




# An analysis about the accuracy of geographic profiling in relation to the number of observations and the buffer zone

Ugo Santosuosso<sup>1</sup> · Alessio Papini<sup>2</sup> 

Received: 9 October 2020 / Accepted: 28 March 2022 / Published online: 9 June 2022  
© The Author(s) 2022, corrected publication 2022

## Abstract

Geographic Profiling (GP) attempts to reconstruct the spreading centre of a series of events due to the same cause. The result of the analysis provides an approximated localization of the spreading centre within an area (often represented as a red red), where the probability of finding it is higher than a given threshold (typically 95%). The analysis has as an assumption that the events will be likely to occur at very low probability around the spreading centre, in a ring-shaped zone called the buffer zone. Obvious examples are series of crimes perpetrated by an offender (unwilling to perpetrate offences close to home), or the localities of spread of an invasive species, where the buffer zone, if present, depends on the biological features of the species. Our first aim was to show how the addition of new events may change the preliminary approximate localization of the spreading centre. The analyses of the simulated data showed that if  $B$ , the parameter used to represent the radius of the buffer zone, varies within a range of 10% from the real value, after a low number of events (7–8), the method yields converging results in terms of distance between the barycentre of the red zone and the “real” user provided spreading centre of a simulated data set. The convergence occurs more slowly with the increase in inaccuracy of  $B$ . These results provide further validity to the method of the GP, showing that even an approximate choice of the  $B$  value can be sufficient for an accurate location of the spreading centre. The results allow also to quantify how many samples are needed in relation to the uncertainty of the chosen parameters, to obtain feasible results.

**Keywords** Geographic profiling · Criminal geographic targeting algorithm · Modelling · Centre of origin · Buffer zone · Crimes mapping

**JEL Classification** C13 C15 · C31 Q56

---

✉ Alessio Papini  
alpapini@unifi.it

Extended author information available on the last page of the article

## 1 Introduction

Geographic Profiling (GP) is a data analysis concept that aims to identify the origin of a series of events that can be represented on a map. It was firstly proposed in criminology as Rossmo's Criminal Geographic targeting algorithm -CGT- (Verity et al. 2014; Gorski 2021) as a method to deal with serial killers (Butkovic et al. 2019) and later for the spreading of populations of invasive species, for targeting writers (Hauge et al. 2016) and also in epidemiology (Papini et al. 2013, 2017a; Santosuosso and Papini 2016).

The model requires that there is only one spreading centre and that this centre is surrounded by a buffer zone (Rossmo 2000), that is an area around the spreading centre itself where the probability of finding an event is extremely low. In criminology, but still more in biological invasions, the existence and the extension of the buffer zone should be checked case by case.  $B$  (the parameter used to represent the buffer zone around the offender's home (see Fig. 1S in supplementary material), is a number that is between zero and the maximal distance of an event from the spreading point, otherwise the analysis would not produce any result. Moreover, in the model the probability of finding an event tends to decline proportionally to the increase in distance of the event from the offender's home.

A GP uses coordinates of points (corresponding to the events of interest) to calculate a probability surface called a geoprofile (Rossmo 2000). The geoprofiles provide areas of different priority on the map with a varied probability density (Rossmo 1993, 2000). The geoprofile consists in an approximated localization of an area (often represented as a red area) containing the spreading centre, where the probability of finding it is higher than a given threshold (typically 95%). After its first use in criminology, GP was applied to biological problems such as the targeting of an infectious disease (Papini and Santosuosso 2016), the prediction of nest locations of bumble bees (Suzuki-Ohno et al. 2010), animal foraging (Le Comber et al. 2006; Raine et al. 2009) and shark hunting patterns (Martin et al. 2009). The obtained results can be compared to those of other analytical methods of mapping the higher or lower probability of crimes occurrence in a given area, as in Quick (2019).

More recently, GP was used to guess the source of an invasion by alien organisms using the positions of their current populations (Stevenson et al. 2012; Cini et al. 2014; Papini et al. 2017a). In this case, in the place of the offender's home in criminology, the spreading centre will be the first place of introduction of an alien organism with invasive capability (Papini et al. 2013). Such data are increasing with respect to the past, also due to the diffuse monitoring activity on the territory, as with the Citizen science initiatives (Baker et al. 2019). This analysis is useful, since the knowledge of the source of events can suggest hints about the offender's home or give ideas about the control methods of the invasion of alien plants and animals (Cini et al. 2014). Recently, further refinements of the method were proposed to a) improve the reliability of GP from the point of view of possible presence of multiple waves of invasion, rather than from a single starting point (Cini et al. 2014), b) to increase the robustness of the results with a jackknife procedure (Papini et al. 2017b), c) to give different weights to data on a quantitative basis (for instance

on the basis of the population dimension in a given point in the case of an invasion) or several methods of data partitioning (Santosuosso and Papini 2018; Cini et al. 2019). GP may even be associated with other methods that could record the presence of events to be considered outliers of the distribution, as with the Isolation Forest method as defined by Liu et al. (2012). This method was implemented by Santosuosso et al. (2020) on biological invasions. In order to overcome some of the pitfalls of GP analysis, some authors proposed alternative approaches, such as Dirichlet process mixture -DPM- (Verity et al. 2014), the Topological Weighted Centroid -TWC- (Buscema et al. 2018a,b) or the O'Leary's simple Bayesian model (2009, 2010, 2012), even if GP remained the method of choice for many analyses aiming to find a centre of a series of events. Stevenson et al. (2012) showed also that GP gave better results compared to other techniques in 52 of the 53 data sets explored for invasive species spreading analysis in Great Britain.

