CrossMark

ORIGINAL PAPER

Methods for geographic profiling of biological invasions with multiple origin sites

U. Santosuosso[1] · A. Papini[2]

Abstract Geographic profiling is a method that proved to be useful also in order to investigate the point of origin of a biological invasion. K-means clustering and Voronoi diagrams can partition a data set of geographic positions of populations invading a defined area and are, therefore, useful in cases in which an invasion had more introduction events as points of origin. One critical point of the method is to identify the right number of clusters in which to divide the starting data set formed by groups of points on a map. The *Silhouette* method proved to be capable of identifying the best number of subsets (clusters) of the general set of observations by providing different values for different subdivisions of the set of observations in clusters. For each cluster, the corresponding Voronoi tessellation was built on the starting map. To test the method, we did a simulation of clusters of data (points) on a map and we verified whether the proposed methods worked efficiently with the simulated data set with hundred repeats and using a varying number of clusters on the same map. The used techniques revealed to be efficient in finding the highest probability area of the map that would include the starting points for each cluster. A case study consisted in a known data set, that is, the spreading pattern of *Caulerpa racemosa* var. *cylindracea* (sea grapes), that was compatible (highest probability) with an original point of introduction in southern Italy and long distance (thousands of kilometers) secondary spreads via anthropic dispersal. The proposed techniques may also be applied to other kinds of data sets of biological data distributed on a map or in general on a geometrical surface.

✉ A. Papini
  alpapini@unifi.it

1 Department of Clinical and Experimental Medicine, Largo Brambilla, 3, 50134 Florence, Italy

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

Introduction

Geographic profiling (GP) can be defined as an analytic tool aiming to identify the geometrical origin of linked events, for instance crimes by a serial killer in criminology or spreading populations of an alien species (Rossmo 2000). GP uses coordinates on a map of linked events (events produced by the same cause) to create a probability surface to superimpose on the original map to produce the so-called geoprofile (Rossmo 2000). Such geoprofiles do not provide the exact origin of the events, but rather prioritize geographic points, or suspects in the case of crimes (Rossmo 2000). The prioritization produces decreasing probability density starting from the highest to the lowest probability of finding the offender's home, or anyway the source of the points drawn on the map (Rossmo 1993, 2000).

The analysis is based on two functions: a distance-decay function (a function that assign lower values in relation to the distance from the center), so built to make the probability of an event (crime or spreading of an alien species) tending to

🦋 ✌ Springer

2038

Int. J. Environ. Sci. Technol. (2016) 13:2037–2044

drop with increasing distance from the events origin; and a buffer zone, within which the events probability increases with distance (Rossmo 2000), meaning that spreading cannot occur too close to the starting point. The distance-decay function is justified because departing from the origin involves costs (Stevenson et al. 2012), in economical (for humans) or energy (for invasive species, even if transported by humans) terms. The buffer zone can be defined as a part of the plane close to the point of origin of the investigated events, where such events occur only with very low probability. Such a zone is quite easy to understand for the necessity, by criminals, to avoid locations too close to their home. In natural biosystems, the existence and the dimension of a buffer zone should be checked case by case. Evidence of a buffer zone was observed for bumblebees foraging (Dramstad 1996; Saville et al. 1997) and invasive species spread (Stevenson et al. 2012; Papini et al. 2013; Cini et al. 2014). Stevenson et al. (2012) also studied the possible values of the buffer zone radius (if any) for a range of species, from woody trees to marine invertebrates.

GP was also recently applied on the targeting of an infectious disease (Le Comber et al. 2011), animal foraging (Le Comber et al. 2006; Raine et al. 2009), hunting patterns of sharks (Martin et al. 2009) and the prediction of multiple nest locations (Suzuki-Ohno et al. 2010) of bumblebees.

One difficult case for GP is that when a spreading pattern of an invasive species is due to more than a single starting point. In such cases, the Rossmo's formula for the GP applied on the whole data set may identify the first introduction origin, but would not help to identify the following dispersal events.

To face problems of biological invasions due to multiple points of origin, we tested analytical tools capable of dividing a data set in homogeneous clusters (a subset of the general set containing the observations), as the K-means method (defined in Jain 2010) and to divide the geographic territory with respect to this clusters into Voronoi cells (defined in Aurenhammer 1991). The subsequent application of the GP analysis on the subset of data would allow to identify the origin of each cluster of observations.

