EXTRACTING POINT OF INTERESTS FROM MOVEMENT DATA USING KERNEL DENSITY AND WEIGHTED K-MEANS

M. Malekzadeh [1,*], R. Javanmard [1], F. Karimipour [1]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran - (miladmalekzadeh, reyhanejavanmard, fkarimipr) @ut.ac.ir

KEYWORDS: Weighted K-means, Trajectory, Movement data, Kernel Density, Point of Interest

ABSTRACT:

Development in spatial data acquisition techniques, facilitate the process of analyzing movement characteristics and removed the lack of spatial data challenge. Annually, an enormous amount of spatial data are produced, and interpretation of this volume of data has become a major challenge. In this study, the movement data of 157 users in Geneva, Switzerland, were used and attempted to analyze their movement patterns. After the pre-processing stage, in order to investigate the dense areas, Kernel Density is calculated for each point for its neighborhood. The size of each cell of the output raster is approximately 100 meter. Afterward, in order to find the point of interests in the Geneva city, Weighted K-means is used for clustering of the raster. The kernel value of each cell is considered as the weight of the cell. Subsequently, the centroid of each final cluster has reflected the point of interest. As a final point, with the intention of assessing the results, the land use of the area is compared to each point of interest. Eventually, an interpretation is given.

1. INTRODUCTION

Movement can be considered as one of the fundamental elements of every creature on this planet or even the beleaguered atmosphere of the universe. Movement can create different states for an entity. It has temporal and spatial dimensions that can occur on different scales. Various studies have been conducted to analyze and interpret the movement with different approaches. The term of computational movement analysis has been introduced by (Gudmundsson, 2012) which represent a research trend which includes different areas. One of the most prominent examples is to study the patterns of animal migration over time (Holyoak, 2008). One of the considerable fields in the movement studies is analyzing and interpreting the human trajectories in order to improve transportation planning (Dodge, 2014) & (Gonzalez, 2008). Even in public health studies, movement analysis plays a significant role in modeling disease spread and pollutions (Tehophilides, 2006). The development of spatial data acquisition technologies, such as global positioning systems, provides a large amount of data every day for a variety of uses. The downward trend in prices of such devices, empowered everybody to produce movement data without high investment. Besides GPS, as the primary source of movement data, many technologies help movement studies with their generated movement data. Bluetooth, Wi-Fi positioning, and even communication systems can be named as systems which help the data acquisition process. Although the growth of volume of spatial data has made significant progress in the accuracy of movement studies, it also increases the computational cost of the analyses. Many studies have been conducted to extract the movement pattern of entities to reduce valuable data volume (Torrens, 2012), (Ewing, 2001) & (Maat, 2005). Data acquisition had been a cost-effective stage of movement analysis. Nonetheless, this stage has been minimized by the growth of technology; in return, the computational cost has risen. There are many methods to extract valuable information from raw movement data. From pattern detection in specified periods to analyzing each trajectory, the methods very.

The first step of the study is the pre-processing of the data. After the pre-processing stage, it can be claimed that the popularity of a cell has a direct relationship with the number of a cell meeting. In order to calculate the density of cells, Kernel Density is calculated for each point for its neighborhood. The next step is to cluster the data according to the kernel value of each cell, which is considered as its weight. The centroid of each cluster is considered as the point of interests. The final step is to interpret the result with regard to land use of the area.

2. DATA SET

Formerly, many movement studies were based on data acquisition by questionnaires (Yamamoto, 1999). For example, at the National Census in the United States, information about the workplace and the place of living of thousands of people was obtained. This census was conducted every ten years (Becker, 2013). Data acquisition with this method was omitted because of high cost, high time-consuming, low accuracy, and lack of repeatability. In recent years, with the increase in the number of mobile users, radio receivers, transmitter and receiver stations, surveillance cameras and a variety of sensors such as global positioning system, gyroscope, accelerometer, camera, and microphone, people's lifestyle has been changed and a large amount of high-precision movement data are being produced. The availability of mobile data has played an essential role in the study of various fields such as geography, transportation, economics, advertising, urban modeling, and air pollution. Researchers are using large-scale mobile data to analyze behavioral issues and social sciences, and because of the availability of this type of data set, the number of studies in the field of understanding movement behaviors, Communication and the pattern of interactions is increasing.

