Determining the Geographical Origin of a Serial Offender Considering the Temporal Uncertainty of the Recorded Crime Data

Marie Trotta, Benoît Bidaine, Jean-Paul Donnay
Geomatics Unit
University of Liège, Belgium
{Marie.Trotta, B.Bidaine, JP.Donnay}@ulg.ac.be

***Abstract*—Since the days the investigating officers used "pin maps" to locate and to think about crime events, crime mapping has become widespread thanks to spatial analysis mainly supplied by GIS-like software. In particular these methods suit well to geographic profiling devoted to crime series characterised by a single offender and hence limited space and time variability. Although spatial techniques are now regularly performed to delineate an offender's area of residence, the temporal dimension is underemployed due to the wider uncertainty of time records. This paper proposes a methodology based on a least-squares adjustment in order to cope with this temporal issue for determining the most probable offender's residence. Moreover, a chi-square test is described to check the significance of the solutions suggested by the method. The process is carried out on the real road network which has been discretised (rasterised) for computing convenience. Three simulations show the validity of the reasoning. Finally the main time and speed assumptions introduced in the model are discussed paving the way for further research.**

***Keywords*-crime mapping; geographic profiling; temporal data quality; least-squares adjustment; raster diffusion process; error propagation**

I. Introduction

Among the domains where the quality of spatial and temporal data especially affects operational performance analysis, crime mapping is probably one of the most intriguing to the general public. Its premises date back to the 19th century, but the analysis offered by GIS allowed geographic profiling to expand only in recent decades. Geographic profiling is defined as a methodology of investigation that uses the locations of a series of connected crimes to determine the criminal's most probable area of residence [1]. In addition to psychologist profiling, it is nowadays widely used in serious crime investigations.

The temporal dimension of geographic profiling has been underexploited while recent criminal literature underlines the importance of simultaneously addressing both spatial and temporal aspects of crimes (several references in [2]). Uncertainty in temporal data largely explains their underutilization. The failure to capture temporal details by the police [3], the practical and/or psychological inability of the victim to specify the time of a crime, and the variability of the offenders' behaviour are the main causes of the low reliability of temporal data. If temporal information is processed in conjunction with spatial information, errors may appear on space (sought-after criminal's place of residence, crime location), on time (starting time of the criminal from his/her residence, time of the crime) and on speed (travel from criminal's residence to crime site).

This paper precisely deals with the modelling of the inaccuracies of the recorded times of crime for providing geographic profiling with a new method exploiting spatio-temporal convergence. In order to address this issue, the paper is organised as follows. The context, Section II, describes the situation in which the offender's anchor site can be determined thanks to temporal information. It also specifies the temporal, behavioural and spatial assumptions required to limit the variation of spatial and temporal parameters. In Section III, we present the basic isotropic methodology used to locate the criminal's most probable area of residence. Then we analyse the least-squares solution for the generalization of the diffusion process on an anisotropic space. Subsequently a chi-square test is proposed in order to assess the validity of the improved method. The reasoning is then illustrated in Section IV by simulation processes generating variances in the data sets.

The consequences of the assumptions introduced in the previous process are discussed in Section V and methodological solutions are considered in order to raise these hypotheses. Finally, the conclusion in Section VI summarises the bringing-in of this article.

II. Context

The context chosen for this research is dictated by real facts relating to multiple rapes perpetrated by the same author, at different dates and for several years within a reduced spatial and temporal variability: crimes spread across an area limited to a radius of a few tens of kilometres, and are perpetrated in a winter time slot ranging between 7.00 and 7.30 AM. The narrow time slot during which the crimes were committed suggests a strong but likely assumption, i.e. a constant departure time from a single anchor point. By modelling the travel time from each crime site, it should then be possible to identify, back in time, the location with similar departure time. This one is assumed to be the offender's anchor point.

Concerning the offender's spatial behaviours, two kinds of scenarii could be explored. According to the criminal profile and the crime type, the offender can take advantage of an opportunity during his daily activities (theory of routine activity in environmental criminology [4]) or, conversely, he prepares the crime beforehand by a precise location.