Data clustering (cluster analysis) aims to identify the natural grouping of a set of objects (Jain 2010) to discover the underlying data structure. A definition of clustering can be: given n objects, find K groups on the basis of a measure of resemblance such that the resemblance of objects within a given group is high(er), while the resemblance between objects belonging to different groups is low(er) (Jain 2010). On this basis, a cluster is a somehow subjective entity, dependent on arbitrary choices of parameters (Jain 2010). The most used algorithm to partition data sets is K-means (Jain 2010). As said above, the main problem with K-means is to identify the "right" number of clusters (Jain and Dubes 1988; Chiang and Mirkin 2010), defining real clusters as the

sets in which the observations can be divided as a result of real natural events. The natural events in the case here described are clusters of data formed by more points of origin of a biological invasion, producing one cluster for each starting point of the invasion. For this purpose, various analytical methods were proposed as a mean for the automatic recognition of the right number of clusters. As a matter of fact, this is still a field of active research, also with many possible applications. A quite intuitive approach is that of *Silhouettes* (Rousseeuw 1987), defined as a graphical method where each cluster is represented by a so-called *Silhouette*, providing a graphical representation about how a given cluster shows tightness (inside) and separation [these terms in the sense used by Rousseeuw (1987)] with respect to the other clusters. It shows which objects lie strictly within their cluster and which ones are positioned in a more uncertain position between clusters. Comparing the cluster *Silhouettes* for situations with a varying number of clusters, it is possible to see (by eye) which solution (number of clusters) shows a partition with the lowest number of objects out of place and hence an assessment of the quality of the clustering (Lleti et al. 2004; Leisch 2006).

Considering a number of points in a plane, the Voronoi diagram divides the plane with the nearest-neighbor rule: Each point is associated with the region of the plane closest to it (Aurenhammer 1991). For each cluster of observations on a plane, the Voronoi diagram shows how the plane can be partitioned between the various clusters.

Our hypothesis is that clustering, followed by GP of the single data sets, can also be used effectively in biological invasions analysis, and we provide here a series of techniques discussed theoretically and implemented with open-source computer programs that may be employed for data sets represented as points in a plane and hence by many other users in environmental research. We test the methods both on simulated data sets and on a data set containing the observations about the spreading of *C. racemosa* in the Mediterranean obtained from Piazzi et al. (2005) coded as points on a geographic map. Our study organism belongs to genus *Caulerpa* (Caulerpales, Chlorophyta) (Meinesz et al. 2001). Two species of *Caulerpa* J. V. Lamouroux, *C. taxifolia* (Vahl) C. Agardh and *C. racemosa* (Forskål) J. Agardh (sea grapes) are among those spreading more extensively in the Mediterranean. Data about the distribution of the investigated species in the Mediterranean were obtained from Piazzi et al. (2005). Papini et al. (2013) applied the GP method on *C. taxifolia*, since the origin of its spread is known, that is, an accidental release from the aquarium of Monte Carlo in 1984 (Meinesz and Hesse 1991; Meinesz et al. 2001), while *C. racemosa* var. *cylindracea* diffusion in the Mediterranean may not have occurred from a single point of origin, but rather first through some long distance (thousands of kilometers) spreading from the general starting origin (possibly via commercial ships) and

later through secondary short distance (tens/hundreds of kilometers) spreads (possibly via tourist boats). For these reasons, this data set is appropriate for testing the here-proposed methods for dividing data sets in subsets of points on a geographic map.

Materials and methods

The efficiency of the K-means method was assessed by creating an increasing number of clusters from 3 to 18. For each number of clusters, 100 random replicates of the process of clusters simulation were done. For this purpose, we wrote the Python script simularepet_1.3.py. The clusters were formed with the Python program bubble_1.1.py (production of simulated clusters). Each cluster was composed of 15 points with a variable dispersion around the centroid of the cluster. kmeans0_57.py was used in order to reconstruct the simulated clusters (given in input the right number of clusters). The resulting data were analyzed with Geoprof2_0_5csv.py and voronoi2.0.py to create a tessellation on a simulated map of 512 * 512 pixels corresponding to the space subdivision among the clusters. All the programs were controlled by the script simula_1.1.py. kmeans_sil_0_0_2.py was used to assess the right number of clusters. With this method, each cluster is graphically represented by a band called *Silhouette*, which is based on the comparison of the tightness and separation of the cluster itself, considered a representation of the quality of the clusters, in the sense that high quality corresponds to low within cluster dissimilarity if compared to the between-clusters dissimilarity (Rousseeuw 1987). The *Silhouette* shows how well the items lie within their cluster and which are not well inserted in it. The comparison is possible since the *Silhouettes* are shown together in the same plot permitting to assess the general quality of the clustering. The average width of the bands (*Silhouette*) can be used to assess the validity of a given number of clusters in which a set of data is divided. The control of the bandwidth variation can hence be used to select the relative best number of clusters (Rousseeuw 1987). The Voronoi tessellation was applied using the given number of clusters and the centroids of each cluster calculated by K-means. While K-means partitions the clusters of items dispersed on the map, the Voronoi tessellation method partitions the map on the basis of the cluster centroids (Aurenhammer 1991).

