udsu

disclaimer

Article:

Produit T, Lachance N, Strano E, Porta S, Joost S, A network based Kernel Density Estimator applied to Barcelona economic activities, in Taniar et al. (eds.), ICCSA 2010: Part I.

Disclaimer:

A Network Based Kernel Density Estimator Applied to Barcelona Economic Activities

Produit Timothée[1], Lachance-Bernard Nicolas[1], Strano Emanuele[2], Porta Sergio[2], and Joost Stéphane[1]

[1] Laboratory of Geographic Information Systems, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[2] Urban Design Studies Unit, Department of Architecture, University of Strathclyde, 131 Rottenrow, Glasgow, G4 ONG, United Kingdom

Abstract. This paper presents a methodology to compute an innovative density indicator of spatial events. The methodology is based on a modified Kernel Density Estimator (KDE) that operates along road networks, and named Network based Kernel Density Estimator (NetKDE). In this research, retail and service economic activities are projected on the road network whose edges are weighted by a set of centrality values calculated with a Multiple Centrality Assessment (MCA). First, this paper calculate a density indicator for the point pattern analysis on human activities in a network constrained environment. Then, this indicator is modified to evaluate network performance in term of centrality. The methodology is applied to the city of Barcelona to explore the potential of the approach on more than 11,000 network edges and 166,000 economic activities.

1 Introduction

Most of the conventional indicators and spatial interpolation techniques use Euclidean distances for space characterization [21]. In the urban environment, these approaches do not take into account the road network constraint and its influence on the location of spatial events.

KDE is widely used as a spatial smoothing technique in many fields such as geography, epidemiology, criminology, demography, hydrology and others (Anselin [1]; Borruso [4]). In a recent work, Porta *et al.* [18] extended the use of planar KDE to examine whether the variation of street centralities was reflected in the intensity of land uses. Recent developments of the same idea have stressed the network constrained nature of some classes of point events such as crime occurrences or car accidents, therefore exploring the advantages of computing density on the network (Network based Kernel Density Estimation, or NetKDE) rather than in the planar space (see Ch. 1.3).

Like these recent works, this paper examines a network oriented approach to density, but still operated over the 2D space, like the conventional Kernel Density. The research was carried out by computing NetKDE on two kinds of spatial entities within the city of Barcelona: firstly, retails and services represented as points; secondly, the road network represented as polylines weighted by a set of

D. Taniar et al. (Eds.): ICCSA 2010, Part I, LNCS 6016, pp. 32–45, 2010.

centrality values resulting from a Multiple Centrality Assessment (MCA) of the same road network (see Ch.1.2).

This paper is divided in three parts. First, the theoretical background of KDE, of network centrality and of the novel NetKDE is presented. Secondly, the methodology to implement NetKDE is explained. And finally, the results of its application to the city of Barcelona case study are presented and discussed.

1.1 Kernel Density Estimation

KDE is a statistical process used for spatial smoothing and/or spatial interpolation [25]. This paper uses KDE to transform the road network MCA measures and the distribution of activities to a common spatial unit (the raster grid). This allows subsequent visual correlation analysis. It is also recognized that the function of density is a means to present analysis and illustrations of complex and technical data in a clear and understandable way to the non-mathematicians [20].

KDE is a well known tool in urban studies. Anselin [1] used the KDE for spatial analysis of crimes to visually simplify their location and to examine the complex characteristics of criminal incidents. Gatrell [13,14] analyzed spatial first-order variation in disease risk with kernel functions. Borruso [4] showed that the KDE applied to an urban system (a density analysis of addresses) allowed a better representation of the phenomenon. He also found out that KDE is less sensitive to size, position and orientation than grid density estimation (GDE). Thurstain-Goodwin and Unwin [22] applied KDE to zip codes data to obtain a continuous surface of density. Associated with indicators of centrality, the kernel density allowed calculation of a composite urban centrality indicator applicable to cities.

KDE is a function balancing events accordingly to their distances and required two parameters. The first is the bandwidth, which is the distance of influence. The choice of the bandwidth has a great impact on the results, some authors used a least-squares cross validation to select the bandwidth [19,24]. Others as Brundson [5] proposed an adaptive KDE with a bandwidth changing accordingly to the cloud of points structure. The second parameter is the weighting function K, which is most often a normal function. Authors agreed that the choice of this function is less critical than the choice of the bandwidth [13].

The technology development of the last decade results in new opportunities in GIS research. Miller [15] pointed out that the hypothesis of a continuous space is too strong for the analysis of events which take place in a one-dimensional sub-space created by a network. Similarly, Batty [2] showed that GIS prevent Euclidean space from being distorted by the constraint of the road network. He also noted that though road network representation is no longer a challenge, further developments on GIS have to take into account this constraint.