One of the main concern with GP is that some parameters must be provided by the users, such as the buffer zone dimension (the above mentioned  $B$  parameter). This parameter represents an area around the spreading point where the probability of finding an event is low, is easy to be understood in criminology, but less in the case of a biological invasion. Nevertheless, evidence of buffer zone in many cases of spreading of invasive species have been recorded and calculated for many organisms, both animal, plants and algae (Stevenson 2012). In these cases the  $B$  value must be evaluated case by case on the basis of the biology of the involved organism. It is important to understand how the approximation in providing this parameter (that must be chosen ad hoc, depending on the type of investigated events) can influence the precision of the results. Moreover, it must be kept in mind that the use of the method always produces a result, independently of the correctness of the field data (observations/events) and the precision of the parameters used for the analysis.

This investigation is not intended as a validation of GP, that has already been object of several contributions, but it aims, rather, to answer an important question: is it possible to calculate how the choice of the  $B$  parameter influences the result? Or in other words: is it possible to evaluate if the inevitable inaccuracy of the  $B$  parameter can anyway produce feasible results? And after how many observations/events? In this article we faced the problem by simulating the data sets with a variable error to be added to the  $B$  parameter (known a priori, since it was chosen during the simulation itself) and calculating how the growing inaccuracy of the  $B$  parameters affected the results, together with an increase in the number of observations. The idea is to find an evaluation about the lowest possible number of observations/events that can still provide a reliable geoprofile, together with the highest allowed variation (inaccuracy) of the  $B$  parameter. The question is relevant, since the number of available events is sometimes relatively low, independently of the researcher's will, and the evaluation of the  $B$  parameters not completely understood, depending on the biological properties of the organism from the point of view of propagation.

The research question is hence: how "wrong" (inaccurate) can be the parameters chosen for the analysis, in order to obtain still a reliable result? And how many cases/events do we need to obtain a reliable result?

## 2 Materials and methods

In order to generalize the results to other fields outside criminology, we will use in the next lines the terms of cases/objects/events instead of crimes; spreading centre or centre of origin instead of “anchor points” as previously used by other authors, depending on the type of data we are talking about. The model requires that all the observed events derive from a single spreading centre and that around this spreading centre there is a belt (the buffer zone), where the probability of finding an event is extremely low. The buffer zone is inserted in the model as parameter  $B$ . Another requirement is that the probability of finding an event decreases with the increase in distance from the spreading centre (Rossmo 2000). To describe this function the model requires parameters  $f$  and  $g$  that control the shape of the decay function (Rossmo 2000; Stevenson et al. 2012).  $f$  and  $g$  are typically set to 1.2 as values originally used in criminology (Stevenson 2013) and later evaluated as giving reliable results also in biological invasions (Stevenson et al. 2012). The model was shown to be much more dependent on variation of the  $B$  parameter, rather than of  $f$  and  $g$  (Le Comber et al. 2006; Raine et al. 2009; Butkovic et al. 2019) and this is the reason why we focused on  $B$ . However, for testing if the choice of the parameters influence too much the approximation of the results, we performed also an analysis on a solved case famous in criminology to check possible strong deviation from the known position of the offender’s home (see later).

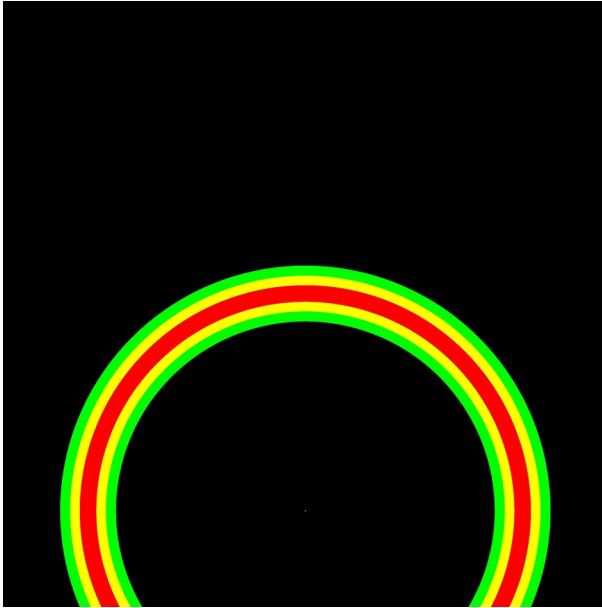
The analysis based on a single event will always produce a single ring around the event of maximum probability of finding the spreading centre, with  $B$  as the radius (Fig. 1). The map should be chosen wisely, large enough, but not too large, to avoid part of the results being drawn outside of it (as shown, as an example, in Fig. 1).

Rossmo’s formula as modified by Papini and Santosuosso (2016):

$$P(X_{i,j}) = \sum_{n=1}^{\text{Number of Events C}} W_i \left( \frac{\varphi}{d(X_{i,j}, C_n)^f} + \frac{(1 - \varphi) \cdot (B^{f-g})}{(2B - d(X_{i,j}, C_n))^g} \right)$$

where  $B$  is the radius of the buffer zone,  $i$  and  $j$  are variable indexes varying from one to  $n$  (numbers of events) and representing the coordinates on the map. The parameters  $f$  and  $g$  control the shape of the distance-decay function on either side of the buffer zone radius, that is how fast the probability of finding an event increases or, respectively, decreases, moving away from the radius.  $f$  and  $g$  hence represent the reduced probability of dispersal within the buffer zone and the fact that dispersal probability declines with distance and their user provided values are relevant mainly very close to circumference formed by the radius.

