Localising People During Surveys : A Versatile Strategy

Arnaud BANOS,

ThéMA, CNRS UMR 6049, Besançon / KEOLIS, Rueil-Malmaison
arnaud.banos@univ-fcomte.fr

Abstract. With the development and spreading of GIS tools, the localisation of geographical objects becomes more and more straightforward. The accurate localisation of people responding to surveys is that way a natural aspiration, leading to powerful new insights in the interpretation of survey results. Anyway, several problems remain unsolved, that may reduce dramatically the production of such an accurate information. The strategy we propose here is based on the idea that we usually need more accurate than exact spatial information. This strategy has proved efficient on two occasions, during transportation surveys led in France. The main results of these tests will be exposed, as well as resulting analysis that can be led from this kind of spatial information, involving both interactive graphical analysis and automated spatial processes.

1. INTRODUCTION

With the development and spreading of geographic information systems, the localisation of geographical objects becomes more and more straightforward. The accurate localisation of people responding to direct surveys is that way a natural aspiration, leading to powerful new insights in the interpretation of survey results. Anyway, several problems remain unsolved, that may reduce dramatically the production of such an accurate information. First, professionals leading surveys can seldom produce, what is more in a repeated way, what can be still seen as a time consuming information. Then, it becomes more and more difficult to obtain information about people's address, in what can be called a general distrust leaning. Finally, it may be recognised that such personalised information is of really sensitive use, particularly in a public perspective. The strategy we propose here is based on the idea that we usually need more accurate than exact spatial information.

First, we'll investigate some of the benefits resulting from producing such a desegregated spatial information, while leading transportation studies. Then, we'll expose the basis of an alternative procedure, before reporting some simulation results, showing its efficiency. Finally, this alternative strategy has proved efficient on two occasions, during transportation surveys led in France. The main results of these tests will be exposed, as well as resulting analysis that can be led from this kind of spatial information, involving both interactive graphical analysis and automated spatial processes.

2. WHY BOTHERING WITH SPATIAL DESEGREGATED INFORMATION ?

Due to the progresses of geographical information technologies, our capacities to produce and handle accurate spatial information have never been so wide. Transportation surveys derived large benefit from this progresses, exploiting these still improving capacities. Locating people during surveys, using their postal address, is a rather straightforward procedure…as long as the spatial database is available. From this kind of information, both desegregated and aggregated analysis can be led

2.1 Towards An Interactive Graphical Exploration Of Fully Desegregated Spatial Information

Amongst the miscellaneous identifiable flow generators, high speed train stations are of a peculiar interest. Indeed, the spread of high speed trains, in France, increased daily trips between distant cities. Railway stations have then become major flow generators, covering schedules seldom served by public transport services (early in the morning and late in the evening). Nevertheless, this kind of travel implies serious peculiar difficulties, closely related to the temporal dimension of the trip. First, any public transport solution imagined can be seen as the first link of a rather long and complex trip chain. Accordingly, reliability and punctuality will play a major role, leading to demanding solutions. Second, operating early in the morning and late in the evening imposes rather competitive trip times, especially when private vehicle is the main concurrent. As a result, it is likely that only personalised and versatile solutions, dedicated to "niche marketing" identified first, will be able to pick up the gauntlet.

In order to identify needs and expectations of targeted customers, a survey combining both revealed and stated preferences (Ortuzar and Willumsen, 1994) was led in Besançon (120 000 inhabitants). The exact location of 476 individuals is therefore known through their postal addresses, allowing for visual explorations to be led, within an interactive environment such as Xlisp-Stat (Tierney, 1990). The map on the upper left corner of figure 1 shows the location of people in Besançon. Using conjoint analysis techniques, two groups of individuals are identified : red squares are for people with high probabilities of using the innovative service proposed, while blue crosses are for people that don't seem interested in this kind of product. The black rotating plot shows the three first axes of a factorial analysis led on the survey data, while the scatterplot-

