CO-281

AREA-UNIT DENSITY ESTIMATION FOR AGGREGATED MASS LOCATION DATA

MURPHY C.
Technische Universität München, MUNICH, GERMANY

ABSTRACT

This research presents an approach for analysing and visualising aggregated mass location data. An increasing amount of location data have been collected in the past by geo-sensor networks. A major part of location data are aggregated and collected with a spatial uncer-tainty. Based on Kernel Density Estimation (KDE) and existent methods for reducing map complexity, continuous density estimates are made for area-unit data to enhance the under-standing of event patterns. The performance of the presented approach is tested on a mobile phone location dataset. For this dataset every outgoing mobile phone call was logged and as-signed to the location of a base station. The records are therefore spatially aggregated within network cells. A choropleth map derived from Voronoi polygons is used as a simple mobile network propagation model to assign the events to area-units. After a grid is laid onto the area-units a kernel function is placed over every grid point. The function weighs the area-unit values within a set bandwidth depending on spatial similarity. The output is a continuous in-tensity surface that visualises the density of occurring events. This has four main benefits: (1) the choropleth visualisation is transformed to an isopleth map, which enhances information communication; (2) the density values are detached from uncertain but prior strictly defined boundaries; (3) outliers are adjusted according to their spatial neighbourhood; and (4) the smoothening effect of the approach enables to identify hotspots and overall trends. The smoothening effect takes account on the spatial uncertainty of the propagation model, on the with uncertainty connected occurrence of outliers and on the high number of diverse area-units due to being mass data.

The approach is designed for large amounts of spatially aggregated event data which generate a high number of area-units. It is also suitable for choropleth maps with a high number of faces when area-units are standardised by surface area. Area-Unit based KDE takes account on the spatial uncertainty of events, by considering mean values weighted by area to meet the assumption of the continuous phenomenon of mobile phone events. In contrast the statistically proven point pattern analysis method of KDE analyses solely the locations in which the data is collected, but disregards the possibility of values occurring in between conceptual points. The Area-Unit based KDE provides analysis of aggregated mass location data as a phenomenon presumed to be smooth and continuous. This helps to understand trends and identify spatial patterns as well as hotspots and therefore facilitates assessment, planning and decision making based on mass location data.

BACKGROUND AND OBJECTIVES

To consider any random distribution one soon comes across density estimation. Density esti-mation is the construction of an estimate of an applied density function (Silverman, 1986). Density estimation is also a graphical technique for analysing location data. Therefore it can be placed into the field of visual analytics. Keim et al. (2008) state that visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets. The exploration and presentation of location data becomes more challenging with highly increasing numbers of events. A low quantity of locations, for instance the density of trees in a garden, might not require density estimation. But things change with large numbers of events over vast areas. The improvement of technical devices and lower costs of geo-sensors allow monitoring people and objects worldwide. This has lead to the storage of huge amounts of complex location data sets. These huge amounts of complex location data, in this paper re-ferred to as mass data, are collected with a wide range of spatial accuracy. For instance, mass data measured with Global Navigation Satellite Systems (GNSS) can have a spatial accuracy of a few meters or better. Mass data in terms of mobile phone location data collected from Cell Ids (as used in this paper) have a spatial uncertainty as large as the cells themselves. They are therefore spatially aggregated and must be dealt with as area-unit data.

Large sets of aggregated location data can be visualised as choropleth maps (figure 1). The colour coded enumeration units enable to visually assess spatial patterns. The interpretation will vary from person to person and can be misled by outliers or chance factors. For these reasons, it makes sense to compute a numerical measure of spatial pattern (Slocum et al., 2005).

