Enhancing Mobility Behavior Analysis Using Spatial Interactive Tools and Computer Intensive Methods

Arnaud Banos

THEMA, UMR 6049 du CNRS
Université de Franche-Comté, 32 rue Mégevand, 25030 Besançon Cedex, France

Abstract

Since several decades, French urbanization evolves in a way that can hardly be seen favorable to public transportation. Innovation is needed in this field, if the individual automobile is to be faced on its own territories. This need for innovation concerns the miscellaneous stages of the public transport production, including the preliminary analysis ones. Indeed, many work remains to be done to understand mobility behavior, what is more if we are to propose versatile and viable alternatives to current trends.

The aim of this paper is to underline the relevance of exploratory spatial analysis strategies in a context of operational research. As a spatial phenomenon, mobility might be studied from a spatial point of view. Interactive visual analysis is a simple but powerful way of analyzing complex information. Moreover, desegregated but still synthetic approaches have to be privileged. Then, local numeric indicators are found to be much more useful than global ones. An example of such a strategy, led using the statistical package Xlisp-Stat, is exposed below, so as to support these assertions.

I. INTRODUCTION

Since the sixties, French towns grow by their outskirts. The spread of the individual automobile within the whole society explains much of this trend, which has major spatial impacts. This diffused and often spontaneous urbanization, based on individual strategies, leads to social dilemmas, including mobility related ones. Indeed, massive daily flows generated by these multiple sources, tend to concentrate on convergent axes, towards less numerous destinations.

Moreover, if travel needs still increase, in France, a qualitative turn seems to be reached : the rate of daily journeys to work decreases, while the rate of leisure trips augments. This development has to be connected with the changing way of life, and will probably be amplified by the decrease of working time. Polarized flows towards towns and town centers are not as much representative as before of the spatial pattern of daily flows, as fringe flows take more and more importance.

Public transport has to manage these trends, and must evolve to fulfill its mission. It cannot remain on its traditional target, that is daily journeys to work during rush hours, within densely populated areas of towns. Nevertheless, the will to compete with the individual automobile, what is more on its own territories, implies to take up many challenges. As a matter of fact, it is likely that only versatile solutions will be able to pick up the gauntlet.

This major characteristic involves the need for innovations in the miscellaneous stages of the public transport production. Much more accurate knowledge of travel needs of individuals on a given territory is then necessary. Space, time and purpose dimensions have to be managed together, within a framework that combine both user oriented and automatic methods.

The aim of this paper is to present a work in progress, which intends to follow this track. It results from a collaboration kept up between the French National Institute of Scientific Research (CNRS) and the society KEOLIS, one of the leaders of terrestrial public transport in this country.

Based on a concrete example, three successive steps will be tackled. In a first part, the importance of localized information will be underlined. Then, the usefulness of dynamically linked graphics will be exposed : visual capacities of the user are supposed to be too important to be forsaken. Computer intensive tools, such as the LOWESS and the bootstrap, will be used to assess confidence in the visualizations obtained. Finally, an example of a local space-time analysis strategy, based on local indicators, will be presented.

II. I-LOCALISING INFORMATION AS MUCH AS POSSIBLE

Saint-Renan is a small town (8000 inhabitants) in the north-west of the country. Its localization in the narrow vicinity (15 km) of a much bigger town (Brest, 150 000 inhabitants) produces important daily flows towards this attractive center : more than 45 % of the total daily trips in Saint-Renan are directed towards Brest. Furthermore, if Saint-Renan is close enough from Brest to be under its direct influence, it remains outside the urban transport perimeter of this town, which means there is no public transport alternative to the private car. In a spatial context so favorable to the private car, it may seem hardly possible to imagine a public transport alternative to this domination.

In co-operation with KEOLIS, a census was led in Saint-Renan.

1082-4006/00/0701-35$5.00

As it seemed vital in this peculiar context, a rough map of the city was joined to the questionnaire (Figure 1). Individuals were simply asked to draw a mark on the map, in the cell of the grid corresponding to their address.

This rough strategy was adopted for two main reasons : first, the French legislation wouldn't have allowed us to use the exact address of individuals. Second, the rate of non-response would have been certainly greater in such a small town.

