The Cartographic Journal *Vol. 48 No. 4 pp. 309–316 Cartographies of Fictional Worlds - Special Issue November 2011*

REFEREED PAPER

Improved Density Estimation for the Visualisation of Literary Spaces

Hans Rudolf Bär and Lorenz Hurni

Institute of Cartography, ETH Zurich, CH-8093 Zurich, Switzerland
Email: baer@karto.baug.ethz.ch

The elements that constitute the literary space can mainly be described by categories such as settings, zones where actions take place and routes along which characters move. Apart from the presentation of the specific locations and spatial distribution of literary places, we are also interested in the spatial pattern such places form, the boundaries that separate the literarily populated from the void regions and the varying density of the literary space. From a GIS point of view, the elements of the literary space correspond to point, line and area data types. However, the established method used to calculate the spatially varying density – the method of density estimation – is restricted to point data only. In this paper, we will present an improved method that is able to estimate the density regardless of the underlying data type. Our approach aims at adopting the typically radially symmetric kernel function to approximate also linear and areal features. We claim that this method treats point, line and area data in a consistent way by taking equivalent density contributions into account. The different steps of the improved method can visually be examined by the accompanying map examples.

Keywords: density estimation, kernel function, literary geography, cartographic visualisation.

INTRODUCTION AND MOTIVATION

The 'Literary Atlas of Europe' is an interdisciplinary research project that aims at providing literary thematic maps and analytical tools for literary experts (Literary Atlas of Europe, 2007; Piatti *et al.*, 2009). Currently, work on three distinct model regions is in progress: the urban area of Prague, the countryside of Northern Frisia and the alpine scenery between Lake Lucerne and Mount Gotthard. Information about literary places is extracted from the relevant texts by literary experts and entered into a structured database. The project basically follows two methodological approaches, the first focussing on the representation of individual literary spatial entities, while the latter on the visualisation of statistical quantitative analyses. The work as presented in this contribution covers one part of the statistical approach of this Literary Atlas.

Mapping literature will confront literary editors as well as cartographers with a number of challenges (Reuschel and Hurni, 2011): settings which are hard to localize, zones of action often without precise borders and mostly fragmentary routes of literary characters. Also, on the part of cartography, there is rarely a consensus about how to visualize such vague and imprecise places (McEachren *et al.*, 2005).

Figure 1 shows a visualisation of a section from the literary database limited to settings acting as 'thematic scenery'. The dominating red colour shades are used to indicate that actions take place at these settings (as opposed for instance to places of mere imagination). Small squares are used to depict single buildings, fuzzy dots and lines refer to places and streets, and fuzzy semi-transparent shapes denote zonal settings using fuzziness to indicate a degree of vagueness of such places. Fields of radially emanating rays, also called 'moiré patters' (Reuschel and Hurni, 2011), provide an impression of locations that can only vaguely be assigned to a region.

From a cartographic point of view, several problems are becoming evident, though the map only shows a section from the database. The main points will be:

- clusters of point symbols partly obscuring each other;
- superimposition of linear elements;
- density by blending of semi-transparent areas difficult to estimate;
- partly invisible base map covered by symbols.

Texts associated with literary places are omitted in the example above but would cause additional rivalry around the spare map space. In the following, density maps are considered as an alternative to such overloaded maps depicting individual objects.

The portrayal of individual literary places on a map is appropriate as long as the focus is on a manageable section of the database such as a single text or the texts of a particular author. Being also interested in visually examining large datasets, statistical methods will be more appropriate. Instead of asking for the location of specific literary

DOI: 10.1179/1743277411Y.0000000022

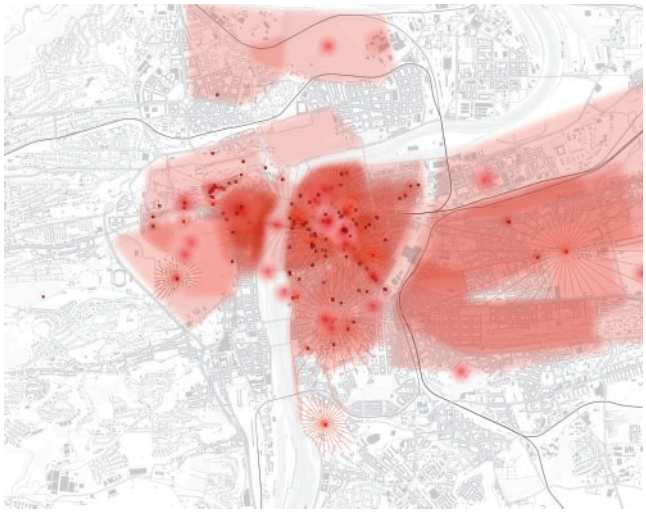

Figure 1. Literary places in Prague, single-object representation. Map visualisation courtesy of A. K. Reuschel, Literary Atlas of Europe

places, we will then ask for the number of places within a region, the densities these locations form and the spatial pattern such densities describe.