smoother relates the binary variable to the second component of the factorial design. A LOWESS (Cleveland, 1979) with its non-parametric confidence band, obtained through a bootstrap simulation procedure (Efron and Tibshirani, 1993), is used to reveal the relationship between these two variables.
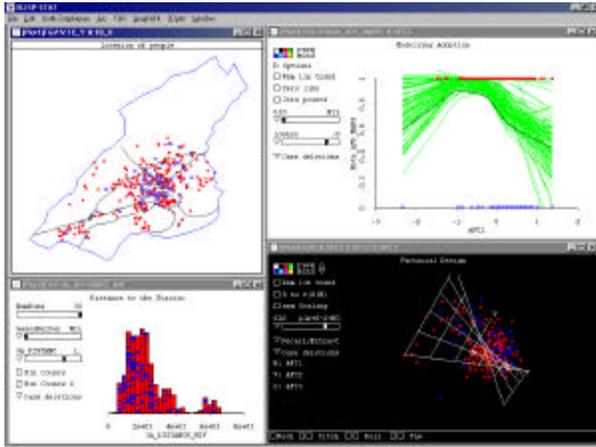

Figure 1: Interactive analysis using Xlisp-Stat dynamic graphics

The existence of a "hot link" between the different windows allows for interactive multidimensional analysis to be led, as every individual or group of individuals selected on one of the graphs is instantaneously highlighted on the other graphs.

2.2 Aggregating Information In Order To Provide Space-Time Map Animation

It is often worth aggregating information, in order to provide additional insights. Aggregating punctual information, such as the location of people, can be usefully realised using a kernel density algorithm (Bailey and Gatrell, 1995 ; Brunsdon, 1991). 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 centre "X" of the window :

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

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

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

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 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" (Sain, 1992) may be expressed as follow :

$$r(X_i) = r\left(\frac{\tilde{I}_g}{\tilde{I}(X_i)}\right)^a \quad (3)$$

This fully automatic process allows much more accurate estimations of the pattern at work. The densities estimated this way are then interpolated by kriging (Isaaks and Srivastava, 1989) producing the smoothed surface shown on the map. This computer intensive procedure allows for map animation to be made, as figure 2 points out[1].
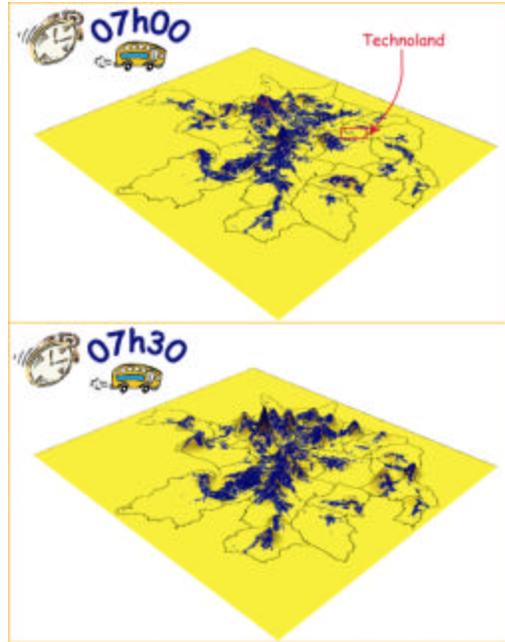

Figure 2: Map animation of people commuting to Technoland, in Montbéliard

This kind of animation is of great help during the design phase of a public transport service, as it helps public transport planners understand that we'd better learn to manage complexity, as we cannot always reduce it.

Nevertheless, if desegregated spatial information proves to be of great interest, several problems remain unsolved. First, professionals leading surveys can seldom produce, what is more in a repeated way, what can be still seen as a time consuming information. Then, it becomes more and more difficult to obtain information about people's address, in what can be called a general distrust leaning. Finally, it may be recognised that such personalised information is of really sensitive use, particularly in a public perspective. The question we then have to answer is : do we really need an "exact" spatial information ? It is the idea defended below that we usually need more accurate than exact spatial information.

---

[1] The complete animation can be viewed at the following address : http://thema.univ-fcomte.fr/BANOS/Banos-Animation2.htm

3. PROPOSITION OF A VERSATILE ALTERNATIVE

The main idea is to avoid as much as possible asking people for their address, as it may be quite embarrassing – particularly in small towns or rural areas.

3.1 A "Localise-at-random" Strategy

A solution that has proved efficient is to join to the questionnaire a map of the city, with superimposed grid. Therefore, people are simply asked to draw a mark on this attractive support, in the cell of the grid corresponding to their address. From this point, a strategy of simulation can be adopted.