From this point, a strategy of simulation was adopted. The Figure 2a points out the geographic information we are really in possession of : each dot corresponds to the center of a cell where at least one individual live at present. This information could be aggregated, leading to a spatial differentiation between cells. Nevertheless, we can go much further, looking for a spatial differentiation within cells as well.

From this statement, a "localize-at-random" strategy called "*jittering*" was used : within each cell of the grid, individuals were localized randomly, preventing them from overlapping each others. The Figure 2b shows the result of this manipulation. Each of the 1500 individuals who replied to the survey is now visible.

At least two arguments can be put forward to bear out such a procedure. First, from the jittered Figure obtained, the dynamic graphic capacities of Xlisp-stat [13] are fully available (see part two). Second, the spatial desegregation of the information allows a better knowledge of the spatial pattern of the population under study. The Figure 3, obtained from the application of a kernel density algorithm [1] to the jittered co-ordinates, underlines the heterogeneous spatial pattern at work. A moving three-dimensional window of a chosen radius "r", scans the studied area, counting the events "Xi" included in its circular area, and weighting them according to their distance to the center "X" of the window :

$$\hat{\lambda}(X) = \frac{1}{r^2(X_i)} \sum_{i=1}^{n} k\left(\frac{(X - X_i)}{r(X_i)}\right) \qquad (1)$$

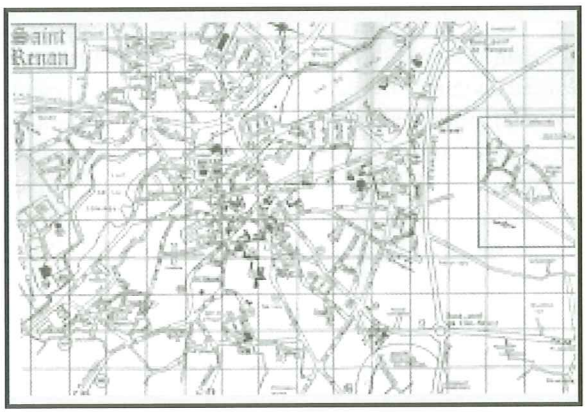


Figure 1. The map used for the survey

The function k(d), which is the kernel, may be defined in several ways. Here, a bi-square function is used :

$$k(d) = \begin{cases} \frac{3}{\pi}\left(1 - d^2\right)^2 \text{ if } d \le 1 & \text{with } d = \left(\frac{(X - X_i)}{r}\right) \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Furthermore, the radius of each moving window is locally adapted to the density of events. Indeed, small radius are used where the density is high, and large radius are preferred where the density is low. The criterion used computes the geometric mean (the numerator of the equation 3) of a first density estimation, based on fixed radius, and uses it to assess a more adapted radius for each window. This "balloon estimator" [12] may be expressed as follow :

$$r(X_i) = r\left(\frac{\tilde{\lambda}_g}{\tilde{\lambda}(X_i)}\right)^{\alpha} \qquad (3)$$

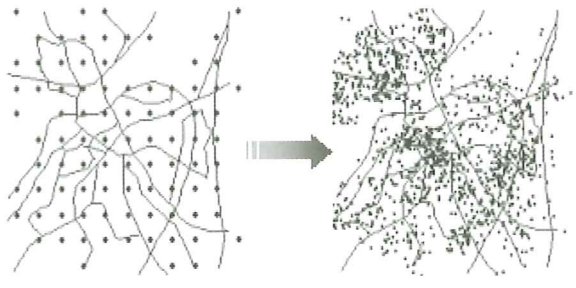
This fully automatic process allows much more accurate estimations of the pattern at work. The densities estimated this way are then interpolated by kriging, producing the smoothed surface shown on the map. The Figure 3 can then be seen as a smooth representation of the Figure 2b, allowing for spatial patterns to be detected easily.

Finally, the hypothesis of random pattern within the cells of the grid is not really restricting. More, it seems more useful to use the Figure 2b under this hypothesis, rather than the Figure 2a preventing oneself from any assumption.
Using the information now available, multidimensional visual explorations of the data set can be performed, using Xlisp-stat [13] visual capacities.