The method typically used for the calculation of spatially distributed densities from a set of data points is *density estimation* (Silverman, 1986; Gatrell, 1994). Density estimation can be considered as a generalized method of counting objects within a unit of area. Central part is the *kernel function*, a typically radially decreasing function, which defines the fraction a data point contributes to the density with respect to the distance of observation. The region within the influence of the kernel is called the *bandwidth* or the *window width* and essentially controls the smoothing of the kernel function. The surface the kernel function describes is expected to cover equal volumes.

Although the method of density estimation does not compensate for vague locations and fragmentary routes, it solves or at least circumvents the problem with imprecise borders: the process of density estimation removes sharp breaks such as distinct boundaries and produces a smooth surface. Density constitutes an important analytical measure since it allows the comparison of different spatial regions and puts the human impression of density on an objective, predictable and reproducible basis. Density maps provide an excellent overview of a spatially distributed phenomenon but, as with every statistical procedure, come at the cost of missing relations to individual locations.

Density estimation is closely related to spatial interpolation, especially to the method of inverse distance weighting (Shepard, 1968). Both method rely on the assumption of a larger influence of nearby data points, but unlike with spatial interpolation, density estimation starts from point locations which are not associated with a data value. In density estimation, we are interested in the *sum* of the contribution of each data point; in spatial interpolation, we consider the *mean* of the contribution of all data values.

The method of kernel density estimation has found its way into various disciplines in order to answer questions on a statistical level of abstraction. Most papers about density estimation refer to the work of Silverman (1986) who to our knowledge first gave a comprehensive overview of non-parametric density estimation (i.e. estimation not based on assumptions about the theoretical distributions and their parameters). First geographic applications of density estimation, i.e. the investigation of *spatial* point patterns, have been described in the disciplines of ecology (Worton, 1989), economy (Donthu and Rust, 1989) and epidemiology (Gatrell *et al.*, 1996).

Few works have considered extensions to the original method. Downs and Horner (2007) and Borruso (2008) have addressed point pattern analysis related to a network. Borruso argues that the assumption of a homogenous and isotropic space is inappropriate for network-based activities and proposes to modify Euclidean distances by the shortest path calculations. Although Borruso explains his approach in terms of the kernel density estimation method, we would rather consider it as an improved moving window method, since it abandons the initial idea of a monotonically decreasing and volume-preserving kernel function. Although based on a network, these approaches describe point patterns. To our knowledge, density estimation for patterns of higher order (linear or areal) has not yet been a subject of research.

IMPROVED DENSITY ESTIMATION

In order to meet the visualisation needs of the Literary Atlas, we consider the following extensions and requirements as essential:

- extension of the density estimation method to include point, line and area data;
- good approximation of shapes to achieve a mostly accurate density estimation;
- every object regardless of its type should contribute the same amount to the overall density;
- efficient calculation that enables interactive applications.

We consider to meet these requirements with the following approach:

- extension of circular kernels to fit polygonal shapes;
- kernel function that continuously decreases from the centroid to the shape's border;
- force the kernel function to keep the covered volume constant for all objects it describes;
- decomposition into triangle sets to enable hardware-accelerated calculations.

The following sub-chapter first presents the application of the existing method by reducing the spatial entities to single point locations. We will then proceed with improved methods for elliptical and polygonal kernel shape approximations.

Density estimation for point representations

The spatial application of the density estimation method assumes that the input data consist of a set of spatially distributed point locations. In contrast, literary places, as collected in the database of the Literary Atlas, may not only include point locations, but also show linear or areal

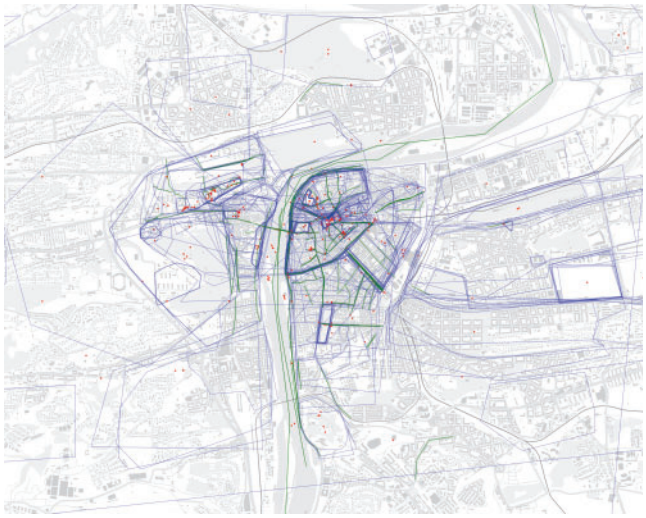

Figure 2. Raw data from the literary database, section of the centre of Prague. Points are shown in red, lines in green, and areas in blue