The first part of figure 3 points out the geographic information we are really in possession of, that is for each cell the number of marks drawn by people during the survey. This simple aggregated information can be used as is, leading to a spatial differentiation between cells.
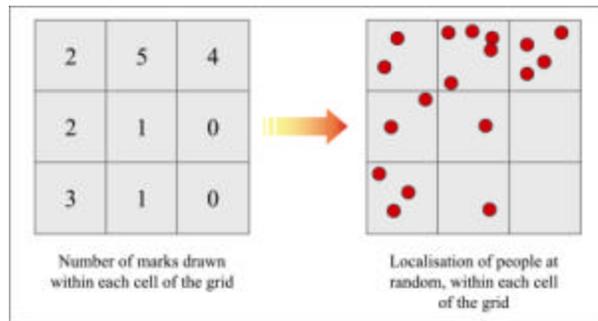

Figure 3: A "localise-at-random" strategy

Nevertheless, we can go much further, looking for a spatial differentiation within cells as well. From this statement, a "localise-at-random" strategy called "*jittering*" can be used : within each cell of the grid, individuals can be localised randomly, preventing them from overlapping each others. The second part of figure 3 shows the result of this manipulation : each of the 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 are fully available. Second, the hypothesis of random pattern within the cells of the grid is not really restricting. More, it seems more useful to use figure 3b under this hypothesis, rather than figure 3a preventing oneself from any assumption.

Of course, a main thought that comes to mind when looking at figure 3, concerns the size of the cells we should use. Indeed, it seems obvious that as the size of cells increases, the simulated spatial pattern moves away from the real underlying one. A simulation procedure can then be used, to evaluate this bias.

3.2 Cells' Size Based Simulation

The simulation we propose here is based on the comparison, for different sizes of cells, of a given known point pattern, with its different "jittered" patterns associated. The algorithm can be detailed as follow :

Step 1 : store the co-ordinates N (X,Y) of the reference point pattern ;

Step 2 : set the minimum, maximum and increasing values of the size of cells ;

Step 3 : input the number of simulation loops n ;

Step 4 : generate a sorted vector of cells' sizes S ;

Step 5 : select a size of cells Si ;

Step 6 : generate a grid covering (X,Y), with cells of size Si :

> {for each cell i :
>
>> {repeat n times :
>>
>> - count the number of points Ni within cell i
>>
>> - generate Ni random co-ordinates (Xi,Yi)
>>
>> - compute and store the Euclidean and Manhattan distances between the (X,Y) and (Xi,Yi) co-ordinates
>>
>> }
>
> }

Step 7 : compute 5%, 25%, 50%, 75% and 95% quantiles of the two distributions of distances.

The two distances used belong to the Minkowski's family :

$$d_{ab} = \left[\sum_{i=1}^{n} \left(|x_{ai} - x_{bi}| \right)^p \right]^{1/p} \quad (4)$$

When p=1, the distance is called "Manhattan", as it reproduces a rectilinear pattern, while the Euclidean distance is obtained when p=2. The first one is thought to be here of a peculiar interest, as it may allow to introduce a pseudo network component, without dealing with this complex geographical objet.