The first behaviour can use several types of travel between the criminal's place of residence and the site of crime (circuit or zigzag for example). The second behaviour, on the other hand, strongly favours the direct travel between his residence and the site of crime i.e. the use of the shortest path on the road network. In our context, the crime locations and the times they were committed suggest that the criminal has made a very specific location prior to each crime corresponding to the second behaviour. It is therefore possible to favour a "star" pattern in the journeys between the criminal's anchoring site and the sites of crimes.

Regarding the spatial assumptions, the extending area of crime is too large to consider a pedestrian behaviour. Assuming car as mode of transportation, the journeys-to-crime are limited to the road network and only crimes carried in its vicinity are considered. All the points of the network are then candidates to host the offender's anchor point.

III. Determining the geographical origin of a serial offender

Our research question is defined as follows: is it possible to determine the geographical origin of a serial offender taking into account the temporal uncertainty of the crime data? The analysis is based on the temporal, behavioural and spatial assumptions described in the context.

This section analyses firstly the research question with a simplified reasoning. Then it introduces gradually the complexity of real cases in order to clearly identify the various issues relating to the problem.

A. Isotropic diffusion case

1) Definition of the problem: n criminal events occurred at places with known coordinates (x_i, y_i) and at recorded times t_i (with $i = 1$ to n). The purpose of the analysis is to determine the offender's origin, assumed to be his/her residence, and his/her constant time of departure. The problem contains three unknowns: x and y the coordinates of the offender's residence and t, the time of departure from this residence. In the specific case of isotropic diffusion and constant velocity v, the problem is analytically formulated as follows:

$$t + \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{v} = t_i \tag{1}$$

In the case described above, we may identify a solution for this problem according to the least-squares adjustment (LSA) in order to take into account the variability of the t_i. This variability has multiple sources: a bad estimation given by the victims, some waiting time before acting for the offender, small variations in his time of departure.

2) Least-squares formulation: Residuals ν_i were added to the n observation equations:

$$t + \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{v} = t_i + \nu_i \tag{2}$$

According to least-squares theory [5], [6], these equations are linearised:

$$t + \frac{D_i^o}{v} + \frac{(x^o - x_i)}{vD_i^o}\,(x - x^o) + \frac{(y^o - y_i)}{vD_i^o}\,(y - y^o) = t_i + \nu_i \tag{3}$$

where (x^o, y^o) is an approximation of the residence coordinates and D_i^o the corresponding distance to the crime i location. These n equations can be written in matrix form:

$$A\ \underline{x} - \underline{\nu} + \underline{w} = \underline{0} \tag{4}$$

where

$$A = \begin{pmatrix} 1 & \frac{x^o - x_1}{vD_1^o} & \frac{y^o - y_1}{vD_1^o} \\ \vdots & \vdots & \vdots \\ 1 & \frac{x^o - x_n}{vD_n^o} & \frac{y^o - y_n}{vD_n^o} \end{pmatrix} \tag{5}$$

$$\underline{x} = \begin{pmatrix} t \\ x - x^o \\ y - y^o \end{pmatrix} \tag{6}$$

$$\underline{w} = \begin{pmatrix} \frac{D_1^o}{v} - t_1 \\ \vdots \\ \frac{D_n^o}{v} - t_n \end{pmatrix} \tag{7}$$

The Jacobian matrix A contains the equations partial derivatives with respect to the unknowns. The vectors $\underline{x}$, $\underline{\nu}$ and $\underline{w}$ gather respectively the unknowns, the residuals and the independent terms.

The least-squares solution is then given by

$$A^T A\ \underline{\hat{x}} = -A^T \underline{\hat{w}} \tag{8}$$

$$\underline{\hat{x}} = -(A^T A)^{-1} A^T \underline{\hat{w}} \tag{9}$$

where $\underline{\hat{x}}$ is the estimation of x constituting the solution in the least-squares sense .

