Santosuosso *et al. J Big Data* (2020) 7:14
https://doi.org/10.1186/s40537-020-00288-8

Journal of Big Data

RESEARCH Open Access

Tracing outliers in the dataset of *Drosophila suzukii* records with the Isolation Forest method

Ugo Santosuosso[1], Alessandro Cini[2] and Alessio Papini[3*]

*Correspondence:
alpapini@unifi.it
[3] Department of Biology, University of Florence, Via Micheli 3, Florence 50121, Italy
Full list of author information is available at the end of the article

Abstract

The analysis of big data is a fundamental challenge for the current and future stream of data coming from many different sources. Geospatial data is one of the sources currently less investigated. A typical example of always increasing data set is that produced by the distribution data of invasive species on the concerned territories. The dataset of *Drosophila suzuki* invasion sites in Europe up to 2011 was used to test a possible method to pinpoint its outliers (anomalies). Our aim was to find a method of analysis that would be able to treat large amount of data in order to produce easily readable outputs to summarize and predict the status and, possibly, the future development of a biological invasion. To do that, we aimed to identify the so called anomalies of the dataset, identified with a Python script based on the machine learning algorithm "Isolation Forest". We used also the K-Means clustering method to partition the dataset. In our test, based on a real dataset, the Silhouette method yielded a number of clusters of 10 as the best result. The clusters were drawn on the map with a Voronoi tessellation, showing that 8 clusters were centered on industrial harbours, while the last two were in the hinterland. This fact led us to guess that: (1) the main entrance mechanisms in Europe may be the wares import fluxes through ports, occurring apparently several times; (2) the spreading into the inland may be due to road transportation of wares; (3) the outliers (anomalies) found with the isolation forest method would identify individuals or populations that tend to detach from their original cluster and hence represent indications about the lines of further spreading of the invasion. This type of analysis aims hence to identify the future direction of an invasion, rather than the center of origin as in the case of geographic profiling. Isolation Forest provides therefore complimentary results with respect to PGP. The recent records of the invasive species, mainly localized close to the outliers position, are an indication that the isolation forest method can be considered predictive and proved to be a useful method to treat large datasets of geospatial data.

Keywords: *Drosophila suzuki*, Geoprofiling, Invasive species, Invasion, Isolation Forest, K-Means

Introduction

Invasive alien species (IAS) are one of the leading threats to native wildlife, human health and food safety/production [42, 43]. The ongoing increase in worldwide trade is facilitating the spread of IAS, causing significant ecological and economic impact. Understanding the spatio-temporal spread of invasion of IAS is crucial to allow prevention

and management of these species of concern. The amount of data about IAS is always increasing, producing huge data sets whose analysis is becoming always more demanding. The so produced geospatial data is one of the main source of Big Data and one of the less investigated sources. "Big Data" is a definition referring to data sets that are not only simply large, but also of difficult management with traditional analysis methods [40]. The analysis of big data is, in general, a fundamental challenge for the current and future stream of data coming from many different sources [13].

Drosophila suzukii (Matsumura, 1931) (Diptera: Drosophilidae) (also known as vinegar fly and as Spotted Wing Drosophila, hereafter DS) is an invasive insect species that recently colonized many countries outside its native range [9]. DS attacks mainly thin exocarp fruits such as soft and stone fruits, particularly belonging to family Rosaceae, such as cherries and apricots [9]. DS can lay eggs in fruit and the developing larvae feed on the fruit flesh [33]. DS can reach 15 generations per year causing possible further damage by secondary infections by bacteria and fungi, with a potential damage to crops up to 30–100% of the total yield [44]. The estimated economic loss amounted to more than 500 million dollars every year in the USA after [4]. As for comparison, even in the small traditional fruit production area of Trento Province (Italy), about 500,000 EUR were lost in 2010 and 3 million in 2011 [11].

DS is endemic of South East Asia [9] and was first recorded in California, Spain and Italy in year 2000 [8], later colonizing large part of USA, Canada and Europe [9, 44]. South America and Central America [15, 20]. Despite more than ten years passed since DS first report in Europe, the spatial spreading of the DS invasion is still to be cleared. Cini et al. [8] investigated the spread pattern of DS in Europe by means of a spatial analysis technique (Probabilistic geographic profiling, hereafter PGP) to understand the possible spreading centre/s in European countries. This approach provided fruitful insights, depicting the South of France as a possible centre of spread [8] and also strongly suggesting that the most likely pathway of introduction of DS e is the trade of fruit, with eggs or larvae being transported unnoticed in fruits sea-traded from South East Asia [33].