1.2 Network Centrality

The urban network has been the object of numerous studies. Its origins can be traced back to Leonhard Euler's solution of the Knigsberg bridges problem,

after which the theory of graphs and complex network has been rapidly growing (Euler [11]). Freeman [12] was one of the first researchers to define sets of indices to measure how central is a node with respect to all others. Nowadays, because of this virtually unlimited capacity to represent relationships in most diverse real systems, networks are used in a variety of fields such as ecology, genetics, epidemiology, physics, communications, computing, urban planning and many others (Costa *et al.* [6]).

The specific category of geographical networks is characterized by nodes with well defined coordinates in an embedded space, like street networks. The geography of centrality on such networks has been explored by means of a Multiple Centrality Assessment, as defined by Porta *et al.* [17]. The main characteristics of MCA are: (1) a 'primal' format for the street network; (2) a metric computation of distances along the real street network; (3) an attribution to each street of diverse centrality values (Crucitti *et al.* [7]). This paper presents only one of those computed indices, namely betweenness centrality (BetC). BetC is based on the idea that a node is more central when it is traversed by a larger number of shortest paths connecting all couples of nodes in the network. BetC is defined as:

$$C_i^B = \frac{1}{(N-1)(N-2)} \sum_{j,k \in N; j \neq k; j,k \neq i} \frac{n_{jk}(i)}{n_{jk}} \tag{1}$$

where n_{jk} is the number of shortest paths between nodes j and k, and $n_{jk}(i)$ is the number of these shortest paths that contain node i.

1.3 Network Based Kernel Density Estimation

NetKDE use similar mathematical formula as KDE, but uses distances measured along a network rather than Euclidean distances to compute density values. NetKDE is applicable to polylines and points. Both are balanced by the distance between an event and the point at which the density is estimated.

A study on road accidents demonstrated "*the risks of false positive detection associated with the use of a statistic (K-function) designed for planar space to analyze a network constrained phenomenon*" [27].

Borruso [3] suggested an analysis of density in the space created from the urban network. His network constrained density indicator, named NDE (Network Density Estimation) was applied to the cities of Trieste (IT) and Swindon (UK) for activities related to banks and insurances. He showed up that "*the difference between the KDE and the NDE was not very high, but NDE seems to be more proficient in highlighting linear clusters oriented along a street network*". Borruso also pointed out that the NDE performed better than KDE for the identification of linear patterns along the network. However, the calculated density does not take into account a distance weigthing function such as that of standard KDE.

Xie and Yan [26] came back to an investigation of car accidents as a champion case of network constrained point events: they developed a methodology to reduce the planar 2D KDE to a linear 1D measurement on the basis of a linear unit of conventional length, or "lixel". This reduction however also reduced the

real meaning of applying a kernel function, as a simple evaluation of the number of events per linear unit provides far more precise results. Moreover, this study concluded that planar 2D KDE overestimated density as compared with their network KDE. While this conclusion is brought over figures, these are hardly comparable being in different space units (respectively pixel and lixel).

Finally, Okabe [16] developed a KDE to estimate the density of points in a network in order to detect traffic accident hot spots. Three kernel functions were proposed, among which two were unbiased and successfully explained mathematical properties. These functions were implemented in a GIS and are available in the SANET extension. Nevertheless, the calculated density values are attributes of the edges and are not generalized to the entire area of the network.

The methodology of this paper, develop more further the one proposed by Porta [18]. It calculates a NetKDE for both point and linear features. However, the NetKDE operates by extending the bandwidth distance along the street network rather than linearly across the space. This methodology is close to the one used by Downs and Horner [9,10] in traffic accident and animal home range analysis.

2 Methodology

The methodology developed has to deal with very large datasets and has to produce results within reasonable laps of time (around a day for the prototype). Network calculations are time consuming. One simple way to increase the efficiency of scripts is to store data in a specialized relational database management system (RDMS). A RDMS has optimized SQL tools and functions to improve searching and editing of spatial elements and attributes. For the current project, the standard *ESRI* shape files (.shp) were converted to a *PostGIS* format (*PostgreSQL*) using the translating function *shape2pgsql*. The recovery of spatial objects is made within *PostGIS*, external *Python* scripts are used for different calculations, and interactions between data and scripts relied on the *Psycopg API*. (Figure 1).

