Geographic Monitoring for Early Disease Detection (GeoMEDD)

Andrew Curtis
Case Western Reserve University
Jayakrishnan Ajayakumar
Case Western Reserve University https://orcid.org/0000-0001-9564-7728
Jacqueline Curtis (jacqueline.curtis@case.edu)
Case Western Reserve University https://orcid.org/0000-0001-6046-6476
Sarah Mihalik
University Hospitals Health System
Maulik Purohit
University Hospitals Health System
Zachary Scott
University Hospitals Health System
James Muisyo
University Hospitals Health System
James Labadorf
University Hospitals Health System
Sorapat Vijitakula
University Hospitals Health System
Justin Yax
University Hospitals Health System
Daniel W. Goldberg
Texas A&M University

Method Article

Keywords: COVID–19, spatial clusters, syndromic surveillance, geography, geographic information science

Posted Date: July 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-39862/v1

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2020. See the published version at https://doi.org/10.1038/s41598-020-78704-5.

Abstract

Identifying emergent patterns of coronavirus disease 2019 (COVID-19) at the local level presents a geographic challenge that is not ideally addressed using traditional spatial clustering approaches. The need is not only to identify statistically significant spatial patterns, but also develop a surveillance system to identify small numbers of cases emerging in vulnerable settings. This paper presents an approach that has been used to provide near-real time assessments of emergent disease to guide a hospital system's intervention strategy: Geographic Monitoring for Early Disease Detection (GeoMEDD). Through integration of a spatial database and two types of clustering algorithms, GeoMEDD has the flexibility to scale in terms of required minimum members, connecting distance between cases, and time frame under consideration. GeoMEDD, has proven effective in revealing different cluster influencers and accelerators that give insight as to why the cluster exists where it does, and why it expands.

Introduction

Geographic aspects of coronavirus disease 2019 (COVID from this point forward) have been widely reported in popular and scientific media with maps visualizing areas experiencing the greatest intensity, patterns of risk, or characteristics of diffusion (e.g., Boulos and Geraghty 2020; Dong, Du, and Gardner 2020; Fowler 2020; Jia et al. 2020; New York Times 2020; Oliver et al. 2020). At the federal and state levels, county or zip code maps of positive tests, hospitalizations and mortality inform professionals and the public alike using spatially enhanced dashboards. However, the map is not just a means of communication, it can also be used as a tool for response. For example, locations of positive cases in conjunction with different types of congregate housing can give a vital first impression of where to mobilize resources. Making sense of those locations in an efficient manner is the challenge—a challenge that is not overcome with traditional spatial analytical approaches.

Disease data are often aggregated by zip codes, counties, or other larger units to protect patient privacy. Sometimes, for more granular data, hot spot mapping is employed (Anselin 1995; Ord & Getis 1995; Kulldorff 1997), usually utilizing a geographic information system (GIS) to tease out patterns in the surveillance. These analytical methods continually evolve through conceptual advances such as incorporating context or developing new field methods and the associated spatial software (e.g., Kwan 2012, Curtis et al. 2016). However, COVID presents a novel spatial health challenge because it not only transmits at a high rate, it also can have a long pre-symptomatic time period in which it can spread. As such, it significantly impacts vulnerable populations such as those that live in congregate housing, the elderly, and people with underlying comorbidities. These vulnerabilities help hasten the propagation of the disease quickly. Without immediate intervention, this propagation can rapidly tax capacity and resources of local hospitals. In order to intervene and reduce the propagation of the disease, patient location details are required at the level of buildings and street sections, and by temporal aggregation of no more than a day so that interventions can be quickly and effectively targeted. Because in most western societies data sources are rich in detail, even though there remain challenges posed by uneven testing practices (e.g., Hooper, Nápoles, & Pérez-Stable (2020)), a quick and effective response to developing hotspots is

achievable. Indeed, there are no mechanical reasons why positive Covid tests cannot be analyzed in real time, or at least as quickly as test results are uploaded to hospital systems. Ideally, COVID positive tests should be analyzed as quickly as possible once resulted. Severity of the associated symptoms, previous medical histories of the patient, and the background neighborhood "risk" can all be added in near real-time to give better perspective on the overall risk of any individual.