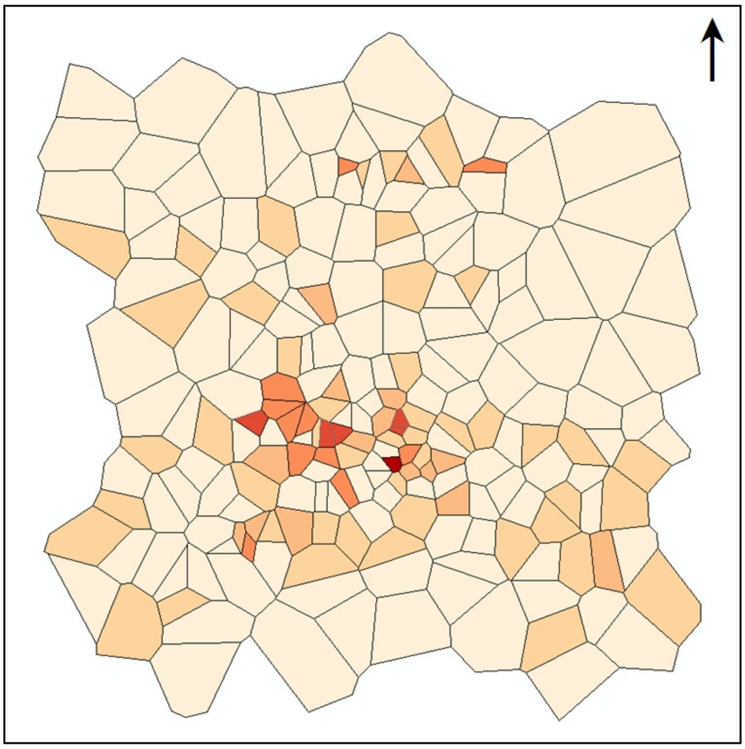

Fig. 1: Aggregated event data: here as a mobile phone call census in Munich visualised as a choropleth map

Descriptive statistics of polygons or areas are needed for the analysis of area-unit data. There has been research for assessing spatial patterns in land cover maps referred to as Landscape Pattern Analysis (Johnson and Patil, 2006). Landscape Pattern Analysis is a primary research tool in landscape ecology that contributes to understanding spatial ecological dynamics (Fu and Chen, 2000). Though, as this research field is based on satellite imagery it deals with areas in an ordered lattice. Since aggregated event data mainly has area units in diverse sizes and shapes Landscape Pattern Analysis cannot be adopted.

Another approach coming into consideration for the analysis of aggregated mass location data is used for simplifying choropleth maps. Herzog (1989) developed an algorithm that enhances the pattern recognition of area unit data. For this approach an adjustment is made for every area unit value. In general an area unit´s value is a function of its own value plus the neighbouring area units' values. The original area unit value has the highest influence in Herzog´s approach. The neighbouring values are weighted depending on the length of the shared boundary. The output is a smoothed value surface choropleth map. Since this approach reduces the map complexity it should be performed on choropleth maps with a high number of faces.

For describing and comparing large location datasets Point Pattern Analysis is a powerful tool. Point Pattern Analysis provides the detection of hotspots and trends within the data set and can test whether there is a significant difference to a random spatial point pattern. A popular Point Pattern Analysis method is Kernel Density Estimation (KDE). For KDE a series of estimations are made over a grid of the study area. A kernel is then placed onto every grid point. The user can choose between different kernel types and has to set the bandwidth which defines the range of the kernel. Each estimation shows the point density at a certain location. For KDE the user receives an intensity value at all places. This gives the advantage of a throughout covered study area. An extensive introduction of KDE is found in Diggle (2003). KDE shall be adopted in this paper to find an appropriate method for the analysis of aggre-gated events. Although KDE is only used for point data the approach for determining the intensity of aggregated events is quite similar. This paper states that KDE can be used to ana-lyse spatial patterns of in distinct boundaries aggregated mass location data.

INPUT DATA

As case study data serves a mobile phone location dataset from Vodafone. The outgoing phone calls broadcasted by base stations were collected taken from an approximately 7 x 7 km boundary centred over Munich city centre during one week. Every mobile phone call was logged and assigned to the location of a base station. A resulting 1.5 million events derived from 216 base stations were stored. For privacy issues all records were stored anonymously by removing the formal identifiers, such as phone number, phone ID, etc..

In the event of a person making a call with their mobile phone, the phone connects to the base station with the strongest radio signal reception. The connected link between mobile phone and base station is in the majority of cases the nearest located base station. The radio signal between mobile phone and base station spreads under electromagnetic wave propagation rules but is affected by obstacles. The height of the base station placement is a major influence on the availability of phone connection. But reflection and inflection effects on objects like buildings or transport or remain to a certain degree unpredictable. Theoretical investigations can only consider a simplified model of the environment (Lüders, 2001). The case study data is therefore modelled by Voronoi diagrams to keep the propagation model simple. The Vo-ronoi diagram partitions space into polygons that contain one generating point (the base sta-tion) each. Every point in the area of a given polygon is closer to its generating point than to any other. Voronoi diagrams have been formerly used to model wireless networks for instance by Fanimokun and Frolik (2003) as well as Meguerdichian et al.(2001). The Voronoi diagrams are also an effective tool for visualizing cell coverage areas. Despite the clear boundary definition through Voronoi cells the mobile phone call locations are not always true. A phone call log connected to one base station could be made from an adjacent cell. However, the Vo-ronoi areas will mostly contain phone activity from inside the boundaries.