III. II-MULTIDIMENSIONAL VISUAL EXPLORATION... WITH A MOUSE

IIA-Unleashing the power of dynamic graphics

Linking maps and graphics, in a graphical environment, allows



What we really know : centroides coordinates

What there is behind : localization at random around zones centroides

Figures 2. The "localize at random" strategy adopted

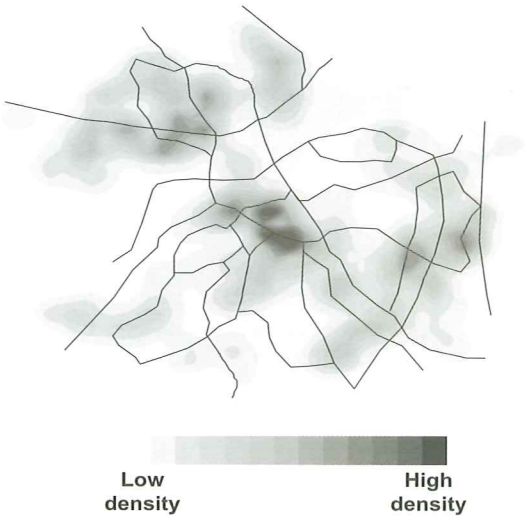


Figure 3. Adaptive kernel density estimation of the people who replied to the survey (initial radius 50 meters)

to explore spatial data visually, in a very simple way [4], [7], [11]. The idea is to dynamically link several figures and to see how the phenomenon they represent interact. For example, it may be instructive to link the Figure 2b, showing the origin of the inhabitants of Saint-Renan who replied to the survey, with a figure showing their destination to work. Then, by selecting an individual from the first map, on the screen of a computer, we would see instantaneously its localization on the second map. In fact, it may be useful to link this figure to any other one : then, a multidimensional visual exploration of the data set would be possible.

The Figure 4 illustrates this idea, using four different variables and a subset of the sample : the origin, destination, work start times and age of people who exercise a daily professional activity.

The first map shows the possible location of people in Saint-Renan, as we obtain it through the "localize at random" strategy described before. Then, the second map shows their destination to work in the whole region of Brest. The main town is divided into 10 districts, while other communes are only known by their centroid coordinates. Saint-Renan is one of the small communes in the north-west of Brest, and is easily identifiable by the vast number of people working there. Two histograms are also linked to these two maps, underlining the age of individuals as well as their work start time. The interval 7 to 7-30 am is selected on this last histogram, so as in each other figure, the individuals beginning to work at this time are highlighted and appear in black.

From selecting to focusing, little work remains to be done. Indeed, we can easily focus on the interval 7 to 7-30 am : the subset remaining can then be explored in the same way as the whole data set. It is worth noticing that complex nested SQL requests can be addressed this way to the data set, in a few mouse-clicks.

Dynamic linked graphics are powerful tools to investigate one's data set. Nevertheless, in many situations, it is worth looking deeper at a relationship underlined by these graphics. For example, we may be interested in looking at the mobility mode of individuals, and relating it to some interesting variables.

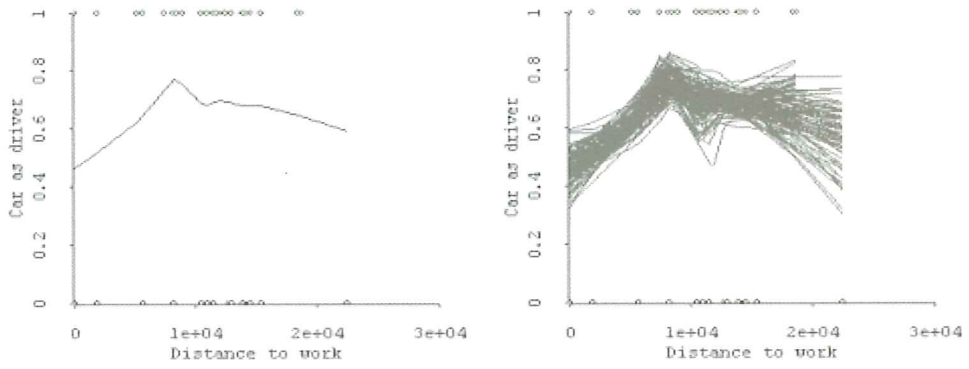
IIB-Assessing relationships using scatterplot smoothing

Scatterplot smoothers, described in [8] allow to identify easily any meaningful trend. This method is very useful, especially in the binomial case. The idea here is to code the "mode" variable as a dichotomous one. Individuals who daily drive to work in their private car may be coded as one, the others as zero. We can then relate this new variable – the proportion of individuals who daily drive to work in their private car – to a predictor one, for example the distance to work or the age of individuals.