By definition, this solution minimises the sum of the squares of the residuals (SSR): $\underline{\nu}^T \underline{\nu}\ min$

B. Diffusion on the network

The previous section shows that the simplified problem can be solved with the LSA. However, we assumed an isotropic diffusion on a continuous space. In real situations the road network conditions the path followed by the offender. Therefore we need to generalise the problem to any diffusion process, which from a practical standpoint, is treated by a discretisation of space. Equation 2 becomes:

$$t + d_i(x, y) = t_i + \nu_i \tag{10}$$

where $d_i(x, y)$ are the travel times or the delays between the origin (x, y) and the crime i. The travel time or equivalently the potential starting time at any point of the network $t_{o,i}(x, y)$ is estimated numerically. The solution will be obtained semi-analytically. Let us denote by $\bar{t}(x, y)$ the mean of the potential starting times between homologous locations (x,y):

$$\bar{t}(x, y) = \frac{\sum_{i=1}^{n} t_{o,i}(x, y)}{n} \tag{11}$$

then the least-squares solution is the triplet $(x, y, \bar{t}(x, y))$ minimising the SSR:

$$\underline{\nu}^T \underline{\nu} = \sum_{i=1}^{n} (\bar{t}(x, y) - t_{o,i}(x, y))^2 \ min \tag{12}$$

C. Statistical validation

According to least-squares hypotheses, residuals follow a normal distribution. Consequently, the SSR follows a χ^2 distribution with $n - p$ degrees of freedom (with n the number of observations and p, the number of unknowns). Therefore the following χ^2 test was built to determine the upper bound of the SSR below, which the area can potentially contain the offender's residence.

$$\underline{\nu}^T \underline{\nu} < \ (n - p) \ \sigma^2 \chi^2_{n-p} \tag{13}$$

σ^2 is the a-priori variance chosen according to the uncertainty attributed to the recorded t_i. In first approach, we postulate that the uncertainty is similar for all the observed times (t_i). Indeed, this uncertainty influences the trust we attribute to the identified solution.

IV. Simulations

We chose to work in raster mode as all the pixels of the road network could potentially be candidates for the anchor point. Indeed, the raster mode is suitable to represent a spatially continuous phenomenon. Besides, this mode is compatible with parallel developments explained in the perspectives. Three simulations, implemented on ArcGis 9.3 (Spatial Analyst) are performed to illustrate the reasoning with the t_i obtained from a randomly chosen residence and starting time.

The first simulation considers the t_i without uncertainty while the two others introduce two different precisions corresponding to variance of 2 and 5 on the t_i.

Spatial data concerning the road network and the sites of crimes are available with high precision (5 m). By contrast, there is no information on the network load and therefore on the varying vehicle speeds on the different parts of the network at crime times. In order to isolate the temporal uncertainty coming from the times of crime, a constant speed of 50 km / h was set for this application. It is worth noting that a constant velocity could correspond to a motorist looking for potential targets as well as, for a lower speed, to a pedestrian movement.

A. Evaluation of times corresponding to the crime events (t_i)

The first step of the simulation is to determine four plausible times of crime at four distinct locations respecting the assumption of a constant departure time from an unique location.

- We first arbitrarily choose a location for the offender's residence and a unique time of departure corresponding to 7 AM.
- We build a cost surface corresponding to the crossing time at an assumed constant speed of 50 km/h. The surface is generated in raster mode with a 10-m pixel resolution. A cost of 0.072 is assigned to every pixel of the rasterised network, corresponding to the time required to travel one meter at this speed of 50km/h.
- The cost distance from the offender's residence d is calculated in each cell of the network as a sum of costs per cell c_j. We computed the cost distance using the function "Cost Distance" in ArcGis based on the following algorithm [7].

  $$c_j = c * r \tag{14}$$

  with c the unit cost and r the cell resolution

  $$a_j = \frac{c_j + c_{j+1}}{2} * k \tag{15}$$

  where a_j is the accumulative cost to move from cell j to cell $j + 1$ (cf. Fig. 1) and

  $$k = \begin{cases} 1 & \text{if the movement is horizontal or vertical} \\ \sqrt{2} & \text{if the movement is diagonal} \end{cases} \tag{16}$$

  $$d = \sum_{j=1}^{m} a_j \tag{17}$$

  where j is the current pixel on the way between the origin and the destination and m is the number of pixels on this same way.
- The departure time is added to the calculated cost distance to generate the crossing time of each cell in the network.
- Four crime sites are randomly selected on the network and their crossing time is considered as the crime time ($n = 4$).