A PGP is an analysis method aiming to identify the origin of linked events on a map. It was firstly proposed for crimes done by a serial killer in criminology and later for the spreading populations of invasive species and epidemiology [5, 6, 25, 28, 35]. PGP uses coordinates of events on a map to calculate a probability surface called geoprofile [14, 32]. The geoprofiles does not provide exactly the center of origin of the event, but rather provide areas at different priority on the map with a variable probability density [31, 32]. After its first use in criminoloy, PGP was applied to biological problems such as the targeting of an infectious disease [27], the prediction of nest locations of bumble bees [39], animal foraging [21, 30], sharks hunting patterns [24] and even the distribution pattern of V2 bombing [3]. More recently, PGP was used to guess the source of an invasion by alien organisms using the positions of their current populations [8, 28, 38]. This analysis is useful, since it can suggest control methods and give an idea about the gateway of the invasion [8]. Recently, further refinements of the method allowed to improve the power and reliability of PGP, in particular by allowing (a) to clarify if a distribution pattern is caused by more waves of invasion, rather than from a single starting point [8], (b) to evaluate the robustness of the results with a jackknife procedure [26], (c) to give different weights to data on a quantitative basis (on the basis of the population dimension) or

new methods of data partitioning [10, 36]. The PGP was applied on the DS distribution in Europe, since it does not require any a priori knowledge about the invasion routes, while the biological justifications of the used parameters can be evaluated also by testing several parameters values.

The Isolation Forest method is a machine learning algorithm belonging to the algorithms family based on the "Random Forest" and decision trees used to identify anomalies in big data data sets. The term "isolation" means the separation of one of the items of the data set from the rest of the other items on the basis of one or more given rules [22, 23]. The items (called "instances" by Liu et al. [23]) that remain separated from the rest of the data are the outliers or "anomalies". The main usage of the Isolation Forest method was for data mining in general [7] and particularly for monitoring of networks and genomic analysis [41]).

While geographic profiling tries to identify the source of a series of events and hence "something" that occurred in the past, it does not provide an idea about the future development of the series of events. On the contrary, the Isolation Forest method attempts to find the points of the geospatial distribution of events apparently less related to the main bulk of data and so, in our interpretation, to give an idea about the future directions of the spreading itself.

Here we compare the results obtained with PGP on DS distribution pattern with the results obtained with the Isolation Forest (IF) method, with the aim to identify, with the latter method, observations that are in particular positions among the general distribution (outliers), in order to evaluate them as possible beginning secondary invasions/spreading points. The considered dataset of observations is referred to central-west-south Europe and was the same used in Cini et al. [8], since it is a well checked dataset. While the results about the future development of *Drosophila suzukii* in Europe is of interest per se, we intend to use this dataset as a test for the application of the Isolation Forest to biological invasions data analysis.

The rest of the article is structured as follows: first, we explain the method of the Isolation Forest in detail, then we pass to the other techniques of data partitioning employed in the article. In the results section we show also graphically the type of data produced as output by the Isolation Forest analysis, giving an idea of the robustness of the data so obtained. Then we show the results obtained with data partitioning. In the discussion section, we show the high correspondence between the Isolation Forest and the classical methods of data partitioning here employed, concluding with a discussion about the meaning of the "anomalies" found by the Isolation Forest and how this data can be used to treat problems such as biological invasion or other similar problems.

Methods/experimental

Isolation Forest

The IF needs a preliminary "training" phase to build the model of the system and the decision tree to be used in the following phases of classification, or "test" [23]. Usually, if a data set previously classified (known data) to train the system is not available, a subset of the original data set is selected and used for the training. In our case, the training

subset was 25% of the complete data set. The so built model is then used for the analysis on the test data set.

Replicates randomization

In order to eliminate the influence due to the random choice of the test data set, we resampled the data set randomly. This technique can be considered an analogous to the Jackknife or the Bootstrap [26] resampling methods and to the "Random Forest" [37]. For each resampling we execute a new "training" and a new "test". For each replicate the anomalies are recorded. The aim was to increase the feasibility of the anomalies inferred with a single test, considering as more probable the anomalies that are recorded in more than 50% of the replicates. The other anomalies were considered as possible random fluctuations of the training data set.

In our case we have 91 observations (consisting in geographical coordinates on a map).

Since we used a training set of about 25% of the total data set, we used 22 samples for this purpose. The test data set consisted of 69 samples, while we performed 100 replicates.

K-Means and silhouette

The K-Means method consists in the partitioning of a data set without supervision [18, 19, 35]. It is necessary to provide a priori the number of clusters in which the data set should be divided. If it is not possible, it is necessary to find other criteria to guess the right number of clusters in which the data should be divided. We used as criterion the maximization of the function Silhouette [34] evaluating a number of clusters varying from 2 to 15. When the value of the Silhouette function was optimized for a given number of clusters, we took that number of clusters as the number of clusters for the k-Means analysis.