Figure 4 reports the main results of this simulation, ran on the point pattern shown on figure 1. The different curves are quasi-straight lines, with micro-local fluctuations owed to the inner loop.
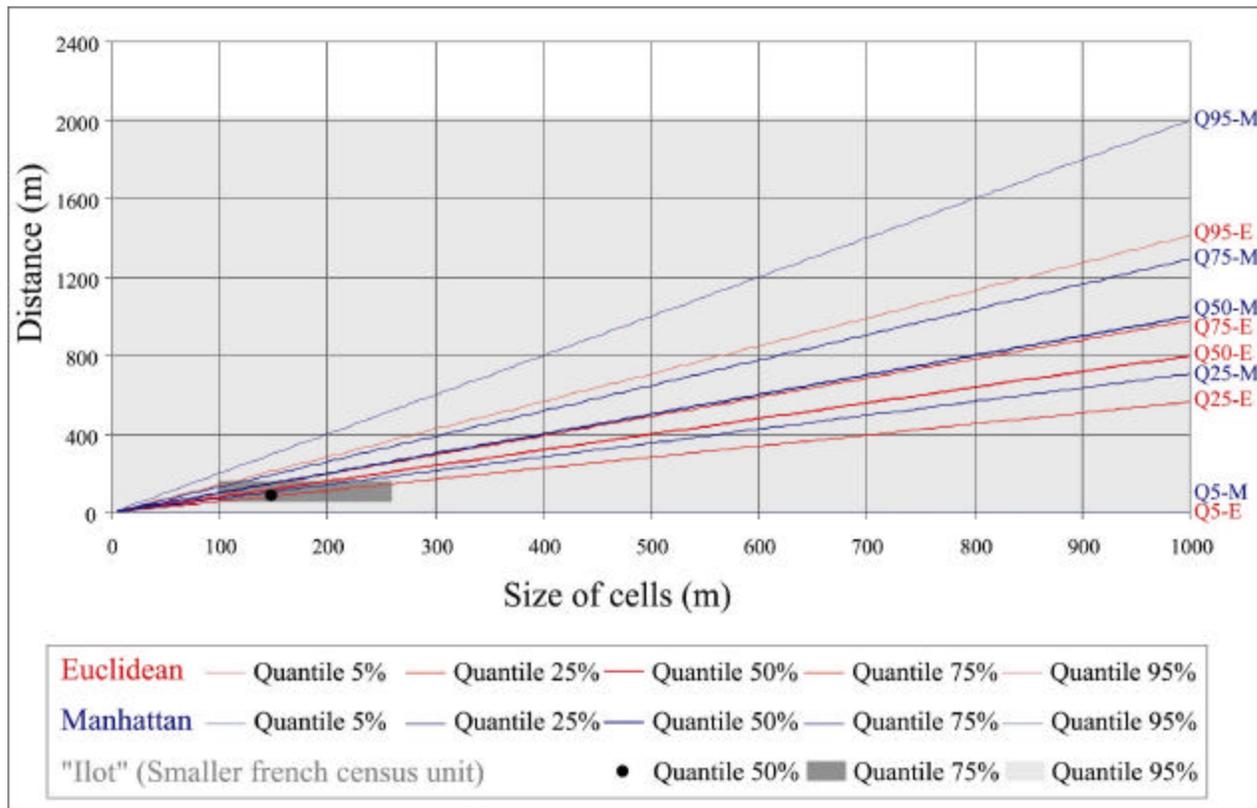
Figure 4: Main results of the simulation procedure, ran for 200 different cells' sizes (5–1000 m), with n=1000 loops

The variability of the two indicators computed increases rapidly as the size of cells increases, but in a more pronounced manner for the Manhattan distance. It may be noticed that the median Manhattan distance and the 75% quantile of the Euclidean one correspond roughly to the *Y=X* line (*y=0,99x+0,008*). As for the Euclidean distance, it shows a lower but still steep slope (*y=0,79x+0,12* for the 50% quantile). This involves that using cells of 200m size, the median Manhattan distance between the points of the reference pattern and the simulated ones equals about 200m, while the Euclidean one is about 160m. As a mean of comparison, the same simulation was ran on the smallest census units available in France ("îlot" census provided by the INSEE). The size of these units shows a high variability, leading to a large variability amongst the distance results.

From this simulation results, we may choose the size of the cells making up our superimposed grid, following the median between-points distance we are ready to assume. Of course, such a decision depends on the context of the study, and on the accuracy we may want to reach. Figure 5 provides additional insights to this problem. During the previous procedure, simulated patterns were chosen at random for different sizes of cells, and were then smoothed using the kernel density algorithm presented before. The succession of maps then shows the progressive dilution of patterns as well as their local shift, as cells' size increases. However, it may be noticed

that up to 200m, the pattern is quite stable, leading to very close maps.

These two results give a reasonable flavour of the expected size bias, even if larger investigations could be led. It would be interesting for example to include the road-network during the "jittering" procedure : within each cell, points could indeed be localised at random solely on the included network. However, the computing effort required for this alternative should be compared with the expected benefits.

This simple strategy was tested twice during transportation surveys we led, in France.

4.   TWO EXAMPLES OF USE

The aim of this last section is to introduce briefly two surveys, during which this strategy has proved efficient. The first one concerns a very small town in the north-west of France (Saint-Renan), while the second one concerns a bigger one, in the western part of the country (La Roche-sur-Yon).