$W_i$  is the weight associated to an event (possibly corresponding to the population dimension in biological invasions, to the number of infections occurring at the same address or the number of crimes occurred in given place): it is one if all the events have the same weight;  $f$  and  $g$  are parameters that control the shape of the distance-decay function on either side of the buffer zone radius as better explained by Rossmo (1995) and Le Comber et al. (2006).



**Fig. 1** Red zone reconstruction based on a single event and a given parameter  $B$ . The red circle (annulus, more exactly) is that where the probability of finding the spreading centre is 95%. The yellow zone is that with 90% of probability, and it would comprise also the red zone; while the green annulus is that with 85% of probability and contains also the yellow and the red zone. Part of the circle of the maximum probability (in red) of centre localization falls outside the map

$D(X_{i,j}, C_n)$  is the distance between  $x_{i,j}$  (point of spread) and  $C_n$  (coordinates of the observed event), that may be calculated through several different metrics (Table 1: 2-dimensions case). The most used distance for criminology (with events most frequently occurring in urban environment), is the Manhattan distance that takes into account the geometry of the streets network calculating a 90 degrees angle in the distance, while for investigating the pattern of distribution of an invasive species, the Euclidean distance is normally preferred. More recently even a distance taking into account the google maps routing system calculation has been proposed as a way to provide more exact models for travelling in urban areas (Stamato et al. 2021). As a matter of fact, the model of distribution of crimes or other events around the centre of origin, is approximated to a circular distribution that is a simplification of reality. The use of the Manhattan distance offers a partial correction in urban areas.

**Table 1** Explanation about the different distances between two points of coordinates, respectively,  $X_{ij} = (x_i, y_j)$  and  $C_n = (x_n, y_n)$

Manhattan distance	$d(X_{i,j}, C_n) =  x_i - x_n  +  y_j - y_n $
Euclidean distance	$d(X_{i,j}, C_n) = \sqrt{(x_i - x_n)^2 + (y_j - y_n)^2}$

The model postulates that the “object” coming out from the origin and reaching its final destination in the recorded event location will reduce its probability of being in a given point on the map when the distance from the origin increases. Correspondingly, the production of one of the event will have a “cost” (from energy or economic point of view) proportional to the extent of its distance from the origin. For crimes committed by an offender, it would simply mean that the farther the offender travels, the more he/it will spend in terms of energy/money. If the distance exceeds a given value, the probability of finding an event tends asymptotically to zero.

The two components of Rossmo’s formula can be then considered as a Pareto’s distribution, that is the best solution/tradeoff or group of solutions that tend to nearing as much as possible the maximization of two (in this case) or more different parameters, with the name of the distribution after the mathematician Vilfredo Pareto (Carapezza et al., 2013). Here the model tends to minimize on the one hand the distance from the origin and, on the other hand, the probability of finding an event very close to the origin (the offender’s home, in criminological terms). A Pareto optimality represents the solution(s) to a problem in which more variables need to be optimized and the optimization of one of them affect the optimization of the other(s). For instance, in biology an organism must perform several conflicting tasks trying to maximize the general fitness (Steuer 1986; Shoval et al. 2012). A simple example for an organism is that of maximizing visibility to enhance probability of mating but trying not to be too visible to avoid predation.

## 2.1 Simulations

An initial algorithm generates a group of data sets consisting in a distribution of points  $(x, y)$  on the map, representing the events in a number varying from zero to  $n$ , with  $n=24$  in our simulations. Then we built the GP map for 2, 3, 4, ..., 24 events (24 is the highest number of events considered in the analysis) in order to see how the shape of the red zone (the area with 95% probability of presence of the spreading centre) changes, by increasing the number of simulated points inserted in the analysis. The minimal number of points to be used in a GP analysis should be five after Rossmo (2000), but we started from two to show how the shape of the red zone changes also starting from extreme situations.

To show the results of the analysis we draw on the map: (a) the red area, that is that with 95% probability of finding the spreading centre and its dimension in number of pixels; (b) the red + yellow area, that is that with 90% probability of finding the spreading centre and its dimension in pixels (it contains the red zone); (c) the red + yellow + green areas area, that is that with 85% probability of finding the spreading centre and its dimension in pixels (it contains the red and yellow zones).

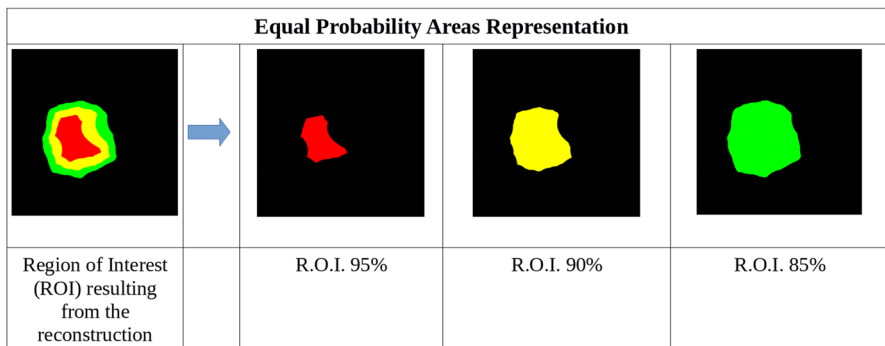
Then we calculated the distance of the known spreading point from the barycentre of the (red) area with 95% probability; the distance to the barycentre of the (yellow) area with 90% probability and, finally, the distance to the barycentre of the (green) area with 85% probability.