In a data set of N items, the minimal number of partitions is 2, while the highest number would be N (only one item for each partition). Apart this case, the best number of clusters should avoid a single clusters having less than a given value and the presence of isolated points, so that each partition should have locally homogeneous features.

After the partitioning of the dataset in subclusters, we divided the space (map) accordingly with the Voronoi algorithm [2] as implemented by Santosuosso and Papini [35].

Used algorithms

The Isolation Forest method

The Isolation Forest algorithm was described by Liu et al. [23]. Replicate randomization to evaluate robustness of data was introduced by Gnerucci et al. [16]. Kmeans clustering and the Silhouette method for assessing the best number of clusters were applied to the analysis of a spreading pattern of an alien species by Papini et al. [26, 28]. The algorithms were applied with procedures written in Python 2.7.14 (default, Sep 23 2017, 22:06:14, www.python.org) programming language. We used the Sk-Learn 0.19.0 library (Pedregosa et al. [29] and MathPlotLib 2.2.2 library [17] (https://matplotlib.org/). The

Python procedure Plt_IF_DATA_0.1.1.py (https://bitbucket.org/ugosnt/al_and_ugo/), was written by the authors.

Results

Anomalies found with the training and the test data sets

In the training set we found n = 24 anomalies with, in total, 300 replicates. In the test set we found n = 46 anomalies with, in total, 1000 replicates. Figure 1 shows the type of data produced by the Isolation Forest analysis during one of the replicates. The anomalies