To test the method, we checked how many of the points of the simulated clusters were correctly assigned to the right clusters by K-means and, since it was done with 100 repetitions, we calculated the mean of the correct cluster assignment with standard deviation for each number of clusters, from 3 to 18.

The model for geoprofiling analysis was described by Rossmo (2000). We used a variant adapted to the use of the

Euclidean rather than the Manhattan Distance (the MD was originally chosen in criminology since it is more fitted to urban areas, particularly North American street patterns). The use of the Euclidean distance rather than the Manhattan Distance was also proposed by Le Comber et al. (2006) and Stevenson et al. (2012). The application of GP on the spreading pattern of algae, including the software by the authors, is described in detail by Papini et al. (2013). The maps were downloaded and adapted from Open Street Map http://www.openstreetmap.org. The settings here used for the GP analysis are the same as in Papini et al. (2013).

To test the geoprofiling analysis on a real data set partitioned with K-means, we used the distribution map of *C. racemosa var. cylindracea*. Data about the distribution of this seaweed were retrieved from Verlaque et al. (2004) and Piazzi et al. (2005).

The programs and two zipped folders containing software and a data set for testing GP applied on the simulated data and the *Caulerpa racemosa* data set, respectively, are available

Table 1 Efficiency of K-means varying the number of clusters

Number of simulated clusters	Mean of the percentage of matching elements in each of the clusters identified by K-means with respect to the simulated clusters (100 replicates)	Standard deviation
3	0.917	0.147
4	0.905	0.159
5	0.923	0.13
6	0.894	0.147
7	0.862	0.169
8	0.875	0.157
9	0.849	0.17
10	0.847	0.172
11	0.824	0.182
12	0.808	0.184
13	0.808	0.182
14	0.793	0.186
15	0.779	0.185
16	0.77	0.192
17	0.761	0.19
18	0.746	0.191

The 100 repetitions of K-means analyses performed on a simulation of a number of clusters from 3 to 18 (15 items for each cluster) showed that the method yielded in average 90 % of correct identification of the items belonging to a single cluster by simulating from 3 to 6 clusters, thereby slowly decreasing its accuracy. 100 simulations of 18 clusters still yielded almost 75 % in average of correct identifications. The first column of the table represents the number of clusters used for each analysis (increasing from 2 to 18). In the second column, the mean of the number of matching elements in each of the clusters identified by K-means with respect to the simulated clusters (100 replicates for each case). The success rate was over 0.8 (80 %) for 3–13 clusters, decreasing to 0.745 with 18 clusters. In the third column, the standard deviation of the mean calculated in column 2