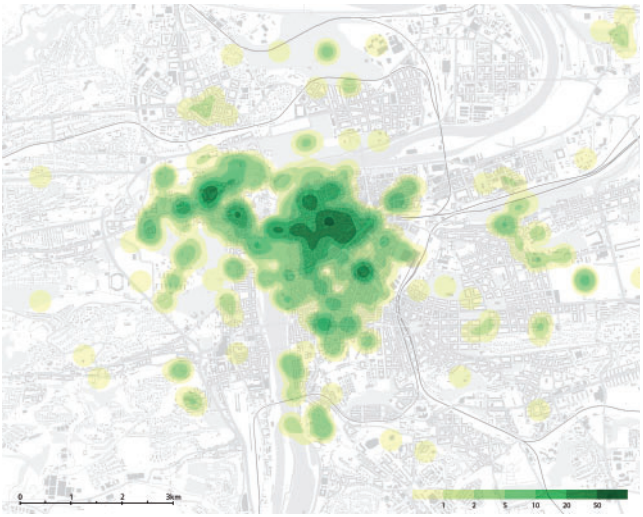

Figure 4. Density estimation for centroids: bandwidth set to 50 pixels (315 m)

dimensions. Figure 2 provides an overview of the currently stored literary entities within the city of Prague. Point locations are shown as red dots, paths as green lines, and areas as blue contours.

In order to apply the method of kernel density estimation to literary places, we either have to restrict the dataset to point locations, collapse linear and areal places down to points or seek to extend the existing methods to process line and area data in a coherent way.

Reducing areal and linear elements to point locations is the method that most straightly opens the door to density estimation. An obvious choice is the centre of gravity or centroid, although, for concave shapes, the centroid of an area is not guaranteed to be contained within the given area, and the centroid of a path is generally not a point on the line. Figure 3 shows the centroids of the Prague dataset with colours denoting the type of geometry they represent (red for points, green for lines and blue for areas).

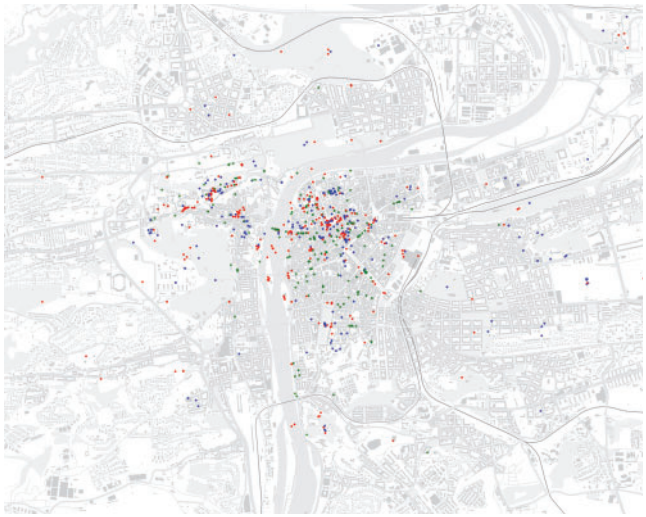

Figure 3. Data reduced to centroids. Original point locations in red, centroids of paths in green and centroids of areas in blue

Density estimation first requires the selection of an appropriate bandwidth. Many suggestions are found in the literature for optimally choosing this parameter (Turlach, 1983), but some authors also state that the selection of the bandwidth is a matter of trial and error (Tukey, 1977). We will currently follow the second approach for a first visual inspection, keeping in mind that an atlas user should not be forced to select an adequate bandwidth. In Figure 4, the bandwidth has iteratively been determined such that the single locations start to build contiguous areas of equal ranges of densities. For better visualisation, the density values have been divided into seven classes, symbolized by colours ranging from fair yellow to dark green.

In the following, we will call this method the *centroid method for circular kernels*. The legend shown in Figure 4 refers to the number of places situated within the area covered by the kernel. The lower break value of each class corresponds to the minimal number of places encountered within the bandwidth of the kernel. Unfortunately, there is no way to specify a maximum number due to the fractions the kernel functions contribute to the density.

Reducing the bandwidth of the kernel dissolves the contiguous areas into a pattern of separated spots, thus making it more difficult to provide an overall impression of the spatial variability of the density (Figure 5).

Otherwise, if the bandwidth is further increased, local details tend to disappear in favour of the 'large forms' (Figure 6).

Considering users of the Literary Atlas who usually do not have a background in statistics, it would be desirable to start with a reasonable suggestion for the bandwidth, provided that for the comparability of maps such as time series, a user should be guaranteed to work with identical parameters. At this stage of research, however, we do not consider an automatic bandwidth selection as of foremost priority.

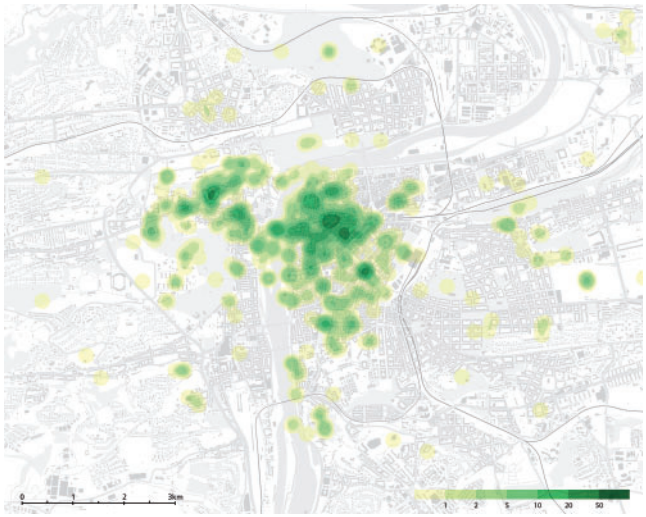

Figure 5. Density estimation for centroids: bandwidth set to 35 pixels (220 m)

In the following, we will be looking for methods that take the specific shapes of linear and areal places into account.

