Kernel density estimation: An innovative exploratory tool for geomarketing

Michael LEITNER and Petra STAUFER-STEINNOCHER

Dieser Beitrag wurde nach Begutachtung durch das Programmkomitee als „reviewed paper" angenommen.

Abstract

This article discusses two specific application areas in geomarketing using the kernel density estimation approach. The first application area relates to the spatial variations of the dominance in the retail market in the city of Vienna. The second area identifies regions within the same market area with an abundance (shortage) of supply of retail market products relative to the local customer base. Since, to the best knowledge of the authors, this is the first time that the kernel density estimation approach has been used in geomarketing, further research is warranted in order to identify additional areas in geomarketing, where this approach or other point pattern analysis (ppa) techniques can be used as an important analysis tool.

1 Introduction

In general, the term 'geomarketing' covers address- and location-based analysis of business marketing data utilizing geographic information systems [GIS] technology and spatial analysis techniques to support business management decisions (FISCHER & STAUFER-STEINNOCHER 2001). One important group of spatial analysis techniques, the analysis of spatial point patterns, have so far been ignored in geomarketing. This is very surprising, since such techniques have been already successfully applied to other address- and location-based analysis, such as ecology, epidemiology and criminology. Like crime incidents or cases of disease, store locations, for example, can also be represented by a point pattern with attribute data, such as the number of employees, or earnings per store attached to them. Store locations can further be referenced to other socioeconomic and geodemographic variables (e.g., customer/population characteristics), to find out which effect these variables might have on the stores' spatial distribution.

The contribution of this article is thus to discuss how one specific technique in point pattern analysis, namely kernel density estimation could be applied to the area of geomarketing. Section 2 provides a brief introduction into what constitutes a spatial point pattern. Section 3 presents the kernel density estimation approach and Section 4 illustrates how this approach could be successfully applied to geomarketing. This is demonstrated using data on the locations of the outlets of the major retail chains and the spatial distribution of the population at the 'Zählsprengel' level in the city of Vienna. It is shown

that the kernel density estimation can be used to delimit specific areas within the city of Vienna, where one retail chain dominates the retail market. The same technique can further be applied to identify regions within the city, where there is an abundance (shortage) of supply of retail products. Section 5 concludes this article by summarizing the results and suggesting future research in this field.

2 Spatial point patterns

Kernel density estimation is a member of procedures that have been designed to analyse spatial point patterns. A major reason to analyse point patterns is the assumption that data represent one source of evidence that may be helpful in learning more about the phenomenon represented and the processes responsible for generating it (FISCHER, LEITNER & STAUFER-STEINNOCHER 2001). A point pattern map contains two types of components: the points representing the events (the *point pattern*) and the geographical area in which they are located (the *study region*). A spatial point pattern is a simple example of spatial data because the data comprise only coordinates of events, at least at the most basic level. Thus, a point pattern data set consists of a series of point locations $\{s_1,..., s_N\}$ in some study region R. We use the term 'event' in a very general sense since the events in question could relate to a wide variety of spatial phenomena that can be regarded to occur at point locations (e.g., the location of a crime incident, the site of an industrial polluter, the locations of the outlets of a retail chain, etc.). One advantage of referring to observations in a spatial point pattern as events is that often we will need to distinguish between observed occurrences and other arbitrary locations in the study region. For these locations we will reserve the term points (FISCHER, LEITNER & STAUFER-STEINNOCHER 2001).

One of the most obvious characteristics of a point pattern is its *size*. This is simply the number of points, *N*, in the pattern. The study region may be represented by features of various dimensions. In this contribution we limit our attention to [study] regions of *two dimensions*. Two-dimensional regions may be bounded in various ways. We shall refer to the figure enclosed by the boundary of the region as the *shape* of the study region (FISCHER, LEITNER & STAUFER-STEINNOCHER 2001).

The behaviour of spatial phenomena is often the result of a mixture of both *first order effects* (global or large scale trend) and *second order effects* (local or small scale effects). First order effects my be described in terms of the intensity λ ($\boldsymbol{s}$), which is the mean number of events per unit area at point $\boldsymbol{s}$. Second order effects, or spatial dependence, involve the relationship between pairs of events in R. Some point pattern analysis techniques such as quadrat methods and kernel density estimations are more concerned with analysing first order effects, others such as nearest neighbour distances and the K-function look at the possibility of spatial dependence or second order effects.