An example of the results of the different zones of probability of finding the centre of origin of the events is shown in Fig. 2: the yellow zone (90% probability of containing the spreading centre) is a subset of the green zone (85% probability), while the red zone (95% probability) is a subset of the yellow zone. We executed 57 randomized simulations of the data set. For each of the 57 simulations we produced 23 data sets, each containing 2, 3, 4, ..., 24 items (we omitted the case of a set composed by a single event) corresponding to the number of events (represented as points on the map) considered in each simulations, for a total of 1311 data sets and for each data set we calculated a geoprofiling with six different values of  $B$ , chosen as explained below for a total of 7866 GP reconstructions, each one taking about one minute with our data set. The relatively low number of simulations is due to limitations in computational power. We omitted the green zone in the results, limiting the analysis to the two more stringent conditions.

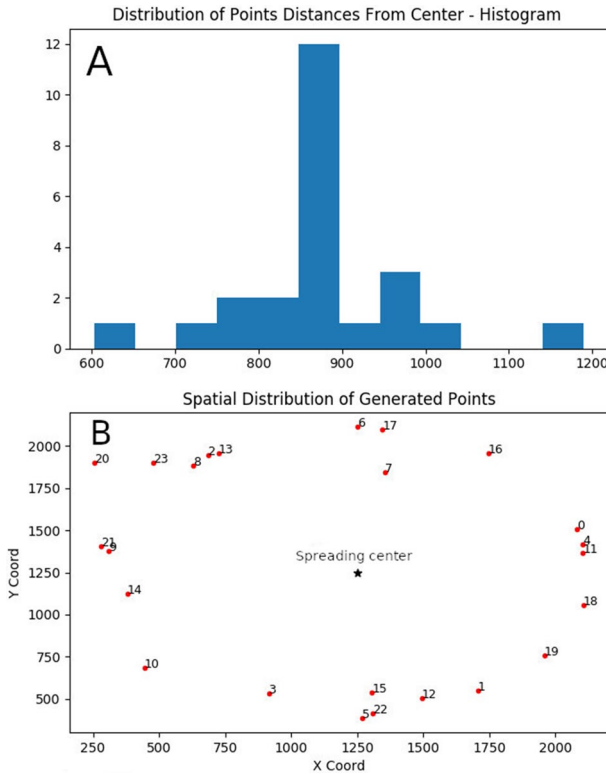
### 2.2 Data simulation

We developed a procedure generating data sets on which the probability map reconstruction should be carried out on the basis of the following criteria: (1) the radial distribution of data should follow the Pareto's envelope (general group of solutions) whose maximum has a distance of  $B$  from the spreading centre; (2) the angular distribution of points should be uniform. To do that we used Python 2.7.14 (programming language) and the libraries Numpy 0.19.0 and Scikit-learn (Pedregosa et al. 2011) for the randomization routine in order to obtain 15 data sets. We used Matplotlib 2.2.2 (Hunter 2007) for the graphical outputs. The result was the generation of fifteen linear sequences of points with the required distributions. An example of distribution of the points is shown in Fig. 3. In Fig. 3A the distribution of the points in terms of distance from the centre and in Fig. 3B the distribution of the points on the map. The procedure, written in Python, was named `my_pareto_1.py` (see below).

The same procedure deals with the interpretation of the generated data as polar coordinates and their conversion into cartesian coordinates with the well known trigonometric formula. The obtained distribution is centred in 0 (0, 0), while the



**Fig. 2** Representation on the map of the regions containing the points with normalized probability in the intervals (0.95, 1] (red zone); (0.9, 1] (yellow zone); (0.85, 1] (green zone)



**Fig. 3** **A:** radial distribution of the simulated events: 0 is the position of the centre; the  $B$  distance falls between 800 and 900 in the simulation, while the probability outside  $B$  tends to fall to zero towards right (increasing distance). **B:** spatial distribution of the generated simulated events

algorithm GPF v3.0.2, described in Papini et al. (2013) and Santosuosso and Papini (2018), works on data projected on a map, whose origin is the upper left point of the map. For this reason, we translated the data in order to have a coincidence between the centre of the map and the spreading centre. The procedure also adapts the data to the dimension of the map with the criteria of optimizing the occupation of the map, maintaining a minimal distance of the points from the map edges corresponding to 5% of the highest values (of x and y coordinates).

The procedure will generate a sequence of points on the map with a Pareto distribution with randomized polar coordinates multiplying the radius in relation to the map dimension in order to occupy it (a sort of normalization of the data on the map).

### 2.3 Parameter modifications

One critical aspect of GP for the reconstruction of the areas with highest probability of finding the spreading origin is the use of a  $B$  parameter that can be as



close as possible to the real  $B$  value ( $B_{\text{exact}}$ ). We left  $f$  and  $g$  parameters at given values, corresponding to those normally used in biological invasions and criminology, that is 1.2 (Stevenson et al. 2012).