The Figure 5a shows the scatterplot obtained this way. Since Y is binary, the mean proportion E(Y/distance to work) represents the proportion of people in the subset, working at a given distance from Saint-Renan, who drive to work using their own car. A LOWESS [5] is then used to estimate this proportion.

As the curve points out, this proportion increases up to 9 kilometers, reaches a peak at this distance, then decreases slowly. We may then be interested in testing the significance of the design of this curve. A bootstrap procedure is adopted here, following Efron's work [6].The idea is to draw many bootstrap samples (that is with replacement) from the subset, then to fit a LOWESS to each of it. The different curves are plotted on the same graph, as the Figure 5b points out . These curves then produce a non-parametric confidence band. It may be seen from this simulation that a higher variability occurs during the downward trend.

Figure 4. Interactive space-time analysis using Xlisp-Stat dynamic graphics



Scatterplot smoothing : visualization of the relationship using a LOWESS

Computing a non-parametric confidence--interval band using bootstrap simulation

Figures 5. Visualizing the relationship between the modal choice « private car as driver » and the distance to work, using scatterplot smoothing

Sometimes, it is worth using simple smoothers, such as a logistic regression with a quadratic term. The Figure 6 underlines this idea, showing the mean proportion of people of a given age in the subset, who drive to work using their own car. The confidence band was obtained with a bootstrap procedure, involving repeated fits of bootstrap samples.

Scatterplot smoothers are thus very powerful tools, especially when they provide visual indicators of confidence. Bootstrap simulations indeed beware us of over-interpreting the relationships underlined.

The approach presented here is above all pragmatic : the idea is to explore one's data set, using the data available instead of rough global indicators. Nevertheless, scatterplot smoothers may as well be seen as useful tools allowing to construct versatile discrete choice models. From this point, our objective is to try to add the power of generalized additive modeling [8] to the wide probabilistic theory of discrete choices [3]. We believe indeed that such versatile and pragmatic approach may enhance greatly this theory, somewhat difficult to apply to practical problems.

Anyway, this "user oriented" approach may not be an end in itself. The interactive space-time analysis sketched out previously can be enriched with relevant local indicators.

IV. III-LEADING PRAGMATIC SPACE-TIME ANALYSIS, MAPPING LOCAL INDICATORS

What is the most favorable space-time configuration for a public transport service in a small town like Saint-Renan ? It may be argued that a configuration where all individuals working in the same area would leave the same origin at the same moment, could pretend to this glorious title. At the opposite, if all individuals of the sample leave their very scattered home at different hours, to very scattered work destinations, then the public transport alternative may be seen as an illusion.

Versatile but pragmatic approaches to this problem can be imagined, as we intend to show below, using a very simple example : the subset of people who exercise a daily professional activity.

Identifying space-time clustering

The Figure 7 shows the space-time variability of the subset. Different colors were affected to individuals, according to their work start time. Pockets of homogeneity (close work start times) can be hardly identified from this kind of representation. A local indicator of spatial association (LISA) can then be used to assess significantly homogeneous areas, in term of work start times.