Density estimation for approximated elliptical shapes

Representing linear elements by circular kernels is generally a coarse and inappropriate approximation. We will now consider the *centroid method for elliptical kernels*, a method that uses elliptical shapes to describe equivalent volumes in order to approximate the given shapes more precisely. For this purpose, we calculate the *minimum area rectangle*, which is similar to the bounding box, but occasionally rotated. This rectangle is then used to describe an ellipse with the axes parallel and proportional to the main axes of the rectangle.

In the following, we will first focus on linear elements only. Figure 7 shows the literary routes along which literary characters move.

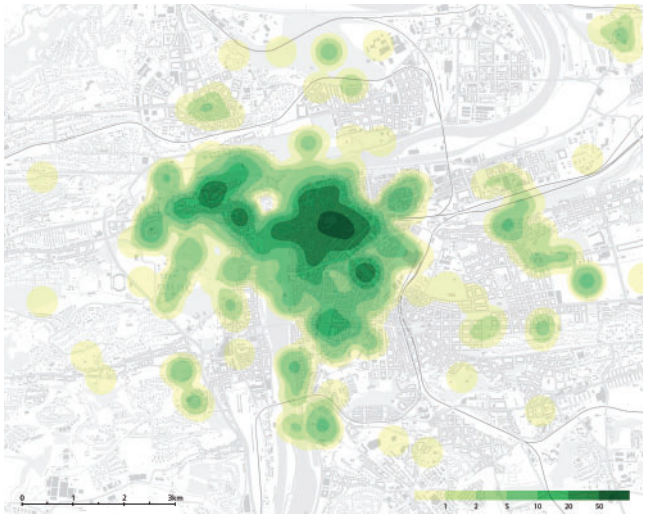

Figure 6. Density estimation for centroids: bandwidth set to 70 pixels (440 m)

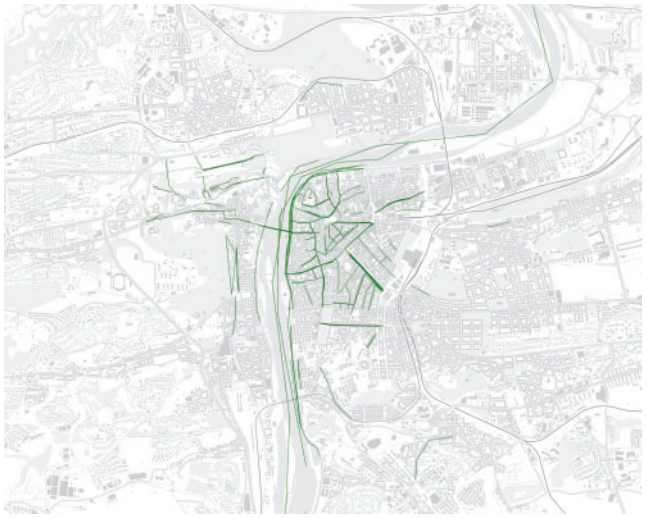

Figure 7. Raw paths as stored in the database

What might be anticipated is a problem with single straight line segments. In this case, the resulting minimum area rectangle will collapse to a line element and the constructed ellipse will end up with a minor axis of zero. As a remedy, we decided to restrict the ratio of the axes of the ellipse to a minimal value.

Figure 8 shows the effect of a mean bandwidth of 50 pixels (315 m) and a maximum ellipse axes ratio of 1 : 5. These values have been chosen to prevent ellipses from growing beyond the original extent but also to provide an impression of the density induced by the path elements.

Reducing the mean bandwidth to a clearly lower value and admitting also lower ellipse axes ratios will accentuate the linear character of the elements as shown in Figure 9.

Larger mean bandwidths and higher ellipse axes ratios increasingly suppress the observable linear structure which gradually approaches the circular kernel shape (Figure 10).

Figure 8. Density estimation for paths approximated by ellipses: a mean bandwidth of 50 pixels (315 m) and a minimal ellipse axes ratio of 1 : 5

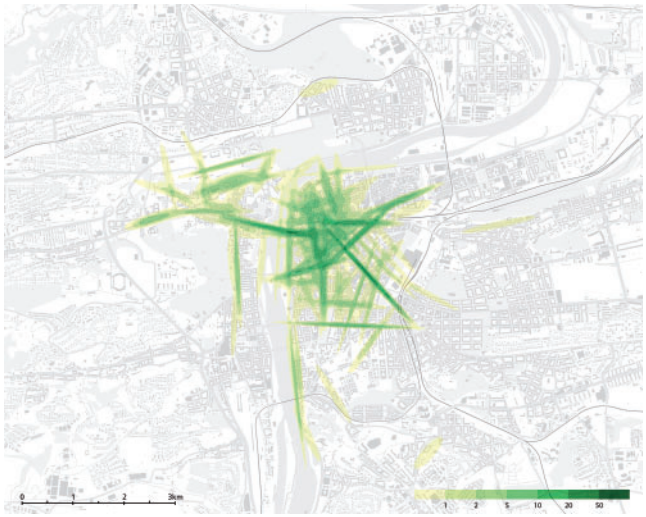

Figure 9. Density estimation for paths approximated by ellipses: a mean bandwidth of 35 pixels (220 m) and a minimal ellipse axes ratio of 1 : 10