In our case we know  $B_{\text{exact}}$ , since we provided it during the procedure of generation of the events (represented as points on the map). For this reason, we executed a series of reconstructions based on the same data set, with  $B$  values corresponding to six values chosen as 90%, 80% and 60% of approximation of  $B_{\text{exact}}$  (that is real value of  $B_{\text{exact}}$ ,  $B_{\text{exact}} \pm 10\%$ ,  $B_{\text{exact}} \pm 20\%$  and  $B_{\text{exact}} \pm 40\%$ ) in order to verify how the use of incorrect values of  $B$  affects the identification of the spread centre and in what amount. For this purpose it is needed to investigate the shape and dimension of the red zone. The total number of reconstructions necessary for the analysis corresponds hence to the number of simulated data sets\*number of  $B$  values tested =  $1311 \times 6 = 7866$  reconstructions of a geographic profiling in this article. Each GP reconstruction requires five minutes of calculation in average.

The choice of the above-mentioned  $B$  values variations (simulated error) allows testing of  $B$  values that are not too far from the real value, since no objective criterion is currently known to define a range of “feasible”  $B$  values only on the base of the events distribution. In the real world, the  $B$  value should be evaluated case by case, analysing the type of event and cause (for instance, the biology and propagation mode of an invading organism).

Currently the choice of  $B$  values is justified either on the basis of a priori knowledge of the spreading mechanism, the events and the cause (spreading centre) or on the basis of previous analogous studies. For this reason, the choice of  $B$  appears to be crucial and can affect the feasibility of the GP reconstruction. In order to verify the effect of  $B$  variation on our replicates of simulated sets of data, we increased and decreased progressively the dimension of  $B$ .  $B$  is expressed here in pixels (relative to the chosen map), and the variation is shown in Table 2.

GPF v3.0.2 and the other programs (particularly `my_pareto_1.py`, also provided together with the supplementary material file) here used, were written by the authors (Santosuosso and Papini 2018) and are released under GPL license and available on Bitbucket ([https://bitbucket.org/ugosnt/al\\_and\\_ugo/](https://bitbucket.org/ugosnt/al_and_ugo/)). Supplementary material is available at [http://www.caryologia.unifi.it/supplementarymaterial\\_geoprof.pdf](http://www.caryologia.unifi.it/supplementarymaterial_geoprof.pdf).

**Table 2**  $B$  values (expressed in pixels) by modification of the “real” a priori value (100%). The so created artificial error ranged from 60 to 120% of the “real” value

$B$	Percentage (%)
517	60
690	80
776	90
862	100
948	110
1034	120

### 3 Results and discussion

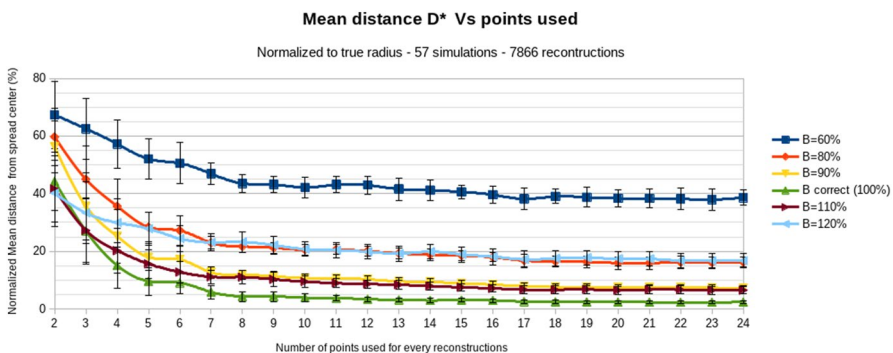
The results obtained by executing the GP analysis on the simulated data sets with the different  $B$  values shown in Table 2 produced the following results: if the number of cases/events increases with time, as it happens normally by waiting (think about the examples of criminal events or invasion sites of an invasive species), the average distance of the barycentre of the (95%) red zone from the real spreading centre tends to decrease asymptotically (Fig. 4), as expected if the spreading model is correct. In Fig. 2S (supplementary material) the same calculation is represented, but taking into consideration the barycentre of the red + the yellow zone, that is the area containing the 90% of probability to find the spreading centre (containing the red zone).

After seven points (cases/events), the average distance of the barycentre from the spreading centre is less than 10% of the  $B$  value chosen for the reconstruction (Fig. 4).

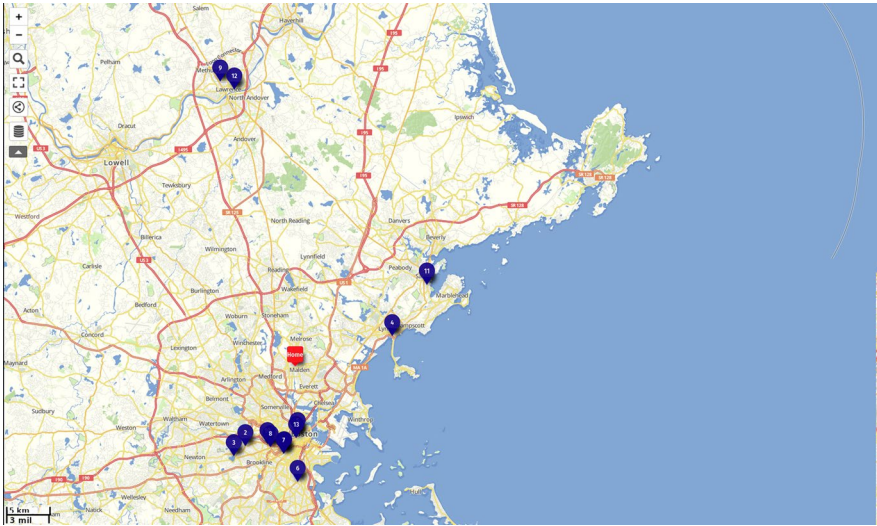
By increasing or decreasing the  $B$  value used for the reconstruction from the real  $B$ , the curves behave in a similar way. In other words: if we use  $B$  values that are  $\pm 20\%$  of the real  $B$  value ( $B = 120\%$  of the real  $B$  or  $B = 80\%$  of the real  $B$ ), the average curves overlap with an error decreasing proportionally to the number of points (cases/events), see Fig. 4. These results confirm that increasing the number of events we are allowed to reduce the precision of the  $B$  value.