Techniques in spatial point pattern analysis can be used in both *exploratory* and *explanatory* (i.e., model-driven) fashion. Spatial exploratory techniques aim to search for

data characteristics such as spatial trends, spatial patterns and spatial outliers and are usually applied in a preliminary step in spatial analysis. This step is important when the data are of poor quality, when the data sets are large, or genuine a priori hypotheses are lacking. Model-driven analysis of spatial data relies on testing hypotheses about patterns, utilising a range of techniques and methodologies for hypothesis testing, the determination of confidence intervals, estimation of spatial models, simulation, prediction, and assessment of model fit (FISCHER 2001). In this contribution exploratory spatial data analysis using the kernel density estimation will be the focus of discussion.

3 Kernel density estimation

Kernel density estimation is an interpolation technique that is appropriate for individual point locations, which provides density estimates for all parts of a region (i.e., at any location) (SILVERMAN 1986, HÄRDLE 1991, BAILEY & GATRELL 1995, BURT & BARBER 1996, BOWMAN & AZALINI 1997). The kernel density estimation involves a symetrical surface, or three-dimensional function {the kernel $\kappa(.)$} that 'visits' each point $\boldsymbol{s}$ on a fine grid superimposed over R. Distances to each observed event $\boldsymbol{s}_i$ that lies within the region of influence (as controlled by bandwidth, b), are measured and contribute to the intensity estimate at $\boldsymbol{s}$ {i.e., $\hat{\lambda}_b(\boldsymbol{s})$} according to how close they are to $\boldsymbol{s}$. We may then use a suitable contouring algorithm, or some form of raster display, to represent the resulting intensity estimates as a continuous surface showing how intensity varies over R (BAILEY & GATRELL 1995) (Fig. 1).

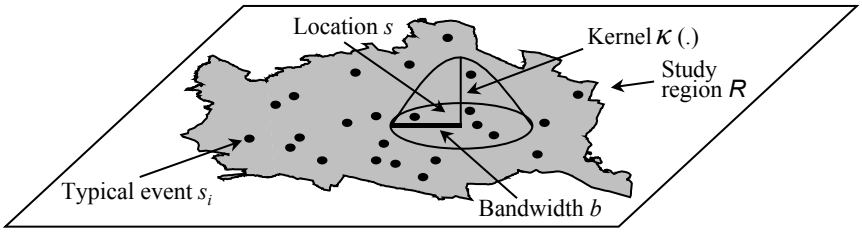


Fig. 1: Kernel density estimation of a point pattern [adapted from BAILEY & GATRELL (1995)]

The use of a symetrical kernel $\kappa(.)$ assumes isotropic conditions around each point $\boldsymbol{s}$ within bandwidth b, which, of course, is a very questionable assumption, especially when anisotropic conditions exist within the study region R (e.g., valley in an Alpine region). The choice of an appropriate bandwidth or smoothing constant $b > 0$ is of primary concern when estimating $\lambda(.)$ (SILVERMAN 1978). The effect of increasing the bandwidth, b, is to stretch the region around $\boldsymbol{s}$ within which observed events influence the intensity estimate at $\boldsymbol{s}$. For very large b, $\hat{\lambda}_b(\boldsymbol{s})$ will appear flat and local features will be obscured. If b is small, then $\hat{\lambda}_b(\boldsymbol{s})$ tends to a collection of spikes centered on the $\boldsymbol{s}_i$. The choice of the kernel function $\kappa(.)$ is of secondary importance, since as long as the kernel is symetrical, any reasonable choice gives close to optimal results (SILVERMAN 1986). There are a number of different kernel functions that have been used, including the normal distribution function

(KELSALL & DIGGLE 1995), the quartic function (BAILEY & GATRELL 1995) and the triangular function (BURT & BARBER 1996). The normal distribution function [1], which is the most commonly used kernel function, has the following functional form:

$$\hat{\lambda}_b(\boldsymbol{s}) = \sum_{i=1}^{N} \left\{ [W_i] \left[\frac{1}{b^2 2\pi} \right] e^{-\left[\frac{d_i^2}{2b^2}\right]} \right\} \qquad [1]$$

where d_i is the distance between the point $\boldsymbol{s}$ and the observed event location $\boldsymbol{s}_i$, b is the standard deviation of the normal distribution (the bandwidth) and W_i is the weight at the event location. This function extends to infinity in all directions and, thus, will be applied to any point in the region.