2040

Int. J. Environ. Sci. Technol. (2016) 13:2037–2044
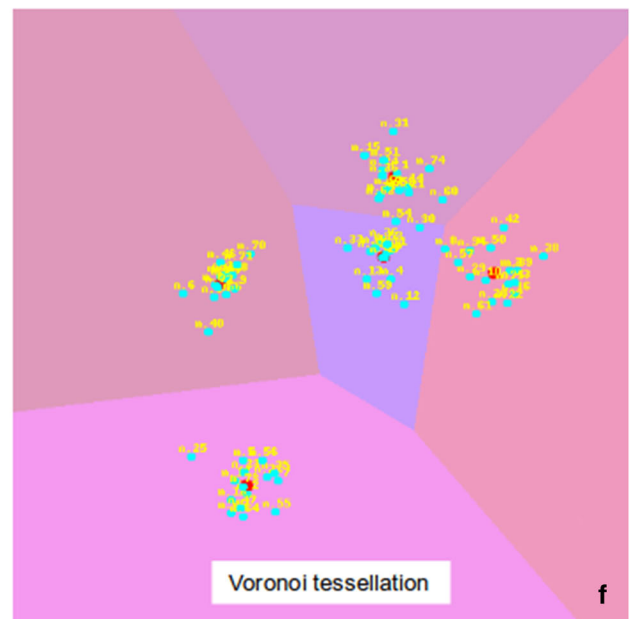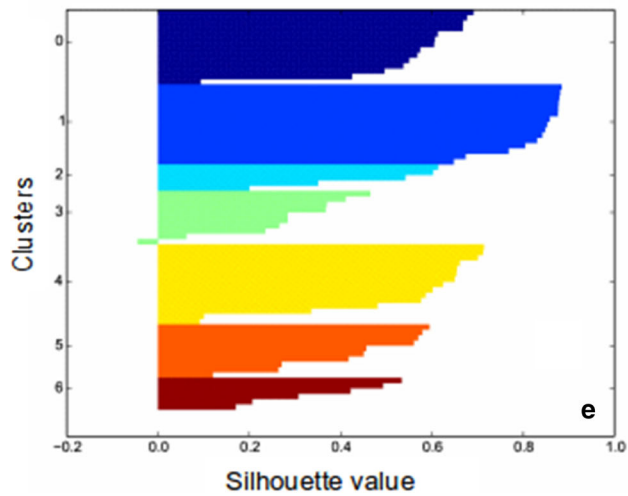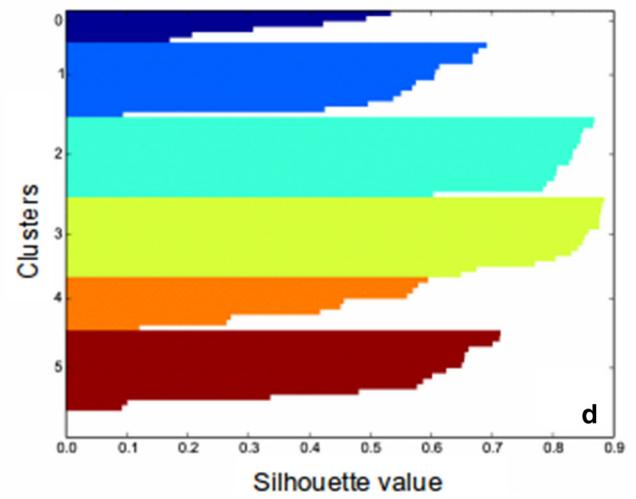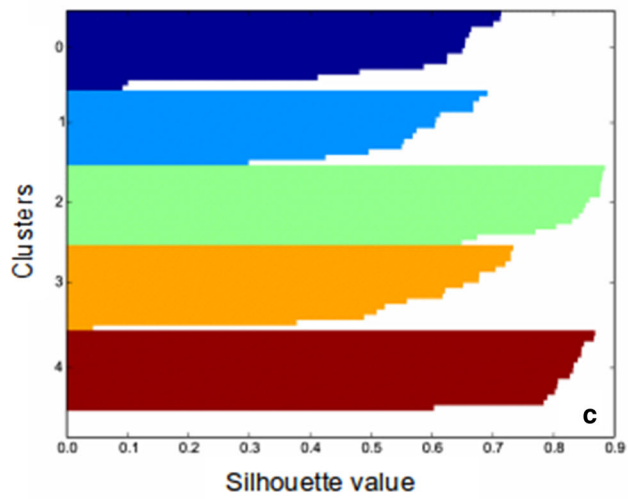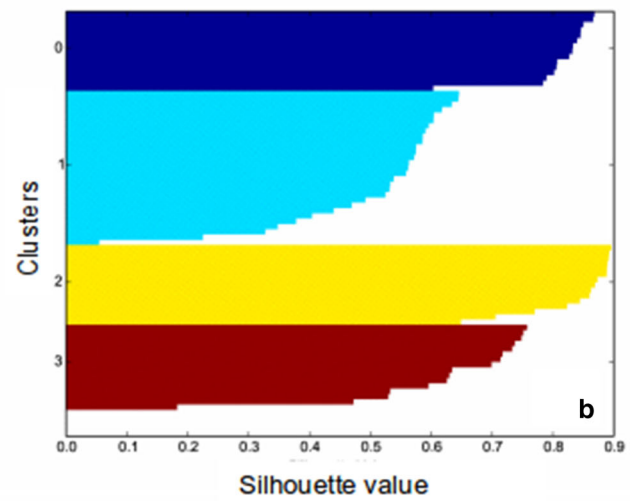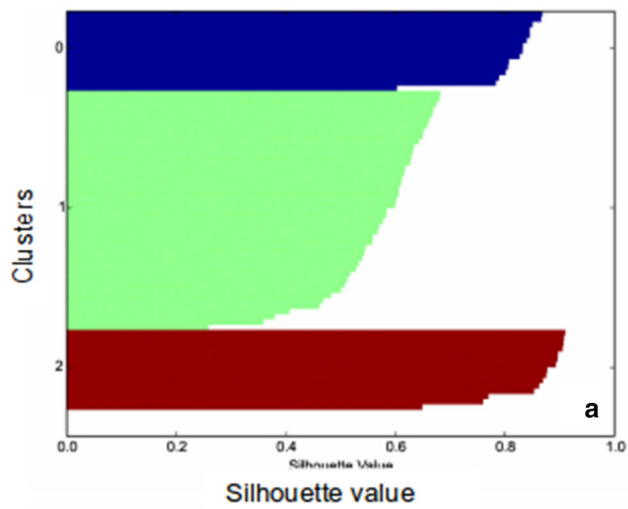
◄**Fig. 1** An example with a simulated data set composed by five subsets of data and obtained with the Python script bubble_1.1.py is shown. The *Silhouette* of each cluster is shown for: **a** 3, **b** 4, **c** 5, **d** 6 and **e** 7 clusters. The five clusters in **c** are those with the lowest variation in bandwidth. **a**–**e** the geoprofiling for each of the four clusters as divided by K-means is shown. The area with the highest priority of geoprofiling (in *red* and *yellow*) contained the simulated starting point of each cluster; **f** the simulated data set, here represented as points drawn on the map, divided into the five clusters calculated by K-means on the basis of the *Silhouette*. The space is also divided with a Voronoi tessellation