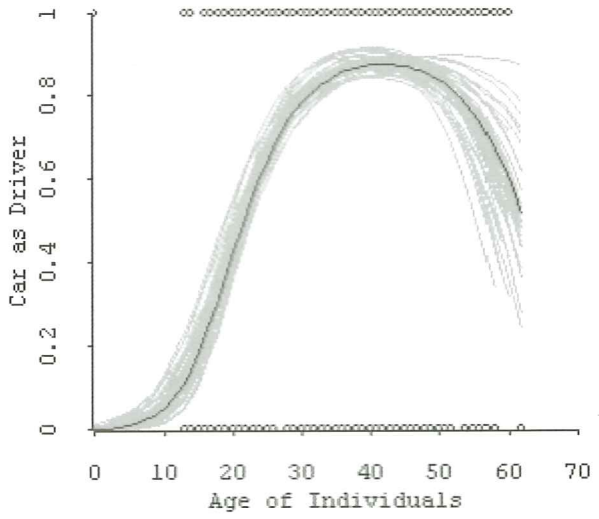


Figure 6. Visualizing the relationship between the modal choice « private car as driver » and the age of individuals, using a bootstraped logistic regression

The idea of Anselin [1] is to decompose global indicators, such as Moran's I, into local indicators. This recent work was adapted here to the specific nature of the data set (point pattern), in order to maintain its richness.

For each individual *i*, neighbors are selected, within a distance specified by the user. Each neighbor is then weighted according to its distance to the individual *i* at which the value is estimated. Proximity between neighbors, in term of work start times, is then compared to the total variance, that is :

$$Moran_{Local} = \frac{(X_i - X') \times \sum_j w_{ij}(X_j - X')}{\sum_i \frac{(X_i - X')^2}{N}} \quad (4)$$

with *"Xi"* the work start time of individual *i*, *"X'"* the mean work start time of the subset, *"N"* the size of the subset and *"w_{ij}"* the weight of individuals, estimated from a bi-square function, defined by :

$$\begin{cases} w_{ij} = \left(1 - \frac{d_{ij}^2}{d_{max}^2}\right)^2 & \text{if } d_{ij} < d_{max} \\ w_{ij} = 0 & \text{otherwise} \end{cases} \quad (5)$$

This local indicator is positive when positive spatial autocorrelation exists (i.e. spatial clustering of similar values), negative when negative spatial exists autocorrelation (i.e. spatial clustering of dissimilar values), and null when no spatial autocorrelation exists at all.

A Monte Carlo permutation test is then used to assess the pseudo significance level of each value of the indicator. The Figure 8 shows the results obtained from the application of this indicator to the subset studied, with a specified distance of 25 meters. Values insignificant to the level of 5% do not contribute to the smoothed surface shown on this map.

A few pockets of similarity come into view, which can be seen as areas where close people start to work at very close times. It is worth comparing this map with Figure 3, as some of these pockets concern areas with high densities of people, which is a serious advantage when designing public transport networks.

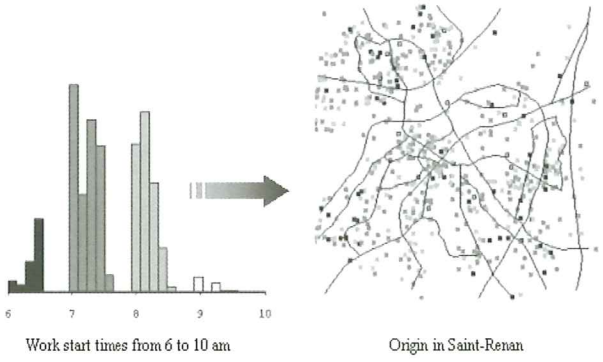


Figure 7. Work start times of individuals who work daily

Anyway, we may want to introduce more complexity : the destination to work is a major factor we can't ignore, if we are to propose a real indicator of "public transport relevance".

Towards an indicator of "public transport relevance"

The idea here is to imagine an indicator of "public transport relevance", based on the combination of three variables : where people live, where they work and at what time. As said before, we may assume that a "perfect" combination of these three factors could be : all people live at the same location and start to work at the same time, at the same location. Of course, if such homogeneous groups could also distribute gradually all day long, for public transport to avoid off-peak hours, this would be really perfect ! This "perfect state" can then be seen as a theoretic state, never attained, in other words a pragmatic normative model.

The approach we propose here is based on now classical literature about space-time modeling. It extends the global indicator of Knox [9] and Mantel [10], providing a local indicator. More, a third dimension is added, allowing for our problem to be handled with more accuracy.