In January 2009, the Nokia Research Center, in collaboration with the Swiss Academy of Sciences and EPFL, decided to produce large-scale mobile phone data [8]. For this purpose, the location-dependent data types (spatial positioning system, global communication system), motion (acceleration), proximity

* Corresponding Author

(Bluetooth), communications (call and message), media (camera, media player) of about 200 volunteers (including 62% of women and 38% of men, aged between 22 to 33 years) were captured over a year in the Geneva Lake area. (Table 1)

Data type	Quantity
Calls (in/out/missed)	240227
SMS (in/out/failed/pending)	175832
Photos and Videos	40091
Application events	8096870
Calendar entries	13792
Photobook entries	45928
Location points	26152
Unique cell tower	99166
Accelerometer samples	1273333
Bluetooth observations	38259550
Unique Bluetooth devices	498593
WLAN observations	31013270
Unique WLAN access points	560441
Audio samples	595895

Table 1. Data types and their quantity

3. IMPLEMENTATION

3.1 Pre-Processing

Segmentation and filtering are two crucial steps of the movement studies (Laube, 2014). In this study, firstly, noises of the trajectories are removed by applying Douglas-Peucker line simplification. Since the data may contain some periods which the entity is not moving while the tracker records the location, each trajectory segmented into sub trajectories in the next step. Pseudo movement can be recognized by considering a circle with a specific radius around each point of the trajectory (Laube, 2011) and calculate the number of points the circle covered.

3.2 Kernel Density Estimation

Generally, Kernel Density can be defined as an interpolation or smoothing technique that estimates the probability density of points in the area of the cells in the network where they are located. Kernel Density Estimators is from Non-Parametric density estimators' class. Despite parametric estimators which have a defined functional structure and this structure is the only facts we need to store, Non-Parametric estimators depend upon all the data which are used in estimation and not only a fixed structure. This interpolation can result in the calculation of the value of any point, cell, or a subset of the investigated environment. Kernel density allows to simultaneously examine the effects of the position and the number of visits. The initial process of calculating the kernel density is to transform the discrete-space into the continuous density space. Kernel density has been used in various studies. (Kwan, 2000) & (Buliung, 2001) have used kernel density to interpret big data. Three-dimensional visualization of Spatio-Temporal data is one of the capabilities of Kernel Density which is taken into consideration in the modern spatial science.

The first step of implementing kernel density on the data is to define a Kernel function. Usually, a Kernel function is a smooth unimodal function, but there are various kernel functions such as Triangle, Epanechnikov, Quartic, Triweight, Sigmoid function, Gaussian, and Cosine. In this study, a quartic kernel function is considered which is defined as below (Equation 1):

$$K(u) = \frac{15}{16}(1-u^2)^2 \quad (1)$$

Where: K = Kernel Density
u = Value

The kernel function is implanted on each point of the data. The bandwidth may vary according to the level of smoothing. The bandwidth can be defined as search radius of the Kernel function. Defining a precise and accurate bandwidth is a challenge. It reflects the maximum distance which activity locations can have spatial effects on each other. The value of bandwidth should be logically acceptable. In order to define the bandwidth, the mean travel distance which is approximately 20000 meters is considered. The final value of each cell is the summed value of the overlapping densities. Kernel density formula can be defined as below (Equation 2) in a grid structure:

$$\lambda(p) = \sum_{di<\tau} K\left(\frac{di}{\tau}\right) \quad (2)$$

Where: λ = Density
τ = Bandwidth
K = Kernel Function
di = Distance between grid point of p and the ith point of the data

Figure (1) shows the visualization of the Geneva movement data by implementing Kernel Density using the quartic function.