by the site www.unifi.it/caryologia/PapiniPrograms.html and were written and executed in Python 2.7.3 (http://www.python.org/) and run on a Linux Ubuntu 12.04 LTS (http://www.ubuntu.com/) Operating System, Linux kernel 2.6.32. The Python (≥2.6 version) programs need NumPy (http://www.numpy.org/), SciPy (http://www.scipy.org/), Matplotlib (http://matplotlib.org/), Scikit-learn (http://scikit-learn.org) and Python Image Library—PIL (http://www.pythonware.com/products/pil/) libraries installed, all of them open source and under GPL distribution licence.

Results and discussion

Analysis with K-means and *Silhouette*

The 100 repetitions of K-means analyses performed on the same number of simulations for each of the different situations caused by varying the number of clusters from 3 to 18 (15 items for each cluster) showed that the method yielded in average 90 % of correct identification of the items belonging to a single cluster by simulating from 3 to 6 clusters, thereby slowly decreasing its accuracy. The 100 simulations with 18 clusters still yielded almost 75 % in average of correct identifications (Table 1).

An example with a simulated data set composed by 5 subsets of data and obtained with the python script

bubble_1.1.py is shown in Fig. 1. The *Silhouette* analysis with kmeans_sil_0_0_2.py showed that the graph with the "right" number of clusters (that is, the number corresponding to the clusters produced by the simulation itself) appeared to be more regular (as bandwidth) with respect to the graphs with the wrong number of clusters (Fig. 1), where "artificial" clusters are penalized by narrow *Silhouettes* [as in the treatment by Rousseeuw (1987)]. An example of a Voronoi tessellation of the map on the basis of the five clusters calculated by K-means after the *Silhouette* analysis of Fig. 1c is shown in Fig. 1f.

