Geo-profiling: beyond the current limits.
A preliminary study of mathematical methods to improve the monitoring of invasive species

Ugo Santosuosso[1] and Alessio Papini*[,2]

[1]Department of Clinical and experimental Medicine, University of Florence, Largo Brambilla, 3 – 50134 Firenze, Italy

[2]Department of Biology, University of Florence, Via Micheli 3, 50121 Firenze, Italy

*Corresponding author alpapini@unifi.it

Abstract

The Geographic Profiling (GP) is a data analysis tool that has great potential. Presently, it is used only minimally, and is almost always used "as it is", independently on other analysis or data processing methods. GP was initially created as a forensic tool, to find the origin of a series of events (crimes) done by a single actor. However, using this method in integration with others, it is possible to enlarge the opportunities of geographical data analysis. The promising results of this method in integration with others, even if some of them are quite well known methods since many years - and thus well tested - show a number of further possible applications. Here we treat data clustering and partitioning with Kmeans and Dbscan methods; space partitioning (Voronoi tessellation) and a method to assign weights to the events constituting the data set.

The software used in this review was written in Python, was released under GPL license and is available on Bitbucket (https://bitbucket.org/ugosnt/al_and_ugo/).

Introduction

In the last few years, the problem of invasive species has become increasingly

relevant and is also felt as a result of globalization of the exchange of people and goods (Meyerson and Mooney 2007).

Species endemic of other continents have begun to appear in Europe and North America, sometimes with harmful or unpredictable effects on native fauna and flora, becoming a major threat of extinction on indigenous species (Cini et al. 2014; Papini et al. 2013; Sansosuosso and Papini, 2016; Vitousek *et al*. 1996; Wilcover *et al*. 1998), also altering the abiotic environment and spreading pathogens in the territory (Strayer *et al*. 2006; Ricciardi & Cohen 2007; Stevenson *et al.* 2012). Human intervention is not a negligible factor in this "migration." These types of organisms are defined "invasive species", especially if their vital and reproductive success allows them a fast spread. We intend here invasive species as a biogeographical concept as proposed by Colautti and Macisaac (2004).

Problems that arise are many, including the loss of biodiversity in some regions and damage caused to agriculture (Paini et al. 2016; Pimentel et al. 2005) or the presence of new predators that do not have competitors in the regions where they settle (Gagliardo et al. 2016).

To reduce the damage, besides various methods of reducing the spread of invasive species, recently various methods have been applied to discover the location of the first invasion site and monitoring the progress in the territory of these species (Papini et al. 2013). These methods are often derived from other fields of science.

Probabilistic Computed Geoprofiling (or, from now on, simply “Geoprofiling” [GP]) is one of these. Initially it was developed by Rossmo (1993; 2000) to analyze geographic data of interconnected criminal events with the purpouse to identify the area with the maximun probability where the “subject perpetrator” of these acts lived or was based. Later, Geoprofiling was applied to identify the points of origin of a series of events, always interconnected but not necessarily due to criminal activities, such as the spreading of invasive plant and animal species on a territory (Cini et al., 2015 ; Papini et al., 2013; Stevenson et al., 2012) and the place of origin of an epidemic event (Papini and Santosuosso, in press).

The computational expression of this probability is given at point P_{ij} by the following

formula:

$$P_{ij} = k * \sum_{n=1}^{c} \left[\frac{\phi}{\left(\sqrt{(x_i - x_n)^2 + (y_i - y_n)^2}\right)^f} + \frac{(1-\phi) * B^{(f-g)}}{(2B - \sqrt{(x_i - x_n)^2 + (y_i - y_n)^2}\ )^g} \right]$$

where:

$$\phi = \begin{cases} 1, & \text{if distance} > B \\ 0, & \text{otherwise} \end{cases}$$

Rossmo's Formula (1)