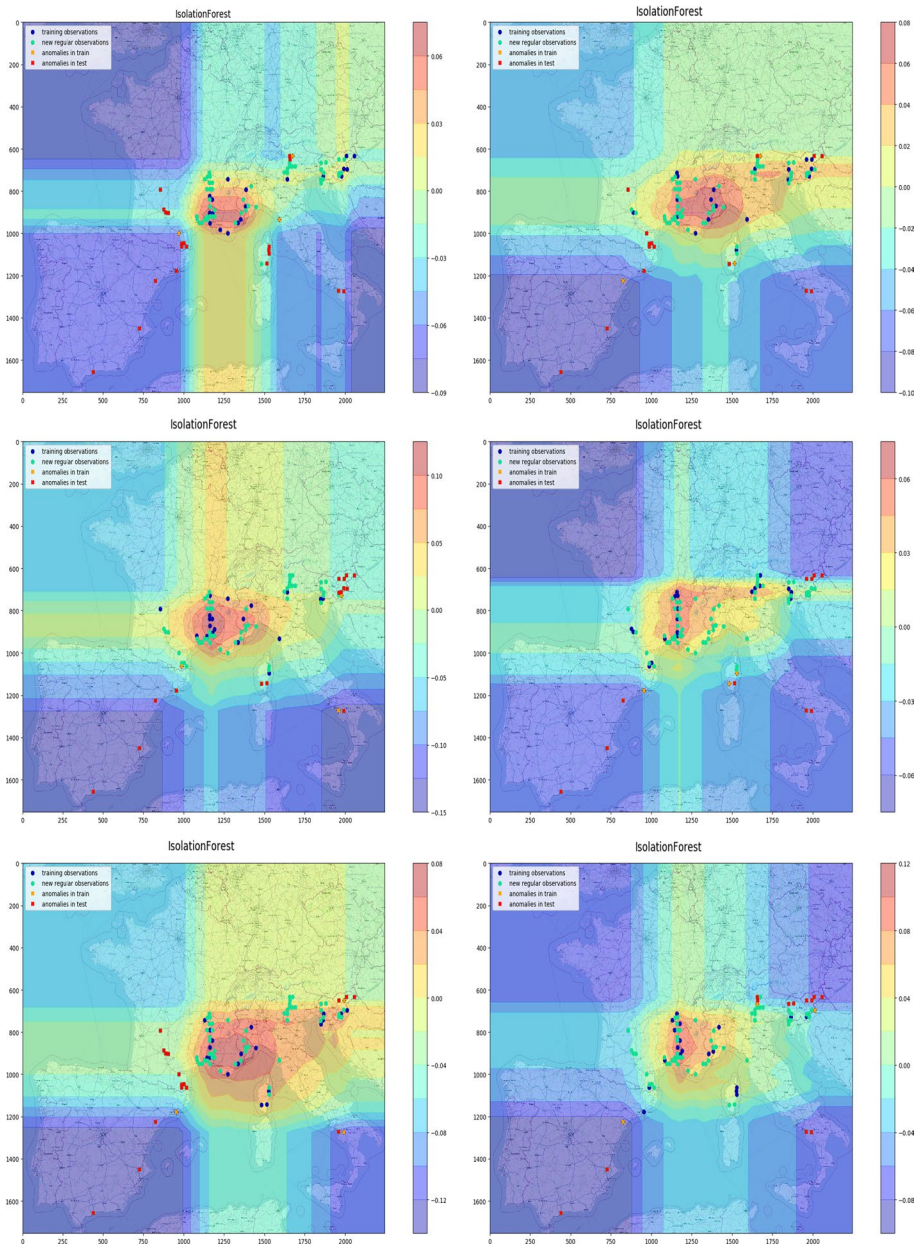

Fig. 1 Some examples of the results of replicates of the execution of IF analysis. The anomalies are reported: those obtained from the training set as yellow squares and in red those obtained from the test set

were those observations (points in the map) that were less coherent with respect to the other observations. The anomalies were reported on the maps in Fig. 1 separating those obtained with the training set from those obtained from the test set in the shown replicates. The average number of occurrences in the training phase was = 24 [12.5 ± 1165 (mean ± standard deviation)]. The average number of occurrences in the test phase was 46 [occurrences = 2774 ± 2306 (mean ± standard deviation)].

The total number of anomalies observed was 49. It corresponds to the average number of occurrences (train + test) = 49.

The set composed by "train + test" is the union (in the set theory sense) of the items of the two sets and not by the mathematical sum, since some items may appear in both sets. There is not a significant difference in frequency of anomalies between the training and the test sets (the f-ratio value was 3.38299 and p-value 0.070238. Hence, the result was not significant at $p < 0.05$). Even between the test set and the whole set of the anomalies the f-ratio value was 1.05112 and the p-value was 0.307966. However, a difference was found between the training set and the total set (the f-ratio value is 5.06563 and p-value 0.027595). Hence, the result was significant at $p < 0.05$. In conclusion, the training and the test data sets do not present significant differences from the point of view of the anomalies.

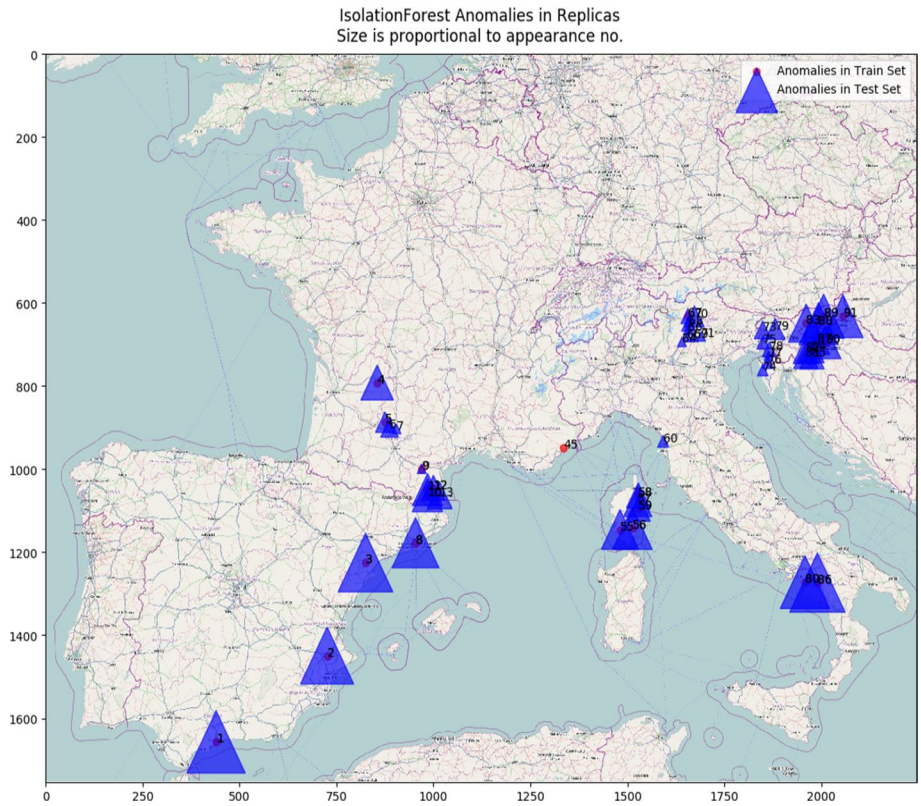

Fig. 2 Robustness of the found anomalies. Only the anomalies appearing in more than 50% of the replicates are recorded. The dimension of the triangle is proportional to the number of replicates in which the anomaly appears