The implemented algorithm uses three main input files. The first one is the road network with centrality values calculated by a MCA. This network has a clean topology and each node has a unique identifier (ID). The second one is constituted of the activities in the studied area, and the last one is the grid of points used by the KDE and the NetKDE. The KDE of activities and the KDE of edges centrality are calculated with *ArcGIS*. Secondly, on the basis of a *PostGIS* database, different *Python* scripts project activities on network edges, compute shortest path tree (SPT), and compute NetKDE of activities and NetKDE of edge centrality. SPT is a set of connected road network edges that are accessible from a specified location within a maximum trip cost. In our case, this cost is the considered bandwidth. During the development of the methodology several bandwidth have been used (100m, 200m, 400m). These choices were done in respect with city's human scale (walking space), computing resources and the ratio between the raster grid resolution and the bandwidth.

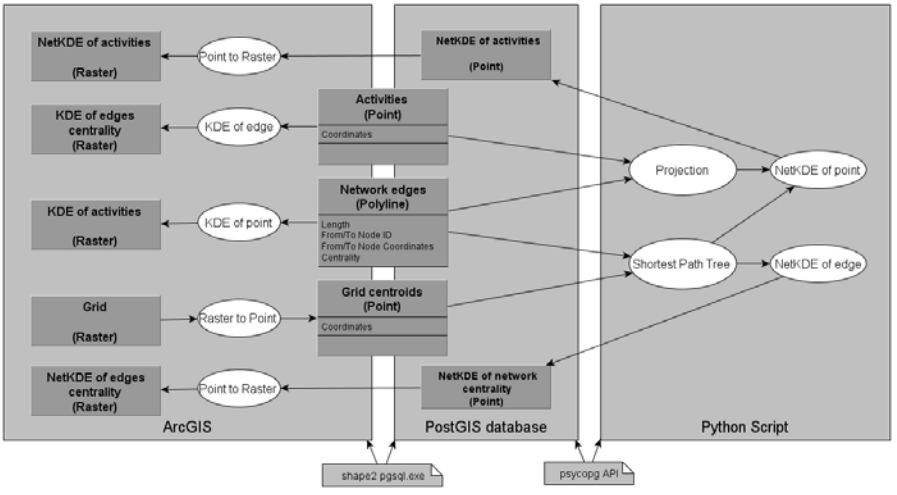

Fig. 1. Algorithm and interaction between the software

2.1 KDE Approach

A sliding window is used over the dataset to estimate the density of events. For the KDE, this sliding window is defined by a 3D function to weight the events according to their distance from the grid point for which the density is evaluated. With x_j being a location vector over the field R and $x_1 \ldots x_n$ the location vectors of the n events, the intensity estimation $f(x_j)$ in x_j is [20,13,14]:

$$\hat{f}_h(x_j) = \sum_{i=1}^{n} \frac{1}{h^2} K(\frac{x_j - x_i}{h}) \tag{2}$$

$d_{ij} = x_i - x_j$ is the Euclidean distance between the grid point x_j and the event n_i, h being the bandwidth. Actually, several kernel functions are implemented in different GIS. *ESRI ArcGIS Spatial Analyst* module allows only one kernel function for point and line density, known as quadratic or Epanechnikov [21]:

$$K(x_i) = \{ \begin{matrix} \frac{1}{3\pi}(1 - t_i^2)^2 \text{ if } t_i^2 < 1 \\ 0 \text{ otherwise} \end{matrix} \tag{3}$$

with $t_i = d_{ij}/h$.

The value at each point of the grid j at a distance d_{ij} of the event n_i is obtained from the sum of the individual kernel functions ($K(x_i)$) of the points belonging the bandwidth h. Any activity beyond the bandwidth h from the considered grid point does not contribute to the summation. In case of weighted events, the weight is used as a value of population. Namely, if a point has a value w, the algorithm takes this point into account as if there are w points for this distance d_{ij}. The same approach is applied to edges for which weight is a centrality value as does Porta [18]. Figure 2 shows the Euclidean approach of KDE.

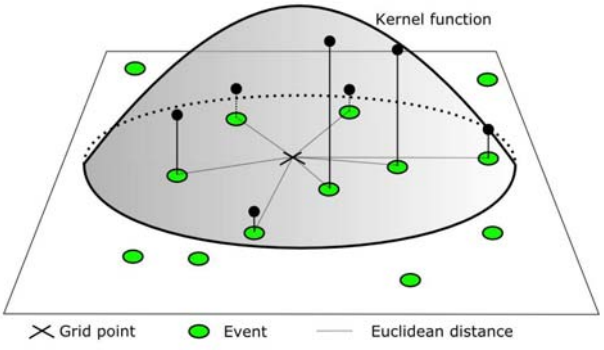

Fig. 2. Kernel function of the Kernel Density estimation

2.2 NetKDE Approach

The NetKDE needs the distance between the grid points and the activities projected onto the road network. The edges of the road network are stored in a database as polyline objects. The attributes of these edges are: From Node ID (FN), To Node ID (TN), FN and TN coordinates, polyline length and centrality values (Figure 1).

