Smart Statistics for Smart Applications

Book of Short Papers SIS2019

SIS 2019
SMART STATISTICS FOR SMART APPLICATIONS

SIS Società Italiana di Statistica

Editors: Giuseppe Arbia, Stefano Peluso,
Alessia Pini and Giulia Rivellini

Published by Pearson

www.pearson.com

Giugno 2019 ISBN 9788891915108

Preface

Section 1. Plenary Sessions and Round Table

Section 2. Invited Papers

Section 3. Contributed Papers

Section 4. Posters

Density-based Algorithm and Network Analysis for GPS Data

Algoritmi di Cluster e Reti per lo studio di dati GPS

Antonino Abbruzzo, Mauro Ferrante, Stefano De Cantis

Abstract The use of advanced global positional system (GPS) trackers has emerged as a novel technology in data collection of units movements. GPS data contain a large amount of information since the signals of the units are recorded almost in real time. The analysis of GPS data can be carried on several aspects of the spatial movements. In this study, we focus on statistical methods for the identification of points of interests and the analysis of the network of movements for GPS data. In particular, a density cluster-based algorithm is applied to summarize the vast amount of information and to find the most relevant points of attractions. A directed network synthesizes the individual unit path by using the latter information. Finally, we aggregate the unit paths in a weighted directed network which is studied through network analysis. We apply the proposed approach to a case study on cruise passengers' movements in an urban context.

Abstract *La diffusione dei sistemi di localizzazione GPS offre numerose opportunità per la raccolta di dati di movimento. I dati GPS presentano diversi elementi di complessità derivanti anche dall'elevato dettaglio temporale e territoriale. Numerosi sono gli aspetti che possono essere presi in esame per tale tipologia di dati. Il presente studio propone un approccio statistico basato sull'identificazione dei punti di attrazione e sullo studio dei network. In particolare, viene proposto un algoritmo di identificazione di cluster, sulla base della densità di punti, che vengono sintetizzati in un network che riassume il comportamento individuale. In un secondo step, i movimenti complessivi sono aggregati ed analizzati tramite la network analysis. L'approccio proposto è applicato allo studio dei movimenti di croceristi in contesti urbani.*

Key words: Spatial-Temporal Data, Cluster-Based Method, Tourists' Behaviors, Destination Management.

Department of Economics, Business and Statistics
University of Palermo, Viale delle Scienze, building 13, 90128 - Palermo e-mail: antonino.abbruzzo@unipa.it; stefano.decantis@unipa.it

Department of Culture and Society
University of Palermo, Viale delle Scienze, building 15, 90128 - Palermo e-mail: mauro.ferrante@unipa.it

1 Introduction

Nowadays, GPS devices have become of small size, not bulky, equipped with significant autonomy and, what matters most, once activated manage to memorize the geographical coordinates in which the statistical unit is at a given moment. Once the experience is over, it will be possible to download the data and to analyze the route taken by the unit. These devices have several advantages and give the opportunity to collect high-quality data, that are very accurate both in terms of temporal (seconds) and spatial (meters) resolution. Moreover, this type of data provides information on the unit's movement not influenced by units' perceptions or other issues (e.g., recall bias) which generally affect traditional survey instruments (e.g., diaries or questionnaires).

The analysis of GPS data can be conducted on several aspects of the spatial movements. In this paper, we focus our attention on statistical methods for the identification of points of interests (POIs) and the units' networks of movements. Moreover, we aggregate these units' movements networks to produce a directed weighted network which can be studied through network analysis. Various techniques have been proposed to detect POIs and visualize patterns, including density estimation grid-based aggregation, spatial movement sequence, network analysis [4] and algorithms for clustering spatial data [5].

We introduce a statistical approach based on density based-cluster algorithm (DBSCAN [2]) and networks analysis. The clustered approach on the GPS data recovers the "points of interests. The network analysis summarizes units density cluster-based data to access to the most relevant characteristics of the units movements. Finally, we apply the proposed approach to the case of cruise passengers' mobility in the city of Palermo.

2 Statistical Methods

In this section, we first describe the DBSCAN algorithm which summarizes a large amount of information and find the most relevant points of interests. Secondly, we define the directed network approach to summarize units density cluster-based data to access to the most significant characteristics of the units movements. Finally, we derive a weighted direct graph which summarizes the behaviors of all the sampled units.

