Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
COVID-19 in Italy and extreme data mining

Paolo Massimo Buscema\textsuperscript{a,b,}\textsuperscript{*}, Francesca Della Torre\textsuperscript{a}, Marco Breda\textsuperscript{a}, Giulia Massini\textsuperscript{a}, Enzo Grossi\textsuperscript{a}
\textsuperscript{a}Semeion Research Center of Sciences of Communication, via Sersale, 117, 00128 Rome, Italy
\textsuperscript{b}University of Colorado at Denver, Department of Mathematical and Statistical Sciences, Denver, CO, USA

\textbf{ARTICLE INFO}

\textbf{Article history:}
Received 24 April 2020
Available online 25 July 2020

\textbf{Keywords:}
Topological weighted centroid
COVID-19
Geographic profiling
Artificial intelligence
Epidemics
Adaptive systems

\textbf{ABSTRACT}

In this article we want to show the potential of an evolutionary algorithm called Topological Weighted Centroid (TWC). This algorithm can obtain new and relevant information from extremely limited and poor datasets. In a world dominated by the concept of big (fat?) data we want to show that it is possible, by necessity or choice, to work profitably even on small data. This peculiarity of the algorithm means that even in the early stages of an epidemic process, when the data are too few to have sufficient statistics, it is possible to obtain important information.

To prove our theory, we addressed one of the most central issues at the moment: the COVID-19 epidemic. In particular, the cases recorded in Italy have been selected. Italy seems to have a central role in this epidemic because of the high number of measured infections. Through this innovative artificial intelligence algorithm, we have tried to analyze the evolution of the phenomenon and to predict its future steps using a dataset that contained only geospatial coordinates (longitude and latitude) of the first recorded cases.

Once the coordinates of the places where at least one case of contagion had been officially diagnosed until February 26th, 2020 had been collected, research and analysis was carried out on: outbreak point and related heat map (TWC alpha); probability distribution of the contagion on February 26th (TWC beta); possible spread of the phenomenon in the immediate future and then in the future of the future (TWC gamma and TWC theta); how this passage occurred in terms of paths and mutual influence (Theta paths and Markov Machine). Finally, a heat map of the possible situation towards the end of the epidemic in terms of infectiousness of the areas was drawn up. The analyses with TWC confirm the assumptions made at the beginning.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

On January 9th, 2020, the World Health Organization (WHO) stated that Chinese health authorities have identified a new strain of coronavirus never before identified in humans named COVID-19. On January 30th, the Italian Superior Institute of Health (Istituto Superiore di Sanità) confirmed the first two cases of COVID-19 infection in Italy [1].

COVID-19 is producing a tremendous impact in Italy from sanitary and economic point of view. According to General Confederation of Italian Industry, better known as Confindustria, the impact of the coronavirus on the GDP of Italy could be more than $-6.0\%$ [2]. At the same time, specific data on the infected cases are hidden because of privacy. Consequently,
without special permissions and authorizations it is not possible to get data about infected individuals like the precise address and other personal and clinic information.

These days it seems like the watchword is “big data”. Instead, we believe that an innovative and different strategy, extrapolating everything possible even from limited dataset, is possible. As a demonstration, we aim to analyze the (few) publicly available data about the Italian areas of infection of COVID-19 in its early stages to obtain estimates of the possible outbreak and the areas and methods of future spread of the virus within Italy.

The algorithm used is based on geographic profiling using a topological approach. One of the advantages of this algorithm, called Topological Weighted Centroid (TWC) \[3–9\], compared to standard methods of analysis of diffusion processes, lies in the fact that it requires very simple data: the coordinates of the places where the events of the process took place without any kind of other assumptions. For this reason, in conditions of poor data availability, it is particularly useful. In this case, the data used correspond to all places (by means of longitude and latitude) of Italy where at least one case of COVID-19 was detected until February 26th, 2020. The TWC has already been successfully applied several times to epidemic cases, in particular to Escherichia coli (Germany 2011), Chikungunya Fever (Italy 2007), Foot and Mouth Disease (UK 1968–1969), Food epidemic (Oahu 2010), Dengue fever (Brazil 2001), Listeria (US 2011) and Ebola (West Africa and Congo 2014–2018) epidemics \[3–5,8,9\].