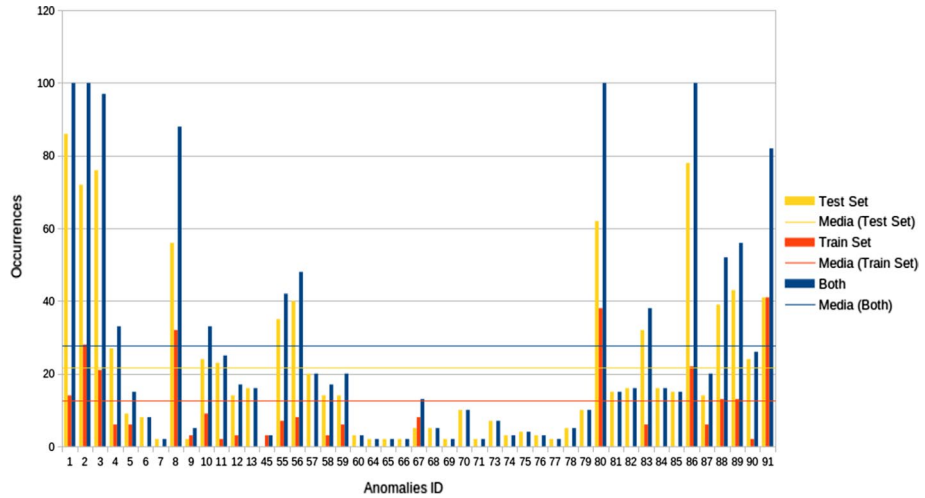

Fig. 3 For each observation (1–91) the height of the histogram represents the number of times of occurrence of that specific observation among the anomalies recorded among the replicates

The robustness of the results is represented in Fig. 2 as proportional to the dimension of the blue triangles. The triangles correspond to the cases in which the single anomaly appears in more than 24 replicates (50%). Of the 24 items resulting from the IF-training analysis phase, only 4 appeared in the replicates more than 24 times (value higher than the mean of the sample + standard deviation = 3.57). Of the 46 items resulted from the IF analysis-Test phase, 6 appeared more than 45 times.

The histogram in Fig. 3 shows the number of times of occurrence of each specific observations among the anomalies during the replicates. The height of the column is proportional to the number of occurrence.

Of the 29 results of the IF test, 6 "anomalies" appeared in the replicates more than 45 times (value higher than the mean of the sample + standard deviation (i.e. 3.57). Considering the union of the two subsets of data ("train + test"), only 7 points presented an occurrence frequency higher than the mean of the sample + standard deviation (i.e. 3.57).

K-Means and Silhouette analysis

The Silhouette values for the varying number of possible clusters are drawn in Fig. 4.

In our results, best number of partitions N = 10 (Fig. 5), we had 2 clusters with relatively few items, but the IF algorithm had already indicated such points as anomalies. This reason, together with the best value of the Silhouette function (Fig. 4), induced us to accept this value as the correct number of clusters. We represented the distribution of the clusters on the map by partitioning space with Voronoi tessellation in Fig. 5, where on the left we indicate both the numerical consistency of the cluster and the Silhouette value for each cluster. Of the 10 clusters, 8 resulted centered on industrial ports on the coast, while the last two are located in the inland.

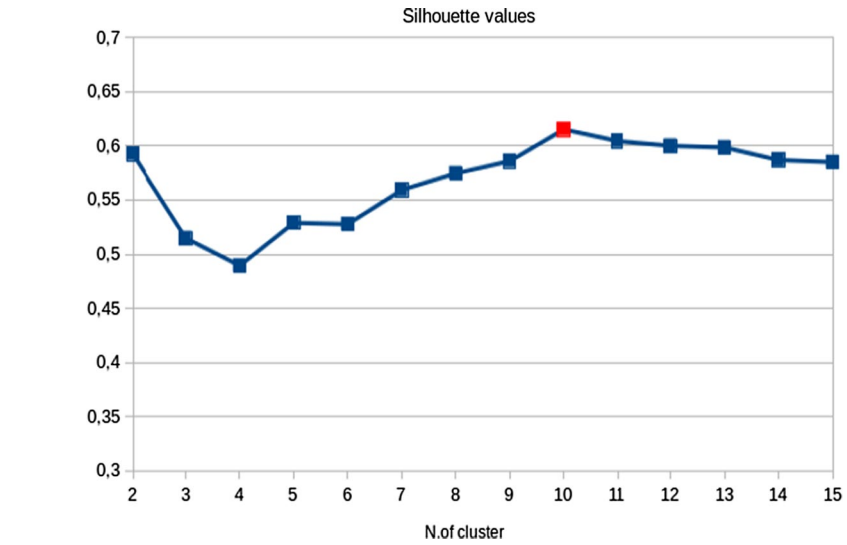

Fig. 4 Values of Silhouette. The best value is for 10 clusters (in red)