4.1   The Case Of Saint-Renan

Saint-Renan is a small town (8 000 inhabitants) in the north-west of the country. Its localisation in the narrow vicinity (15 km) of a much bigger town (Brest, 150 000 inhabitants) produces important daily flows towards this attractive centre : 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 favourable 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, so as to identify any possible "niche markets".
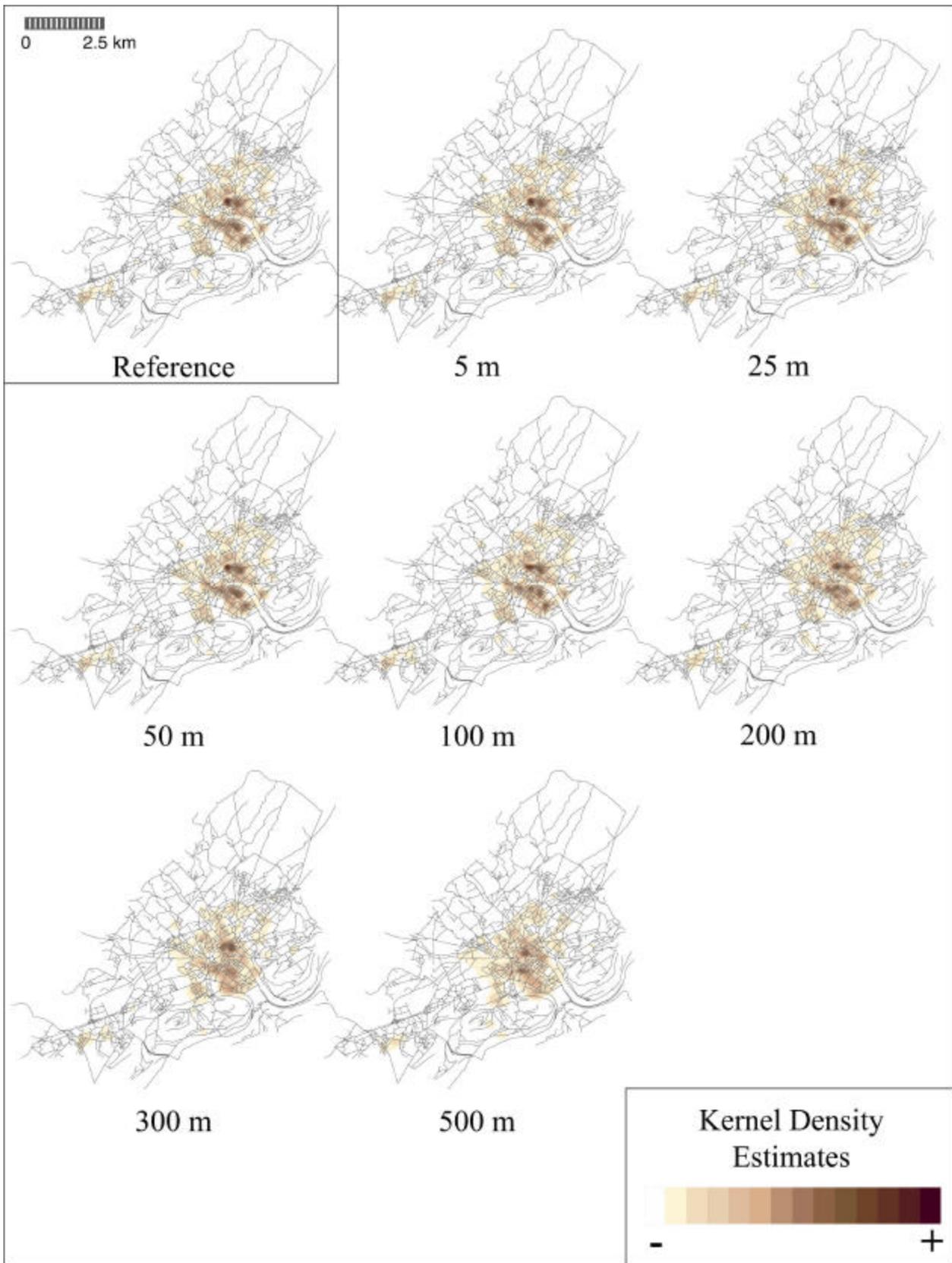
Figure 5: Comparison of simulated patterns extracted at random for different sizes of cell

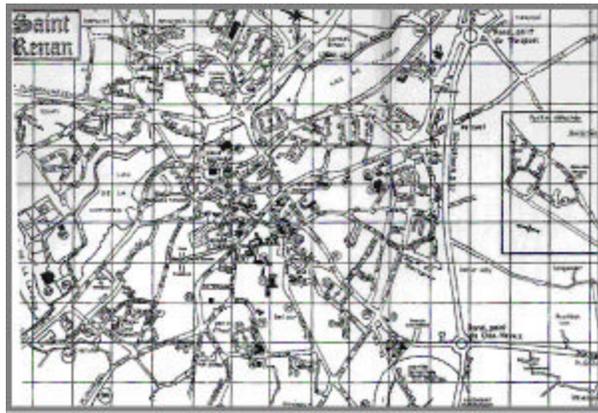Figure 6 shows the map that was joined to the self-filled questionnaire.