On the other hand, a quartic kernel function [2] has a circumscribed radius, b, which, at the same time, is also the bandwidth. The quartic kernel function is, therefore, applied to a limited area around each event. Its functional form is:

$$\hat{\lambda}_b(\boldsymbol{s}) = \sum_{i=1, d_i \le b}^{N} \left\{ [W_i] \left[\frac{3}{b^2 \pi} \right] \left[1 - \frac{d_i^2}{b^2} \right]^2 \right\} \qquad [2]$$

where d_i is the distance between the point $\boldsymbol{s}$ and the observed event location $\boldsymbol{s}_i$, b is the radius of the search area (the bandwidth) and W_i is the weight at the event location.

For the problem of edge effects, that can occur, different solutions have been proposed. One way to avoid the problem is to construct a guard area inside the perimeter of R. Kernel estimates are computed only for points inside R which are not in the guard area but events in the guard area are allowed to contribute to such kernel estimates (BAILEY & GATRELL 1995). SILVERMAN (1986), HÄRDLE (1991) and VENABLES & RIPLEY (1997) for example, suggest variations of the size of the bandwidth with various formulas and criteria. Generally, bandwidth choices can either be fixed or adaptive (variable) (KELSALL & DIGGLE 1995, BAILEY & GATRELL 1995). With the adaptive choice, sub-areas in which events are more densely packed than others and where more detailed information on the variation in the intensity is available are 'visited' by a kernel whose bandwidth is smaller than elsewhere, as a means of avoiding smoothing out too much detail (BAILEY & GATRELL 1995). In the fixed choice, the bandwidth always stays the same whether events are more densely packed or not.

If the kernel density estimation is applied to one variable, then this is referred to as a single density estimate, if it is applied to two variables, then we call it a dual density estimate. In the latter case, a kernel density is estimated for each variable in turn and then the two density estimates are related with each other through some simple algebraic operation, such as the sum, the difference or the quotient. The following section will show how both the single and the dual density estimates could be applied to exploratory analysis in geomarketing.

4 Application to geomarketing

To illustrate the use of kernel density estimation for geomarketing we use two different point data sets representing the locations of the outlets of two groups of retail chains in the city of Vienna. One group represents the largest Austrian retail chain group, BML, consisting of the Billa, Merkur, Mondo and Bipa retail chains, which, in May 2000, possessed a total of 400 retail locations. Because the Bipa retail locations market predominantly perfume products and not food products like the other retail chains, the 110 Bipa retail locations were not included in the analysis, leaving the BML group with 290 retail locations. The second group represents the main retail competitors for the BML group and includes the Zielpunkt, Meinl and Spar retail chains, with a total of 279 retail locations (May 2000). In the remainder of this article, this latter group will be referred to as the ZMS retail group.