PROBLEM STATEMENT AND APPLIED METHOD

In conventional KDE only point values are considered. When area-units are to be analysed a specific point which represents the unit has to be taken into account. From a number of possi-bilities it would be reasonable to use the centroid of an area-unit or in the case of Voronoi tes-sellation the centre points of every Voronoi region. These value points represent the data value of the area-unit and are referred to as conceptual points. But when the KDE comes into action several issues have to be addressed to. As the kernels are placed onto the grid points, kernels that do not cover any value points will assign zero density to its grid point even though the phenomenon is continuous. On the other hand a kernel that is close to a value point or covers several value points will assign a significantly higher density to its grid point. The KDE output surface will show a high interaction with the value points. Hot spots would rather be detected close to value points and around closely located value points. Larger area-units with fewer value points will tend to result in low densities independently from their area-unit value. This is a phenomenon closely related to the ecological fallacy. Wrigley et al (1996) state that the ecological fallacy involves the inappropriate inference of individual-level relationships from areal-unit-level results. It arises typically, when areal-unit data are the only source available to the researches but the objects of study are individual-level characteristics and relationships. It is obviously not a useful process to focus on the centroids or the centre points of Voronoi regions for analysing aggregated event data. This is a known fact to researchers. For example O´Sullivan and Unwin (2003) state that one important criterion for using Point Pattern Analysis (and therefore KDE) is to use true locations of event data. They further state that the points must be true incidents with real spatial coordinates. A general formulation for KDE is shown in equation (1).

Formula based on Scott (1992):

$$\hat{f}_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \qquad (1)$$

With:

$\hat{f}_h(x)$ = general Kernel Density

K = Kernel function

h = Kernel radius (bandwidth)

n = Number of points within the kernel radius

$x_1, x_2, \ldots, x_n$ = points within the kernel radius

So, apparently KDE rules the analysis of aggregated events out. But a few modifications of the KDE principle combined with elements of Herzog´s approach (1989) can lead to a much better analysis of enumeration units. A closer look has to be taken onto the kernel range. The kernel bandwidth defines an area in which in the Point Pattern Analysis case all events are taken into account that fall into the kernel. All points that lie outside the kernel range are not considered. An Area-Unit based KDE should take every distinct polygon area inside the ker-nel into account, before assigning a density value to the grid point. On the basis of the general formulation of KDE, a calculation is suggested formulated in the following equation (2).

Formula for Area-Unit Density Estimation:

$$\hat{f}_h(A) = \sum_{i=1}^{n} \left(\frac{A_n \cdot W_n}{\pi \cdot h^2} \right) \qquad (2)$$

With:
$\hat{f}_h(A)$ = general kernel Area-Unit Density
h = bandwidth
A_1, A_2, ..., A_n = area portion of area units located within the kernel range
W_n = area unit value (relative)

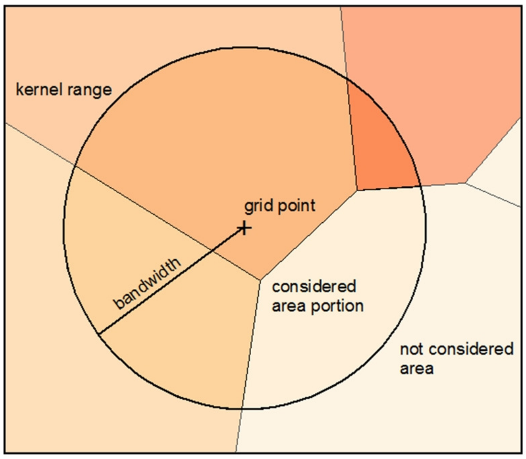

Fig. 2: Area-Unit Density Estimation

Input values of this function are the kernel bandwidth, the partial areas of the area-units lo-cated within the kernel range and the corresponding values of the area units. Note that the area unit values should be relative values. In comparison to the selection of appropriate data for choropleth maps, the area-units must be standardised by surface area. The varying sizes of area-units require values in proportion to the surface area. The main principle of this method is showed in figure 2.

Equation (2) formulates a Density Estimation for Area-Units with a uniform kernel. Every area section within the kernel range has the same influence on the resulting density value. The equation is applied to every grid point. This leads to a wholly covered study area with point density values.