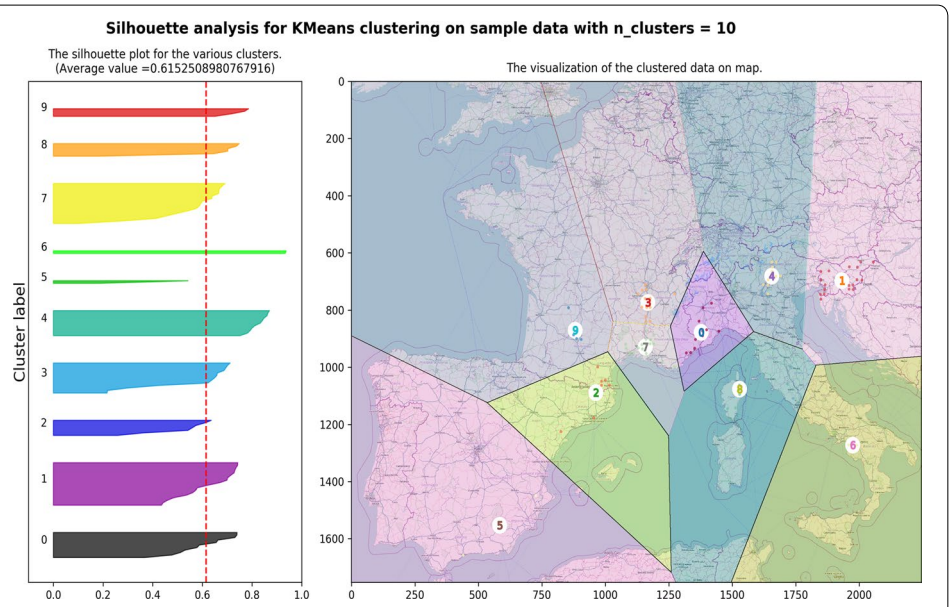

Fig. 5 Silhouette values on the left and Voronoi tessellation on the right

Discussion

The meaning of the anomalies

The anomalies were interpreted as the observations of populations (points in the map) that are tending to become independent with respect to the rest of the populations. The main populations clusters were identified by K-Means. In the case of an invasion, the k-Means identifies the main clusters of individuals or populations observed on the map [35]. The anomalies would then be individuals or parts of populations that are

tending to detach from the cluster to which they belong. The anomalies would then also represent an indication of the future lines of further invasion in a territory.

Since the data set on which we based our analysis was of 2011, we compared the anomalies position of the 2011 database with the more recent observations of *D. suzukii*. Where available, the new observations where often corresponding to areas close to anomalies identified by the Isolation Forest analysis. See, for instance the new records in the region south-east of Bordeaux 2011–2014 in France by Delbac et al. [12]. Also the expansion from Croatia towards Hungary, Bosnia and Serbia [1] corresponded to the position of the outliers in Croatia on the basis of the 2011 data set, while quite unexpected was the record of *D. suzukii* in Poland [1], possibily as a result of a new independent introduction.

Almost all the clusters identified with K-Means (interpreted as main invasion sites) contained anomalies, possibly indicating that the invasion of *D. suzukii* is still in development in Europe.

The method may hence be considered predictive of the lines of future development of the invasion and this knowledge may provide useful suggestion about where an intervention to limit and contain the invasion could be more effective.

Also the DBSCAN [10, 36] clustering method is able to identify outliers but requires a priori knowledge about the maximum distance for which points on a map are considered adjacent and the minimum number of adjacent points that are to be considered for a cluster. In both cases, it is necessary to find a biological justification for these assumptions. The isolation forest is a disruptive method based on a probabilistic approach composed of 2 phases: unsupervised learning on a random sub-sample of the data distribution and subsequent extension to the whole data set of the one learned during the test phase. This method considers the whole data set as a single cluster affected by "anomalies" and tries to recognize and distinguish these anomalies from the rest of the points in order to rule out them from a second phase from the planned processing. In both cases it is necessary to find a biological justification for these assumptions.

Anomalies can be considered in our analysis either all the cases highlighted by the various replicates or, more restrictively, the outliers found in all replicates. We can then apply an analysis method (the PGP in our case) for NON-anomalous cases, while on the anomalous cases a further analysis is made to understand why they are "anomalous" and which information they contain.

Conclusions

The Isolation Forest algorithm was used to identify the anomalies within the dataset containing the localities in Europe where the invasive species *D. suzukii* was recorded. By adding these data to cluster analysis with Kmeans and Silhouette algorithms and by partitioning accordingly the space with Voronoi tessellation we could identify the most probable areas where the invasion of DS is more probable to occur.