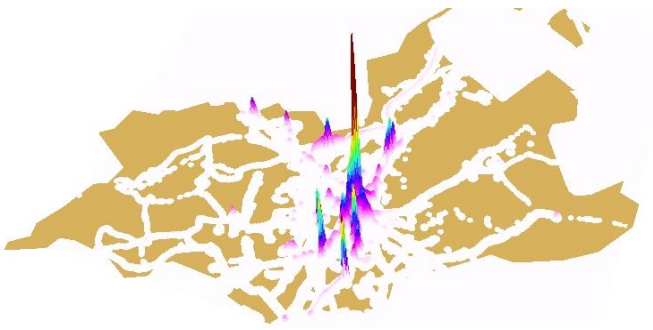

Figure 1. Kernel density visualization of Geneva movement data

3.3 Clustering

One of the challenging steps of this study is to extract the hot spot of the city considering the number of visits by people. Weighted K-Means clustering method is used to cluster the output raster of the kernel density calculation stage. Basic K-Means method has no capacity to handle the clustering of a raster considering the weight of each cell. In order to cluster the rater centroid of each cell is considered as the represented point

of the cell with the weight of the cell, so Weighted K-Means is defined (Equation 3 & 4):

$$\underset{ci \in c}{argmin} = wj * dist\,(ci, xj)^2 \tag{3}$$

Where: c_i = Centroid of the i^{th} cluster
w_j = Weight of the x_j
x_j = Representor point of j^{th} cell

And

$$c_i = \frac{\sum_{x_j \in s_i} w_j x_j}{\sum w_j} \tag{4}$$

Figure (2) shows the clusters with their centroids.

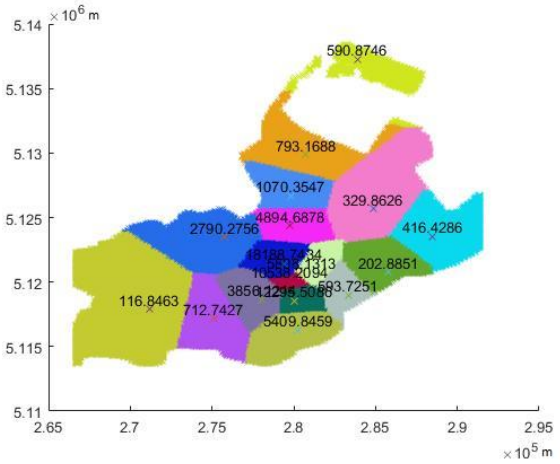

Figure (2) Output clusters of the clustering process

Clustering steps may result in cold spot clusters. To eliminate the cold stop, all of the clusters with the mean of lower than the third quarter of the entire data are removed. The result of the hot spot areas and their centroids are shown in Figure (3).

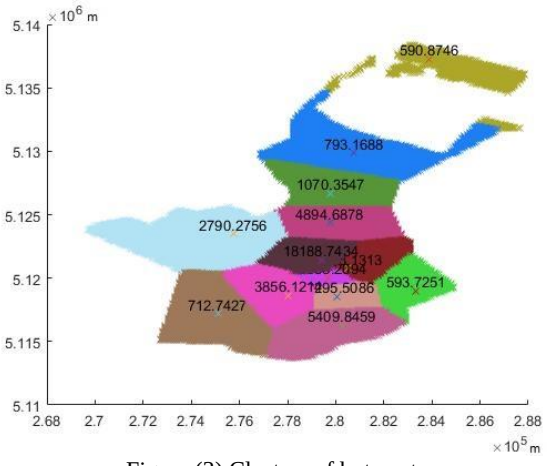

Figure (3) Clusters of hot spots

4. CONCLUSION

Finding hot spots considering movement data is always a challenge according to the high volume of the data. In this paper, a novel approach using Kernel Density and Weighted K-Means Clustering method is stated. To conclude, the approach finds thirteen hot spot locations which are shown in Figure (3). The locations were compared to the land-use of the area. The comparison area is a square with a length of 100 meters in accordance with the size of the cell size of the raster. As an example, the centroid of class number 5 is compared with the land-use around it in Figure (4).

