of our study, we need to find clusters that are representative at each resolution level. For this, we will recursively apply DBSCAN (Ester et al.) to the data and thus obtain a hierarchy of clusters across resolution levels.

Initially eps_0 and *MinPoints* are selected using the *k-dist* plot (Ester et al.) for the whole sample in the corresponding time slice. Then for each iteration of the algorithm, the value for eps_0 is halved while *MinPoints*



(a) K Means.

(b) Recursive DBSCAN.

Figure 1: Comparison of two clustering strategies over the same points sample. Figure (a) shows the K means clusters and Voronoi polygons around the centroids. Figure (b) shows the clusters obtained in each iteration of the recursive application of DBSCAN.

is held constant. This process is repeated until the number of points in the largest cluster obtained falls below a predefined threshold. In Figure 1 we show a comparison between the iterative application of DBSCAN and the flat tessellation obtained with K-Means.

2.2 Spatio-Temporal Averaging

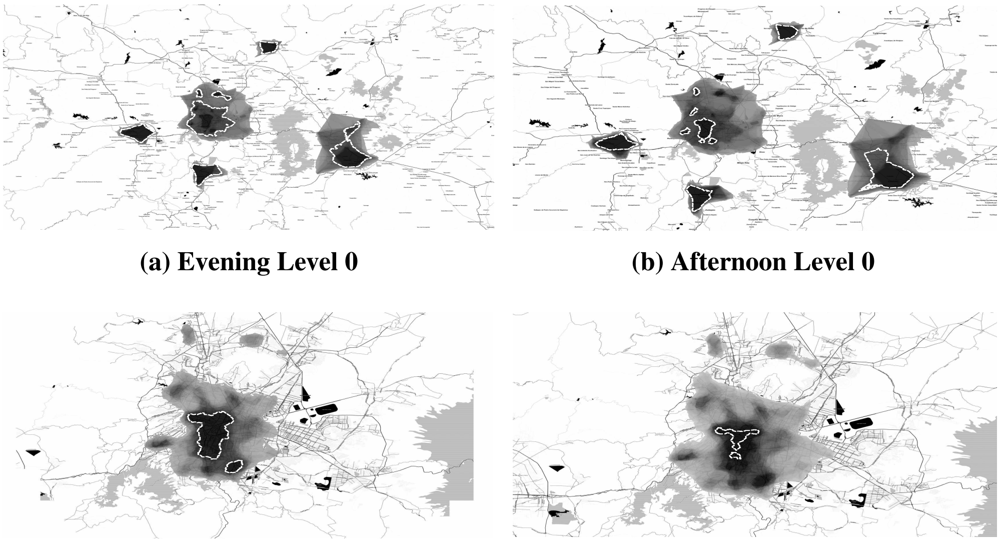
For every day in the input geolocated Twitter feed, a time segmentation similar to those used in Lee et al. and Frias-Martinez and Frias-Martinez will be performed. Then, for each resulting time slice, the recursive clustering strategy described in Section 2.1 will be used. This process is carried out for the whole study period, thus ending with a hierarchical representation for each day.

To average this representations and obtain a single hierarchy representing the whole period for each time slice, the following procedure is carried out:

- A polygon for each cluster is extracted using the *Optimal Alpha Shape* (Edelsbrunner) of the cluster points.
- The polygons obtained are then rasterized to obtain images that have a value of 1, if the pixel belongs to a cluster, or 0 otherwise.
- The resulting images are aggregated to obtain a single image whose values represent the number of days a given pixel has belonged to a cluster.
- The aggregated images are polygonized by cutting them with a threshold value- the number of days a pixel must belong to a cluster in order for it to be considered part of the regular activity. This allows for further characterization of activity zones, such as assigning user counts or other activity measures.

3 Experiment

An experiment was carried out to demonstrate the application of the proposed methodology to the extraction of regular activity patterns around Central Mexico. The database consists of geolocated tweets in the area from October 10 2014 to April 4 2015, (5,415,827 tweets). Prior to the extraction of regular activity zones, we need



(a) Evening Level 0 **(b) Afternoon Level 0**

(c) Evening Level 1 **(d) Afternoon Level 1**

Figure 2: Aggregated activity rasters with threshold cut polygons in dashed lines.

to deal with the *pollution* commonly encountered in the tweeter feed, this means that we must perform some preprocessing to clean up the database. In this case, we filtered out tweets by users that have more than one update within 100 meters of their original location in the same time period. The rationale behind this filtering is that this kind of behavior might be representative of bots, or that it might artificially alter the shape of the clusters without representing the regular activity of the population. The resulting activity patterns for levels 0 and 1 for the Afternoon (14:00 to 18:00 hours) and Evening (18:00 to 22:00 hours) periods are shown in Figure 2.