The second step of the methodology is to use the Dijkstra's SPT algorithm [8] to create trees of all accessible edges within a specified bandwidth from every grid points (Figure 3). First, the grid points are projected to their nearest edge.

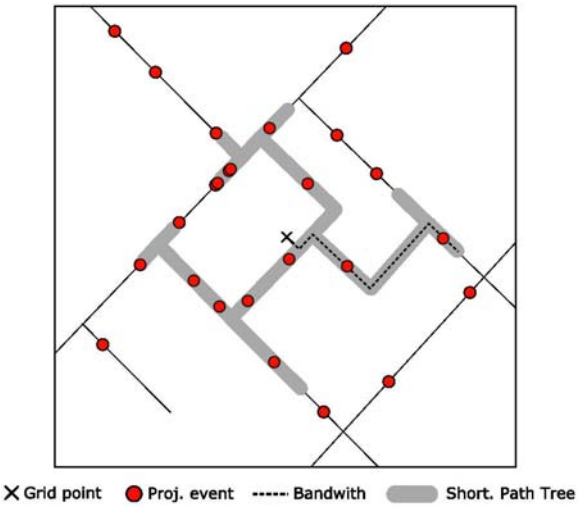

Fig. 3. Shortest Path Tree: The grid point and the activities are projected on the nearest edge

Then, the SPT are calculated from these projected points. All novel indexes presented in this paper use this SPT.

To calculate the NetKDE, the Euclidean distances are replaced by distances measured along the road network graph. For comparison purpose, the NetKDE uses the same KDE formula than *ArcGIS* (Equation 3). Thus, the NetKDE of activities is calculated with:

$$K_{net}(t_{net,i}) = \begin{cases} \frac{1}{3\pi}(1 - t_{net,i}^2)^2 \text{ if } t_{net,i}^2 < 1 \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

with $t_{net,i} = d_{net,ij}/h$ and $d_{net,ij}$ is the distance between the grid point x_j and the event n_i measured along the network. Then, the NetKDE value in grid point x_j is:

$$NetKDE(x_j) = \frac{1}{nh^2} \sum_{i=1}^{n} K_{net}(t_{net,i}) \quad (5)$$

n is the number of events on the SPT for the bandwidth h. NetKDE of edges centrality is calculated in the same way. The edge is reduced to his midpoint and the centrality $BetC_i$ of the edge i is used as a value of population:

$$NetKDE(x_j) = \frac{1}{h^2 \sum_{i=1}^{n} BetC_i} \sum_{i=1}^{n} K_{net}(t_{net,i}) BetC_i \quad (6)$$

This paper makes use of the Epanechnikov kernel function developped for event in an Euclidean space. Nevertheless, for the NetKDE, this function is applied to the events in the non-uniform space created from the SPTs. Our indicator has no unit but refer more to a linear density index than a spatial density index. Okabe [16] proposes kernel function developped for network based analysis. Figure 4 illustrates the NetKDE approach.

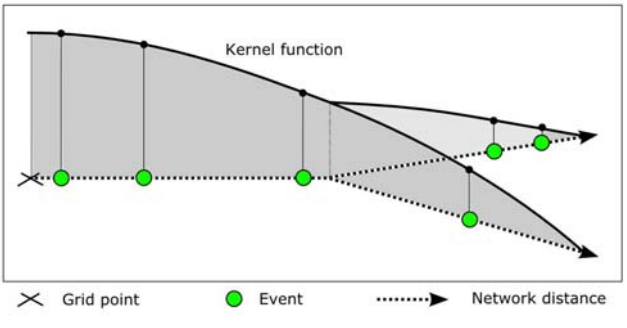

Fig. 4. Kernel function applied to the network approach, kernel function weights the projected events using the distance measured along the network

3 Barcelona Case Study

We applied in two-steps the methodology on the city of Barcelona. Firstly, the methodology algorithms were tested and validated with a sample (500m X 500m) of the Barcelona road network and economic activities and a 10m grid. **For this extract, calculation of the indexes takes less than one hour.** Some of the results from this first exercise are presented in this section. Secondly, once the methodology has been fine tuned and proofed with the data samples, the KDE and NetKDE approaches were applied to the whole city. The complete road network is characterized by 11,222 edges, there are more than 166,311 activities within an area of 92.65km^2, and the 10m resolution raster grid represents 1,890,000 grid points. Activities come from the the *Agencia de Ecologia Urbana* database describing all the economical, public or associational entities in 2002. For the entire data, the calculation takes approximately 2 days with Intel(R) Core(TM)2 Quad CPU, Q950 @ 3.00GHz, 2.99Ghz, 7.83GB of RAM.