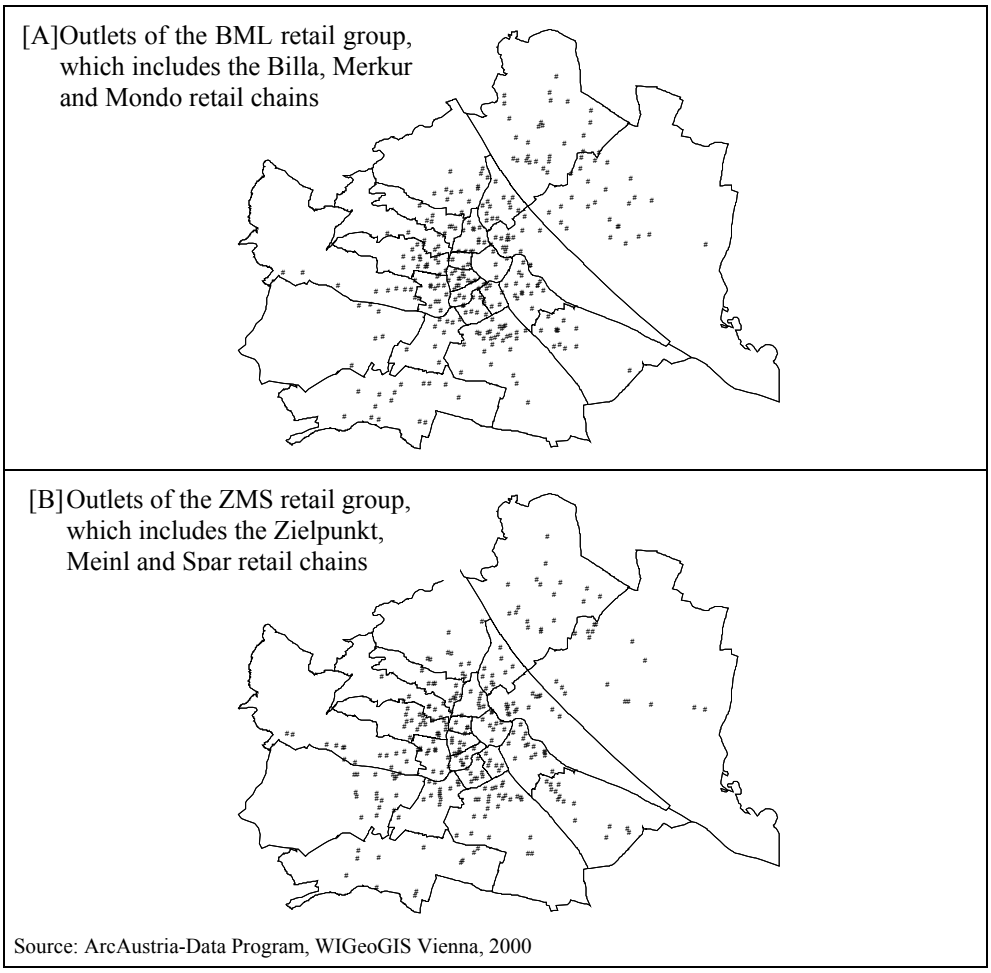


Fig. 2: Spatial distribution of two groups of retail chains

Figure 2 provides a first impression of the spatial distribution of the two groups of retail chains in form of dot maps. Is there a difference in the spatial distribution of the locations of the retail chains between the two groups? Are there regional differences in the retail chain densities between the two groups? In other words, is the BML retail group dominant in some regions of the city, whereas its competitor ZMS is dominant in others? Where are these regions? If we add the spatial distribution of the customers (i.e., the population) to the analysis, do the two retail groups supply their customers equally well? Are there regions in the city, where any one of the two groups faces a shortage (an abundance) of supply? Can we identify these regions? Clearly, such questions are, without quantification, very difficult, if not impossible, to answer. In the subsections that follow, we will show how these questions could be answered using the kernel density estimation approach.

4.1 Spatial variations in the densities of the retail locations

Figure 3 shows the results of a single kernel density estimate for the two groups of retail chains. The estimates are based on a quartic kernel function with a fixed bandwidth, *b*, of 1,000 meters. The density estimates are interpolated to individual cell sizes of 310x308 meters. Edge correction is not used. In both cases, the final classes of the kernel estimates are derived by natural breaks. A relatively large portion of the study region includes raster cells falling outside the bandwidth of 1,000 meters and all these cells are put into a single class. Several test runs that were carried out but are not shown here revealed that a *b*-value higher than 1,000 meters would smooth the distribution too much, while a *b*-value lower than 1,000 meters would give a rather spiky impression of the spatial distribution. A value of b = 1,000 meters gives the most adequate representation of the spatial distribution.

At first glance, the spatial distribution of the density estimates for both groups are somehow similar. The highest densities for both groups can generally be found in the districts or the portions of the districts that are adjacent to the inner city (first district), namely, the second through the ninth district. These highest densities extend further in the south into the tenth district and in the west into the 12th, the 15th, and the 16th district. As far as the differences in the densities between the two retail groups are concerned, it can be said that the BML group covers a larger area in the northeastern portion of the city (the 21st and 22nd district), whereas the ZMS group possesses higher densities in the northeastern portion of the inner city (first district) and has a higher dominance in the western portion of the study region. In order to make a detailed comparison of the density estimates between the two retail groups, for example to delimit regions, where one retail group is dominant over the other (and vice versa), both retail groups need to be related with each other through a simple algebraic operation. This is discussed in the subsection to follow.