Limitations of the GP method

- It is a probabilistic method, while the implementation of the method is deterministic: This means that you always have a result, but it is not sure that this result is significant. For this reason, we studied a validation of the method results using resampling techniques So producing confidence limits.
- It's a retrospective analysis. At the moment of the writing of this article, there are no indications about the quality of the fit of the red zone (that with highest probability) when the number of the examined cases increases.It does not consider the temporal sequence of events. At the moment of the writing of this article, no investigation has been done to know the behavior of GPs in identifying the center of an ongoing phenomenon.
- Data always are on a 2 Dimensional plane.
- The data must be represented on a discrete (pixelated) map , after a shift from the real geographic coordinates. A greater detail involves a larger map and a longer processing time. This time increases in proportion to the number of observations and the area (in pixels) of the map and the square of the linear size of the map.
- GP considers only the behavior in the space of a single person who commits criminal acts in a serial way (at least in the original version).
- The method itself is not predictive: it does not allow to make predictions on the progress of the event.
- The accuracy and the validity of the model is closely correlated to the number of

points, namely the “criminals sites”, identified and available for analysis. The greater the number of points is used for the analysis, more accurate and reliable will be the the analysis outcome. The method allows you to find a "red zone" even with only 2 cases, but the reliability of this result is, at least, weak.

- It is not able to distinguish between several "agents" who are committing separate series of crimes in the same geographical area if they have similar modus operandi. In this case using “classical” GP implementation yields a meaningless result.
- The territory where the events take place has to be considered for the most part "homogenous" and free of barriers, natural o artificial, like lakes and rivers, which may affect the agent's behavior. This is not the case for many natural territories.
- There should be no preferential ways (main roads, railways and similar) that may lead to one direction instead of another as sometimes happens with invasive species (Hulme, 2009).
- Every single event has the value of “1”: i.e. you can not attribute weight to the individual case on the basis of criteria such as the extent of the event or other factors. For example: in the case of a viral infection in a city area it is attributed the same value to an observation, independently if in the same building there was a single case of infection or there have been several.

Despite all these limits, the GP formula works well also in cases where it has not to do with criminal events but rather of purely biological nature. For example GP can be used for the identification of hunting trails of white sharks (Martin et al., 2009) or to identify the place of origin of invasive species (Cini et al., 2014; Papini et al., 2013; Stevenson et al ., 2012).

Some solutions

A method to solve at least one of the points of the list above may be to automatically partition the data by some clustering algorithm in order to highlight any groups generated by more than a single agent, so as to separate events on the basis of different origins. This approach was recently proposed by Santosuosso and Papini

(2016) , who tested it on a data set represented by the known records of presence of invasive algae (*Caulerpa taxifolia*) with a known point of origin of the invasion.

Data clustering

Clustering or cluster analysis (Robert Tryon introduced this term in 1939) is a mathematical method to automatically partition the data based on criteria set out in advance, in order to have homogeneous subsets by type of content data. This selection is made according to the data similarity criteria.

"Non-hierarchical" methods:

There are several ways to aggregate the data in a cluster, and these are classified according to the parameters with which the similarity criteria of the data is chosen. This is also based on how the method performs processing. If it is deterministic – that is, if it takes place in a number, although large but finite, of steps - or if the method is iterative, with a successive stop criterion (otherwise the data processing could go on forever). In this last case, the processing is stopped when the criteria are met, that may when the found solution can be assigned a numerical value that must be higher or lower than a given amount.

The most common stop criteria are:

- reaching the maximum number of iterations computed, that is chosen a priori, in such a way as to limit the running time. This criterion is related also to the available computational power;
- the deviation between the values reached at the N-th iteration and the previous one: if this difference is below a predetermined cut-off value, it is assumed that the optimal solution is differs from the value found by an amount lesser than that value. This method is very similar to the individuation of the wrong solutions (phylogenetic trees) individuated during the first Monte Carlo simulations at the beginning of a bayesian analysis in phylogenetic software such as MrBayes (Huelsenbeck, 2001, Huelsenbeck *et al.*, 2002).