3.1 Results Obtained with Data Samples

This section presents some of the results obtained with the city of Barcelona data samples. Firstly, the KDE of activities and NetKDE of activities are presented.

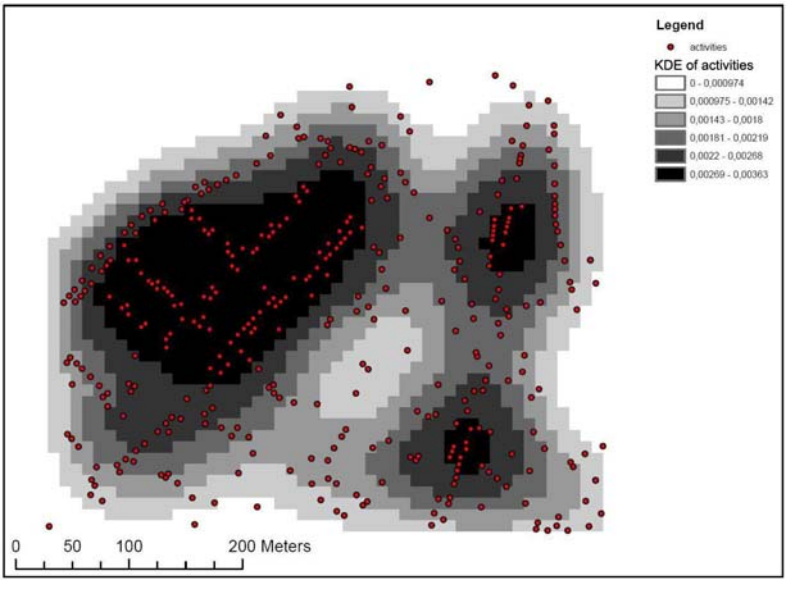

Fig. 5. Kernel Density of activities, bandwidth = 100m, raster cell = 10m, black clusters highlight high density of activities (red dots)

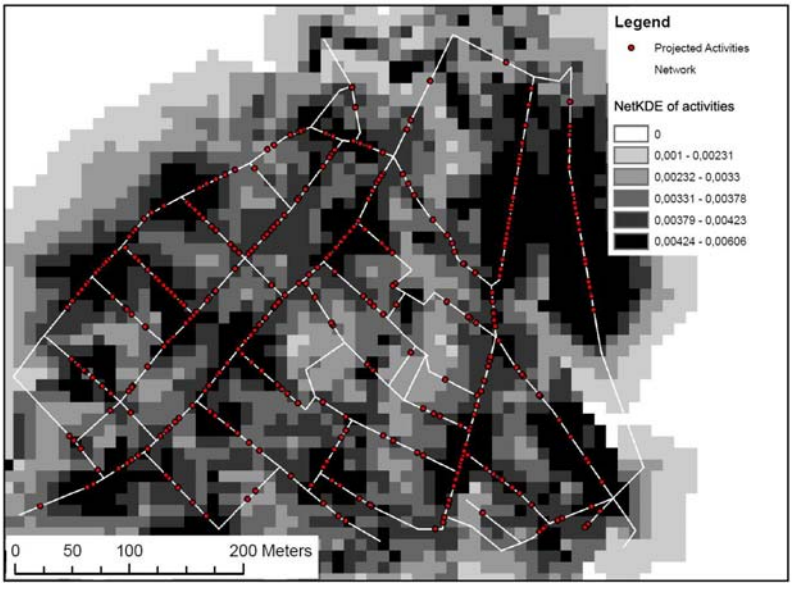

Fig. 6. Network based Kernel Density of activities, bandwidth = 100m, raster cell = 10m, activities are projected on the network, black clusters highlight high NetKDE density of activities

Secondly, the KDE of betweenness centrality and the NetKDE of betweenness centrality are presented.

The figure 5 presents the KDE of activities with a bandwidth of 100m and the figure 6 presents the NetKDE of activities with a bandwidth of 100m. The results show that the NetKDE approach has a less important smoothing effect than the KDE approach for the data samples. We discovered that NetKDE bandwidths have to be larger than those of KDE to produce visual clusters of comparable size; this is no surprise, as the same distance covers a smaller area when extended along the network than when extended linearly as the radius of a circle. Also, visual inspection of results from both approaches shows high density clusters of activities approximately at the same locations. The NetKDE approach is directly linked to the road network geometry and this explains some of the irregular patterns.