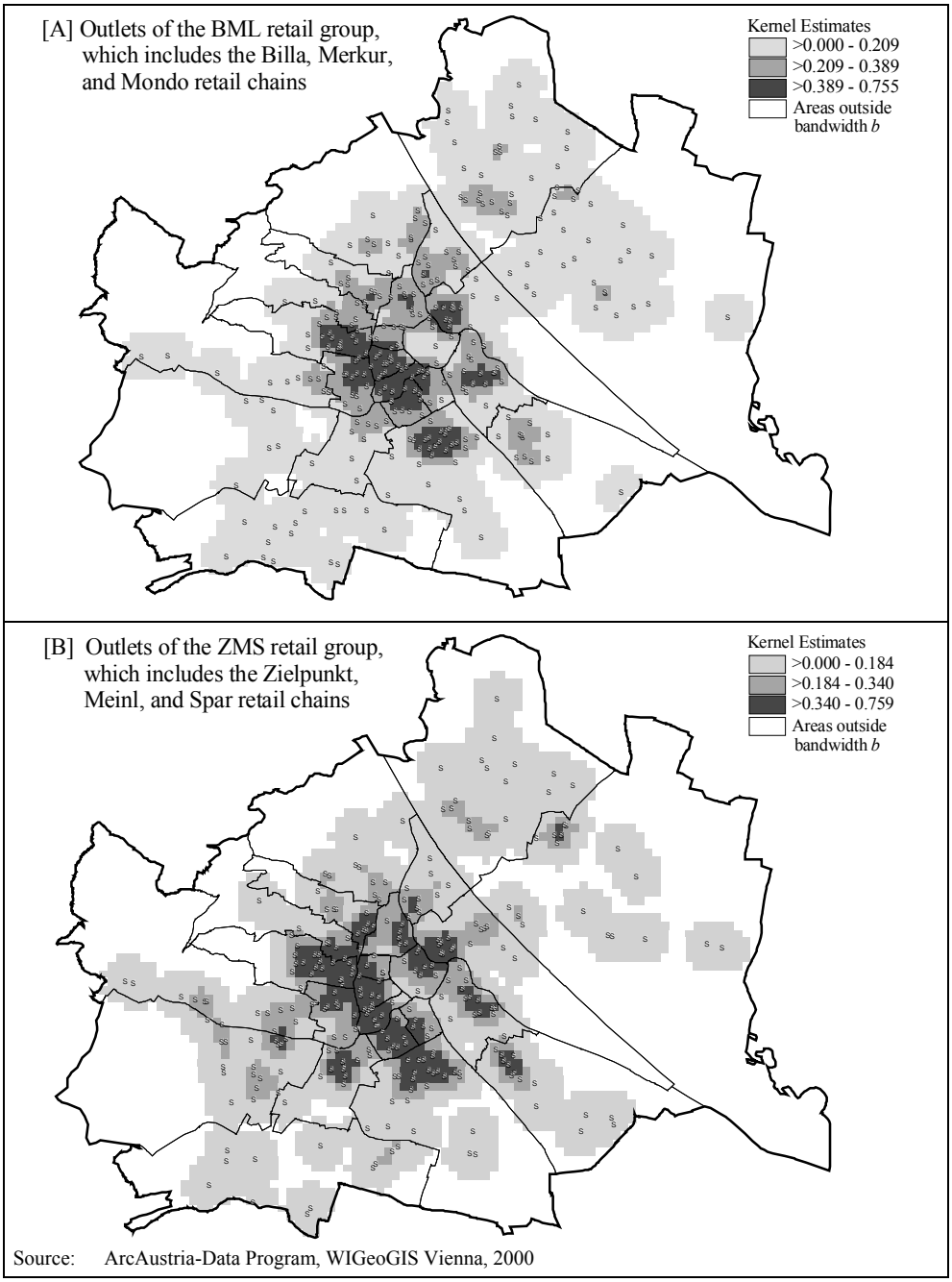
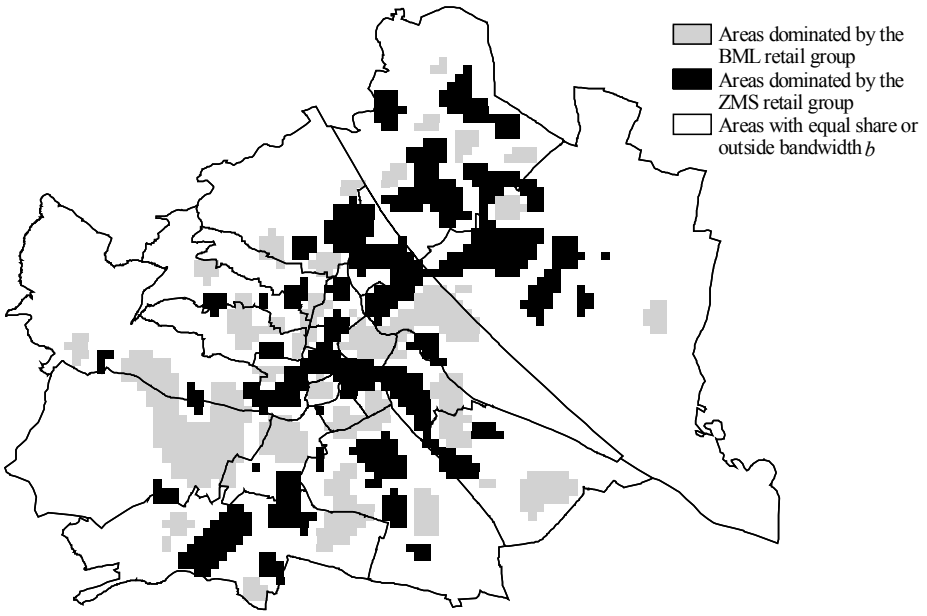