Examples of these two different approaches are:

- the aggregation methods known as "K-means", an iterative Method - (Jain, 2010).
- "D.B.S.C.A.N." [Density-based spatial clustering of applications with noise], which is a Deterministic Method - (Esther et al., 1996).

Below is a comparative table of the specific features of the two methods.

Summary table of the differences between K-Means and DBSCAN.

K-means	DBSCAN
Uses all the points in the dataset	Points that are located away from the other, can be excluded without attributing them to a specific cluster
It allows to subdivide a set of objects into K groups on the basis of their characteristics.	Connects regions containing objects with sufficiently high density.
It requires to know “a priori” the number of clusters	Does not requires to know “a priori” the number of clusters
It does not require other parameters.	it requires to know the minimum distance to a point which is considered to be away from the other and the minimum number of neighboring points to determine the formation of a cluster.
The cluster had an approximately round shape	The cluster can have arbitrary shapes
Has centroids	Centroids are not defined". If necessary, it is possible to calculate

	"medioids" using only the points awarded to a cluster.
Each point is a valid data	It owns the “noise” notion: some data may not belong to a cluster If these points do not meet the necessary requirements (for example: points too far away from the other).
Always gives a result	It may not be able to find any cluster
Uses an iterative algorithm Starting from random centroids places,their position is recalculated at each iteration. The calculation stops when the centroids do not change their position, or the algorithm has exceeded the maximum number of iterations required.	The calculation performs exactly N^2 iterations (where N is the number of the data set points)
The found results may not be the optimal result *(especially in the case if the algorithm stops for exceeding the maximum*	The found result is always optimal (on the basis of the criterion).

number of iterations)	
If the data are not naturally partitioned (structured in cluster) results can be "strange".	The results are under all circumstances "consistent"
According to the presence of centroids and attributed to the cluster of all the points, it is possible partition the data distribution area with a deterministic tiling (like Voronoi)	According to the presence of noise, it is impossible to partition the data distribution area with a deterministic tessellation

As we said earlier, the search for the starting point of an infection or an invasion with the GP technique fails when the real starting points are more than one.
In this case, since the method finds a “focal” point anyway, we can have the following results:

- A “mean” central point that has no real meaning is found .
- Together with the absolute maximum, a local maximum may occur (consisting in a second “red” area in the map), which, however should have a minor probability than the main one and consequently this presence would be difficult to be interpreted.
- The method may detect one of the real points of origin, but the other(s) may be neglected.

Now let's see an example of the last case, on simulated data.

Simulation of a case when Rossmo's Formula fails:

The procedures that perform the simulations were written in Python 2.7, and library routines used for clustering are those of scikit-learn (v.0.14).
We performed a simulation in which we generated two independent clusters and tried then to reconstruct the GP center for the entire dataset and for the partitioned data set with the k-means.

Dataset generation parameters:

- `Image dimension = 512*512 pixels`
- `Num_Clusters = 2`
- `Points_X_Cluster = 20`
- `Standard_Dev = 2.0`
- `Global Standard Dev = 3.95702363988`
- `Center Mean and Standard Dev = 210.25 48.458100458`

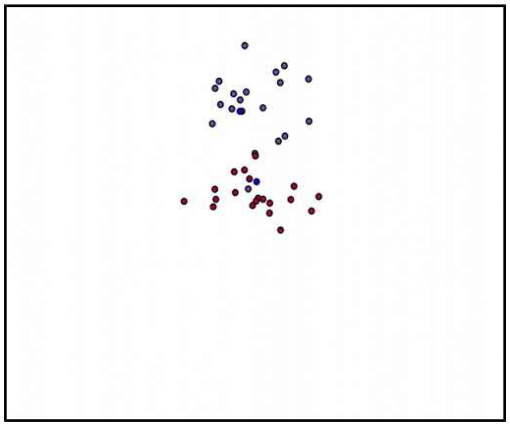

Fig. 1. Simulated dataset:

In blue the cluster centers, red points: cluster 1, purple points: cluster 2.