The results show that the Isolation Forest can be considered a useful tool to identify the anomalies in a data set consisting in records of observation of an invasive species in a territory or, possibly, other datasets referred to positions in the space. The anomalies can

be interpreted as the individuals or populations that are tending to become independent with respect to the rest of the distribution of populations. Such anomalies represent hence an indication of the spreading lines of the invasion. This is a type of analysis not addressed by other methods of spread analysis such as the GPG, that aims rather to the identification of the center of origin.

Abbreviations
PGP: Probabilistic Geoprofiling; DS: *Drosophila suzukii*; IF: Isolation Forest.

Authors' contributions
US ideated the method and wrote large part of the software and part of the article; AP wrote part of the software and of the article and discussed with the first author the biological meaning of the isolation forest; AC arranged the database used for the analysis e participated to the writing of the manuscript. All authors read and approved the final manuscript.

Funding
University of Florence (ricatena2019).

Availability of data and materials
All the material and the software is open source.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Clinical and Experimental Medicine, University of Florence, Largo Brambilla 3, 50134 Florence, Italy. [2] Department of Biology, University of Florence, Via Madonna del Piano 6, 50019 Florence, Italy. [3] Department of Biology, University of Florence, Via Micheli 3, Florence 50121, Italy.

Received: 20 October 2019 Accepted: 9 February 2020
Published online: 05 March 2020

References

1. Asplen MK, Anfora G, Biondi A, et al. Invasion biology of spotted wing Drosophila (*Drosophila suzukii*): a global perspective and future priorities. J Pest Sci. 2015;88:469–94.
2. Aurenhammer F. Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput Surv. 1991;23(3):345–405.
3. Aygin DT, Cox LA, Faulkner SC, Stevens MCA, Verity R, Le Comber SC. Double cross: geographic profiling of V-2 impact sites. J Spat Sci. 2019. https://doi.org/10.1080/14498596.2019.1642249.
4. Bolda M, Goodhue RE, Zalom FG. Spotted wing Drosophila: potential economic impact of a newly established pest. Agric Res Econ Updat. 2010;13:5–8.
5. Butkovic A, Mrdovic S, Uludag S, Tanovic A. Geographic profiling for serial cybercrime investigation. Digit Invest. 2019;28:176–82.
6. Cerri J, Mori E, Zozzoli R, Gigliotti A, Chirco A, Bertolino S. Managing invasive Siberian chipmunks *Eutamias sibiricus* in Italy: a matter of attitudes and risk of dispersal. Biol Invasions. 2019. https://doi.org/10.1007/s10530-019-02115-5.
7. Cheng Z, Zou C, Dong J. Outlier detection using isolation forest and local outlier factor. Proceedings of the Conference on Research in Adaptive and Convergent Systems. 2019; 161–168. Chongqing, China — September 24–27, 2019. ACM New York, NY, USA. ISBN: 978-1-4503-6843-8 https://doi.org/10.1145/3338840.3355641.
8. 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):559–66.
9. Cini A, Ioriatti C, Anfora G. A review of the invasion of *Drosophila suzukii* in Europe and a draft research agenda for integrated pest management. B Insectol. 2012;65:149–60.
10. Cini A, Santosuosso U, Papini A. Uncovering the spatial pattern of invasion of the honeybee pest small hive beetle, *Aethina tumida in* Italy. Rev Bras Entomol. 2019;63(1):12–7.
11. De Ros G, Anfora G, Grassi A, Ioriatti C. The potential economic impact of *Drosophila suzukii* on small fruits production in Trentino (Italy). IOBC-WPRS Bul. 2013;91:317–21.
12. Delbac L, Rouzes R, Rusch A, Thiery D. Geographical area extension of *Drosophila suzukii* (Diptera: Drosophilidae) in Bordeaux vineyards. Integr Prot Prod Viticulture IOBC–WPRS Bull. 2017;128:28–36.
13. Dheyab SA, Abdullah MN, Abed BF. A novel approach for big data processing using message passing interface based on memory mapping. J Big Data. 2019;6:112. https://doi.org/10.1186/s40537-019-0275-3.
14. Faulkner S. Integrating GIS approaches with geographic profiling as a novel conservation tool. PhD thesis, Queen Mary University, London, 2018. https://qmro.qmul.ac.uk/xmlui/handle/123456789/46763.