Fig. 3: Kernel density estimations of two groups of retail chains

4.2 Spatial variations in the dominance of the retail market

The results in Figure 4 were derived by a dual density estimate. First, a single density estimate was derived for the locations of the BML retail group, then the same was done for the ZMS retail group, finally, the ZMS densities were subtracted from the BML densities on a cell by cell basis. A positive result can thus be interpreted as an area with dominance of the BML retail group, a negative result as an area, where the ZMS retail group dominates. Areas, where the two single density estimates for both retail groups are exactly the same will lead to a final cell value of 0. These areas can be interpreted as regions where both retail groups possess an equal market share. Such areas and areas falling outside the bandwidth of 1,000 meters are put into a single class.



Notes: The BML retail group includes the outlets of the retail chains Billa, Merkur and Mondo
The ZMS retail group includes the outlets of the retail chains Zielpunkt, Meinl and Spar

Source: ArcAustria-Data Program, WIGeoGIS Vienna, 2000

Fig. 4: Dominance in the retail market

The results in Figure 4 confirm what was said at the end of the previous subsection. The BML group mainly dominates in the northeastern section of Vienna, especially in the 20th, 21st, and 22nd district, further in the seventh district (this district is adjacent to the inner city in the west) and in the 15th district (this district lies in the west adjacent to the seventh district). A minor dominance can be found in the 23rd district, which is the southernmost district in the study area. On the other hand, the ZMS group clearly dominates in the inner city (the first district) and in the two western most districts (the 13th and the 14th districts).

One important variable that clearly determines the spatial distribution of the oulets of the retail chains is the spatial distribution of their customers (population). Relating the density estimates of the outlets with an underlying density estimate of the population distribution will provide important insights into the supply and demand structure within the study region.

4.3 Spatial variations in the demand and supply structure of the retail market

Until now, we have only discussed the supply structure of the retail market, by looking at the spatial variations in the density estimates of the BML and ZMS retail chains and by delimiting areas within the study region, where either the BML retail group, or the ZMS retail group dominates the market. In the following discussion, we will relate the density estimates of the retail chains with the underlying population density distribution at the level of 'Zählsprengel'. As expected, a high spatial association between the (density estimates of the) retail chains and the distribution of the population density is apparent. Figure 5 shows that in Vienna the population is highly concentrated in the smaller districts that lie immediately adjacent to the inner city (first district) in the west and northwest, further in portions of larger districts in the east or southeast that are also adjacent to the inner city. The highest densities then extend into a second set of larger districts in the western and southern portions of the study area. The peripheral areas possess the lowest population densities.