The problem with degenerated minimum area rectangles also turns up with area elements. Similar measures can be taken to avoid excessive elliptic distortions. Applying the technique of the approximated elliptical kernel to the entire dataset will result in a map as shown in Figure 11. Note that point elements will still be described by circular kernels. Compared to Figure 4, the method using the approximated elliptical kernels tends towards the creation of more contiguous areas and less local maxima.

Unlike the common circular kernel, the use of an elliptical kernel requires additional tests to detect degenerated shapes and limit excessive axes ratios. The limitation of the ellipse ratio is a rather arbitrary choice and its value can only be estimated by visual inspection. An ellipse is doubtlessly a more flexible object to approximate an arbitrary shape more closely, but we currently do not see

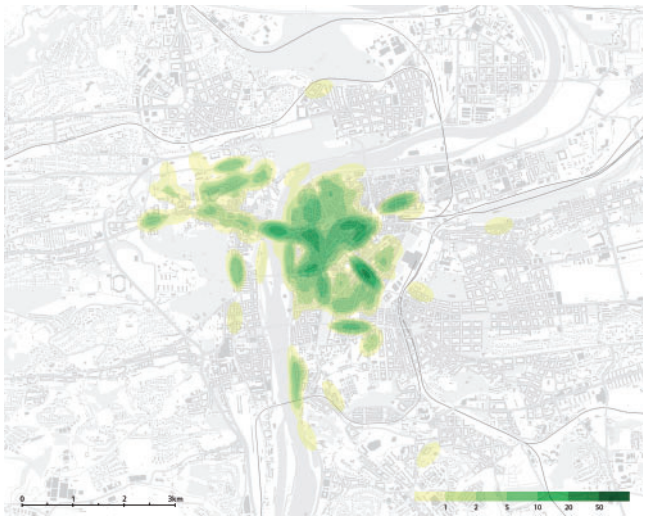

Figure 10. Density estimation for paths approximated by ellipses: a mean bandwidth of 70 pixels (440 m) and a minimal ellipse axes ratio of 1 : 3

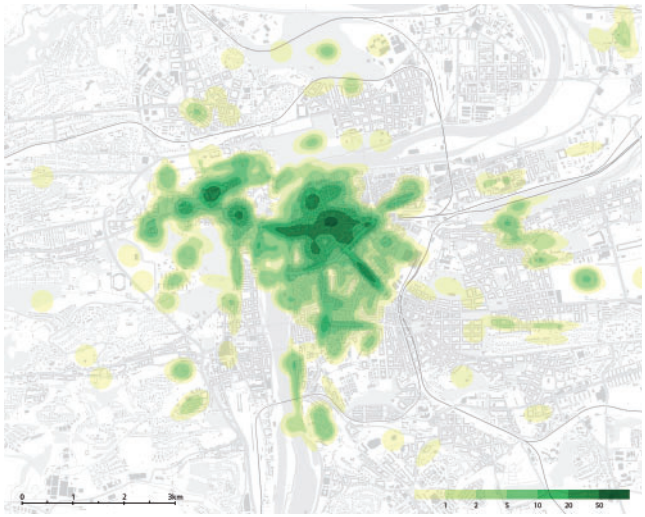

Figure 11. Density estimation for point, line and area elements approximated by circles and ellipses: a mean bandwidth set to 50 pixels (315 m) and a minimal ellipse axes ratio of 1 : 5

a theoretic base for its proper parametrisation. In the following, we will try to consider arbitrary shapes of the literary places more precisely.

Density estimation for approximated convex polygon shapes

In the following, we will present the *convex polygon kernel method*. The elliptical kernel will be replaced by a polygonal structure. The restriction to convex polygons is necessary to avoid problems with centroids lying outside the shapes and to avoid the creation of overlapping triangles, as will be shown later in this section. Since we do not allow any places to have more weight than others, the volume described by the kernel function needs to remain constant. This is in accordance with the idea of the Literary Atlas which does no ranking of literary places. As a consequence, to keep the volume constant, large kernel areas will produce lower ranges of values than small ones.

Using convex polygons as kernels, a smooth transition between point, short line and small area objects needs to be guaranteed. Point objects will still use circular kernels of a given size. Since a circular kernel actually corresponds to a buffer around a point with the distance equal to the bandwidth, we will apply buffers to convex polygons as well. This procedure will also solve the problem with degenerated shapes.