15. Ferronato P, Woch AL, Soares PL, Bernardi D, Botton M, Andreazza F, Oliveira E, Corrêa AS. A phylogeographic approach to the *Drosophila suzukii* (Diptera: Drosophilidae) invasion in Brazil. J Econ Entomol. 2018;112(1):425–33.
16. Gnerucci A, Romano G, Ratto F, Fusi F. Statistical detection of nanoparticles in cells by darkfield microscopy. Physica Med. 2016;32(7):938–43.
17. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90–5.
18. Jain AK. Data clustering: 50 years beyond K-Means. Pattern Recogn Lett. 2010;31(8):651–66.
19. Jain AK, Dubes RC. Algorithms for Clustering Data. New Jersey: Prentice Hall; 1988.
20. Lasa R, Tadeo E. Invasive drosophilid pests *Drosophila suzukii* and *Zaprionus indianus* (Diptera: Drosophilidae) in Veracruz, Mexico. Florida Entomol. 2015;98(3):987–9.
21. Le Comber SC, Nicholls B, Rossmo DK, Racey PA. Geographic profiling and animal foraging. J Theor Biol. 2006;240:233–40.
22. Liu FT, Ting KM, Zhou ZH. Isolation-based anomaly detection. ACM Trans Knowl Discov Data (TKDD). 2012;6(1):3.
23. Liu FT, Ting KM, Zhou ZH. Isolation forests. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), 2008. pp. 413–422.
24. Martin RA, Rossmo DK, Hammerschlag N. Hunting patterns and geographic profiling of white shark predation. J Zool. 2009;279:111–8.
25. 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;15:1613–21.
26. Papini A, Rossmo DK, Le Comber SC, Verity R, Stevenson MD, Santosuosso U. The use of jackknifing for the evaluation of geographic profiling reliability. Ecol Inform. 2017;38:76–81.
27. Papini A, Santosuosso U. Snow's case revisited: new tool in geographic profiling of epidemiology. Braz J Infect Dis. 2016;21(1):112–5.
28. Papini A, Signorini MA, Foggi B, Della Giovampaola E, Ongaro L, Vivona L, Santosuosso U, Tani C, Bruschi P. History vs. legend: retracing invasion and spread of *Oxalis pes-caprae* L. in Europe and the Mediterranean area. PLoS ONE. 2017;12(12):0190237.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
30. Raine NE, Rossmo DK, Le Comber SC. Geographic profiling applied to testing models of bumble-bee foraging. J R Soc Interface. 2009;6:307–19.
31. Rossmo DK. A methodological model. Am J Crim Justice. 1993;172:1–21.
32. Rossmo DK. Geographic profiling. Boca Raton: CRC Press; 2000.
33. Rota-Stabelli O, Blaxter M, Anfora G. Quick guide: *Drosophila suzukii*. Curr Biol. 2013;23(1):R8.
34. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math. 1987;20:53–65.
35. Santosuosso U, Papini A. Methods for geographic profiling of biological invasions with multiple origin sites. Int J Environ Sci Technol. 2016;13(8):2037–44.
36. Santosuosso U, Papini A. Geo-profiling: beyond the current limits. A preliminary study of mathematical methods to improve the monitoring of invasive species. Russ J Immunol Ecol. 2018;49(4):362–70.
37. Shi T, Horvath S. Unsupervised learning with random forest predictors. J Comput Graph Stat. 2006;15(1):118–38.
38. Stevenson MD, Rossmo DK, Knell RJ, Le Comber SC. Geographic profiling as a novel spatial tool for targeting the control of invasive species. Ecography. 2012;35:1–12.
39. Suzuki-Ohno Y, Inoue MN, Ohno K. Applying geographic profiling used in the field of criminology for predicting the nest locations of bumble bees. J Theor Biol. 2010;265:211–7.
40. Swan M. The quantified self: fundamental disruption in big data science and biological discovery. Big Data. 2013;1(2):85–99.
41. Tian P, Che D. GI-IsolationForest: Genomic Island Discovery Using Isolation Forest Algorithm Internatonal Conf. Bioinformatics and Computational Biology| BIOCOMP'18 17–23 ISBN: 1-60132-471-5, CSREA Press, 2018. https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/BIC4116.pdf.
42. Vilà M, Basnou C, Pysek P, Josefsson M, Genovesi P, Gollasch S, Nentwig W, Olenin S, Roques A, Roy D, Hulme PE, DAISIE partners. How well do we understand the impacts of alien species on ecosystem services? A pan-European, cross-taxa assessment. Front Ecol Environ. 2010;8:135–44.
43. Vilà M, Espinar JL, Hejda M, Hulme PE, Jarosik V, Maron JL, Pergl J, Schaffner U, Sun Y, Pyšek P. Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. Ecol Lett. 2011;14:702–8.
44. Walsh DB, Bolda MP, Goodhue RE, Dreeves AJ, Lee JC, Bruck DJ, Walton VM, O'Neal SD, Zalom FG. *Drosophila suzukii* (Diptera: Drosophilidae): Invasive pest of ripening soft fruit expanding its geographic range and damage potential. J Integr Pest Manag. 2011;1:1–7.

Publisher's Note