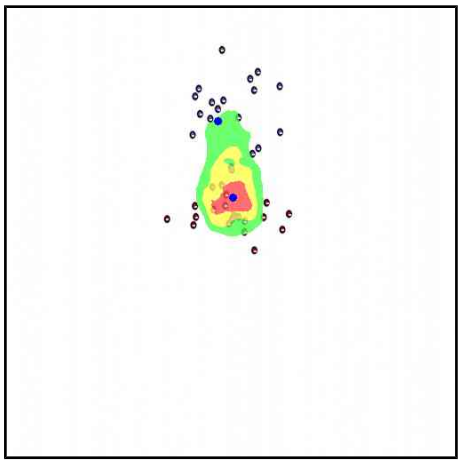

Fig. 2. GeoProfiling on the entire dataset (unpartitioned).

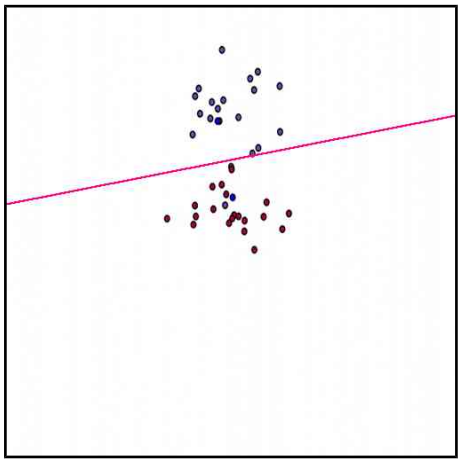

Fig. 3. Clustering with K-Means (and Voronoi tessellation). Some purple points are misclassified.

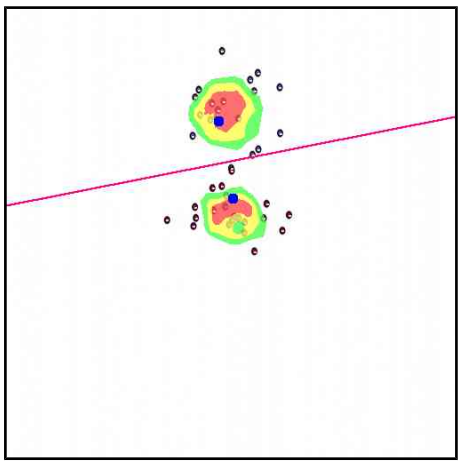

Fig. 4. GeoProfiling performed on both cluster separately.
The blue dot represents the center of the "bubble" in the original dataset.

It is evident that the application of a clustering algorithm also allows to highlight the point of origin that resulted neglected with the Geoprofiling performed on the entire data set.

Apparently, the use of kmeans and of the data partition may lead to an increase of the total area of the maximum-probability (red area) of finding the center of origin of the events (points in the map). In fact, increasing the number of "centers" with data partition with Kmeans or Dbscan, the total number of red pixels, and hence of the total red area size, remains quite low with respect to the whole image (less than 1% of the total pixels of the image), as it can be seen from the following pixels counts:

Counts

Image dimension: 512 * 512 - Area (in pixels): 262144

Map area in pixels	No partitioning	Cluster 1	Cluster 2	Cluster 1+2
Area 95% (red)	1146 (0.4372 %)	837 (0,3193%)	1474 (0,5623%)	2310 (0,8812%)
Area 90% (yellow)	3800 (0,43716%)	2229 (0,8503%)	3300 (1,2588%)	5529 (2,1091%)
Area 85% (green)	7858 (2,9976%)	3634 (1,3863%)	5130 (1,9569%)	8764 (3,3432 %)
Remaining Area	254286 (97.0024 %)	258510 (98,6137%)	257014 (98,0431%)	253380 (96,6568%)

Let us now perform the same procedure with the DBSCAN algorithm (Ester et al. 1996).