Although the epidemic has a worldwide spread, it was chosen to focus on the atypical case of Italy for several reasons. On the one hand, the evolution of the epidemic in Italy has taken on a completely different character from the rest of the world, which leads us to think of a type of propagation of the local phenomenon. Moreover, it was considered that the movements within a single nation had a different structure compared to international movements. Therefore, it was preferred not to mix together information that could be discordant with each other. As an additional motivation, the set of data available on Italy was really small. Only 24 cities were involved until February 26th, 2020, so the dataset consisted of 24 rows and two columns, latitude and longitude.

2. Data

The model we are going to use in this paper is based on highly simple datasets. It considers only the geospatial coordinates of latitude and longitude of the places where the events occurred. In this case, all the Italian provinces in which at least one case of contagion has occurred have been taken into consideration. In fact, the only data publicly known are the cities involved in the COVID-19 infection. The data collected correspond to the episodes of contagion recorded up to February 26th, 2020 (23 provinces and 1 municipality).

The information about the contagions was taken from the official Italian civil protection repository \[10\]. During the course of the epidemic, this repository kept a page constantly updated with information about new episodes of contagion.

For each city we have decided to ignore the frequency of cases and the date. Thus, for every city were COVID-19 is present at least with one case, we only have latitude and longitude. Table 1 shows the name of these 24 locations and the relevant geospatial coordinates. Fig. 1 shows where they are located. We must emphasize that the latitude and longitude of each city is very fuzzy information on the real location of each infected case; the coordinates of Lodi, for example, globally summarize 10 different municipalities spread out around Lodi.

3. Methods

We analyzed the small dataset (48 real numbers) using a geographic profiling approach \[2,11–14\]. In particular, as mentioned above, we used the Topological Weighted Centroid (TWC), a special adaptive system, with a solid theoretical and mathematical background \[3\] and already successfully applied in various fields and published in numerous papers \[4–7\] \[8,9\].

The whole TWC theory can be summarized in five main points: $\text{TWC}(\alpha)$. Represents a spatial estimate of the point or area where the process under examination originated (outbreak); $\text{TWC}(\beta)$. Represents the current likely distribution of the process under consideration; $\text{TWC}(\gamma)$. Represents the likely future evolution of the $\text{TWC}(\beta)$ distribution, considering the system’s self-organizing properties; $\text{TWC}(\theta)$. Represents a further level of evolution over time, developed from $\text{TWC}(\gamma)$ as the communication and interaction between the observed events stabilizes to become highly organized. $\text{TWC}(\theta)$ also provides a directed weighted graph in which a hypothesized flow of communication among the points (cities) is represented; $\text{TWC}(\iota)$. Provides a heatmap representing the infectivity rate of the area.

$\text{TWC}$ is able to extrapolate time from space, bringing to light fundamental information trapped in the coordinates of events. Similar to the core drilling operations in geology, which allow the analysis of the history of the terrain one layer at a time, the $\text{TWC}$ enables the entire process to be analyzed one frame at a time. The main outputs of the different types of $\text{TWC}$ are heat maps. For all the mathematical details of the theory, please refer to the cited literature and to the supplementary material.
Table 1
Coordinates of the 24 infected cities up to February 26th, 2020.

| City          | Longitude | Latitude |
|---------------|-----------|----------|
| Alassio (SV)  | 8.167     | 44.004   |
| Ancona        | 13.219    | 43.480   |
| Roma          | 12.483    | 41.893   |
| Palermo       | 13.352    | 38.111   |
| La Spezia     | 9.691     | 44.238   |
| Bolzano       | 11.230    | 46.656   |
| Pistoia       | 10.869    | 43.974   |
| Firenze       | 11.256    | 43.770   |
| Torino        | 7.682     | 45.068   |
| Rimini        | 12.631    | 43.947   |
| Modena        | 10.936    | 44.538   |
| Parma         | 10.328    | 44.801   |
| Piacenza      | 9.667     | 44.848   |
| Treviso       | 12.206    | 45.807   |
| Venezia       | 12.335    | 45.437   |
| Padova        | 11.873    | 45.408   |
| Sondrio       | 10.258    | 46.323   |
| Brescia       | 10.426    | 45.780   |
| Monza Brianza | 9.279     | 45.640   |
| Milano        | 9.191     | 45.467   |
| Bergamo       | 9.754     | 45.757   |
| Pavia         | 9.138     | 45.037   |
| Cremona       | 10.037    | 45.221   |
| Lodi          | 9.492     | 45.261   |

Fig. 1. Map of the infected Italian cities considered in the analysis of Feb. 26th (yellow circles).