2.1 Density Cluster-Based Model for GPS data

Let $D_j^{(i)} = (t_j^{(i)}, long_j^{(i)}, lat_j^{(i)})$, where $i = 1,\ldots,N$, $j = 1,\ldots,n_i$, $t_{j+1}^{(i)} - t_j^{(i)} = c$ and c is a constant, be the temporal-spatial movement of the i-th unit. Figure 1 shows two temporal-spatial sequences referred to two different units. In this Figure, we may recognize the points of interests for each of the units, by looking at the higher points density in some locations of the path. DBSCAN is a density-based algorithm designed to discover arbitrary-shaped clusters in any database $D_j^{(i)}$ and at the same

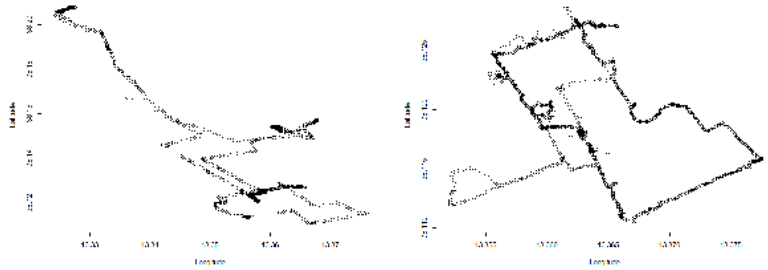

Fig. 1 Two illustrative units' behaviors.

time to distinguish noise points. These clusters are called "points of interests" and they can be interpreted as a set of places a unit visited for a certain amount of time.

Let us introduce some concepts before describing the DBSCAN algorithm. The **ε-neighbourhood** of a point p is defined by $ne_\varepsilon(p) = \{q \in D_{ij} : dist(p,q) \leq \varepsilon\}$, where dist$(p,q)$ is a distance function (e.g., Manhattan Distance, Euclidean Distance). If the cardinality of an ε-neighbourhood of a point p, i.e. $|ne_\varepsilon(p)|$, is at least greater than a minimum number ($MinPts$) then p is a **core point**. A point p is **directly density-reachable** from the object q with respect to ε and $MinPts$ if $p \in ne_\varepsilon(q)$ and $|ne_\varepsilon(q)| \geq MinPts$. A point p is **density-reachable** from the object q with respect to ε and $MinPts$ if there is a chain $p_1, \ldots, p_l$, $p_1 = q$, $p_l = p$ such that p_{l+1} is directly density-reachable from p_l. An object p is **density-connected** to object q with respect to ε and $MinPts$ if there is an object o such that both, p and q are density reachable from o with respect to ε and $MinPts$. A **cluster** C is a non-empty subset of $D_j^{(i)}$ satisfying the following maximality and connectivity requirements:

- $\forall$ p,q: if $q \in C$ and p is density-reachable from q with respect to ε and $MinPts$, then $q \in C$;
- $\forall$ $p,q \in C$: p is density-connected to q with respect to ε and $MinPts$.

An object p is a **noise** object if it is not a core object but density-reachable from another core object.

The algorithm starts with the first point p in the database $D_j^{(i)}$, and it retrieves all the neighbors of a point p with respect to ε and $MinPts$. If p is a core point, this procedure yields a cluster concerning ε and $MinPts$. If p is a border point, no points are density-reachable from p, and the DBSCAN algorithm proceeds in considering the next point of the database.

2.2 Network of a DBSCAN object

The application of the DBSCAN algorithm defined in Section 2.1 produces, for each $\mathbf{D}^{(i)} = \{D_1^{(i)}, \ldots, D_{n_i}^{(i)}\}$, a DBSCAN object, i.e. a set of clusters $C^{(i)} = \{C_1^{(i)}, \ldots, C_{k_i}^{(i)}\}$ corresponding to a set of spatial coordinates of unit i which satisfy the maximality

and connectivity requirements. For each $C_m^{(i)}$, $m = 1 \dots k_i$, we can associate the time spent by the unit i into the cluster m by multiplying the cardinality of the cluster by the constant c.