The concern about the  $B$  value (related to the buffer zone) was well clear since the beginning of the use of the geographic profiling method, with most authors declaring the need for an empirical evaluation of this parameter (Rossmo, 2000; Beaugard et al. 2005; Kent et al. 2006; Stevenson et al. 2012; Papini et al. 2013). These results provide evidence about how the error in this estimation may influence the results of the following geoprofiling analysis, and what an approximation we can use without significantly affecting the results.

The total number of simulations is relatively low (1311 simulated data sets) with respect to more recent analyses, and this limitations to our results is due to computational requirements for the calculations of the geographic profiling for each data



**Fig. 4** Comparison of the distances of the barycentre of the 95% probability zone (red zone) to the real centre (user provided for the simulation) with respect to the  $B$  value variation (expressed as different curves with different colours) and the increase in number of events (crimes, for instance)



**Fig. 5** Albert de salvo case—geolocalization of the crimes (blue dots) and of the offender's home (red square)

set (with 6 different values of  $B$ ). Lerche and Mudford (2005) suggested to evaluate the mean and standard error of an output of interest. The process should be repeated until one determines that the change in the mean value with the increase in number of runs is less than a pre-specified degree of accuracy. Marriott (1979) also considered the estimation of the error a critical evaluation to assess the number of replicates to obtain a reliable result. We observed the convergence of the results already with 15 simulations each with 23 data sets of events, and hence, we consider our results reliable for the parameters here used. A higher number of replicates may lead to a better generalization of our results.

#### 4 Application of the method to a real case

To check the correctness of the previously tested algorithm with simulated data, we applied the method to a real case, that is the crimes committed by the so called Boston strangler (Albert De Salvo, [https://en.wikipedia.org/wiki/Albert\\_DeSalvo](https://en.wikipedia.org/wiki/Albert_DeSalvo)) of whom the identity and home address was known after the trial, together with the locations of the crimes (Kanchan et al., 2015), listed in Table 1S and visualized on a map in Fig. 5 and with a zoomable view in [https://umap.openstreetmap.fr/it/map/de-salvo\\_237784](https://umap.openstreetmap.fr/it/map/de-salvo_237784). We considered  $B$  as a median value of the distance calculated in pixels on the map between the crimes location and the offender's home. A list of the cases is reported in Table 1S (supplementary material), while Table 2S (supplementary material) shows the translation of the position of the crimes to the  $(x,y)$

positions on our map. Table 2S indicates also the distance, in pixels and Km, of the crimes from the offender's home.

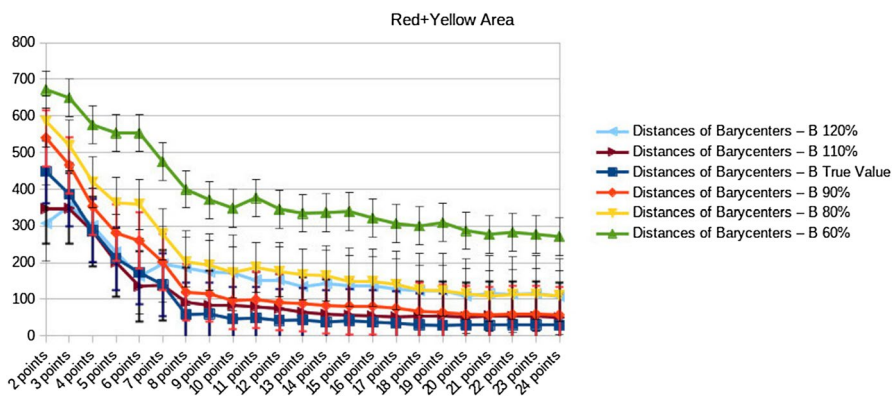
We executed a sequence of reconstructions during the time with several values of  $B$  (expressed in pixels) within a variation of 15% lower than the real value and 29% higher than the real value.

The results are reported in Fig. 6, expressing  $B$  as a percentage of the real  $B$  value. It corresponds to the analysis done with the simulated data.

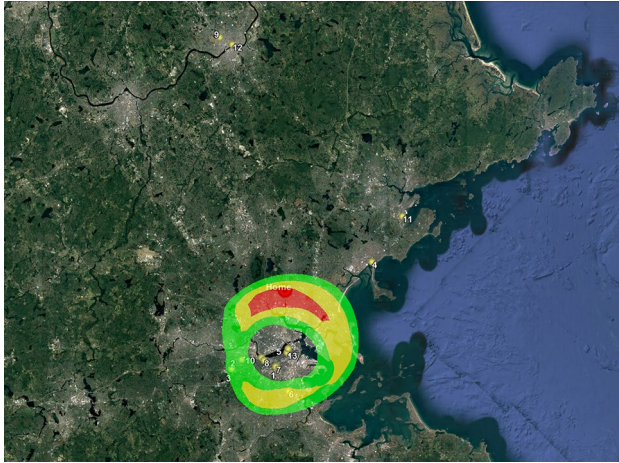
Figure 6 reports the  $B$  values used for the sequence of reconstructions, normalized with respect to the supposedly real  $B$  value. The more that the used  $B$  value gets closer to the real  $B$  value, the more the distance between the red zone and the spread centre decreases.