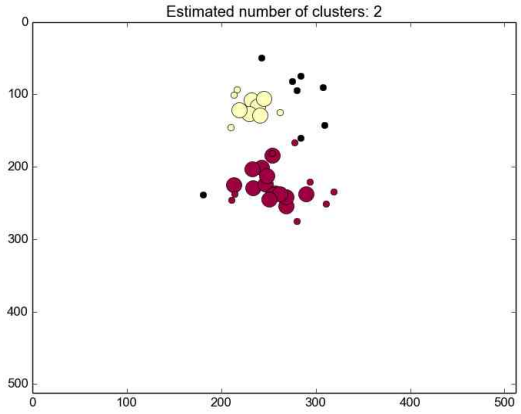

Fig. 5. Clustering with DBSCAN.

Colors - Amaranth: cluster 1, Yellow: cluster 2, black: noise (no attribution).

Processing parameters:

```
out.csv 30.0 5 euclidean auto
Estimated number of clusters: 2
Homogeneity: 0.270
Completeness: 1.000
V-measure: 0.426
Adjusted Rand Index: 0.000
Adjusted Mutual Information: 0.000
Silhouette Coefficient: 0.435
```

Repeating the procedure followed above, with the clusters found by the DBSCAN and overlapping reconstructions of the spread points with the clusters found (and related medioids) we obtain the following image:

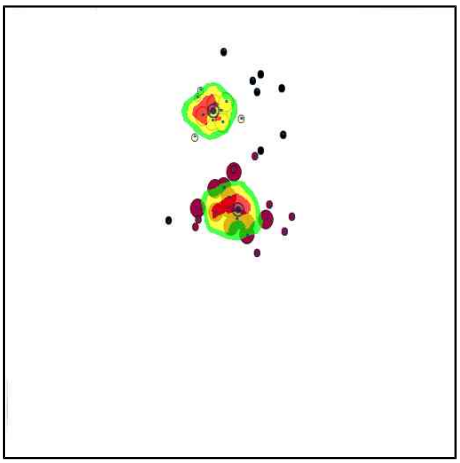

Fig. 6. GP performed on both cluster separately (through BDSCAN algorithm). The blue dot represents the center of the "bubbles" of the original data.

Regarding the extension of the area of maximum probability, the results are comparable to what was seen for the K means. The highest probability of finding the point of origin is fragmented between multiple areas, but the overall number of positive pixels remains almost constant.

Although there are attribution errors of some points, since, while belonging to a cluster, they may be attributed to another one, the red areas covered with good sensitivity the original centers of the clusters. The wrong attribution of some of the points is a typical problem with clustering analysis, in case of (partial) overlapping of the clusters. By varying the overlapping percentage of the clusters, the identification of the centers using the GP is more accurate, increasing sensitivity and with a low number of false positive red pixels, as also verified by Eckes and Orlik (1993) and Gerig et al. (2005) about the classification using clusters.

Similar results arise with data in which there are 3 or more clusters. In this last case, in relation to the different aggregation methods, we can have two different behaviors:

- the method of k-means partitions the original dataset, anyway, in the number of cluster selected. Only with statistical tools such as the Silhouette (Rousseeuw 1987; Santosuosso and Papini 2016) value of the single cluster, it is possible to evaluate if the resulting clustering is a "reasonable" result or not.
- the DBSCAN, a density-based algorithm for discovering clusters (Ester et al. 1996) can find a greater or lesser number of clusters, compared to those identified by the K-means method, depending on the parameters set for processing: Minimum and maximum number of points in the cluster and between-points relative maximun distance.

For these reasons, we do not suggest one method of clustering as “better” than another one, but it is possible to use a method that better fit to the data set, or to use a method to validate the results of the other.