A directed graph for a unit i is defined as $G^{(i)} = (V^{(i)}, E^{(i)})$ where $V^{(i)}$ is a set of nodes which corresponds to the set of clusters or points of interests $\bar{C}^{(i)} = \{1, \dots, k_i\}$, $E^{(i)}$ is a set of links (connections between points of interest for the i-th unit), where $e_{kl}^{(i)} = 1$ if the unit goes from node k to node l and 0 otherwise. The latter information is recovered thanks to the temporal ordering of the spatial movements.

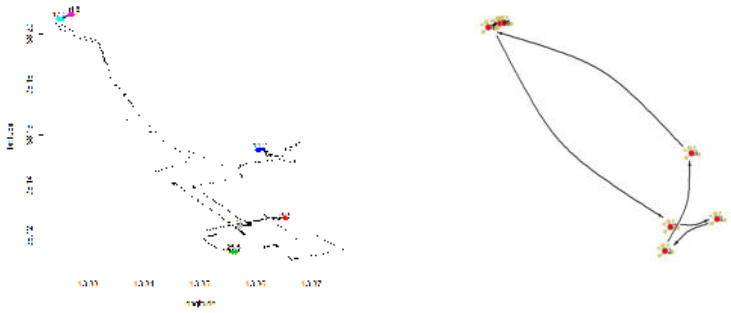

Fig. 2 Raw GPS data and clusters obtained with the DBSCAN algorithm (on the left); the numbers over each cluster represent the times spent by the unit into that cluster. Nodes and links (on the right) represent points of interest and sequence of visit, respectively, for the considered unit.

Figure 2 shows an example of the application of DBSCAN to $\mathbf{D}_1$, i.e. the temporal-spatial data for unit 1 (left part of Figure 2), and the recovered directed network (right part of Figure 2). In this example DBSCAN estimates seven clusters $C^{(1)} = \{C_1^{(1)}, \dots, C_7^{(1)}\}$ with times spent in each cluster that are equal to: $t_1^{(1)} = 5.5, t_2^{(1)} = 25.5, t_3^{(1)} = 14.2, t_4^{(1)} = 17.5, t_5^{(1)} = 9, t_6^{(1)} = 4.8, t_7^{(1)} = 5$ minutes, respectively. The directed network has seven vertices $V = \{1, \dots, 7\}$ and the directed links show the path of the i-th unit. In this example, we can recover from the temporal information that unit 1 starts from node 1 and chooses the following itinerary 1 - 2 - 3 - 4 - 5 - 6 - 5 - 4 - 7 - 1. Note that the positions of each node in the network correspond to the centroid of each cluster. According to this approach, the information about the noise points (i.e., the streets the units followed during their path) is lost, being not of interest for the present study.

2.3 Network Analysis

Let N be the number of collected sample units. So far, we have described an approach to derive a set of clusters and directed networks from the temporal-spatial GPS data $\mathbf{D}^{(i)}$ for each unit i. The application of the DBSCAN to $\mathbf{D}^{(i)}$, $i = 1, \dots, N$

produces the spatial data $\mathbf{C} = \{C^{(1)}, \ldots, C^{(N)}\}$ which allows to reduce the number of points collected for the sample units by discarding the noise points.

In this section, we define a weighted direct graph which summarizes the behaviors of all the sampled units. The weighted directed graph is a triplet $G = (V, E, W)$, where the set V represents the POIs of the collected sample units obtained by applying the DBSCAN algorithm to $\mathbf{C}$. In other words, the DBSCAN algorithm is applied at an aggregate level. Then the set of edges E represents the transitions from a POI to another one. An edge $e_{lk} = 1$ if a unit goes from cluster l to cluster k and zero otherwise. Finally, the weight matrix W represents the number of units that go from a node to another one. Specifically, w_{lk} indicates the number of units that went from the node l to the node k and the diagonal of W_{ll} represents the number of units that spent a certain amount of time into the cluster l

The identification of relevant nodes of the network G is one of the main applications of network analysis. We consider two measures to establish nodes importance: **degree centrality** which count of the number of directed connections of a node, and the **page rank centrality**, a variant form of eigenvector centrality, which has been used to evaluate the influence power of nodes on surrounding nodes [4].

3 Empirical Application

In this section, we apply the proposed methods on a sample of 303 GPS tracks related to cruiser passengers disembarked in the city of Palermo in 2014. Details on data collection procedures and other characteristics of the survey may be found in [1].