Traditional spatial approaches to disease cluster detection often rely on having reference data as a denominator and/or require reaching statistical significance threshold for the cluster to be identified as such. While having minimum denominators is important for retrospective analyses of disease, here the interest is in the *first* case in a post-acute care home, or the temporal pattern of emerging positives in an apartment complex, or how houses in connected streets "emerge" suggesting a local transmission mechanism.

Such cases and patterns can be identified in the typical continuous data stream of health systems and analysis is focused on emerging patterns and telltale signals so that physical distancing strategies (such as quarantine or facility lock downs) can be mobilized to reduce further transmission. It is well known that this virus spreads effectively in congregate housing and to a disastrous effect (Gardner, States, & Bagley. 2020; Yancy 2020). Therefore, knowing when cases emerge inside, or even proximate, can literally be used to activate immediate intervention and prevention strategies to save lives.

Results

A team of spatial scientists in the GIS Health & Hazards Lab in the School of Medicine at Case Western Reserve University developed GeoMEDD. This team then partnered with data scientists, clinicians, and leadership from the University Hospitals Health System in Cleveland, Ohio to provide a location based early warning system for the Northeast Ohio region. GeoMedd consists of three new, but interlinked cluster types: sentinel clusters, micro-clusters, and neighborhood clusters.

New Conceptualizations of Disease Clusters

Conceptually, for the novel cluster criteria, if X members fall inside Y distance then a cluster is identified. These clusters are not static, but can grow when a new positive case is identified within the set distance of a cluster member. For example, for the *sentinel cluster,* here reported at 100m and 2 members, if two positive cases (mapped by their residence) appear within 100m of each other, then a *sentinel cluster* centroid is calculated. If another positive case falls within 100m of either of these cluster members, then the cluster grows. This is repeated until no further members are added. A polygon bounding box is automatically calculated using all the exterior members. This *sentinel cluster* is the first early warning, with results sometimes including members within the same household, living within the same congregate housing, or neighbors on the same or connecting street. While potentially the creation of these clusters could be continuous as is the incoming test result data, the reality of validating the geocoding, and then fully interpreting and prioritizing each cluster means that this happens at least once daily, or in response to other syndromic surveillance leading indicators, such as an increase in ambulance runs. The

interpretation of findings includes whether the cluster coincides with obvious vulnerabilities, such as cases being inside a post-acute care home, or other forms of congregate housing, or being residential but proximate to either of those. The sentinel cluster provides an early warning for geographically targeted intervention, for example in the form of directing intercept teams.

The second type of signal, the *micro cluster,* uses the same conceptual logic but extends the distance to 500m with a minimum membership of 5 cases (Figure 1). The *micro cluster* provides an indication of either a more substantial congregate housing concern, such as in a public housing complex, or the indication of community spread where multiple positives, possibly outside the 100m distance bands of the *sentinel cluster,* show that a section of a neighborhood has become “hot” possibly due to a shared resource such as a grocery store.

Sometimes a pulse of positive tests can result in a simultaneous new *sentinel* and *micro cluster,* usually in the same building where, for example, all members of a family test positive at the same time. In other examples, the complexity of the underlying diffusion pattern is revealed as a single *sentinel cluster* continues to grow, or merges together with others. It is also possible for a *micro cluster* to contain no *sentinel clusters.* When used together, these clusters can provide a vital insight into the granular drivers of an outbreak, and where spread is likely to happen next.

The third signal, the *neighborhood cluster,* again uses the same logic but now captures the next level of the aggregation hierarchy to identify more widespread community spread. The *neighborhood cluster* has a distance bandwidth of 1000m and a10-member minimum. As with the *micro cluster,* this growth can reveal the various diffusions at play, possibly with *micro* and *sentinel clusters* merging, with other positive case addresses not falling into either cluster type but “filling in” the geographic gaps between. For this cluster type, the automatic polygon creation around the exterior members provides a useful visual to be shared with health departments to target a region for specific intervention. By comparing this boundary extent over a series of days, the growth or contraction of community spread can be assessed, with vulnerable housing, the typical comorbidities of those falling inside, locations of pending tests awaiting results, building footprints, parcel and tax assessor information, aerial photography, even measures of social vulnerability all providing additional context as to potential risk and where intervention should be guided. In an example that has repeated multiple times in the Cleveland area, sentinel clusters have grown to form micro clusters which turn into neighborhood clusters containing multiple apartment complexes. Just as before, reading the interactions of the three cluster types, along with reference to Google Earth Imagery and Google Street View, revealed not only where cases were found, but where they were likely to spread to next. Those points of vulnerability, especially proximate post-acute care homes, would be identified and reported.