The evaluation of the kernel function is then straightforward. Starting from the centroid of the buffered shape, a triangle fan can be constructed and a kernel function assigned to each triangle. Figure 12 shows how an arbitrary polygon is approximated by a triangle fan. From the original polygon (concave polygon with red contour line), the convex hull is first created (dark grey area with green contour line). A buffer (light grey area) is then applied to the convex hull. The triangle fan is then spanned between the centroid and each point of the buffer (delimited by the blue contour line). Note that the construction of the buffer will add linearly approximated circular arcs (blue contour line).

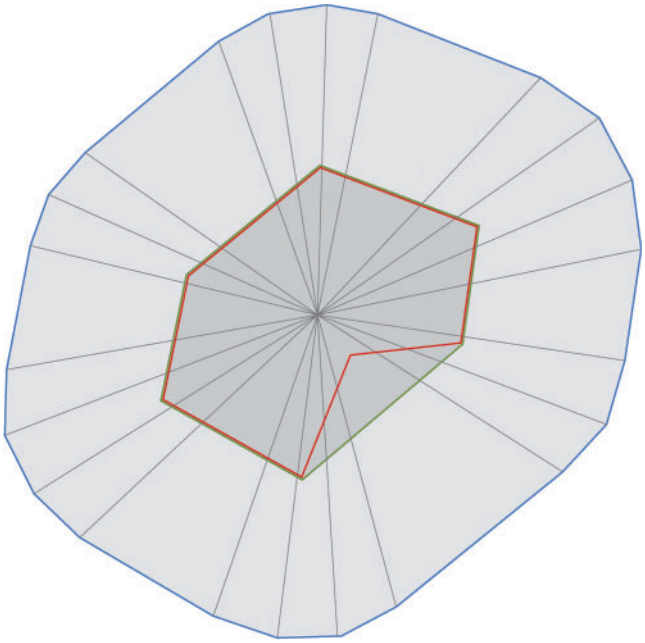

Figure 12. Principle of the transformation of an arbitrary polygon to a triangle fan

Figure 13 shows the result of the application of the convex polygon kernel. As a side effect, adding a buffer will result in a rounding of corners. Given the fact that the person in charge will occasionally only gather rough polygonal approximation of the desired area, a smoothing of the original shape might even be desired.

In order to remain comparable, a threshold for small values has been applied. This threshold is used to remove low densities within the buffer area. Users rather interested in the overall area covered by the literary places might be interested in a lower threshold.

A low threshold value may reveal large areas of very low densities (Figure 14). Being able to adapt this value also gives a user a chance to filter out very large regions. This is also to compensate for some shortcomings of the literary database, since for an editor there is currently no restriction considering the geometric type (e.g. point, line and area) he or she uses and there is no claim about the map scale the digitized geometry should base on.

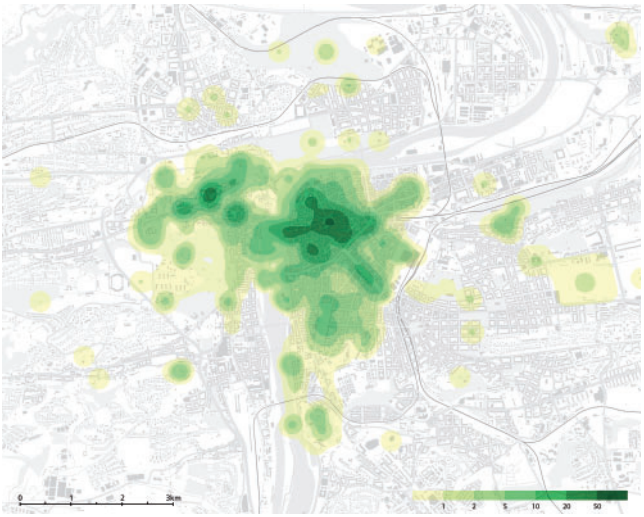

Figure 13. Density estimation for convex area shapes: threshold for small values set to 0.25

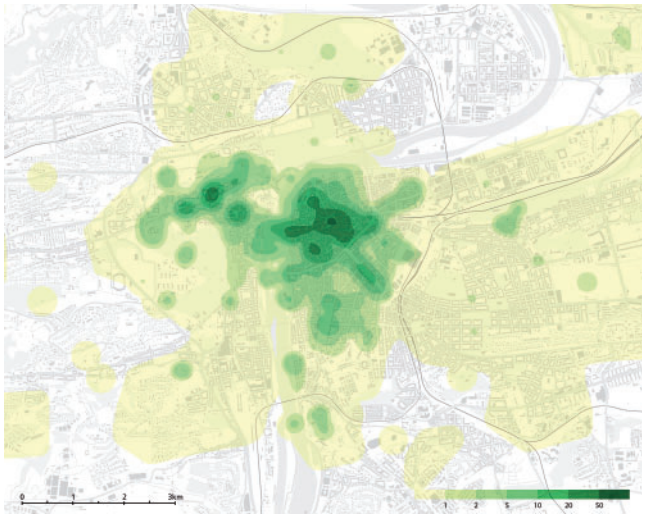

Figure 14. Density estimation for convex area shapes: threshold for small values set to 0.01

COMPARISON OF THE METHODS

Figure 15 provides an overview of the three methods presented above. For better comparison and also to better discover artefacts caused by these methods, an oblique surface shading has been added.