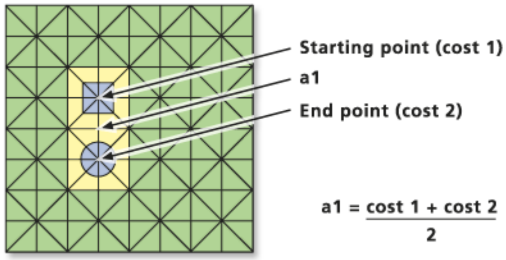

Figure 1. Horizontal and verical node calculations. [7]

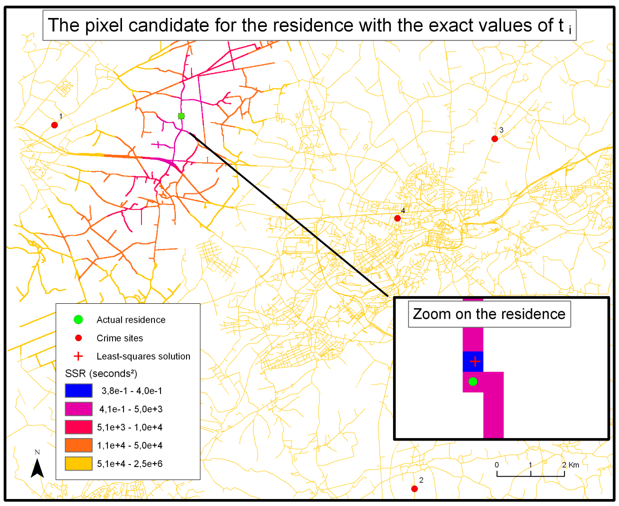

Figure 2. The almost-overlap of the actual residence and the least-squares solution, minimizing the sum of the suared residuals (SSR), illustrates the validity of the least-squares adjustment.

B. LSA for the exact values of t_i

The potential starting time for each observation t_i in every cell of the network $t_{o,i}$ is evaluated by reverse processing (regressive time from the crime sites through the road network using the same cost). This step generates four images that will be used as input for the LSA.

As explained in Section III-B, the pixel average of these images is calculated. This corresponds to the average starting time for the analysed crimes. The residuals consist in the differences between this average $\bar{t}$ and each observation $t_{o,i}$ computed at every pixel of the network. The least-squares solution of the system is the cell that minimises:

$$\sum_{i=1}^{n} (\bar{t} - t_{o,i})^2 \tag{18}$$

and should ideally be equal to 0.

The result is presented in Fig. 2. The obtained minimum is slightly different from 0 and located in the cell just next to the one containing the arbitrary chosen residence. The result can still be considered as valuable despite this difference explained by the algorithm used to calculate the cost distance. Indeed, the cost is calculated from the cell next to the one containing the residence and its value depends on the precision of the cell resolution.

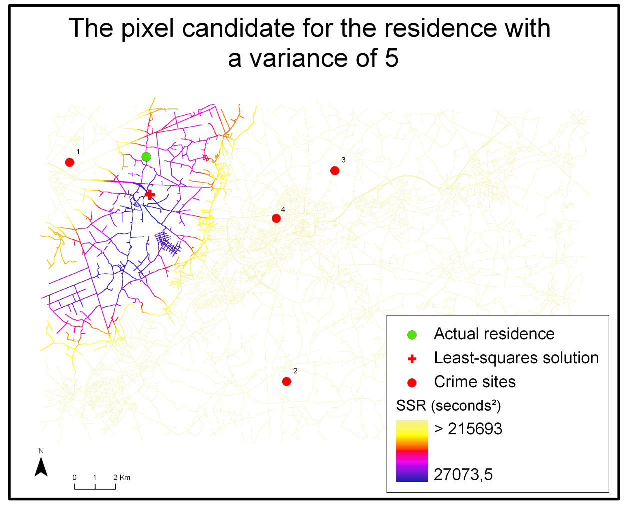

Figure 3. The offender's residence is located inside the area delineated by the χ^2 threshold on the sum of the squared residuals (SSR).