4. Results

4.1. The outbreak

We remind that $\text{TWC}(\alpha)$ represents a spatial estimate of the point or area where the process under examination originated (outbreak). Fig. 2 shows this outbreak point. This area is outside the red zone indicated by the Italian authorities. In fact, the Alpha Point is located about 39 km south east from Codogno, the alleged official outbreak.

Fig. 3 shows the distance between the red zone, identified by the municipalities of Bertonico, Casalpusterlengo, Castelgerundo, Castiglione D’Adda, Codogno, Fombio, Maleo, San Fiorano, Somaglia, Terranova dei Passerini and Vo’ according to the decree law on February 23rd, 2020 [15] and the estimated alpha point. The alpha point is therefore
slightly on the edge of the area defined as central by the Italian administration. Vo’ remains isolated. The distance between TWC alpha point and Codogno, where the Case 1 was, is about 35–40 km.

4.2. TWC heatmaps

Fig. 4 shows the heatmaps generated by the TWC algorithm.

As seen before, TWC(\(\beta\)) represents the current likely probability distribution of the process under consideration (Fig. 4 — top left). The Beta Map shows the current active area of the epidemic until February 26th, 2020. According to this advanced algorithm this area should be considered the real risk area to be monitored.

TWC(\(\gamma\)) represents the likely future evolution of the TWC(\(\beta\)) distribution, considering the self-organizing properties of the system (Fig. 4 — top right). According to the TWC algorithm, the expansion of the diffusion will be contracted in the north and it will expand slightly in the south, in the direction of Florence, and in the west of Emilia Romagna.

TWC(\(\theta\)) represents a further level of evolution over time, developed from TWC(\(\gamma\)) as the communication and interaction between the observed events stabilizes to become highly organized (Fig. 4 — bottom left). The epidemic shows stopping. TWC(\(\theta\)) can also provide a hypothesis on the diffusion path of epidemic in Italy, according to the data processed (Fig. 5). Through the use of a Markov chain it is possible to generate a directed weighted graph of how each city was infected in prevalence from which other city (Fig. 6).

The graph shown in Fig. 6 shows a possible network of influences, reciprocal or not, of contagions occurred in Italy. Although the largest number of cases occurred around the city of Lodi, for the system, Cremona seems to play a more
Fig. 4. Beta Map: The actual risk active area of the epidemics up to February 26th, 2020 (top left). Gamma map: The possible next evolution of the epidemic (top right). Theta map: Hypothesis of long-term evolution of the epidemic (bottom left). Iota map: Hypothesis about the final evolution of the epidemic (bottom right).

Fig. 5. Possible diffusion path of the epidemic in Italy (theta paths).
Fig. 6. Influences among the cities infected. City A → City B means the infection in the City A comes from the City B. In this graph also the hypothesized outbreak is considered (TWC alpha point).

central role. Cremona and TWC alpha, which we remember to be the estimated outbreak, highlighted in red within the graph, mutually affect each other. The epidemic then spreads to other areas of northern Italy. The southernmost areas of Italy, Rome and then Palermo, seem to have been involved through the cases of Rimini and Ancona, which were mutually infected.

According to the hypothesis of the TWC iota (Fig. 4 — bottom right), towards its final stages the epidemic will tend to return to the areas from which it originated but slightly further north.

5. Discussion

TWC alpha map seems to have identified an outbreak zone rather close (about 40 km) to the currently estimated epicenter in the municipality of Codogno (Fig. 2). Also the TWC theta elaborations seem to suggest an area slightly further south-east approaching Cremona. By the way it is not entirely sure that Codogno is actually the main outbreak of the epidemic as the Italian epidemiologists themselves talk about the presence in Codogno of Case 1 while Case 0 has not yet been found. The result seems very interesting especially if you consider that the system only takes into account the coordinates of the events, without knowing the frequency for each coordinate. It should be noted that, until February 26th, the city of Lodi counted more than 100 cases while Palermo only 1.

The Beta map (Fig. 4 — top left), which indicates the probability distribution of contagion episodes at the time the data are collected, shows how the “hot” zone was, as expected, to be mainly located in northern Italy and extending towards the Emilia-Romagna region.

The Gamma map (Fig. 4 — top right), which corresponds to the hypothesis of propagation in the immediate future, seems to indicate a certain stability of the epidemic phenomenon.

The Theta map (Fig. 4 — bottom left) instead would seem to indicate a certain regression of the phenomenon that, in the “anterior” future, could return to concentrate in the Lodi area. This trend seems to be confirmed also by the Iota map (Fig. 4 — bottom right) which still reduces the diameter of the red zone.