For all these cluster types, the distance and minimum cluster membership can be varied to explore appropriate thresholds for different environments. For example, in rural environments physical distance between cases is greater, with neighbors along the same stretch of road being more readily identified using a 500m micro cluster with 2 members. In this example all urban or dense residential areas are

excluded as this is a “rural” *sentinel cluster.* The other consideration is time, as initial cluster mapping involved all cases, but as weeks passed, more emphasis was needed on the current situation, and so only positives tests within a 21-day period were incorporated. The daily analysis of these clusters would reveal that clusters could both grow (new positives added) and lose members (older than 21 days) simultaneously.

The continual daily surveillance of COVID using these new cluster types has created a geographic way of conceptualizing the local components of the epidemic. Is the cluster in a post-acute care home, or other congregate housing, or are there other high density living nearby? If so, is this a new cluster? Do we know about the facility or is it relatively unknown? Are there connecting paths or roads between “hot” buildings? For community spread, are the positive tests at a single residential address or spread across multiple proximate houses? Is there an ongoing temporal pattern to the cluster, or is there just a short burst and then stagnation? While contact tracing remains a vital part of contextualizing disease spread, the clusters described here reveal geographic patterns and aspects of diffusion in near real time—far in advance of other public health monitoring in the region. While more traditional methods of spatial analysis, such as Sat Scan analysis (Kulldorff 1997; Kulldorf et al. 2005) might reveal similar coarser neighborhood patterns, the hierarchical geographic insight gained from the three cluster types has proven translational. Not only will it be a vital tool to monitor emergence during recovery, but the system is ready for any subsequent “wave”. And in the interim, the same clustering can be switched to other hospital data, such as identifying patterns in asthma exacerbations, child injury or overdoses.

Online Methods

Operationalizing from Data Inflow to Intervention

The typical cluster analysis described here involves correcting geocoded addresses in a continuous feed of COVID–19 test results. Once these had been corrected, a set of queries made possible by using a specially developed spatial database, identified *sentinel, micro* and *neighborhood* Agglomerative and Density based clusters (see the subsequent section for definitions) for both the entire period and only for those testing positive in the previous 21 days. Output tables were automatically generated to visualize changes in patterns across the cumulative totals for each day, with different guideline summary statistics added. The most common interrogation of these tables included adding a final column where each cluster defaulted to a “0” if no change had occurred from the daily total to two days previously, a “1” if a new cluster had formed, and “2” if the number of cluster members had increased. A convex hull polygon is automatically created by joining all of the exterior cluster members, with a buffer polygon being used (with a radius of the cluster type) if all members reside at the same address. The centroid of each polygon is also automatically generated with its latitude and longitude being added to the output table. Using a Geographic Information System (GIS), each centroid and bounding polygon is mapped onto any preferred underlay such as a street network, or social vulnerability layer.

Spatial Database Development

In order to achieve this near real time clustering, three operational advances were required. First, a spatial database was built that would ingest real time COVID test results from across the entire hospital system (Figure 2). The spatial database ingests multiple health, emergency responder, social vulnerability data, and different congregate living spatial such as post-acute care homes, correctional facilities, and subsidized housing.

More specifically, the spatial database was built using PostgreSQL, which is a free and open-source relational database management system (RDBMS) along with PostGIS which is an open source software program that adds support for geographical objects in PostgreSQL. Automatic batch jobs were written to pull locational data from the Health System's clinical and operational source systems, including COVID test data, encounter data from the Electronic Medical Record (EMR) system, Emergency Management Service (EMS) dispatches and previous addresses of those with Covid-related comorbidities that could be used to establish background neighborhood health risk. To contextualize the clinical data, various vulnerability measures such as Centers for Disease Control and Prevention (Social Vulnerability Index), and Health Resources & Service Administration (Area Deprivation Index) were also ingested allowing for the automatic spatial joining of layers.