The figure 7 presents the KDE of betweenness centrality with a bandwidth of 100m and the figure 8 presents the NetKDE of betweenness centrality with a bandwidth of 100m. For both figures, values of betweeness centrality are displayed from low centrality in red to high centrality in blue. The raster grid (10m resolution) presents the values of KDE and NetKDE related to road network segment centrality. The density clusters is useful to identify clusters characterized by high or low centrality.

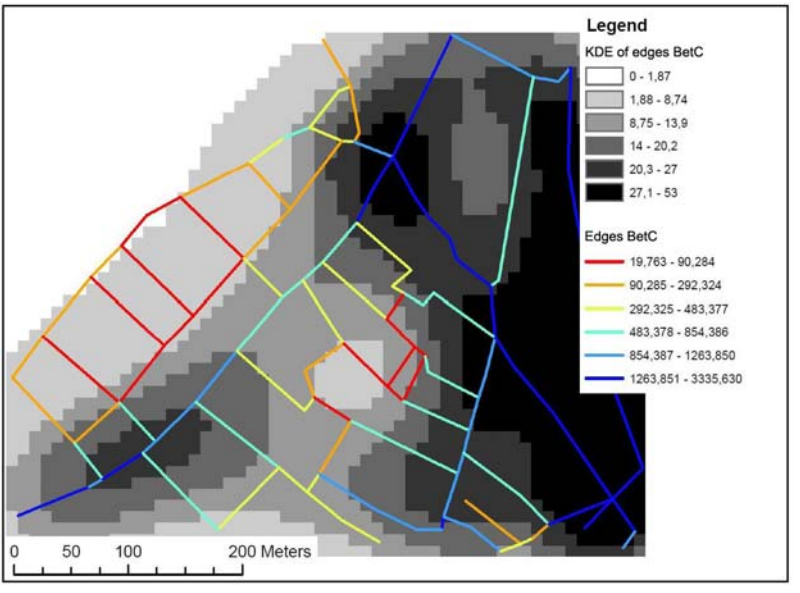

Fig. 7. Kernel Density of Betweenness centrality, bandwidth = 100m. BetC values are calculated from the entire road network.

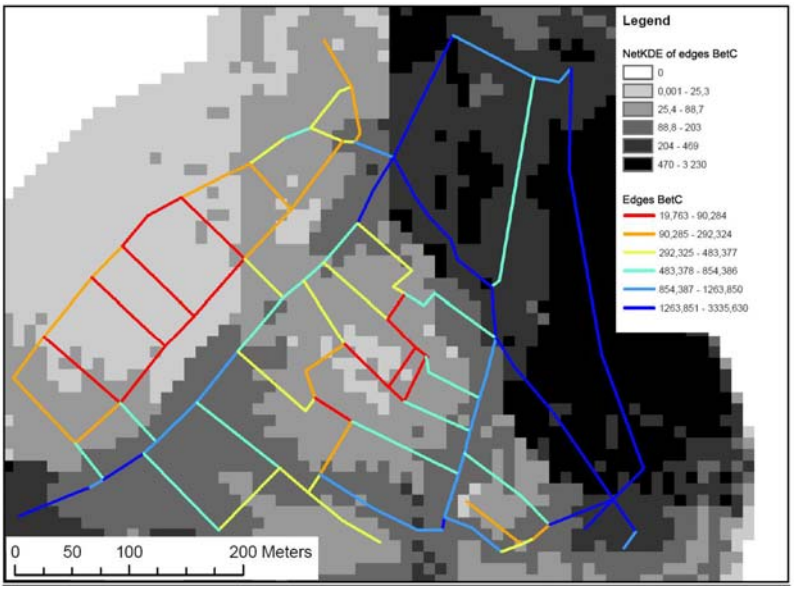

Fig. 8. Network based Density of Betweenness centrality, bandwidth = 100m. BetC values are calculated from the entire road network.

3.2 Preliminary Results for the Whole City of Barcelona

Figure 9 shows the Barcelona road network segments ranked by centrality values (Betweenness centrality), higher centrality values appearing in blue. Among these high centrality segments, it is possible to identify the "Avingunda Diagonal". The particular implantation of this street, planned by Cerda's master plan, contrasts with the rest of the surrounding street patterns, and constitutes a very convenient route in this part of the city. This explains probably its high centrality value. A NetKDE with a 400m bandwidth has been applied to this road network to get the figure 10 shown on the left. On the right, a KDE is applied with a 100m bandwidth. The NetKDE allows characterization of each 10m raster grid point in regard with the surrounding edges centrality values. Clusters resulting from the NetKDE conform to the distribution of the centrality values. Thus, segments highlighted by the global betweenness correspond to regions with a high density indicator. The betweenness centrality highlights specific street areas of the Barcelona road network. Most of these areas correspond to streets with a particular economic role in the city.