Weighted Geoprofiling

In its original form (1), the Geoprofiling Probabilistic method does not take into account the possibility that a case may be “different” from the others. That provides a uniformity of probabilistic value of these cases, or, on a physical level, this is equivalent to considering only the presence or absence in a certain place of the phenomenon under observation. It does not take into account the amount or the importance of the single event itself.

Examples of phenomena that can not be represented with the model "presence / absence" may be:

- diffusion of a non-homogeneous particulate, the traces of which appear in "random" manner within a solvent
- presence of bacterial colonies of variable dimensions within a lake or swamp or a wetland,

- the number of deaths in the same street number address, due to an epidemic event.
  (Snow 1936; Papini and Santosuosso 2016)

The use of the original formula (1) can still be performed, but it leads to less precision in identifying the point of origin, because there is a loss of information. Overlooking the number of deaths in the same place, leads to a reduction in precision proportional to the number of neglected cases. Giving to each individual case a different "weight", the formula (1) is amended as follows (Papini and Santosuosso 2016):

$$P_{ij} = k * \sum_{n=1}^{c} w_n * \left[\frac{\phi}{\left(\sqrt{(x_i - x_n)^2 + (y_i - y_n)^2}\right)^f} + \frac{(1-\phi) * B^{(f-g)}}{\left(2B - \sqrt{(x_i - x_n)^2 + (y_i - y_n)^2}\right)^g} \right]$$

where:

$$\phi = \begin{cases} 1, & \text{if distance} > B \\ 0, & \text{otherwise} \end{cases}$$

w_n= weight relative to the point “n”

Modified Formula (Papini-Santosuosso 2016)

In this way, the original full information is maintained and considered in the calculation. In the specific case of Snow's cholera dataset (Papini and Santosuosso 2016) we will have the following comparative table which illustrates the size (in number of pixels) of the areas found with the various methods:

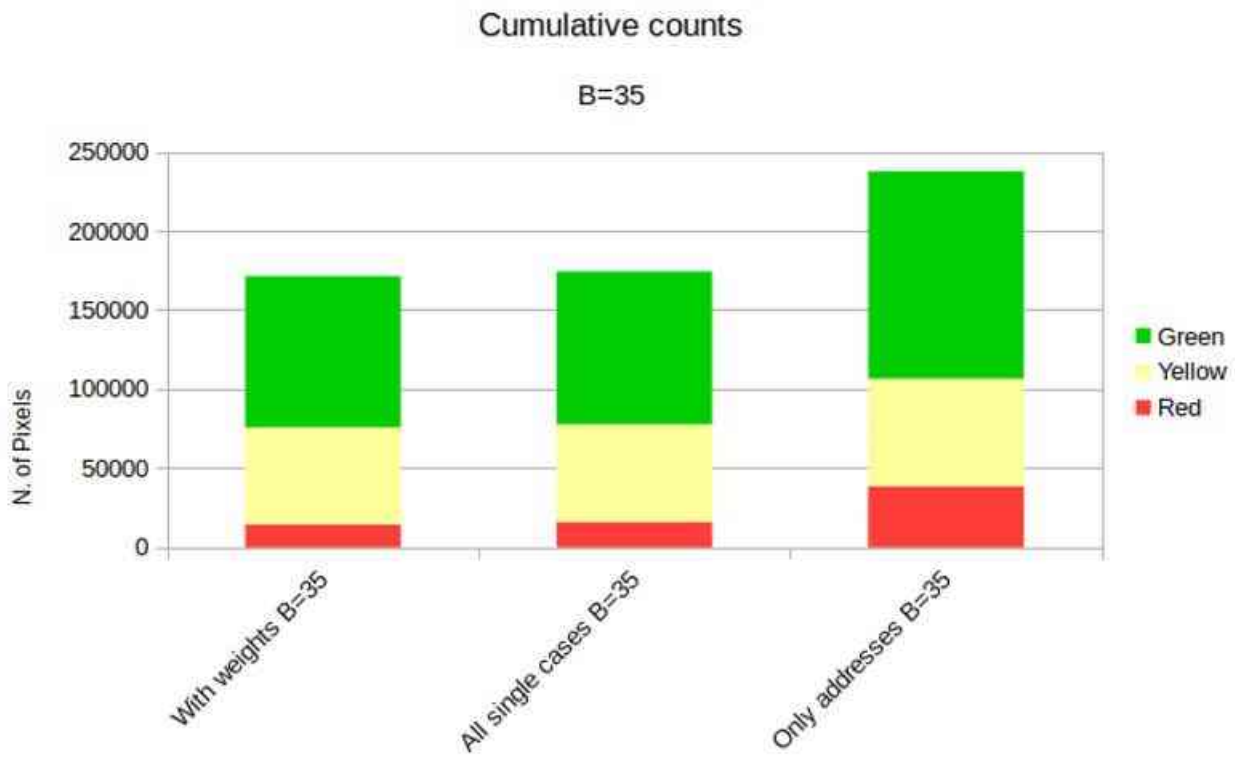

Fig. 7. Pixel counts (image of 512x512 pixels). Red pixels are those with 95% probability of finding the center of origin. Yellow pixels are those with 90% and the green zone that with 85% probability. The lowest number of red pixels, that is the lowest number of false positives, is obtained with the weighted GP analysis.

Legenda	
Mesaures	**Columns**
• **red zone (95% max prob),** • **yellow zone (90% max prob)** • **green zone (85% max prob) in pixel**	• **"With weights" is the area size obtained with the Modified Formula,** • **"All single cases"is the result obtained with the Original Formula, repeating the calcolus, considering the n case at each address as a single/different case** • **"Only addresses"is the result obtained with the Original Formula**

	considering only 1 case for each address (as in Le Comber et al. 2011)

Future developings

As already remarked above, the Geographic Profiling method has, in itself, great limits
that, nevertheless, do not affect its validity or its effectiveness. The authors are currently investigating:

- the statistical validation of the results obtained from the GeoProfiling through resampling methods.
- “Fuzzy” Voronoi (Fuzzy sensu Hüllermeier, 2005) – to be applied when making a clustering using DBSCAN: it differs from the classical Voronoi method (sensu Aurenhammer 1991) because the boundaries of the Voronoi tessellation are not perfectly defined.
- 3D Geoprofiling – in case you are in the presence of events not arranged on a surface but within a volume: the Geoprofiling method was developed to operate on a flat surface or at least approximated by a plane.
- Clustering / classification of data with hierarchical algorithms (like phylogenetic trees) and validation of the results obtained by these methods.
- Identifying the prevailing diffusion directions. This can be useful in some cases of biological invasions along preferential directions both natural (rivers) or built by man such as railways and highways.
- Identification of historical changes in the diffusion center: over the years, the scheme of spread locations on the territory may complicates very much the initial scheme.
- Diffusion coming from a point moving along a path: with the current method it is assumed that the diffusion occurs by a well defined center point and fixed in time and space, whereas this spread can happen through a linear "phenomenon" type the spread of a contaminant that is released to the environment during transport - perhaps as a result of uncontrolled leakage from

a tank on wheels or turbulent dispersion by wind.

- Reconstruction of the distribution and of the spread origin in case of considering a non homogenous territory, due to the presence of natural, artificial or anthropic obstacles.

.

Conclusions:

The Probabilistic Computed Geoprofiling is a data analysis tool that has great potential. At present, it is used only minimally, and is generally used "as it is", with rare use of other analyses or data processing methods. GP was initially created as a forensic tool, and perhaps it was rarely used in different fields. However, using this method in integration with others, may enlarge the opportunities and fields of application of geographical data analysis. The promising results of this method in integration with others, even if some of them are quite well known methods since many years, and thus well tested, showed various prospects of future applications.