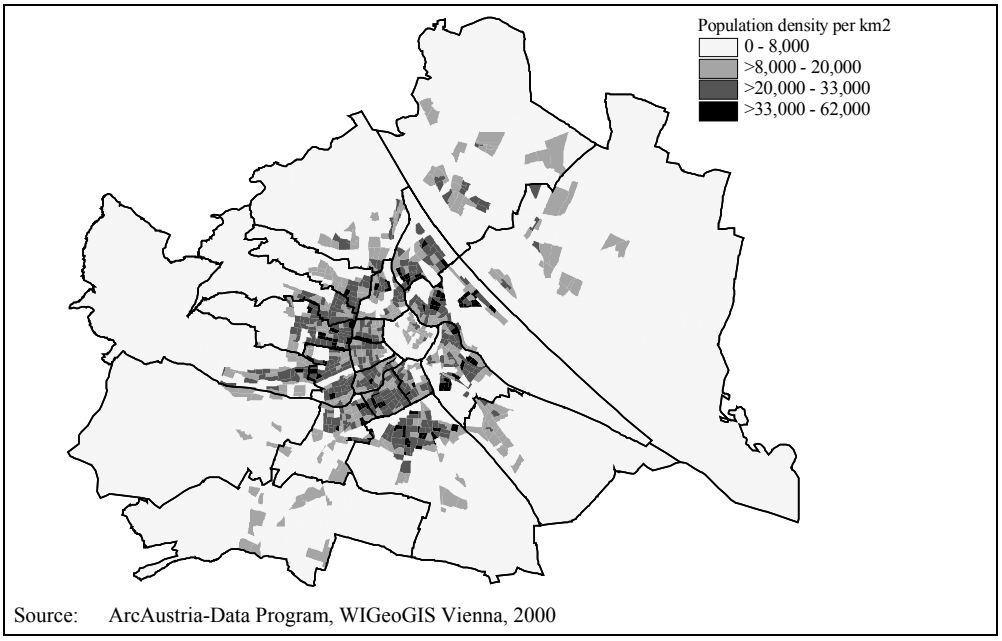


Fig. 5: 1991 population density at the 'Zählsprengel' level

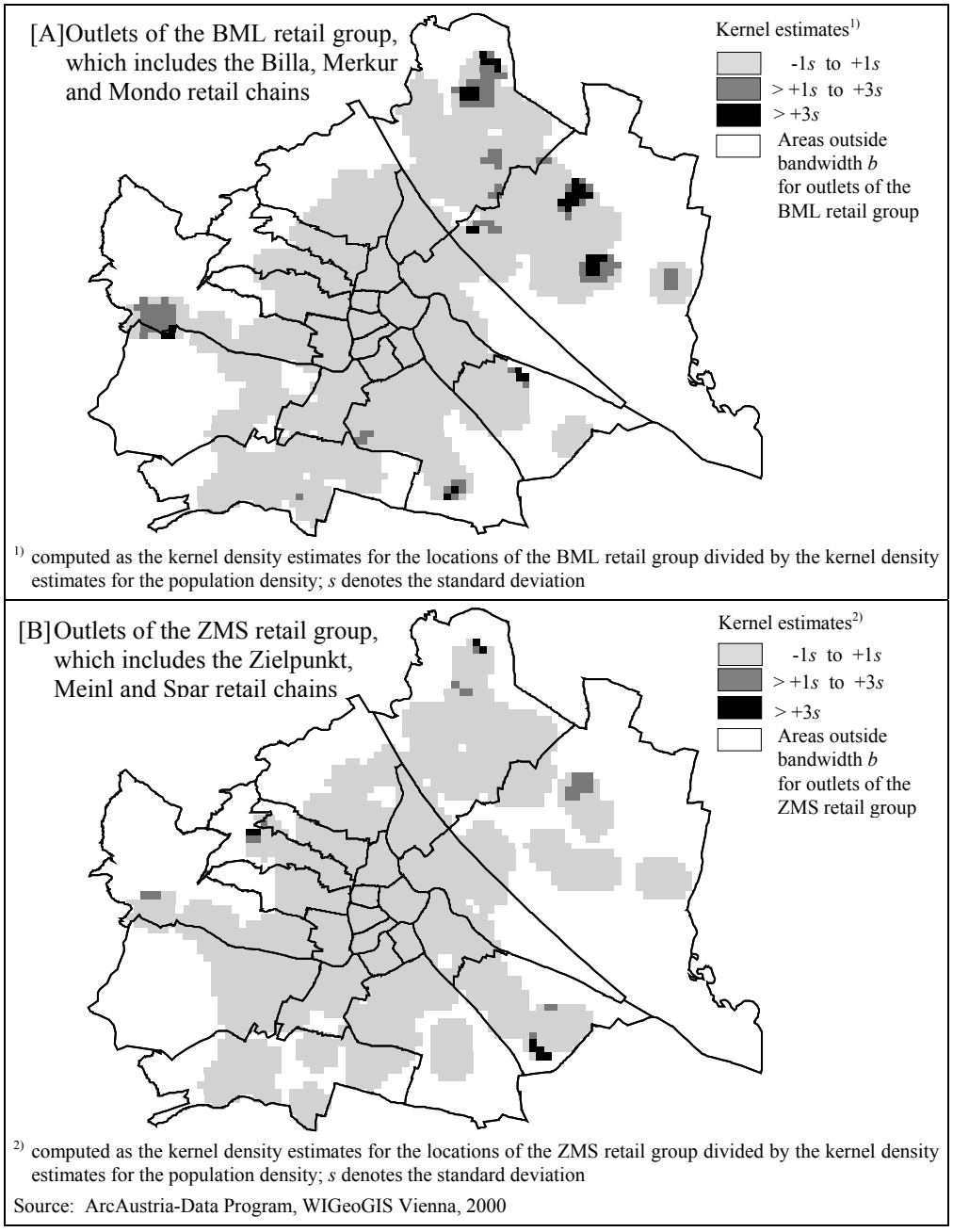


Fig. 6: Demand and supply structure in the retail market

In order to investigate the spatial variations in the demand and supply structure of the retail market, a dual kernel density estimate was derived. It was computed as a quotient of the kernel density estimates for the locations of the BML retail group [numerator] and the kernel density estimates for the population density at the level of 'Zählsprengel' [denominator] (see Fig. 6A). The results in Figure 6B were computed the same way, but using the kernel density estimates for the locations of the ZMS retail group in the numerator of the quotient. The kernel estimates for the population density were derived by using the population density value as the weight, W_i and the centroid of its corresponding 'Zählsprengel' as the event location, $\boldsymbol{s}_i$ [see equation (2)].