C. Introduction of the uncertainty on the t_i

We choose two a-priori variances of 5 and 10 corresponding respectively to uncertainties of 2'15" and 3'10". The new values of t_i consistent with these variances, respectively $t_{i,2min}$ and $t_{i,5min}$, are presented in Tab. I. The potential starting time in every cell of the network is then updated using the methodology previously described in order to evaluate in each cell the new value of SSR. Figures 3 and 4 illustrate the result of the test described in Equation 13 using a probability level of 95 %. The least SSR value, the least-squares solution, does not correspond to the arbitrary offender's residence because of the introduced uncertainties. Nevertheless it is worth noting that this place remains located inside the area delineated by the thresholded value of the χ^2 test.

In addition, the search for the criminal can be prioritised in this area.

V. Discussion of the Assumptions and Perspectives

The previous analyses are based on several assumptions that are not necessarily encountered in real life.

Firstly, it assumes an offender's constant starting time. This is certainly the most restrictive hypothesis as it renders useless any attempt to solve crime series where the crime locations are close to each other while the crime times are very different. Therefore, other hypotheses than a constant departure time of the offender have to be considered. The minimization of the variances of the journey times from

Table I
SIMULATION PARAMETERS

Crime event	Distance (km)	Travel time	Exact t_i	t_i with a variance of 5	t_i with a variance of 10
1	5.11	6'08"	07:06:08	07:08:22	07:10:08
2	15.89	19'04"	07:19:03	07:16:37	07:21:31
3	12.31	14'46"	07:14:46	07:16:46	07:11:46
4	8.50	10'12"	07:10:12	07:07:58	07:07:12
Value of the upper bounds calculated from the χ^2 (seconds^2)				215693	431385

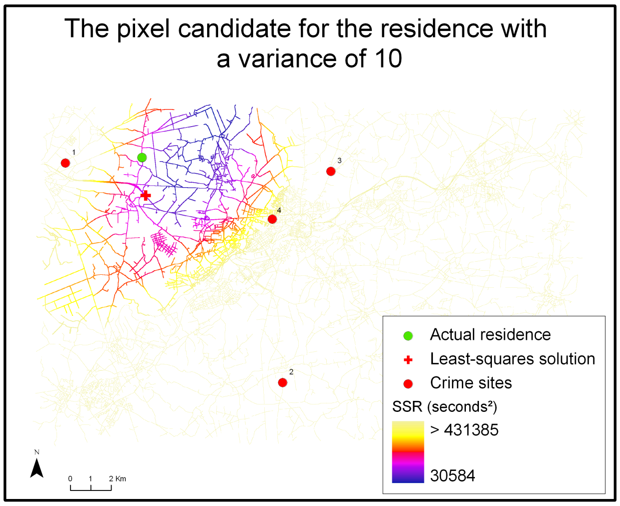

Figure 4. The delineated area widens and moves with the change of temporal uncertainty.

the offender's anchor point could be analysed with a quite similar methodology.

Secondly, an identical time uncertainty is attached to every observation. In practice some events may be recorded with better precision than others: a witness can corroborate the information given by the victim; the time of the event can be constrained by the victim's activity schedule; etc. In order to consider such a variable time uncertainty, a solution could be obtained thanks to a weighted LSA [5]. Indeed the relative confidence attributed to each crime time could be computed as the ratio of the a priori variance and the corresponding time variance. These weights introduced in the resolution process would then modify the estimation of the unknowns as well as the SSR analysed in the χ^2 test.

A third assumption deals with the speed supposed to be constant going through the road network. If this assumption can be considered acceptable for pedestrian journeys, it is not the case for an offender travelling by car. Car cruising speed varies considerably on a network according to a number of parameters depending notably on the spatial environment (urban street, rural road, etc.) and on temporal conditions (time of the day, day of week, etc.). Most of these parameters are well identified in the traffic engineering literature [8]–[10] and may then be exploited by our methodology with very little changes. However the speed at a precise place and moment can only be approximated by the dedicated algorithms (e.g., multi-agent system) [11], [12]. Consequently the errors inherent to speed are added to the time uncertainties on the events and they are merged in the residual values provided by the LSA.