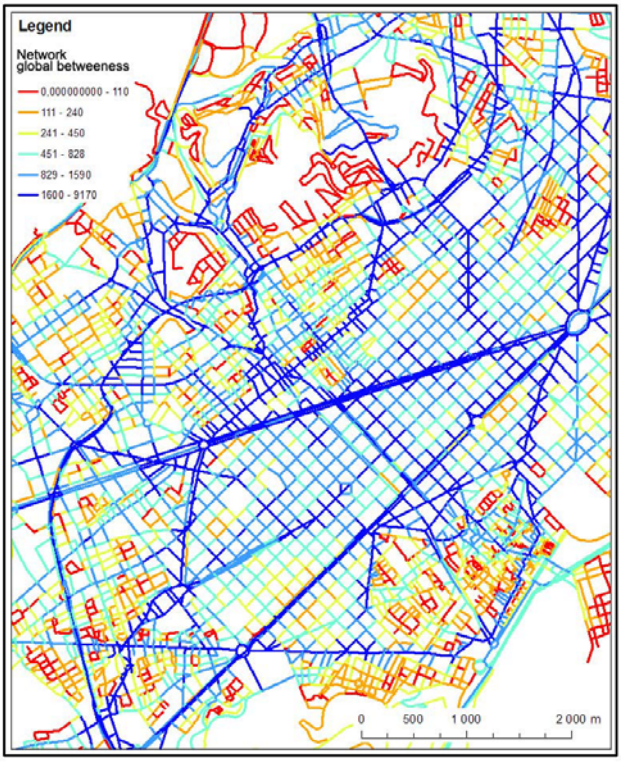

Fig. 9. High value in blue, low value in red: Network of Barcelona ranked by BetC

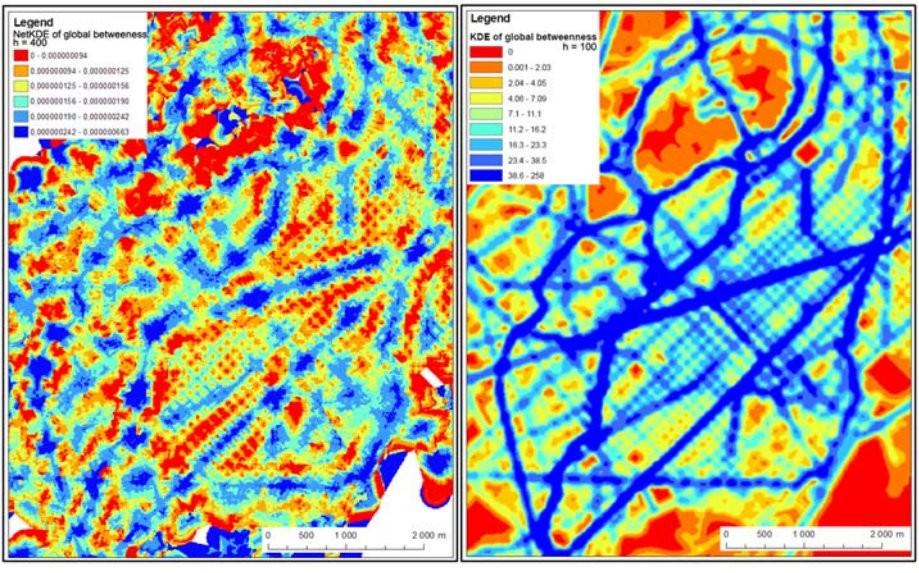

Fig. 10. Density of BetC: High value in blue, low value in red: **Left**: NetKDE of BetC, bandwidth = 400m, raster cell = 10m. **Right**: KDE of BetC, bandwidth = 100m, raster cell = 10m.

4 Discussion

The goal of this research was to improve the analysis of the spatial density of events in urban environment, taking into account network constrains. Firstly, this paper analyzed the distribution of retails and services projected on the road network. Secondly, it analyzed the distribution of edge centralities.

The NetKDE approach works on point events, like economic activities, and linear features, like streets. Instead of calculating linear density of collapsed point events on the network, NetKDE calculates spatial density. The NetKDE approach uses bandwidth constrained by the network rather than Euclidean bandwidth, like previous approaches. This reflects more adequately the reality of urban mobility.