Analysis with geographic profiling and Voronoi tessellation

The application of the GP formula on the data set of *C. racemosa* var. *cylindracea* showed that the general point of origin of this species in the Mediterranean was in West Sicily (Fig. 2). We applied the *Silhouette* technique to assess the best number of clusters to divide the data set with the K-means method. The best results were for 2 and 3 clusters with a preference for 3 clusters (Fig. 3). The best result is defined graphically as that showing the largest *Silhouette* widths in average (Rousseeuw 1987). The *Silhouette* index with two clusters was about 0.51 and 0.53 with 3 clusters. In Fig. 3, it is also possible to observe the effect of varying the number of clusters on the subsequent Voronoi tessellation (on the right part of the figure), on the map of the Mediterranean.
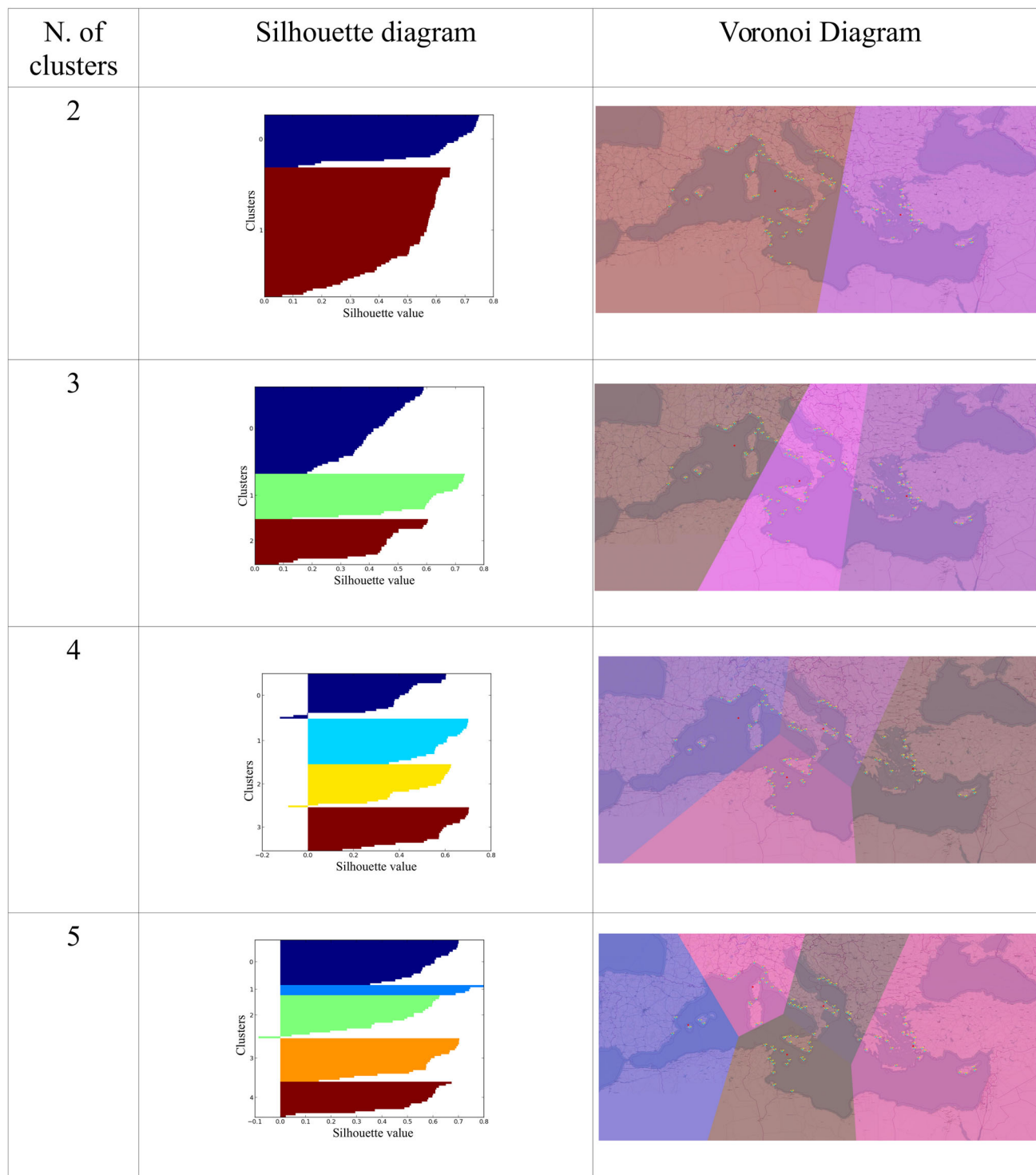
The application of K-means and Voronoi tessellation of the data set in the Mediterranean with subdivision of data into three clusters is shown in Fig. 4. For each cluster, the GP analysis was performed. We mapped the most probable origin of the spreading points for each cluster.

A map of the main commercial and touristic ports of the Mediterranean is shown in Fig. 1S (Supplementary material). Comparing this last map with the pattern of Voronoi tessellation with the sites of the presence of *C. racemosa*, as shown in Fig. 4, we could see a correspondence between

Fig. 2 The original result of a GP on the entire data set of *C. racemosa* var. *cylindracea*. The general point of origin of the invasion resulted to be in West Sicily

2042

Int. J. Environ. Sci. Technol. (2016) 13:2037–2044
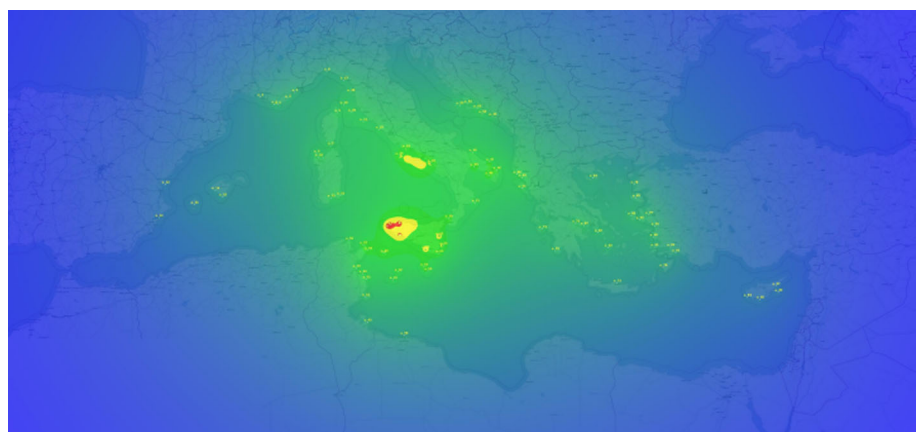
For selecting the number K of cluster used for partitioning this data set we obtained the following results:

N. of clusters	Silhouette diagram	Voronoi Diagram
2		
3		
4		
5		

Fig. 3 *Silhouette* results about the subdivision of the *C. racemosa* data set. The best number of clusters after the *Silhouette* criterion was three

Fig. 4 The application of K-means and Voronoi tessellation of the data set in the Mediterranean with subdivision of data into three clusters is shown
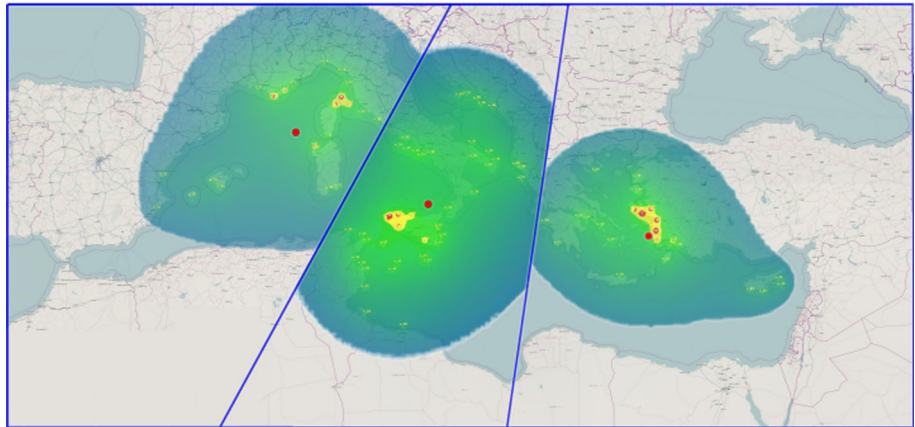
the spreading origin for each cluster identified by the GP (in red) and the presence of large commercial ports, while the further presence of populations of the invading species, largely corresponded to tourists ports.

Discussion of the analysis

The data given in Table 1 prove that K-means can correctly interpret clusters of simulated data with a very good accuracy for 3–6 clusters, while as expected, decreasing its accuracy proportionally to the increase in number of the simulated clusters. Nevertheless, the decrease in accuracy was very slow. Of course, the here-presented results are linked to the map dimension (512 * 512 pixels here) and to the amount of items for cluster (15 here). Increasing density of the points in the map would result in a more difficult clusters separation.

The *Silhouette* analysis with kmeans_sil_0_0_2.py proved to be a useful method to assess the number of clusters. This method may be used as an analysis to assess the quality of the clusters identified by K-means.

The application of the same techniques validated by the simulated data set for the identification of different clusters allows to propose a successive steps model of spreading for the "real" data set of *C. racemosa*.

The technique promises to be useful for treating data sets with secondary diffusion points, that is, not all the points of observation derive directly from the first origin, but, instead, they would derive from secondary points of spreading.

The observed spreading pattern would correspond to long distance diffusion by commercial ships traveling from a commercial port to another for thousands of km, while shorter distance diffusion would occur through touristic boat traveling from a touristic port to another (forming the clusters of data). Natural diffusion via sexual or vegetative propagation would occur only within few kilometers from the starting point(s) corresponding to observations within the single populations (the items of a cluster). The spreading of *Caulerpa* through the transport of the alga by boat anchors or fishing

nets, even over potentially great distances, was already recognized by other authors (Sant et al. 1996; Relini et al. 2000).

The proposal of new analytical techniques to be used together to the GP for the investigation of biological invasions appears promising, in which also quite complex pattern of spreading can be taken into consideration.

The proposal of splitting the data set in clusters was very recently proposed by Faulkner et al. (2015) to locate elusive nocturnal animals. These authors applied a recently developed method for GP analysis, that is, the DPM model of GP (Verity et al. 2014), starting with a preliminary data subdivision in clusters (Verity et al. 2014). We add to this promising method in our here-developed analysis, the use of the *Silhouette* as method to assess the right number of clusters. A first K-means analysis can help to identify subclusters of data, provided that a criterion is given to choose the right number of clusters. The *Silhouette* criterion was able to suggest a number of clusters that, after the Voronoi tessellation, resulted corresponding to approximate large geographic subdivisions of the Mediterranean. We chose 3 clusters, since it was that with the best *Silhouette* index, even if the model with two clusters had a *Silhouette* index slightly higher than 0.5. As a matter of fact, *Silhouette* indices of 0.5 or above are considered evidence of reasonable clustering structure (Yatsunenko et al. 2012). The Voronoi tessellation of the space provides a mathematical prediction about where, more probably, the single clusters will expand in the future.

