Open Peer Review on Qeios

When a Cluster Is a Cluster

Enzo Grossi

Funding: No specific funding was received for this work.
Potential competing interests: No potential competing interests to declare.

Abstract

The study of epidemic spread has generally relied on the description of certain number of cases of an infectious diseases like COVID-19 in relation to time occurrence of disease manifestations rather than to the exact place of occurrence. In recent times, computer generated dot maps have facilitated the modeling of the spread of infectious epidemic diseases either with classical statistics approaches or with artificial ``intelligent systems''. When new cases occur in relatively distant locations, it is very difficult to determine whether they constitute a cluster. The identification of the spatial clustering should be the first step when developing effective policies to manage and control any new epidemic.

Enzo Grossi[*]

Villa Santa Maria Foundation, Tavernerio, Italy

[*]**Correspondence**: Enzo Grossi M.D. Scientific Director, Villa Santa Maria Foundation, Via IV Novembre, 22038 Tavernerio (CO), Tel. 0039031426042, Email: enzo.grossi@bracco.com.

Keywords: Cluster; COVID-19; epidemic; artificial intelligence.

Introduction

The term "cluster" refers to an aggregation of cases grouped in space and/or in time that are suspected to be greater than the number expected, even though the expected number may not be known [1]. In this brief essay I will refer mainly to spatial clusters.

Several breakthroughs and triumphs in infectious disease control have resulted from the epidemiologic evaluation of spatial clusters of cases. Well-known examples include the epidemic of cholera in London in the 1850s [2], the investigation of cases of pneumonia at the Bellevue-Stratford Hotel in Philadelphia in 1976 [3], and the report in 1981 that seven cases of Pneumocystis carinii pneumonia had occurred among young, homosexual men in Los Angeles [4].

The basic premise of spatial epidemiology, is that epidemic spread does not occur randomly. Starting to gather data and tie them down to their location can help to see relationships between things that might not be obvious from looking at graphs or tables. Computerized dot maps constitute an attractive way to achieve this.

In the case of Coronavirus SARS-CoV-2 when a cluster refers to a precise and tight spatial entity like workers in a slaughterhouse, friends attending a party, people hosted in a navy, its identification is straightforward. A significant COVID-19 cluster is generally recognized when there are ten or more cases connected through transmission and who are not all part of the same household. The cluster includes both confirmed and probable cases. An especially concerning cluster was the 2 ½-hour choir rehearsal in Washington State, after which 45 of 60 members in attendance were diagnosed with COVID-19 or had compatible symptoms, including 3 hospitalizations and 2 deaths [5].

On the contrary when the geographical distance among new cases tends to enlarge and a clear connection among cases in missing, cluster definition becomes uncertain.

A huge amount of literature has been produced on COVID 19 with more than 400.000 papers. Only a minority of these concern geographical spatial distribution of the infection, and almost all are coming from China. All papers rely on aggregate data referred to a county, a region, or a state in a federative nation.

Just to give an idea in the early phase of COVID-19 spread in China as number of cases was increasing nearly 1000 cluster cases have been reported [6].

From an epidemiological point of view, understanding the clustering of individual cases in space to identify potential risk factors or outbreak identification, requires methods that are not limited by pre-defined administrative boundaries that may cause a pre-selection bias. This is especially true when the choice of spatial boundaries for examining health risks have important political and economic repercussions [7].

Statistical approaches to cluster recognition

A variety of statistical tests have been developed for the identification of clusters[8][9][10]. These tests can be classified based on whether they detect the presence of clustering or the actual location of clusters. Nearest-neighbor test, autocorrelation, Cuzick-and-Edwards' test, and the spatial scan statistics have been applied in the study of epidemics, especially in veterinary field [11].

There are no completely satisfactory methods for determining the number of population clusters for any type of cluster analysis [12][13][14]. Ordinary significance tests, such as analysis of variance F tests, are not valid for testing differences between clusters. Since clustering methods attempt to maximize the separation between clusters, the assumptions of the usual significance tests, parametric or nonparametric, are drastically violated.

An approximate nonparametric test for the number of clusters has been implemented in the MODECLUS procedure by Sarle and Kuo [15].

This test sacrifices statistical efficiency for computational efficiency. This method has the following useful features:

- No distributional assumptions are required.
- The choice of smoothing parameter is not critical since you can try any number of different values.
- The data can be coordinates or distances.
- The power is high enough to be useful for practical purposes.