Technical note

The software used in this review was written in Python, was released under GPL license and is available on Bitbucket (https://bitbucket.org/ugosnt/al_and_ugo/).

References

Aurenhammer, F. (1991). Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys*, 23(3), 345–405.

Cini A, Anfora G, Escudero-Colomar LA, Grassi A, Santosuosso U, Seljak G, Papini A. Tracking the invasion of the alien fruit pest *Drosophila suzukii* in Europe. J Pest Sci **2014**; 87(4):559-566.

Colautti RI, MacIsaac HJ (2004) A neutral terminology to define ‘invasive’ species. Diversity abd Distributions 10(2): 135-141.

Eckes T, Orlik P (1993) An error variance approach to two-mode hierarchical clustering. Journal of Classification 10(1): 51-74.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, E. Simoudis, J. Han, and U. Fayyad, Eds. AAAI Press, 226–231.

Gallardo B, Clavero M, Sánchez MI and Vilà M (2016) Global ecological impacts of invasive species in aquatic ecosystems. Global change Biology 22(1): 151-163.

Gerig G, Martin J, Kikinis R, Kübler O, Shenton M, Jolesz FA (2005) Automating segmentation of dual-echo MR head data. Segmentation: specific applications. Lecture Notes in Computer Science 511: 175-187.

Huelsenbeck JP, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.

Huelsenbeck JP, Larget B, Miller RE, Ronquist F (2002) Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. Systematic Biology 51(5): 673–688.

Hüllermeier E (2005) Fuzzy methods in machine learning and data mining: Status and prospects. Fuzzy sets and Systems, 156(3): 387-406.

Hulme PE (2009) Trade, transport and trouble: managing invasive species pathways in an era of globalization. Journal of Applied Ecology, 46(1): 10-18.

Jain, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

Le Comber SC, Rossmo DK, Hassan AN, Fuller DO, Beier JC (2011) Geographic profiling as a novel spatial tool for targeting infectious disease control. International Journal of Health Geographics 10:35.

Martin RA, Rossmo DK, Hammerschlag N (2009) Hunting patterns and geographic profiling of white shark predation. J. Zool. 279:111–118.

Meyerson LA, Mooney HA (2007) Invasive alien species in an era of globalization. Frontiers in Ecology and the Environment 5(4): 199-208.

Paini DR, Sheppard AW, Cook DC, De Barro PJ, Worner SP amd Thomas MB (2016) Global threat to agriculture from invasive species. PNAS 113(27): 7575-7579.

Papini A, Mosti S, Santosuosso U. Tracking the origin of the invading *Caulerpa* (Caulerpales, Chlorophyta) with geographic profiling, a criminological technique for a killer alga. Biol Invasions **2013**, 15(7):1613-1621.

Papini A., Santosuosso U. (in press) Snow's case revisited: new tool in geographic profiling of epidemiology. Brazilian Journal of Infectious Disease, in press.

Pimentel D, Zuniga R, Morrison D (2005) Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecological economics, 52(3): 273-288.

Rossmo DK (1993) A methodological model. Am. J. Crimin. Just. 172:1–21.

Rossmo DK (2000) Geographic profiling. CRC Press, Boca Raton, FL.

Rousseeuw, P.J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, 20, 53–65.

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.

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.

Strayer DL, Eviner VT, Jeschke JM, Pace ML (2006) Understanding the long-term effects of species invasions. Trends Ecol. Evol. 21:645–651.

Tryon, R. C. (1939). Cluster analysis. New York: McGraw-Hill.

Vitousek P, D'Antonio CM, Loope L, Westbrooks R (1996) Biological invasions as global environmental change. Amer. Sci. 84:468–478.

Wilcover DS, Rothstein D, Dubow J, Phillips A, Losos E (1998) Quantifying threats to imperilled species in the United States. Bioscience 48:607–615.