Two families of clustering types were utilized for the sentinel, micro and neighborhood clustering:

Agglomerative Clustering

Agglomerative Clustering is a type of hierarchical clustering that builds nested clusters by adding individuals through a bottom up approach starting with each observation point as a seed to which other seeds are added (Müllner 2011). At the beginning each observation is deemed as a cluster (Figure 3). The merging between two seeds (in effect each being a "cluster") is based on a single linkage criterion (Figure 3) which minimizes the distance between the closest pairs of clusters , with a threshold distance (α) acting as a constraint to cluster merging. Clusters can grow (merge together) until the threshold is reached. The algorithm terminates after all points have been accounted for.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

A second more efficient clustering approach, *D*ensity-*B*ased *S*patial *C*lustering of *A*pplications with *N*oise (DBSCAN) utilizes an additional density-based restriction on the clustering algorithm (Ester et al., 1996). Here points are initially grouped that are closely packed together (high density areas) as clusters, with isolated points being identified as "noise". The DBSCAN algorithm uses a *core sample* concept comprised of *core* (points close to each other by a selected distance measure) and *non-core* or *boundary/fringe* members which are close to but not a *core (*Figure 4). The algorithm decides whether a point is a *core, non-core* or an *outlier* for a cluster. The maximum distance parameter (d_{max}) determines the threshold distance up to which two points are considered neighbors, and the minimum sample parameter (α) determines the minimum number of data points required to define observation as core member. Based on d_{max} and α, a *core* can be defined as a point which has at least α number of points (including itself) within a distance of d_{max}, while a *non-core* can be defined as a point that is not a *core* but still reachable

from a *core.* A cluster needs to have least one core member and can have any number of non-core members. An outlier is a point that is not reachable from any core, even though it might be reachable from a non-core. This is the biggest difference between the two cluster types as the agglomerative cluster can grow its boundary through outlier to outlier connections.

Various components of the DBSCAN algorithm are outlined in Figure 4. The thick black arrows indicate distances that are within the maximum distance parameter (d_{max}), while the dashed arrows indicate distances that are greater than d_{max}. The minimum sample parameter (α) is assigned as 5 for this case. The green colored points are core points having at least α neighbors (including itself) which are within a distance of d_{max}. The yellow colored is a non-core point as it is reachable only from a single point. The red point is an outlier as it is not reachable from any core points, and only reachable from a non-core point. All the green points (core members) along with the yellow point (non-core member) form a cluster while the red point is treated as an outlier.

To run the algorithm, a random point is selected, and its neighboring points are identified using the d_{max}. If the point has α number of neighbors within d_{max}, then it is marked as a *core* and the cluster formation procedure is started. All the neighbors for the *core* are combined together to form a cluster and the neighbor points are further checked to determine whether they are also *cores.* If a neighbor point is also a *core* then its neighbors are also added to the current cluster increasing the total membership. After cluster formation another random unvisited point is taken and the algorithm is repeated until all points have been visited. For this and the agglomerative clustering, the date for each cluster member is recorded to allow for a simultaneously constructed longitudinal analysis of the clusters. In this way the current cluster “map” can be compared against the previous Covid situation.

Declarations

Code Availability

A Scipy based working implementation for agglomerative and DBSCAN algorithm can be found in https://github.com/JayakrishnanAjayakumar/SyndromicSurveillance

Author Contributions

A. C. created the GeoMEDD framework and lead manuscript development. J. A. developed GeoMEDD code. J. C. contributed to application of GeoMEDD use case and manuscript development. D. W. provided geocoding services for GeoMEDD use case. All other authors contributed equally to application of GeoMEDD in a health system use case and in manuscript development. All authors reviewed the manuscript.

Competing Interests

The authors declare no competing interests.