Another popular test employed in spatial epidemiology is Moran index. In statistics, Moran's*I* is a measure of spatial autocorrelation developed by Patrick Alfred Pierce Moran [16][17]. Spatial autocorrelation is characterized by a correlation in a signal among nearby locations in space. Spatial autocorrelation is more complex than one-dimensional autocorrelation because spatial correlation is multi-dimensional (i.e. 2 or 3 dimensions of space) and multi-directional. The value of *I* can depend quite a bit on the expert assumptions built into the spatial weights matrix. A common approach is to give a weight of 1 if two zones are neighbors, and 0 otherwise, though the definition of 'neighbors' can vary. Another common approach might be to give a weight of 1 to *K* nearest neighbors, 0 otherwise. The need to express an “a priori” assumption can pose problems when a theory behind is lacking. Considering that in presence of sparse cases the nature of potential clusters is vague and not well defined, a fuzzy approach can be of value.

The binary character of partitions in fact is not always a convincing representation of the structure of data. Clusters may not be well separated for a variety of reasons, and in the living systems situations of partial membership occur quite often. With classical crisp clustering methods like K-Means, PAM, and CLARA each object (case) is assigned to exactly one cluster. For instance, an object lying between two clusters must be assigned to one of them.

FANNY for example is a soft clustering algorithm, where each node in the graph is associated with a membership coefficient, indicating degree of belongingness of each node to different clusters [18] and allowing to infer the structure of a cluster using a fuzzy approach.

Managing the uncertainty of the early stages of an epidemic

The value of mapping and Geographic Information Systems (GIS) has gained popularity among public health professionals to help in drawing disease maps more precisely, in understanding how a disease like COVID-19 spreads. Mathematical models applied to disease maps can help in distinguishing the low and high risk areas, as well as in handling of case clusters and formulation of hypothesis about the source of infection and analysis of data.

A special case is when the exact number of cases in a given location is uncertain, as can happen in the very early phase of a pandemic. Through artificial intelligence algorithm it is possible to treat the data with “at least one case approach”. Our group has been active in the last years in applying novel computational methods to a variety of epidemic contexts [19][20][21]

As a demonstration, our group analyzed the 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 [22].

The algorithm used is based on geographic profiling using a topological approach. One of the advantages of this algorithm, called Topological Weighted Centroid (TWC) [23][24][25][26], 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 algorithm produced the coordinates and heat map of the area considered to be around the outbreak of the epidemic. This area resulted very close to the location of Codogno (the centroid of the heat map was only 40 km from Codogno), a city which has been considered the starting point of the epidemic in Italy. Through other functions of TWC it was possible to build a prediction for the near future and the future of the future. According to these parameters the maximum level of infectivity resulted to correspond to northern Italy without spreading too much to the rest of the Italian peninsula. This finding was confirmed three months later, when the paper, submitted in February 2020, has been accepted for publication.

In summary, today several statistical methods allow to establish the existence of geographical clusters, some of them with a bottom-up data driven approach.

In the last three years we have been overwhelmed by curves, graphs, histograms focusing temporal trends in epidemic growth. Almost never we have seen little dots on a map pointing the exact location of new cases, an information particularly important in the early and in the late phase of an epidemic.

The effectiveness of surveillance systems in differentiating a forthcoming outbreak from just sporadic cases could have important implications for analytical studies that assume independence among cases.

The identification of the spatial clustering should be the first step when developing effective policies to manage and control any new epidemic. The big question is why scientific community is ignoring the fundamental value of cases exact location.

References

1. ^*https://www.cdc.gov/mmwr/preview/mmwrhtml/00001797.htm*
2. ^*Snow J. Snow on cholera. Hafner: New York, 1965.*
3. ^*Fraser DW, Tsai TR, Orenstein W, et al. Legionnaires' disease: description of an epidemic of pneumonia. N Engl J Med 1977;297:1189-97.*
4. ^*CDC. Pneumocystic pneumonia--Los Angeles. MMWR 1981;30:250-2. 4. Waxweiler RJ, Stringer W, Wagoner JK, et al. Neoplastic risk among workers exposed to vinyl chloride. Ann N Y Acad Sci 1976:271:40-8.*
5. ^*Waldrop T, Toropin K, Sutton J. 2 Dead from coronavirus, 45 ill after March choir rehearsal. Published (updated) April 2, 2020. Accessed May 28, 2020. www.cnn.com/2020/04/01/us/ washington-choir-practice-coronavirus-deaths/*