Figure 6: The map used in Saint-Renan

A very common map, produced by the local tourist information centre, was thought to be very adapted, as people responding to the survey had to fill the questionnaire at home, on their own. The sides of cells are about 60m long, which give quite an accurate spatial information. It is besides our belief this strategy helped reducing the rate of non-response, which would have been certainly greater in such a small town, if we had asked people for their address. Figure 7 illustrates the "jittering" procedure used.
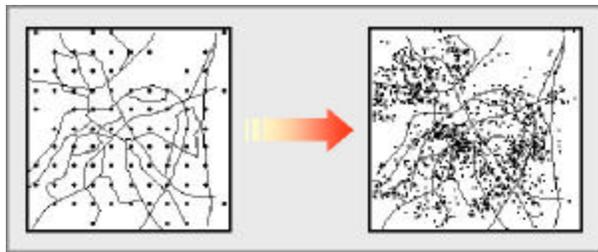

Figure 7: Illustration of the localise-at-random strategy

The first part of figure 7 points out the geographic information we are really in possession of : each dot corresponds to the centre of a cell where at least one individual live at present. Within each cell of the grid, individuals were then localised randomly, preventing them from overlapping each others, an making visible each of the 1500 individuals who replied to the survey. From this point, the dynamic graphic capacities of Xlisp-stat are fully available, as may be seen on figure 8. The first map shows the possible location of people in Saint-Renan, as we obtain it through the "localise 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 co-ordinates. 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. Several histograms are also linked to these two maps, illustrating some variables of the survey. Here, as an example, people using their private car to go to work are coloured in red.
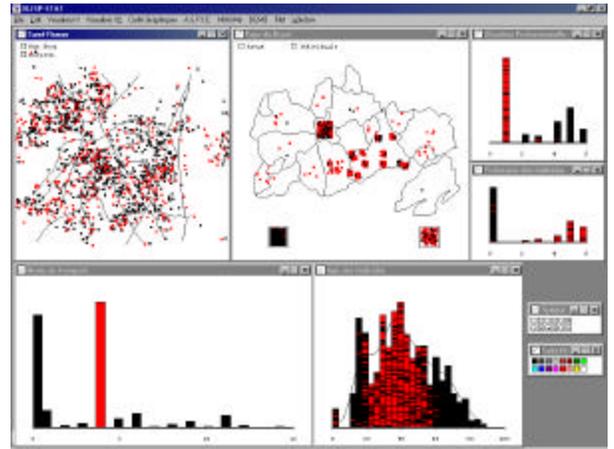

Figure 8: Taking advantages of our visual capacities

4.2 La Roche-Sur-Yon

La Roche-Sur-Yon is a 50 000 inhabitants town in the western part of France. A leisure centre was planned in the outskirts of the town, and we were asked to evaluate the interest of designing a specific public transport service dedicated to this market. The survey was realised in *vis-à-vis* with people an in this context, it was found useful to exploit the much detailed map of the local bus network. Figure 9 shows a very small part of this map, with a superimposed grid of 350m sized cells.