Passing the real  $B$  value, the distance of the barycentre of the red zone from the offender's home tends to rapidly increase again. Analogous results can be obtained for the dimension of the 90% zone. Hence, values close enough to the real  $B$  value or lower are able to approximate well the results. Increasing the number of cases, the distance between the curves tend to reduce and to overlap with the real value curve (Fig. 6). It means that the approximation of  $B$  can be compensated by increasing the number of events. We see also that passing 6–7 cases, the distance between the real values curves and  $\pm 7\%$  appear to be low (Fig. 6).

This distance does not tend to zero as in the simulations, since the cases are not spread homogeneously around the spread centre, but are rather grouped in three different clusters. The results of the geoprofiling are shown in Fig. 7. The calculated red zone included the offender's home (as from the trial). This result confirms that the choice of the  $f$  and  $g$  parameters (that are empirically chosen) did not affect too much the results of the analysis that is always to be considered as an approximation of the offender's position.



**Fig. 6** Effect of the variation of  $B$  expressed as a percentage of the real  $B$  value (from 85 to 129%, each chosen value originating a curve with different colour with respect to the others). On the  $X$  axis the number of cases and in  $Y$  the distance in pixel between the position of the spreading centre (offender's home) and the barycentre of the red zone pixels. The dotted lines of the same colour as the continuous lines are the corresponding logarithmic curves



**Fig. 7** Result of the geoprofiling analysis of the cases shown in Fig. 6. The red area is that where the probability of finding the offender's home is 95%. It includes the offender's home (after the trial result), indicated in the map

## 5 Conclusions

The GP method is sensitive to variations of the  $B$  parameter. Hence, it is necessary to find criteria allowing the use of a  $B$  value, with an approximation of 5–20% in order to obtain feasible results with a number of observations lower than ten.

Our results give an answer to the question through data simulation and show that it is possible to ensure feasible results with geographic profiling analysis also with an amount of inaccuracy of the  $B$  parameter choice and how much this inaccuracy can be.

If the chosen  $B$  parameter is not more than 5–20% different from the real one, the method converges quite quickly producing feasible results by a low increase in the number of needed observations. Our analysis shows that with 6/8 events/cases, the distance between the barycentre of the red zone and the real spreading centre is lower than the standard deviation.

In other words, if we have a sequence of events/cases in the time, with only one cause, when we reach eight cases/events there are no significant variations in the identification of the spread centre for an error in the estimated  $B$  value up to 20%.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10109-022-00379-5>.

**Acknowledgements** The article is dedicated to Steven C. Le Comber, who sadly passed suddenly away on 14 September 2019. We co-authored an article with Steven and we had the pleasure to meet him personally at a conference in Pisa (the Internet Festival of Pisa) where we had a nice dinner together. We

realized that, apart his well renowned scientific skills, he was an exceptional person with many interests in life. Ciao Steven.

**Funding** Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement.


**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Baker E, Jeger MJ, Mumford JD, Brown N (2019) Enhancing plant biosecurity with citizen science monitoring: comparing methodologies using reports of acute oak decline. *J Geogr Syst* 21:111–131
- Beauregard E, Proulx J, Rossmo K (2005) Spatial patterns of sex offenders: theoretical, empirical, and practical issues. *Aggress Violent Behav* 10(5):579–603
- Buscema M, Massini G, Sacco PL (2018a) The Topological weighted centroid (TWC): a topological approach to the time-space structure of epidemic and pseudo-epidemic processes. *Physica A* 492:582–627
- Buscema M, Sacco PL, Massini G, Della Torre F, Brogi M, Salonia M, Ferilli G (2018b) Unraveling the space grammar of terrorist attacks: a TWC approach. *Technol Forecast Soc Chang* 132:230–254
- Butkovic A, Mrdovic S, Uludag S, Tanovic A (2019) Geographic profiling for serial cybercrime investigation. *Digit Investig* 28:176–182
- Carapezza G, Umeton R, Costanza J, Angione C, Stracquadanio G, Papini A, Lio', P. and Nicosia, G. (2013) Efficient behavior of photosynthetic organelles via pareto optimality, identifiability and sensitivity analysis. *ACS Synthetic Biol* 2(5):2784–3288
- Cini A, Anfora G, Escudero-Colomar LA, Grassi A, Santosuosso U, Seljak G, Papini A (2014) Tracking the invasion of the alien fruit pest *Drosophila suzukii* in Europe. *J Pest Sci* 87(4):559–566
- Cini A, Santosuosso U, Papini A (2019) Uncovering the spatial pattern of invasion of the honeybee pest small hive beetle, *Aethina tumida*. *Italy Revista Brasileira De Entomologia* 63(1):12–17
- Górski M (2021) The accuracy of geographic profiling methods based on the example of burglaries in Warsaw. *Probl Forensic Sci* 125:51–65
- Hauge MV, Stevenson MD, Rossmo DK, Le Comber SC (2016) Tagging Banksy: using geographic profiling to investigate a modern art mystery. *J Spat Sci* 61(1):185–190. <https://doi.org/10.1080/14498596.2016.1138246>
- Hunter JD (2007) Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9:90–95
- Kanchan T, Krishan K, Kharoshah MA (2015) DNA analysis for mysteries buried in history. *Egypt J Forensic Sci* 5(3):73–74
- Kent J, Leitner M, Curtis A (2006) Evaluating the usefulness of functional distance measures when calibrating journey-to-crime distance decay functions. *Comput Environ Urban Syst* 30(2):181–200
- Le Comber SC, Nicholls B, Rossmo DK, Racey PA (2006) Geographic profiling and animal foraging. *J Theor Biol* 240:233–240
- Lerche I, Mudford BS (2005) How many monte carlo simulations does one need to do? *Energy Explor Exploit* 23(6):405–427
- Liu FT, Ting KM, Zhou ZH (2012) Isolation-based anomaly detection. *ACM Trans Knowl Disc Data (TKDD)* 6(1):3
- Marriott FHC (1979) Barnard's monte carlo tests: how many simulations? *Appl Stat* 28(1):75–77