Conclusion

K-means clustering and Voronoi diagrams can partition a data set of geographic positions of populations of an alien species invading a determined geographic area.

In the article, the identification of a known right number of clusters (known as right, since it was simulated) of data for the following K-means analysis was executed using the

2044

Int. J. Environ. Sci. Technol. (2016) 13:2037–2044

Silhouette technique (Rousseeuw [1987]). The correspondence of the simulated clusters with the clusters found by *Silhouette* + K-means was quantified, showing that even with 18 clusters in a fixed space of 512 * 512 pixels assigned correctly more than 75 % of the data set.

The proposed algorithms and bioinformatic tools (in Python language) could also be applied on other biological phenomena, where the main data consist in geographic positions that may be usefully partitioned in more homogeneous observations.

References

Aurenhammer F (1991) Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput Surv 23(3):345–405

Chiang MM, Mirkin B (2010) Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. J Classif 27(1):1–38

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 Manag 87(4):559–566

Dramstad W (1996) Do bumblebees (Hymenoptera: Apidae) really forage close to their nests? J Insect Behav 9:163–182

Faulkner SC, Stevenson MD, Verity R, Mustari AH, Semple S, Tosh DG, Le Comber SC (2015) Using geographic profiling to locate elusive nocturnal animals: a case study with spectral tarsiers. J Zool 295:261–268

Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recogn Lett 31(8):651–666

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice Hall, New Jersey

Le Comber SC, Nicholls B, Rossmo DK, Racey PA (2006) Geographic profiling and animal foraging. J Theor Biol 240:233–240

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

Leisch F (2006) A toolbox for K-centroids cluster analysis. Comput Stat Data Anal 51(2):526–544

Lleti R, Ortiz MC, Sarabia LA, Sánchez MS (2004) Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the *silhouettes*. Anal Chim Acta 515(1):87–100

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

Meinesz A, Hesse B (1991) Introduction et invasion de l'algue tropicale *Caulerpa taxifolia* en Mediterranée Nord occidentale. Oceanol Acta 14(4):415–426

Meinesz A, Belsher T, Thibaut T, Antolic B, Karim Ben Mustapha K, Boudouresque CF, Chiaverini D, Cinelli F, Cottalorda J-M, Djellouli A, El Abed A, Orestano C, Grau AM, Ivesa L, Jaklin A, Langar H, Massuti-Pascual E, Peirano A, Tunesi L, de Vaugelas J, Zavodnik N, Zuljevic A (2001) The introduced green alga *Caulerpa taxifolia* continues to spread in the Mediterranean. Biol Invasions 3:201–210

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

Piazzi L, Meinesz A, Verlaque M et al (2005) Invasion of *Caulerpa racemosa* var. *cylindracea* (Caulerpales, Chlorophyta) in the Mediterranean Sea: an assessment of the spread. Cryptogam Algol 26:189–202

Raine NE, Rossmo DK, Le Comber SC (2009) Geographic profiling applied to testing models of bumble-bee foraging. J R Soc Interface 6:307–319

Relini G, Relini M, Torchia G (2000) The role of fishing gear in the spreading of allochthonous species: the case of *Caulerpa taxifolia* in the Ligurian Sea. ICES J Mar Sci 57:1421–1427

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

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

Rousseeuw PJ (1987) *Silhouettes*: a graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math 20:53–65

Sant N, Delgado O, Rodriguez-Prieto C, Ballesteros E (1996) The spreading of the introduced seaweed *Caulerpa taxifolia* (Vahl) C. Agardh in the Mediterranean Sea: testing the boat transportation hypothesis. Bot Mar 39:427–430

Saville NM, Dramstad WE, Fry GLA, Corbet SA (1997) Bumblebee movement in a fragmented agricultural landscape. Agric Ecosyst Environ 61:145–154

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 KD, Nichols RA, Le Comber SC (2014) Spatial targeting of infectious disease control: identifying multiple, unknown sources. Methods Ecol Evol 5:647–655

Verlaque M, Carrillo JA, Gil-Rodriguez MC, Durand C, Boudouresque CF, Le Parco Y (2004) Blitzkrieg in a marine invasion: *Caulerpa racemosa* var. *cylindracea* (Bryopsidales, Chlorophyta) reaches the Canary Islands (north-east Atlantic). Biol Invasions 6:269–281

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI (2012) Human gut microbiome viewed across age and geography. Nature 486:222–227