Figure 9: A small part of the map used in La Roche-sur-Yon

As the size of cell is quite large, we chose to take advantage of the dynamic capacities of the Arc extension (Cook and Weisberg, 1999), as figure 10 points out. A slider on the left part of each graph is indeed used to expand points co-ordinates in a versatile way. Therefore, it's up to the user to exploit the benefits of this kind of interactive adds-on, which enhance in a significant way the strategy we propose.
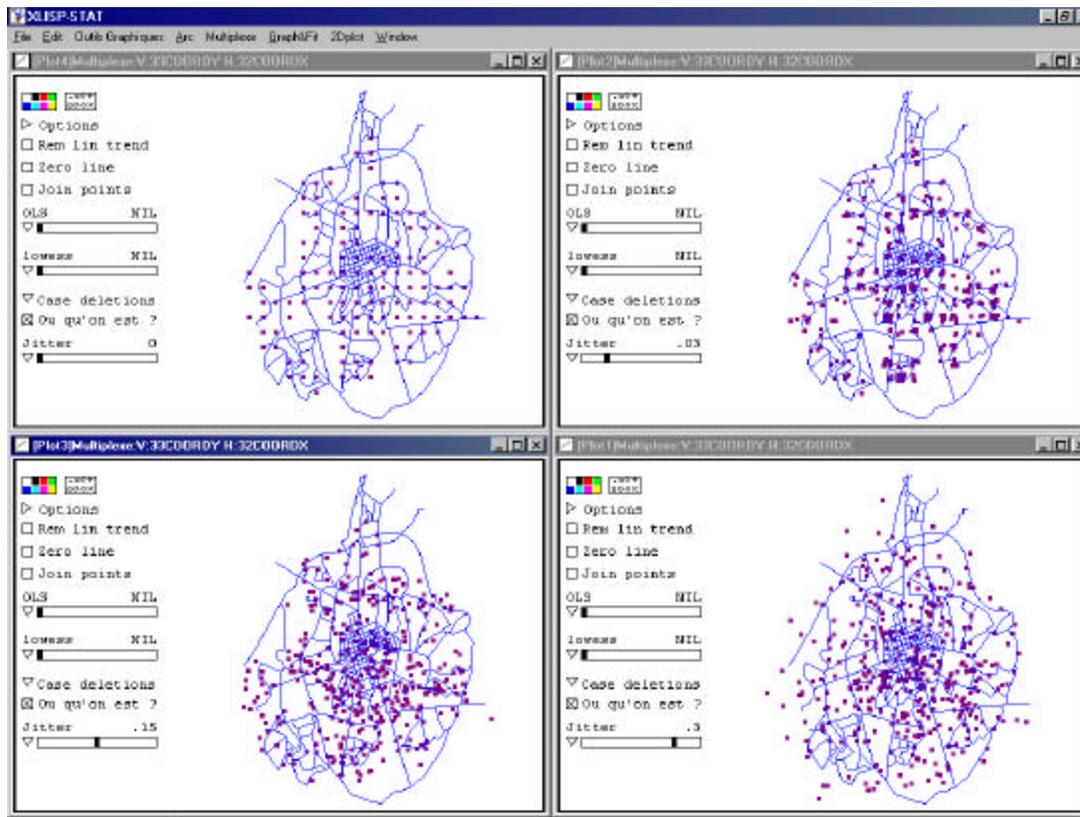
Figure 10: Towards an interactive "jittering" procedure

5. CONCLUSION

The aim of this paper was to propose a versatile alternative to the exact localisation of people by means of their address. Indeed, this objective is often hardly reached in pragmatic context, for different reasons related to either lack of time, lack of information, or legislation sensitivity. Therefore, easily realised maps with superimposed grid can be used during any geomarketing oriented surveys. People are then simply asked to draw a mark on this attractive support, in the cell of the grid corresponding to their address. From this point, a jittering procedure can be used : within each cell of the grid, individuals are localised at random – possibly on the local road network –, preventing them from overlapping each others. A simulation was ran to evaluate the inevitable bias introduced by the size of cells. It showed that up to 200m, the point pattern is fairly stable, when a significant trend to spatial shift and dilution emerges from this size. Finally, this strategy has proved efficient on two occasions, during transportation surveys led in France.

REFERENCES

Bailey, T., Gatrell, A. (1995) *Interactive Spatial Data Analysis*. Longman Scientific and Technical : London, 413 p.

Brunsdon C. (1991) "Estimating probability surfaces in GIS : an adaptive technique", *EGIS '91 Proceeding*, Second European Conference in GIS, Brussels, Belgium, April 2-5, vol. 1, EGIS Foundation, pp. 155-164

Cleveland, W. (1979) "Robust locally weighted regression and smoothing scatterplots", *Journal of the American Statistician Association*, n° 74, pp. 829-836.

Cook D., Weisberg D. (1999) *Applied regression including computing and graphics*, John Wiley & Sons : New York, 593 p.

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

Isaaks E., Srivastava R. (1989) *An introduction to applied geostatistics*, Oxford University Press : Oxford, 561 p.

Ortuzar J., Willumsen L. (1994) *Modelling transport*, John Wiley & Sons : Chichester, 439 p.

Sain S. (1994) *Adaptive kernel density estimation*, Thesis : Rice University, Houston, Texas, 128 p.

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