Although the methods greatly differ, there is no map among the three versions that would give reason for a considerably different interpretation. On all maps, there will be a consensus on where the main centre of the literary places is located, and also the density of this place appears in the same category. There might also be an agreement on the order of the local centres though less clearly. Considering Figures 4–6, it becomes obvious that a differing bandwidth will contribute more to the creation of local maximums than the chosen method.

Not surprisingly, the use of circular kernels also tends to produce circular patterns of densities. The use of elliptical shapes reveals that the approximation by ellipses occasionally misses the original shapes and creates visible elliptical artefacts that do not well represent linear shapes. Since a specific limitation of the axes ratio cannot be based on a theoretical reason, its use remains restricted. The authors tend to prefer the last method that is based on convex polygon kernels. It not only provides the most balanced appearance from the underlying structures, but also bases on sound principles.

The maps above have shown that reasonable methods exist for an appropriate visualisation of the density of the literary space based on the contents of the literary database. Point, line and area data can be intermixed for density calculation, but a user should always be aware of the different type of objects put together.

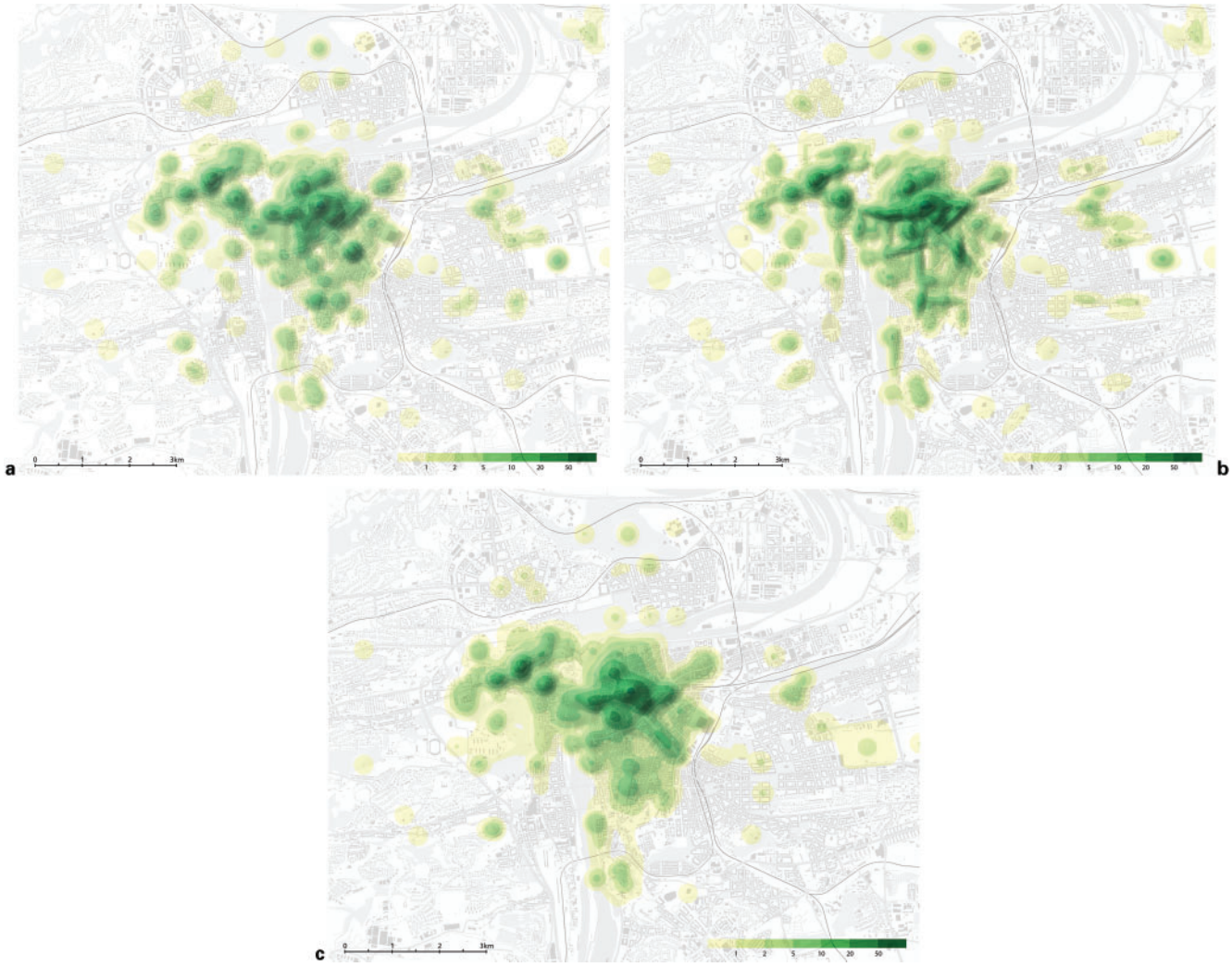

Figure 15. Comparison of the proposed methods: (a) centroid method for circular kernels, (b) centroid method for elliptic kernels, (c) the convex polygon kernel method

CONCLUSIONS AND OUTLOOK

We have presented three methods for density estimation that are applicable to point, area and line data likewise. Although the methods have been exemplified in the context of literary geography, they should also be transferable to other topics, keeping the current limitations in mind.