TWC Iota shows some very interesting aspects. The areas with the most intense red color show a peak of infectivity in the northernmost area of Italy.

6. Conclusion

Even considering only very few data such as the latitude and longitude of cities where at least one case of COVID-19 was detected, it was possible to make analyses with interesting results. This shows that, even in the presence of extremely poor data, if you have sufficiently powerful tools, it is possible to carry out real data mining by actually determining new useful information.

In this case, it was possible to obtain the coordinates and heat map of the area considered to be the outbreak of the epidemic. This point appeared less than 40 km from the location of Codogno, currently considered central for the spread of the virus in Italy. Through the TWC beta, gamma and theta, it was possible to build a prediction for the near future and the future of the future. What we do not know yet is the “delta time” between one prediction and another. For this reason, the monitoring activity continues over time. TWC Iota shows how the maximum level of infectivity seems to correspond
to northern Italy without spreading too much to the rest of the Italian peninsula. Indeed, at the time of the submission, i.e. April 2020, there was no massive expansion of contagions in the central-southern areas of Italy.

CRediT authorship contribution statement

Paolo Massimo Buscema: Ideation of the mathematics of the algorithms used, Software implementation, Data management and processing, Application of artificial adaptive systems, Analysis of results, Manuscript preparation. Francesca Della Torre: Data collection, Data preparation, Analysis of results, Manuscript preparation. Marco Breda: Analysis of results, Manuscript preparation. Giulia Massini: Analysis of results, manuscript review. Enzo Grossi: Analysis of results, manuscript review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2020.124991.

References

[1] https://www.iss.it/coronavirus, 2020.
[2] Confindustria report, 2020, https://www.confindustria.it/wcm/connect/01f1ad4b-d609-4728-b975-9e2d4a184e01/Italian+economic+outlook+2020/textunderscoresummary+and+main+conclusions+310320\textunderscoretextunderscoreconfindustria.pdf?mod=ajperes&cacheid=rootworkspace-01f1ad4b-d609-4728-b975-9e2d4a184e01-n4nczg.
[3] M. Buscema, G. Massini, P.L. Sacco, The topological weighted centroid (TWC): A topological approach to the time-space structure of epidemic and pseudo-epidemic processes, Physica A 492 (2018) 582–627.
[4] M. Buscema, E. Grossi, A. Bronstein, W. Lodwick, M. Asadi-Zeydabadi, R. Benzi, F. Newman, A new algorithm for identifying possible epidemic sources with application to the german escherichia coli outbreak, ISPRS Int. J. Geo-Inf. 2 (1) (2013) 155–200.
[5] A.C. Bronstein, M. Buscema, A. Esfahani, W.A. Lodwick, Grossi, locating the source of public health events using intelligent adaptive systems: 2011 United States listeriosis outbreak linked to whole cantaloupes, Clin. Toxicol. 51 (7) (2013) 625–626, Informa Healthcare.
[6] M. Buscema, P.L. Sacco, G. Massini, F. Della Torre, M. Brogi, M. Salonia, G. Ferilli, Unraveling the space grammar of terrorist attacks: A TWC approach, Technol. Forecast. Soc. Change 132 (2018) 230–254.
[7] P.M. Buscema, G. Ferilli, C. Gustafsson, P.L. Sacco, The Complex Dynamic Evolution of Cultural Vibrancy in the Region of Halland, Sweden, International Regional Science Review, 2019, 0160017619849633.
[8] P.M. Buscema, F. Della Torre, Novel applications of spatial mapping to chemical or biological outbreaks, in: Chemical Health Threats, 2018, pp. 64–95.
[9] M. Buscema, M. Asadi-Zeydabadi, W. Lodwick, A. Nde Nembot, A. Bronstein, F. Newman, Analysis of the ebola outbreak in 2014 and 2018 in west africa and congo by using artificial adaptive systems, Appl. Artif. Intell. (2020) 1–21.
[10] https://github.com/pcm-dpc/COVID-19 2020.
[11] P.J. Brantingham, P. L. Brantingham (Eds.), Environmental Criminology, Sage Publications, Beverly Hills, CA, 1981, pp. 27–54.
[12] P.J. Brantingham, P.L. Brantingham, Patterns in Crime, Macmillan, New York, 1984.
[13] Rossmo D.K., Geographic Profiling, CRC press, 1999.
[14] M. O’Leary, A new mathematical technique for geographic profiling, in: The NJ Conference, Washington DC. pp. 17–19.
[15] http://www.trovanorme.salute.gov.it/norme/dettaglioAtto?id=73461.