- Martin RA, Rossmo DK, Hammerschlag N (2009) Hunting patterns and geographic profiling of white shark predation. *J Zool* 279:111–118
- O’Leary M (2009) The mathematics of geographic profiling. *J Invest Psychol Offender Profiling* 6:253–265
- O’Leary M (2010) Implementing a Bayesian approach to criminal geographic profiling. In: First international conference on computing for geospatial research and application, June 21–23, Washington, DC
- O’Leary M (2012) New mathematical approach to geographic profiling. National Institute of Justice, Washington, D.C.
- Papini A, Santosuosso U (2016) Snow’s case revisited: new tool in geographic profiling of epidemiology. *Braz J Infect Dis* 21(1):112–115
- Papini A, Mosti S, Santosuosso U (2013) Tracking the origin of the invading *Caulerpa* (Cauleriales, Chlorophyta) with Geographic Profiling, a criminological technique for a killer alga. *Biol Invasions* 15:1613–1621
- Papini A, Rossmo DK, Le Comber SC, Verity R, Stevenson MD, Santosuosso U (2017a) The use of jack-knifing for the evaluation of geographic profiling reliability. *Eco Inform* 38:76–81
- Papini A, Signorini MA, Foggi B, Della Giovampaola E, Ongaro L, Vivona L, Santosuosso U, Tani C, Bruschi P (2017b) History vs legend: retracing invasion and spread of *Oxalis pes-caprae* L in Europe and the Mediterranean area. *PLoS ONE* 12(12):e0190237
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. 12, 2825–2830
- Quick M (2019) Multiscale spatiotemporal patterns of crime: a Bayesian cross-classified multilevel modelling approach. *J Geogr Syst* 21:339–365
- Raine NE, Rossmo DK, Le Comber SC (2009) Geographic profiling applied to testing models of bumblebee foraging. *J R Soc Interface* 6:307–319
- Rossmo DK (1993) A methodological model. *Am J Crim Justice* 172:1–21
- Rossmo DK (1995) Geographic profiling: target patterns of serial murderers unpublished doctoral dissertation. Simon Fraser University, Canada
- Rossmo DK (2000) Geographic profiling. CRC Press, Boca Raton, Florida
- Santosuosso U, Papini A (2016) Methods for geographic profiling of biological invasions with multiple origin sites. *Int J Environ Sci Technol* 13(8):2037–2044
- Santosuosso U, Papini A (2018) Geo-profiling: beyond the current limits: a preliminary study of mathematical methods to improve the monitoring of invasive species. *Russian J Ecol* 49(4):362–370
- Santosuosso U, Cini A, Papini A (2020) Tracing outliers in the dataset of *Drosophila suzukii* records with the Isolation Forest method. *J Big Data* 7:14. <https://doi.org/10.1186/s40537-020-00288-8>
- Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K, Alon U (2012) Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* 336:1157–1160
- Stamato SZ, Park AJ, Eng B, Spicer V, Tsang HH, Rossmo DK, (2021) Differences in geographic profiles when using street routing versus manhattan distances in buffer zone radii calculations. 2021 IEEE international conference on intelligence and security informatics (ISI), pp 1–6. <https://doi.org/10.1109/ISI53945.2021.9624736>
- Steuer RE (1986) Multiple criteria optimization: theory, computation, and application. Wiley, New York
- Stevenson MD (2013) Geographic profiling in biology. PhD thesis Queen Mary University of London, London
- Stevenson MD, Rossmo DK, Knell RJ, Le Comber SC (2012) Geographic profiling as a novel spatial tool for targeting the control of invasive species. *Ecography* 35:1–12
- Suzuki-Ohno Y, Inoue MN, Ohno K (2010) Applying geographic profiling used in the field of criminology for predicting the nest locations of bumble bees. *J Theor Biol* 265:211–217
- Verity R, Stevenson MD, Rossmo DK, Nichols RA, Le Comber SC (2014) Spatial targeting of infectious disease control: identifying multiple unknown sources. *Methods Ecol Evol* 5(7):647–655

## Authors and Affiliations

Ugo Santosuosso<sup>1</sup> · Alessio Papini<sup>2</sup> 

Ugo Santosuosso  
ugo@unifi.it

<sup>1</sup> Department of Clinical and Experimental Medicine, University of Florence, Largo Brambilla 3, 50134 Florence, Italy

<sup>2</sup> Department of Biology, University of Florence, Via Micheli 3, 50121 Florence, Italy