The methods allow for the calculation of an overall density regardless of the data type. The use of a buffer guarantees that there is a smooth transition of the kernel function from single points to lines and areas. Points, lines and areas always contribute the same amount to the overall density since the volume described by the kernel function remains constant. This is well in accordance with the idea of the Literary Atlas, since there is no preference or ranking of literary places.

However, an integration of different data types introduces side effects a user should be aware of. The restriction to convex shapes may possibly enlarge a given area. This is especially true for long and winding character paths. Also, the calculated position of the centroid might differ from the intention of the literary expert. Furthermore, the use of a buffer will introduce rounded corners which will result in a smoothed appearance of the given area. This effect, however, might also be welcome since the digitized areas are often rough approximations of a shape by straight-line segments.

A user will be able to control the density map mainly with a single parameter, the bandwidth. An additional parameter can be used to filter off very small density values which is crucial for example to delimit the populated literary space from the literary void space.

So far, map projections have not been considered. Density estimation does not take local distortions into account. Although the problem is rather relevant for small-scale maps, the visualisation of the density might deviate from the underlying base map.

The paper does not discuss any implementation details of the presented methods for density estimation. The use of hardware-accelerated rendering and programmable graphics processor will possibly follow in another publication.

In order to be used as an analysis tool by literary experts, a graphical user interface still needs to be designed that simplifies the access of the database and supports the user for an optimal visualisation of the contents of the database. User interaction needs to be used to provide the basic

information hidden behind the statistical display, and sure enough, perspective views of the statistical surfaces will attract the attention of potential users of the Literary Atlas.

BIOGRAPHICAL NOTES

Hans Rudolf Bär received his PhD in Geography from the University of Zurich. In 1995, he joined the Institute of Cartography at ETH Zurich, where he first started work with interactive atlases. Since then, he was in charge with the conception and programming of a number of interactive atlases such as the national *Atlas of Switzerland*, the statistical atlas of the European Union (Statlas) and a web-based school atlas, the *Swiss World Atlas*.

ACKNOWLEDGEMENT

The initial literary analysis and data entry (of 73 texts) has been carried out by Marie Frolíkova and Eva Markvartová, Charles University, Prague.

REFERENCES

Borusso, G. (2008). 'Network density estimation: a GIS approach for analysing point patterns in a network space', **Transaction in GIS**, 12, pp.377–402.

Donthu, N. and Rust, R. T. (1989). 'Estimating geographic densities using kernel density estimation', **Marketing Science**, 8, pp. 191–203.

Downs, J. A. and Horner, M. W. (2007). 'Characterising linear point patterns', in **Proceedings of the Geographical Information Science Research UK Conference**, ed. by Winstanley, A. C., National University of Ireland, Maynooth, pp. 421–424.

Gatrell, A. C. (1994). 'Density estimation and the visualisation of point patterns', in **Visualization in Geographical Information Systems**, ed. by Hearnshaw, H. J. and Unwin, D. J., pp. 65–75, John Wiley & Sons, New York.

Gatrell, A. C, Bailey, T. C., Diggle, P. J. and Rowlingson, B. S. (1996). Spatial Point Pattern Analysis and its Application in Geographical Epidemiology, http://www.dpi.inpe.br/cursos/ser301/referencias/gattrel_paper.pdf (accessed 21 March 2011).

Literary Atlas of Europe. (2007). Towards a Geography of Fiction, http://www.literaturatlas.eu/index_en.html (accessed 21 March 2011).

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M. and Hetzler, E. (2005). 'Visualizing geospatial information uncertainty: what we know and what we need to know', **Cartography and Geographic Information Science**, 32, pp. 139–160.

Piatti, B., Bär, H. R., Reuschel, A. K., Hurni, L. and Cartwright, W. (2009). 'Mapping literature: towards a geography of fiction', in **Art and Cartography**, ed. by Cartwright, W., Gartner, G. and Lehn, A., pp. 177–192, Springer, Berlin/Heidelberg.

Reuschel, A. K. and Hurni, L. (2011). 'Mapping literature: visualisation of vague literary places', **The Cartographic Journal**, **48, 4**, pp. 293–308.

Shepard, D. (1968). 'A two-dimensional interpolation function for irregularly-spaced data', in **Proceedings of the 1968 23rd ACM National Conference**, pp. 517–524, ACM, New York.

Silverman, B. W. (1986). **Density Estimation for Statistics and Data Analysis**, London, Chapman & Hall.

Tukey, J. W. (1977). **Exploratory Data Analysis**, Addison-Wesley, Reading, MA.

Turlach, B. A. (1983). Bandwidth Selection in Kernel Density Estimation: A Review, http://www.stats.adelaide.edu.au/people/bturlach/psfiles/dp9317.ps.gz (accessed 21 March 2011).

Worton, B. J. (1989). 'Kernel methods for estimating the utilization distribution in home-range studies', **Ecology**, 70, pp. 164–168.