index.html

6. ^*Special Expert Group for Control of the Epidemic of Novel Coronavirus Pneumonia of the Chinese Preventive Medicine Association. An update on the epidemiological characteristics of novel coronavirus pneumonia (COVID-19). Chin J Epidemiol. 2020;41(2):139-144*
7. ^*D. L. Pearl, M. Louie, L. Chui, K. Doré, K. M. Grimsrud, D. Leedell, S. W. Martin, P. Michel, L. W. Svenson, S. A. Mcewen. The use of outbreak information in the interpretation of clustering of reported cases of Escherichia coli O157 in space and time in Alberta, Canada, 2000–2002 Epidemiol Infect. 2006 Aug; 134(4): 699–711.*
8. ^*Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Statistics in Medicine 1995; 14: 799–810.*
9. ^*Kulldorff M, Hjalmars U. The Knox method and other tests for space-time interaction. Biometrics 1999; 55: 544–552.*
10. ^*Bell BS. Spatial analysis of disease – applications. Cancer Treatment Research 2002; 113: 151–182.*
11. ^*Ward MP, Carpenter TE. Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. Preventive Veterinary Medicine 2000; 45: 257–284.*
12. ^*Everitt, B.S. (1979), "Unresolved Problems in Cluster Analysis," Biometrics, 35, 169–181.*
13. ^*Hartigan, J.A. (1985), "Statistical Theory in Clustering," Journal of Classification, 2, 63–76.*
14. ^*Bock, H.H. (1985), "On Some Significance Tests in Cluster Analysis," Journal of Classification, 2, 77–108.*
15. ^*Sarle, W.S and Kuo, An-Hsiang (1993), The MODECLUS Procedure, SAS Technical Report P-256, Cary, NC: SAS Institute Inc.*
16. ^*Moran, P. A. P. (1950). "Notes on Continuous Stochastic Phenomena". Biometrika. 37 (1): 17–23. doi:10.2307/2332142. JSTOR 2332142.*
17. ^*Li, Hongfei; Calder, Catherine A.; Cressie, Noel (2007). "Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model". Geographical Analysis. 39 (4): 357–375. doi:10.1111/j.1538-4632.2007.00708.x.*
18. ^*Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, Inc.*
19. ^*Buscema M, Grossi E, Breda M, Jefferson T. Outbreaks source: A new mathematical approach to identify their possible location. Physica A 388 (2009) 4736-4762*
20. ^*Buscema, M., Grossi, E., Bronstein, A., Lodwick, W., Asadi-Zeydabadi, M., Benzi, R., & Newman, F. (2013). A new algorithm for identifying possible epidemic sources with application to the German Escherichia coli outbreak. ISPRS International Journal of Geo-Information, 2(1), 155-200.*
21. ^*Bronstein, AC; Buscema, M; Esfahani, A; Lodwick, WA; Grossi, Locating the source of public health events using intelligent adaptive systems: 2011 United States listeriosis outbreak linked to whole cantaloupes, CLINICAL TOXICOLOGY, 51,7,625-626, INFORMA HEALTHCARE 2013.*
22. ^*Buscema PM, Della Torre F, Breda M, Massini G, Grossi E. COVID-19 in Italy and extreme data mining. Physica A. 2020 Nov 1;557:124991. doi: 10.1016/j.physa.2020.124991. Epub 2020 Jul 25. PMID: 32834435; PMCID: PMC7382358.*
23. ^*Buscema, M., Massini, G., & Sacco, P. L. (2018). The Topological Weighted Centroid (TWC): A topological approach to the time-space structure of epidemic and pseudo-epidemic processes. Physica A: Statistical Mechanics and its Applications, 492, 582-627.*

24. ^*Buscema, M., Grossi, E., Bronstein, A., Lodwick, W., Asadi-Zeydabadi, M., Benzi, R., & Newman, F. (2013). A new algorithm for identifying possible epidemic sources with application to the German Escherichia coli outbreak. ISPRS International Journal of Geo-Information, 2(1), 155-200.*
25. ^*Bronstein, AC; Buscema, M; Esfahani, A; Lodwick, WA; Grossi, Locating the source of public health events using intelligent adaptive systems: 2011 United States listeriosis outbreak linked to whole cantaloupes, CLINICAL TOXICOLOGY, 51,7,625-626, INFORMA HEALTHCARE 2013.*
26. ^*Buscema, M., Sacco, P. L., Massini, G., Della Torre, F., Brogi, M., Salonia, M., & Ferilli, G. (2018). Unraveling the space grammar of terrorist attacks: A TWC approach. Technological Forecasting and Social Change, 132, 230-254.*