4 Results Discussion

At the smaller scale (Level 0 in Figure 2), our technique is able to detect the greater metropolitan areas of Mexico City, Puebla, Toluca, Pachuca and Cuernavaca, in the Central Mexico region. As we increase the resolution, the point density in the smaller cities (Puebla, Toluca, Pachuca and Cuernavaca), does not allow for the recursive algorithm to detect larger scale activity. The opposite is true for Mexico city, where we are able to detect up to three scales within the city (not all are shown in the figure).

By comparing the patterns found for the Afternoon and Evening intervals, we see that in the latter, the activity is more dispersed and Level 0 (Figure 2a) shows activity peaks in the northern low-income housing suburbs. On the other hand, the activity for the Afternoon segment (Figure 2b) is more concentrated around the Central Business District (CBD) of the city. This results are consistent with known activity patterns for Mexico City. For example, Suarez and Delgado (Suarez and Delgado) performed a study in the Job-Housing ratio and found the same T-shaped pattern for the CBD. From the same study, we can see that the job to housing ratio of the northern low-income suburbs is very low, which means it is mostly a residential area. This is in line with our results that show activity peaks for those areas only at the Evening intervals.

5 Conclusions

The main improvements of the proposed technique over the available methods are:

1. The ability to detect patterns at different scale levels. This allows us to detect both major urban areas and activity zones within those areas which have a high enough activity density.

2. Using the Alpha Shapes to polygonize the clusters allows us to account for the shape of the activity zones. This represents an improvement compared to the use of Voronoi tessellations.

Qualitative analysis of the regular activity zones obtained, show great accordance with the known activity patterns for the study area, mainly with the spatial distribution of the Job-Housing ratio.

6 Further Work

The next step is quantitative validation of the results obtained. For this, activity data for the study area is essential. One approach would be coupling Job-Housing ratio maps with mobility patterns extracted from Origin-Destination surveys to disaggregate the latter to the scales needed for our analysis.

Also, a more tractable way of setting parameter values (such as eps_0 and $MinPoints$ at each iteration) is needed. Ground truthing against measured activity distributions would provide basis for a calibration-validation approach. In the same line of thought, it would be important to test different clustering algorithms, such as HDBSCAN, which does not introduce additional parameters or STDBSCAN, which is purely spatio-temporal.

Finally, the multi-scale regular activity zones could be used to detect unusual crowd activity at various scales. The rationale behind this is that unusual events also exhibit scale differences. For example, it is known that important large scale events, such as the Super Bowl or the Arab Spring, produce a general increase of messages in the social networks, while localized small-scale occurrences, such as festivals, demonstrations or accidents, produce small clusters of messages around the locations affected.

References

Elsa Arcaute, Carlos Molinero, Erez Hatna, Roberto Murcio, Camilo Vargas-Ruiz, Paolo Masucci, Jiaqiu Wang, and Michael Batty. Hierarchical organisation of Britain through percolation theory. URL `http://arxiv.org/abs/1504.08318`.

Farzindar Atefeh and Wael Khreich. A Survey of Techniques for Event Detection in Twitter. 31(1):132–164. ISSN 1467-8640. doi: 10.1111/coin.12017. URL `http://onlinelibrary.wiley.com/doi/10.1111/coin.12017/abstract`.

Herbert Edelsbrunner. Smooth surfaces for multi-scale shape representation. In P. S. Thiagarajan, editor, *Foundations of Software Technology and Theoretical Computer Science*, number 1026 in Lecture Notes in Computer Science, pages 391–412. Springer Berlin Heidelberg. ISBN 978-3-540-60692-5 978-3-540-49263-4. doi: 10.1007/3-540-60692-0_63. URL `http://link.springer.com/chapter/10.1007/3-540-60692-0_63`.

Martin Ester, Hans-peter Kriegel, Jörg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press.

Vanessa Frias-Martinez and Enrique Frias-Martinez. Spectral clustering for sensing urban land use using Twitter activity. 35:237–245. ISSN 0952-1976. doi: 10.1016/j.engappai.2014.06.019. URL `http://www.sciencedirect.com/science/article/pii/S0952197614001419`.

Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Discovery of unusual regional social activities using geotagged microblogs. 14(4):321–349. ISSN 1386-145X, 1573-1413. doi: 10.1007/s11280-011-0120-x. URL `http://link.springer.com/article/10.1007/s11280-011-0120-x`.

M. Suarez and J. Delgado. Is Mexico City Polycentric? A Trip Attraction Capacity Approach. 46(10):2187–2211. ISSN 0042-0980. doi: 10.1177/0042098009339429. URL `http://usj.sagepub.com/cgi/doi/10.1177/0042098009339429`.