RESULTS

The here developed Area-Unit Density Estimation was performed on the case study data set. A bandwidth of 300 m was used. After setting a high resolution of 234,000 grid points the result is visualised as an isometric map (see figure 3). Isolines respectively borders are class limits between different values. Nine colour coded intervals were classified by standard deviation. Darker colours represent a higher mobile phone call density. A layer enclosing the buildings of Munich inside the study area has been added to allow orientation within the map.

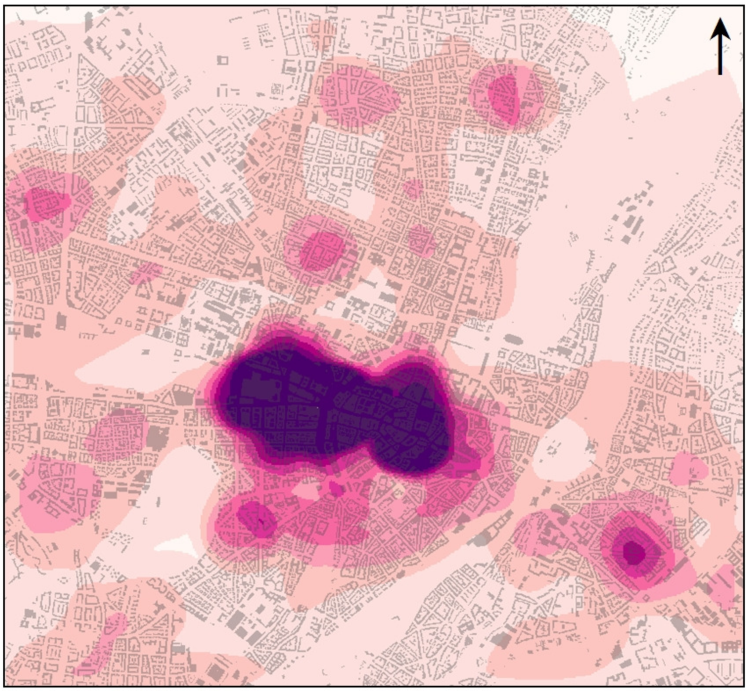

Fig. 3: Mobile Phone Density in Munich from Area-Unit Density Estimation.

Figure 3 shows a very high density in the middle of the study area which is the city centre. The perhaps not surprising phenomenon of a high mobile phone activity in city centres has been surveyed before, for instance by Ahas et al. (2010). Compared to the choropleth map of figure 1 it becomes much easier to identify major hotspots along Munich´s main shopping street between the main station on the western border of the highest value class and the town hall of Munich (more to the east), but also minor hotspots like the ones by the “Ostbahnhof” (East Station) and the ”Münchener Freiheit” (famous square in the city of Munich). At the same time regions with low mobile phone traffic can easily be depicted. In contrast to the mo-bile phone calls visualised as a choropleth map in figure 1 it is clearly visible that density val-ues and spatial borders between adjacent classes are no longer constrained to the Voronoi polygons. This makes sense if one keeps in mind that the polygon borders are only the result of a simple propagation model and the assignment of locations to an area-unit are made with a degree of spatial uncertainty. On this account it can be said that no detail information is lost. One other main difference of the Area-Unit Density Estimation result in comparison to the choropleth visualisation is that the spatial data has been smoothed. Area-Unit Density Estima-tion considers a number of values within the kernel. Subject to the bandwidth this leads to a continuously smooth appearance. This is no setback following the propagation model´s spatial uncertainty.

CONCLUSIONS AND DISCUSSIONS

As demonstrated within the case study the Area-Unit Density Estimation output is a continu-ous intensity surface that visualises the density of occurring events. The original choropleth map is transformed to an isopleth map. This enhances the usability of the visualisation, be-cause choropleth maps are consistently perceived as more complex than isopleths maps made from the same data (MacEachren, 1982a).

The approach is designed for large amounts of spatially aggregated event data which generate a high number of area-units. The major contribution compared to the point based KDE is that the whole continuous study area is considered and not only conceptual points. The distribu-tion of density values over the study area are no longer bound to the area-units. The approach is adapted for the special task of analysing aggregated mobile phone location data and as a result disentangles the output map from the zonal configuration of area-units. Depending on the set bandwidth it smoothes the density result and diminishes outliers.