The indicator of relevance is expressed as follow :

$$PTR_L = \frac{1}{n} \sum_{i \neq j} \sum_{i \neq j} \sum_{i \neq j} D_{ij} T_{ij} W_{ij} \quad (6)$$

$$\text{with } D_{ij} = \frac{1}{(\beta + d_{ij})}, T_{ij} = \frac{1}{(\beta + t_{ij})} \text{ and } W_{ij} = \frac{1}{(\beta + w_{ij})} \quad (7)$$

"PTR_L" is the local indicator of "public transport relevance", estimated within moving circular windows centred on each individual *i*. *"n"* is the number of individuals within the circular window, "d_{ij}" the distance between individuals *i* and *j* in Saint-Renan, *"t_{ij}"* the corresponding time interval and "w_{ij}" the corresponding distance between work places of individuals *i* and *j*. In order to remove the scale effect implied by the

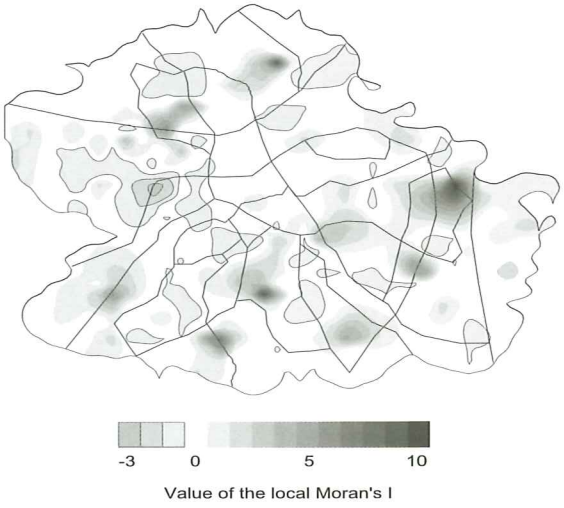


Figure 8. A first local approach to space-time clustering with LISA *(spatial interpolation by kriging)*

combination of three different distances, each of it was previously forced to lay in a [0,1] interval. The transformation used takes the form :

$$\frac{(\text{distance}_{ij} - Min\ \text{distance}_{ij})}{(Max\ \text{distance}_{ij} - Min\ \text{distance}_{ij})} \quad (8)$$

where "distance$_{ij}$" is one of the three distances used, and "*Min*distance$_{ij}$" and "*Max*distance$_{ij}$" the minimum and maximum distances observed for a given metric.

Finally, *"?"* is a constant, allowing for individuals overlapping in any of the above dimensions to be included (here a =1). As can be seen easily, the theoretical state would lead to unity, since the product of the various distances would equal to n. What we are looking for, then, are deviations from this theoretical level.

The Figure 9 underlines the results obtained using circular windows of radius 25 meters. An overlay was realized, allowing for comparisons to be done. The three dimensional surface corresponds to the kernel density estimation mapped in 2D on Figure 3, that is the density of people who exercise a daily professional activity, while the colored scale is for the local indicator of "public transport relevance". This figure confirms then a basic intuition : heterogeneity is the main trend, far from the "perfect state" fixed, as the maximum value is 0.5 (mean=0.357 and standard-deviation=0.038). Nevertheless, few pockets of homogeneity can be identified, that do not necessarily correspond to the main populated areas in Saint-Renan. The main point here is to define whether this apparent conflict corresponds to real pattern, or if there is some instrumental bias behind. Another investigation that would also be worth leading concerns the comparison of this 3D indicator with a 2D one, comparing work start times and distances between work places only for people living within close areas in Saint-Renan. In broader terms, this problem illustrates the kind of conflict that inevitably arise when seeking for "realist" indicators. We have to keep in mind that above all, indicators might be understandable. Realism comes after...

V. CONCLUSION

The aim of this paper was to present examples of pragmatic space-time analysis led in a peculiar context of operational research. Moreover, the miscellaneous approaches developed previously are based on a few simple principles, which may be reminded here.