All kernel density interpolations were based on a quartic kernel function with a fixed bandwidth b of 1,000 meters and on individual cell sizes of 310 x 308 meters. Edge correction was not applied. The final results were classified by standard deviations. Areas outside the bandwidth b = 1,000 meters for both retail groups, were put into a single category and fell almost exclusively into the peripheral areas of the study region.

The majority of the kernel estimation results fall within a range of -1 to +1 standard deviations of the mean. Areas falling into this range may be interpreted as regions, where the demand and the supply of retail products is in an equilibrium. This means that a shortage of supply neither exists for the BML nor for the ZMS retail group. Areas falling into the two categories that are +1 standard deviation above the mean, have higher density estimates for the retail locations compared to the estimates for the underlying population. Such regions may be interpreted as having an abundant supply of retail products. Such regions can be found in the form of isolated clusters in the northeastern, southern, and western peripheral areas of the study region. Retail chains falling into these regions may not only draw customers among the local population, but also customers from across the boundary of the study region. These customers have not been included in this study.

5 Summary and future research

The research in this contribution shows that the kernel density estimation is a useful analysis tool in geomarketing that can be used to identify regions, where store locations from one retail chain dominate the retail market. The same tool can also be used to identify regions, where the large supply of retail products cannot be met by the local customer base alone, but by customers from outside the study area. On the other hand, regions with a shortage of supply could not be identified in this study.

One shortcoming of this study is that the population data are based on the 1991 census information. For this reason the same analysis should be redone using the updated population distribution from the 2001 census. A second improvement to the current analysis would be to calculate the population density on the basis of the permanently inhabited area that is the total area minus all hydrographic features, parks, transportation features, etc. This would provide a more realistic picture of the distribution of the population density.

In this analysis, no distinction is made between the retail stores with respect to their turnover figures, sales performance, number of employees, etc. This is due to the fact that such data were only available for a relatively small group of retail stores. Of course, the inclusion of such variables, if available, would improve the analysis and would provide more realistic results.

A main intention of this contribution is to break some basic groundwork to the application of modern point pattern analysis (ppa) techniques, of which the kernel density estimation is a member of, to the field of geomarketing. We believe that other ppa techniques, including spatial clustering techniques, spatial autocorrelation statistics, spatial descriptive statistics, etc. should also be investigated for its applicability to geomarketing.

6 References

Bailey, T.C. & A.C. Gatrell (1995): *Interactive Spatial Data Analysis*. Longman, Essex, England.

Bowman, A.W. & A. Azzalini (1997): *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford Science Publications, Oxford University Press, Oxford, England.

Burt, J.E. & G.M. Barber (1996): *Elementary Statistics for Geographers*. The Guilford Press, New York.

Fischer, M.M. (2001): Spatial Analysis. In: Smelser, N.J. and Baltes, P.B. (Eds.) *International Encyclopedia of the Social & Behavioural Sciences*. Pergamon, Amsterdam [forthcoming].

Fischer, M.M., M. Leitner & P. Staufer-Steinnocher (2001): Spatial Point Pattern Analysis - Some Useful Tools for Analysing Locational Data. In: Department of Geography, University of Vienna (ed.): *Geographischer Jahresbericht aus Österreich*, Vol. 58, Vienna (in press).

Fischer, M.M. & P. Staufer-Steinnocher (2001): Business-GIS und Geomarketing – GIS für Unternehmen. In: Department of Geography, University of Vienna (ed.): *Geographischer Jahresbericht aus Österreich*, Vol. 58, Vienna (in press).

Härdle, W. (1991): *Smoothing Techniques with Implementation in S*. Springer, New York.

Kelsall, J.E. & P.J. Diggle (1995): Kernel Estimation of Relative Risk. *Bernoulli*, Vol. 1, pp. 3-16.

Levine, N. (1999): *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. Ned Levine and Associates, Annandale, VA and the National Institute of Justice, Washington, DC.

Silverman, B.W. (1978): Choosing the Window Width when Estimating a Density. *Biometrica*, Vol. 65, pp. 1-11.

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

Venables, W.N. & B.D. Ripley (1997): *Modern Applied Statistics with S-Plus*. Springer, New York.