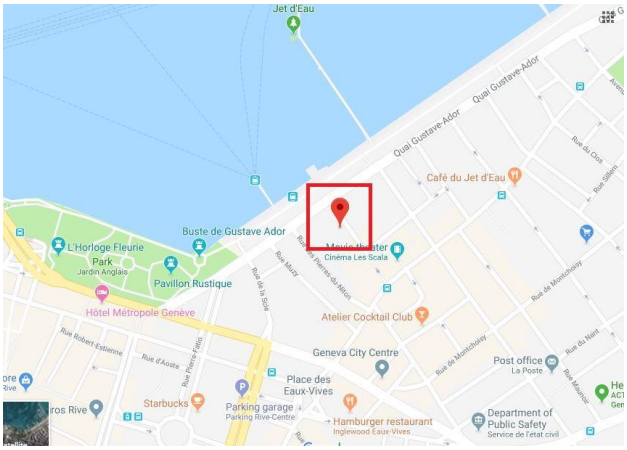

Figure (4) Centroid of Class number 4 and the area around it

The area is so near to the City Center, Movie Theater, many restaurant, and clubs. It is admittedly easy to interpret why the location represents a hot spot area of the city. The other locations as well can be acceptably interpreted. There are four centroids which represent the entertainment venues, two represent Geneva Landmarks, and six represent intersections which two main roads intersect. There was just one class centroid which has no acceptable interpretation.

REFERENCES

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., ... & Volinsky, C. (2013). Human mobility characterization from cellular network data. Communications of the ACM, 56(1), 74-82.

Buliung, R. N. (2001, July). Spatiotemporal patterns of employment and non-work activities in Portland, Oregon. In Proceedings of the 2001 ESRI International User Conference.

Dodge, S., Bohrer, G., Bildstein, K., Davidson, S. C., Weinzierl, R., Bechard, M. J., ... & Wikelski, M. (2014). Environmental drivers of variability in the movement ecology of turkey vultures (Cathartes aura) in North and South America. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1643), 20130195.

Ewing, R., & Cervero, R. (2001). Travel and the built environment: a synthesis. Transportation research record, 1780(1), 87-114.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. nature, 453(7196), 779.

Gudmundsson, J., Laube, P., & Wolle, T. (2011). Computational movement analysis. In Springer handbook of geographic information (pp. 423-438). Springer, Berlin, Heidelberg.

Holyoak, M., Casagrandi, R., Nathan, R., Revilla, E., & Spiegel, O. (2008). Trends and missing parts in the study of movement ecology. Proceedings of the National Academy of Sciences, 105(49), 19060-19065.

Kwan, M. P. (2000). Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. Transportation Research Part C: Emerging Technologies, 8(1-6), 185-203.

Laube, P. (2014). Computational movement analysis. Berlin: Springer International Publishing.

Laube, P., & Purves, R. S. (2011). How fast is a cow? Cross-scale analysis of movement data. Transactions in GIS, 15(3), 401-418.

Maat, K., Van Wee, B., & Stead, D. (2005). Land use and travel behaviour: expected effects from the perspective of utility theory and activity-based theories. Environment and Planning B: Planning and Design, 32(1), 33-46.

Theophilides, C. N., Ahearn, S. C., Binkowski, E. S., Paul, W. S., & Gibbs, K. (2006). First evidence of West Nile virus amplification and relationship to human infections. International Journal of Geographical Information Science, 20(1), 103-115.

Torrens, P. M., Nara, A., Li, X., Zhu, H., Griffin, W. A., & Brown, S. B. (2012). An extensible simulation environment and movement metrics for testing walking behavior in agent-based models. Computers, Environment and Urban Systems, 36(1), 1-17.

Yamamoto, T., & Kitamura, R. (1999). An analysis of time allocation to in-home and out-of-home discretionary activities across working days and non-working days. Transportation, 26(2), 231-250.