First results clearly highlight the high sensitivity of the methodology to capture high resolution density of retail activities and new possibilities to compare it with the network segment density. Nevertheless, this methodology could already be used in urban analysis for any other type of spatial events. Finally, we could quote Tobler's [23] first law of geography, "everything is related to everything else, but near things are more related than distant things" and add "along the street."

References

1. Anselin, L., Cohen, J., Cook, D., Gorr, W., Tita, G.: Spatial analyses of crime. Criminal Justice 4, 213–262 (2000)
2. Batty, M.: Network Geography: Relations, Interactions, Scaling and Spatial Processes in GIS. Re-presenting GIS, 149–170 (2005)
3. Borruso, G.: Network Density Estimation: A GIS Approach for Analysing Point Patterns in a Network Space. Transactions in GIS 12, 377–402 (2008)
4. Borruso, G.: Network Density and the Delimitation of Urban Areas. Transactions in GIS 7, 177–191 (2003)
5. Brunsdon, C.: Estimating probability surfaces for geographical point data: An adaptive kernel algorithm. Computers & Geosciences 21(7), 877–894 (1995)
6. Costa, L.F., Oliveira Jr, O.N., Travieso, G., Rodrigues, F.A., Boas, P.R.V., Antiqueira, L., Viana, M.P., da Rocha, L.E.C.: Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. arXiv, 0711.3199v3 (2007)
7. Crucitti, P., Latora, V., Porta, S.: Centrality measures in spatial networks of urban streets. Physical Review E 73(3, Part 2) (2006)
8. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271 (1959)
9. Downs, J.A., Horner, M.W.: Network-based Kernel Density Estimation for Home Range Analysis. In: Proceedings of the 9th International Conference on GeoComputation, Maynooth, Ireland (2007)
10. Downs, J.A., Horner, M.W.: Characterising Linear Point Patterns. In: Proceedings of the GIS Research UK Annual Conference (GISRUK 2007), Maynooth, Ireland (2007)
11. Euler, L.: Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientiarum Petropolitanae 8, 128–140 (1741)
12. Freeman, L.C.: Centrality in Social Networks: Conceptual Clarification. Social Networks 1, 215–239 (1979)
13. Gatrell, A., Bailey, T., Diggle, P., Rowlingson, B.: Spatial point pattern analysis and its applications in geographical epidemiology. Transactions of the Institute of British geographers 21, 256–274 (1996)
14. Gatrell, A.: Density estimation and the visualization of Point pattern. In: Hearnshaw, H.M., Unwin, D.J. (eds.) Visualization in Geographical Information Systems, pp. 65–75. John Wiley and Sons, New York (1994)
15. Miller, H.J.: Measuring space-time accessibility benefits within transportation networks: Basic theory and computational procedures. Geographical Analysis 31, 187–212 (1999)
16. Okabe, A., Satoh, T., Sugihara, K.: A kernel density estimation method for networks, its computational method and a GIS-based tool. International Journal of Geographical Information Science 23, 7–32 (2009)
17. Porta, S., Crucitti, P., Latora, V.: The network analysis of urban streets: a primal approach. Environment and Planning B: Planning and Design 33, 705–725 (2006)
18. Porta, S., Latora, V., Wang, F., Strano, E., Cardillo, A., Scellato, S., Iacoviello, V., Messora, R.: Street centrality and densities of retail and services in Bologna, Italy. Environment and Planning B: Planning and Design 36, 450–465 (2009)
19. Row, J.R., Blouin-Derners, G.: Kernels Are Not Accurate Estimators of Home-Range Size for Herpetofauna. Copeia 4, 797–802 (2006)

20. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC (1986)
21. Smith, M.J., Goodchild, M.F., Longley, P.: Geospatial Analysis: a comprehension to principles, techniques and software tools. Troubador Publishing (2006)
22. Thurstain-Goodwin, M., Unwin, D.: Defining and Delineating the Central Areas of Towns for Statistical Monitoring Using Continuous Surface Representations. Transactions in GIS 4, 305–317 (2000)
23. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Economic geography 46, 234–240 (1970)
24. Topping, D.T., Lowe, C.G., Caselle, J.E.: Home range and habitat utilization of adult California sheephead, Semicossyphus pulcher (Labriddae), in a temperate no-take marine reserve. Marine Biology 147(2), 301–311
25. Wang, F.: Quantitative methods and applications in GIS. CRC Press, Boca Raton (2006)
26. Xie, Z., Yan, J.: Kernel Density Estimation of traffic accident in network space. Geography, Geology Faculty Publications (2008)
27. Yamada, I., Thill, J.C.: Comparison of planar and network K-functions in traffic accident analysis. Journal of Transport Geography 12, 149–158 (2004)