The effectiveness of this approach is highly dependent on the raw data. Area-unit data is col-lected within the boundaries of the area. Due to the fact that the true locations remain hidden Area-Unit Density Estimation therefore uses average values to represent a certain area-unit. A limitation to this approach is that it does not consider the configuration of area-units. The Area-Unit Density Estimation method as formulated in equation (2) is designed for a uniform kernel set by a fixed bandwidth and a circular shaped footprint. To analyse inhomogeneous sizes and shapes of area-units adaptive bandwidths dependent on the area-unit geometry could be applied. Future modifications of the Area-Unit Density Estimation could also include the use of other kernel functions. For instance, the established Gaussian kernel could weigh values depending on distance. The volume portions under the kernel curve could then be calculated with finite elements. The continuous intensity appearance of Area-Unit Density Estimation is highly dependent on the selected bandwidth. The choice of bandwidth sets the degree of smoothening. To find an appropriate bandwidth for Area-Unit Density Estimation one could apply methods to determine an "optimal" bandwidth (Silverman, 1986) or use a visual based computational tool (Krisp et al., 2009). Nevertheless, the smoothening of the original area-units should follow a rule well known in adjustment theory. The rounding of a numerical value in statistics indicates the accuracy of a computed number. Such rule can be transformed to visualisation: Visualise with no higher grade of accuracy than the raw data is based on. The smoothening of an Area-Unit Density Estimation should therefore indicate the area-units´ level of uncertainty. In the presented case study Area-Unit Density Estimation also reduces the impact of outliers due to propagation modelling errors immensely. This helps to visually explore the study area and to identify spatial patterns and hotspots and therefore proves to be an effective tool for describing and visualising spatially aggregated mass location data.

REFERENCES

Ahas, R., Silm, S., Järv, O., Saluveer, E. and Tiru, M., 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. Journal of Urban Tech-nology, 17(1 April 2010): 25.

Diggle, P.J., 2003. Statistical Analysis of Spatial Point Patterns, 2nd edition. Edward Arnold, London, 159 pp.

Fanimokun, A. and Frolik, J., 2003. Effects of natural propagation environments on wireless sensor network coverage area, IEEE Southeastern Symposium on System Theory, Morgantown.

Fu, B. and Chen, L., 2000. Agricultural landscape spatial pattern analysis in the semi-arid hill area of the Loess Plateau, China. Journal of Arid Environments, 44: 13.

Herzog, A., 1989. Modeling reliability on statistical surfaces by polygon filtering. In: M.F. Goodchild and S. Gopal (Editors), Accuracy of Spatial Databases. Taylor & Francis Ltd, London, UK, pp. 8.

Johnson, G.D. and Patil, G.P., 2006. Landscape Pattern Analysis for Assessing Ecosystem Condition. Environmental and Ecological Statistics. Springer Science+Business Media, New York, 130 pp.

Keim, D. et al., 2008. Visual Analytics: Definition, Process, and Challenges. In: A. Kerren, J.T. Stasko, J.-D. Fekete and C. North (Editors), Information Visualization - Human-Centered Issues and Perspectives. Springer, Berlin, pp. 154-175.

Krisp, J.M., Peters, S., Murphy, C. and Fan, H., 2009. Visual Bandwidth Selection for Kernel Density Maps. PFG - Photogrammetrie Fernerkundung Geoinformation, 5/2009: 10.

Lüders, C., 2001. Mobilfunksysteme. Kamprath-Reihe. Vogel Buchverlag, Würzburg, 357 pp.

MacEachren, A.M., 1982a. Map Complexity: Comparison and Measurement. The American Cartographer, 9(1): 16.

MacEachren, A.M., 1982b. The Role of Complexity and Symbolization Method in Thematic Map Effectiveness. Annals of the Association of American Geographers, 72(4): 19.

Meguerdichian, S., Koushanfar, F., Potkonjak, M. and Srivastava, M., 2001. Exposure in Wireless Ad Hoc Sensor Networks, Proceedings of the Seventh Annual International Conference on Mobile Computing and Networking, Rome, Italy.

O´Sullivan, D. and Unwin, D.J., 2003. Geographic Information Analysis. John Wiley and Sons, Inc., Hoboken, New Jersey.

Scott, D.W., 1992. Multivariate Density Estimation: theory, practice, and visualization. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., Canada, 376 pp.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, 26. Chapman and Hall, London, 176 pp.

Slocum, T.A., McMaster, R.B., Kessler, F.C. and Howard, H.H., 2005. Thematic Cartography and Geographic Visualization. Prentice Hall Series in Geographic Information Science. Pearson Prentice Hall, Upper Saddle River, NJ, 518 pp.

Wrigley, N., Holt, T., Steel, D. and Tranmer, M., 1996. Analysing, modelling, and resolving the ecological fallacy. In: P. Longley and M. Batty (Editors), Spatial Analysis: Model-ling in a GIS Environment. John Wiley and Sons, Inc., New York, pp. 23-40.