First, as mobility is above all a spatial phenomenon, a spatial approach may be privileged. Second, information may remain as much as possible desegregated : even if we are looking for trends, specificity of individuals cannot be erased. Third, interactive graphs are as powerful as simple tools to investigate an information so complex, in all its dimensions (geographical and attribute spaces). Fourth, graph diagnostics are, under certain conditions, much more useful than numeric indicators : the scatterplot smoother and the bootstrap test presented in the second part are examples of such tools. Fifth, local indicators may be preferred to global ones, as they are on the one hand more realistic (the illusive quest of a unique indicator is forsaken) and on the other hand much more operational (local variations are clearly identified).

Much work remains to be done, and many of the developments presented above are still in progress. Nevertheless, we intended to show that such an exploratory spatial analysis strategy can be a relevant contribution to the field of transportation research.

REFERENCES

[1] Anselin, L., 1995, Local indicators of spatial association – LISA, *Geographical Analysis*, 27-2 : 93-115.
[2] Bailey, T., Gatrell, A., 1995, *Interactive Spatial Data Analysis*,

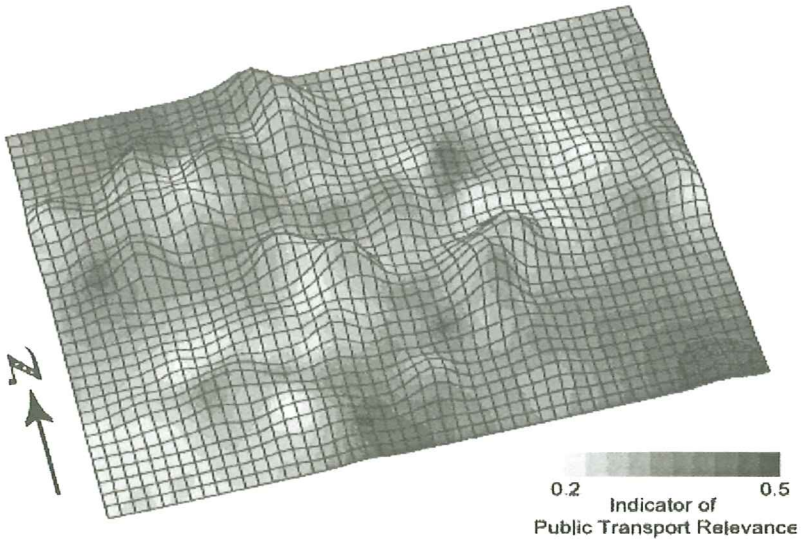


Figure 9. A local indicator of "public transport relevance" *(spatial interpolation by kriging)*

London : Longman Scientific and Technical, 413 p.

[3] Ben-Akiva, M., Lerman, S., 1984, *Discrete choice analysis : theory and application to travel demand*, Cambridge : The MIT Press, 390 p.

[4] Brunsdon, C., 1998, Exploratory spatial data analysis and local indicators of spatial association with Xlisp-Stat, *The Statistician*, 47 : 471-484.

[5] Cleveland, W., 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistician Association*, 74 : 829-836.

[6] Efron, B., Tibshirani, R., 1993, *An introduction to the bootstrap*, London : Chapman&Hall, 436 p.

[7] Haslett J., Bradley R., Craig P., Unwin A., Wills G., 1991, Dynamic graphics for exploring spatial data with application to locating global and local anomalies, *The American Statistician*, 45 : 234-242.

[8] Hastie, T., Tibshirani R., 1991, *Generalized additive models*, London : Chapman&Hall, 335 p.

[9] Knox, E., 1964, Epidemiology of childhood leukaemia in Northumberland and Durham, *British journal of Preventive and Social Medicine*, 18 : 17-24.

[10] Mantel, N., 1967, The detection of disease clustering and a generalized regression approach, *Cancer Research*, 27 : 209-220.

[11] Monmonier, M., 1989, Geographic brushing, enhancing exploratory analysis of the scatterplot matrix, *Geographical Analysis*, 21 : 81-84.

[12] SAIN, S., 1994, *Adaptive kernel density estimation*, Thesis, Rice University, Houston, Texas, 128 p.

[13] Tierney, L., 1990, *Lisp-Stat : an object-oriented environment for statistical computing and dynamic graphics*, New York : John Wiley &Sons, 397 p.