Besides, crimes are generally not committed on the road as considered in this application, but either in discrete locations in the vicinity of the road where the victims are driven, or in the victims'home. The path covered by foot by the criminal on and out of the road can reach tens to hundreds of meters. A research trying to model this kind of trips is under progress, aiming to achieve a cost surface using raster images of land uses at a metric or decametric resolution.

The fact of raising assumptions modifies the assessment of the total time error. The method presented herewith should therefore be part of a more comprehensive approach to error propagation. In this respect the use of a modified Monte Carlo algorithm to train different types of Bayesian neural networks and to estimate uncertainty limits [13], is considered.

VI. CONCLUSION

This study develops a methodology for determining the most probable area for an offender's residence assuming a constant starting time. The process is performed on a rasterised road network crossed at constant speed and the method, which is based on a LSA, is able to include the uncertainty affecting the recorded times of the crimes. A χ^2 test also described herewith allows to check the significance of the value presented by the residence locations (pixels) suggested by the method.

This method allows to test a hypothesis of constant starting time for the offender's spatial behaviour at a constant travelling speed. A geographic profile can only be built on eliminating progressively some hypothesis concerning this behaviour.

The analysis is supported by three simulations: the first one assumes exact values for the crime times, while the others introduce a variability in the crime times corresponding to variances of 5 and 10 respectively. The simulations show that the identified solutions lie in the vicinity of the correct location in a way adequately described by the χ^2 statistic.

Acknowledgment

The researches achieved by M. Trotta and B. Bidaine are funded under F.R.S.-FNRS fellowships (Belgian National Fund for Scientific Research).

References

[1] K. Rossmo, *Geographic profiling*. Boca Raton.: CRC Press, 2000.

[2] T. Grubesic and E. Mack, "Spatio-temporal interaction of urban crime," *Journal of Quantitative Criminology*, vol. 24, no. 3, pp. 285–306, 2008.

[3] P. J. Brantingham and P. L. Brantingham, "Anticipating the displacement of crime using the principles of environmental criminology," pp. 119–148, 2003.

[4] P. Brantingham and P. Brantingham, "Notes on the geometry of crime," in *Environmental Criminology*, B. P.L. and B. P.J., Eds. Beverly Hills: Sage, 1981, pp. 27–54.

[5] C. Ghilani, *Adjustment Computations: Spatial Data Analysis*, 2nd ed. John Wiley & Sons, 2010.

[6] P. Sillard, *Estimation par moindres carrés*, ser. Collection ENSG . IGN. Paris: Hermès Sciences Publications, 2001.

[7] ESRI, "Arcgis resource center. how cost distance tools work," 1999-2010. [Online]. Available: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html. Consulted on the 23th October 2010

[8] W. Homburger, J. Hall, W. Reilly, and E. Sullivan, *Fundamentals of Traffic Engineering*, 16th ed. University of Berkeley, Institute for Transportation Studies, 2007.

[9] H. Miller and S.-L. Shaw, *Geographic Information Systems for Transportation*. Oxford University Press, 2001.

[10] J.-C. Thill, *Geographic Information Systems in Transportation Research.* Pergamon, 2000.

[11] E. Groff, "Characterinzing the spatio-temporal aspects of routine activities and the geographic distribution of street robbery," in *Artificial Crime Analysis System. Using Computer Simulation and Geographic Information Systems*, L. Liu and J. Eck, Eds. New York: Information Science Reference, 2008, pp. 226–251.

[12] F. Balbo and S. Pinson, "Dynamic modeling of a disturbance in a multi-agent system for traffic regulation," *Decision Support Systems*, vol. 41, no. 1, pp. 131–146, 2005.

[13] F. L. Xuesong Zhang, R. Srinivasan, and M. V. Liew, "Estimating uncertainty of streamflow simulation using bayesian neural networks," *Water Resour. Res.*, vol. 45, no. 2, p. 16, 2009. [Online]. Available: http://dx.doi.org/10.1029/2008WR007030