The implementation of the density-based algorithm described in Section 2.1 at the individual level allowed for a synthesis of the main POIs visited by each cruise passengers. Figure 3 shows, on the upper left side, the clusters of the entire city of Palermo which can be classified in three zones: Mondello (a beach place), Monreale (a mountain place) and the city center (historical area). The upper right side of the Figure shows, by zooming at the city center, many POIs. In order to recover the most important attraction points, we applied again the DBSCAN algorithm but at an aggregate level, i.e. on the clusters of Figure 3 upper right side. We obtaing the POIs showed in the left lower part of Figure 3. In the lower right part of Figure 3, we show the weighted directed networks of the main POIs visited by the units of our sample. Note that we cut the edge if $w_{lk} \leq 20$, which means that an edge will be present if more than 20 units go from node l to node k. The size of each POI is proportional to its degree. The colors of the nodes are associated with their measure of page rank centrality. The Figure shows a path that starts from the harbor (node 4) and ends to the Palermo's Real palace (node 8) which is one of the main attraction in the city. Node 8, which represents the Palermo' cathedral, looks like another relevant POI.

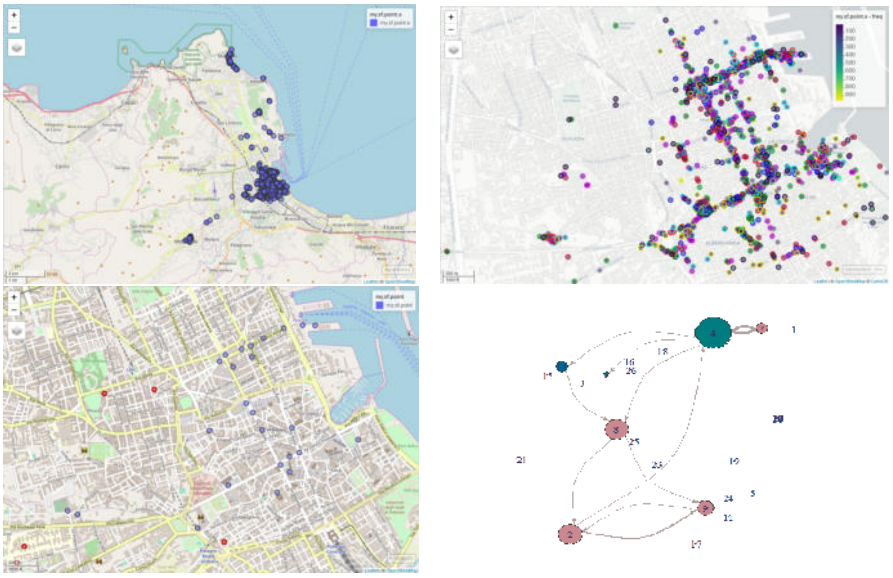

Fig. 3 Centroid Cluster recovered by using Algorithm 1. On the left side the cluster of the entire city of Palermo and on the right side the zoom on the city center.

4 Conclusion

In this paper, we proposed a density-based algorithm to reduce the complexity of GPS spatial-temporal data of a unit by finding the clusters which we interpreted as points of interest. These POIs can be summarized in a network. The information loss is minimal since the network gives the entire path of the unit unless a specific interest on the route followed by the unit is of importance. Moreover, we used the network analysis to merge the different units behaviors. At the individual level, the proposed methodology provides a synthesis of unit behavior in terms of main POIs visited and of the sequence of visits. At the aggregate level, it provides a structure of the main POIs, and on the general chain of visits, as well as on the strength of the links between each POIs in terms of flows of visits. The proposed methodology does not require any prior knowledge for the identification of the POIs, being easily replicable in different contexts.

References

1. De Cantis, S., Ferrante, M., Kahani, A., & Shoval, N. (2016). Cruise passengers' behavior at the destination: Investigation using GPS technology. Tourism Management, 52, 133-150.
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 96(33), 226-231.
3. Sia-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. International Journal of Geographical Information Science, 30(5), 881-906.
4. Sugimoto, K., Ota, K., & Suzuki, S. (2019). Visitor Mobility and Spatial Structure in a Local Urban Tourism Destination: GPS Tracking and Network analysis. Sustainability, 11(3), 919.
5. Varghese, M.,Unnikrishnan, A., & Jacob, K. (2013). Spatial clustering algorithms–An overview. Asian Journal of Computer Science And Information Technology, 3(11), 1-8.