References

1. Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis*, *27*(2), 93-115.
2. Boulos, M. N. K., & Geraghty, E. M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics.
3. Curtis, A., Curtis, J. W., Porter, L. C., Jefferis, E., & Shook, E. (2016). Context and spatial nuance inside a neighborhood's drug hotspot: Implications for the crime–health nexus. *Annals of the American Association of Geographers*, *106*(4), 819-836.
4. Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, *20*(5), 533-534.
5. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & others. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*, 226–231.
6. Fowler, G. 2020. Smartphone data reveal which Americans are social distancing (and not). Washington Post, March 24. Available online: https://www.washingtonpost.com/technology/2020/03/24/social-distancing-maps-cellphone-location/ Last accessed: 06/02/2020
7. Gardner, W., States, D., & Bagley, N. (2020). The coronavirus and the risks to the elderly in long-term care. *Journal of Aging & Social Policy*, 1-6.
8. Goldberg, D. W. (2011). Advances in geocoding research and practice. *Transactions in GIS*, *15*(6), 727-733.
9. Hooper, M. W., Nápoles, A. M., & Pérez-Stable, E. J. (2020). COVID-19 and racial/ethnic disparities. *Jama*. Available online: https://jamanetwork.com/journals/jama/article-abstract/2766098 Last Accessed: 06/04/2020
10. Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, *26*(6), 1481-1496.
11. Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R., & Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS medicine*, *2*(3).
12. Kwan, M. P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers*, *102*(5), 958-968.
13. Jacquez, G. M. (2012). A research agenda: Does geocoding positional error matter in health GIS studies?. *Spatial and spatio-temporal epidemiology*, *3*(1), 7-16.
14. Jia, J. S., Lu, X., Yuan, Y., Xu, G., Jia, J., & Christakis, N. A. (2020). Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature*, 1-5.
15. New York Times, 2020. Coronavirus Map: Tracking the Global Outbreak. Available online: https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html Last accessed:

06/02/2020

16. Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Delataille, S., De Nadai, M., ... & Colizza, V. (2020). Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Available online: https://advances.sciencemag.org/content/early/2020/04/27/sciadv.abc0764 Last accessed: 06/02/2020
17. Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, *27*(4), 286-306.
18. Texas, A. M. (2015). University Geocoder. *College Station, TX*.
19. Yancy, C. W. (2020). COVID-19 and African Americans. *Jama*. Available online: https://jamanetwork.com/journals/jama/article-abstract/2764789 Last Accessed: 06/04/2020
20. Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, *32*(3), 214-232.

Notes

1. CDC Covid Data Tracker: https://www.cdc.gov/covid-data-tracker/; Ohio Department of Health Covid-19 Dashboard: https://coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards/overview; Cleveland Confirmed and Probable COVID-19 Public Dashboard: http://www.clevelandhealth.org/

2. Adding these “risk factors” in real time is possible, but it would take collaboration with a health system to fine tune how this is integrated into workflow. This is a future step for this project.

3. Note that Kulldorf’s (2005) space-time permutation scan statistic that is not reliant on a reference population.

4. Another advantage of the sentinel cluster is that an overreliance on one address or name sometimes means cases are not correctly attributed to a post-acute care home location, the *sentinel cluster* helps remove this “fuzziness”. Geocoding accuracy and precision is, unfortunately, often a neglected aspect of spatial health research (Jacquez 2012). Using the Care Home example – a “campus” might include a variety of names, sometimes associated with different aspects of care. These buildings might also contain different addresses and apartments. The net result is the same general campus might have 10 or more addresses or combinations of address components, including the insertion of names and apartment identifiers. These can be corrected manually with sufficient attention, however the resulting coordinate in either the original or corrected geocode can vary considerably. The challenge is to have a tool precise enough to capture the singular risk posed by and to the campus, while not overly smoothing or aggregating positive cases over a larger area such as a census block group or neighborhood.

5. Due to variations in time frame, both in terms of reporting a day’s activity in the afternoon, or through a fuzziness in how quickly test results are obtained, experience proved that a two day comparison provided a more robust period to describe and report change.

6. The spatial database was also used to automatically join each positive address to the patient's past hospital record existing co morbidities, and whether or not the person was admitted into hospital (part of the hospital census meaning the symptoms were severe). In this way, each cluster can also be characterized by how currently and potentially severe a situation it presented. Likewise, the underlying social vulnerability measure was automatically joined to each address allowing for each cluster to be assessed in terms of neighborhood risks.

7. While some congregate housing layers are available at the state level, such as correctional facilities, certain care home types, and HUD apartments, there are many congregate housing "gaps", with availability (and quality) of these data varying by city and county.

8. There are four different types of linkage criteria including Ward (minimizes the sum of squared differences within all clusters), complete linkage (minimizes the maximum distance between pairs of clusters), average linkage (minimizes the average of the distances between all pairs of clusters), and single linkage (minimizes the distance between the closest pairs of clusters).

Figures

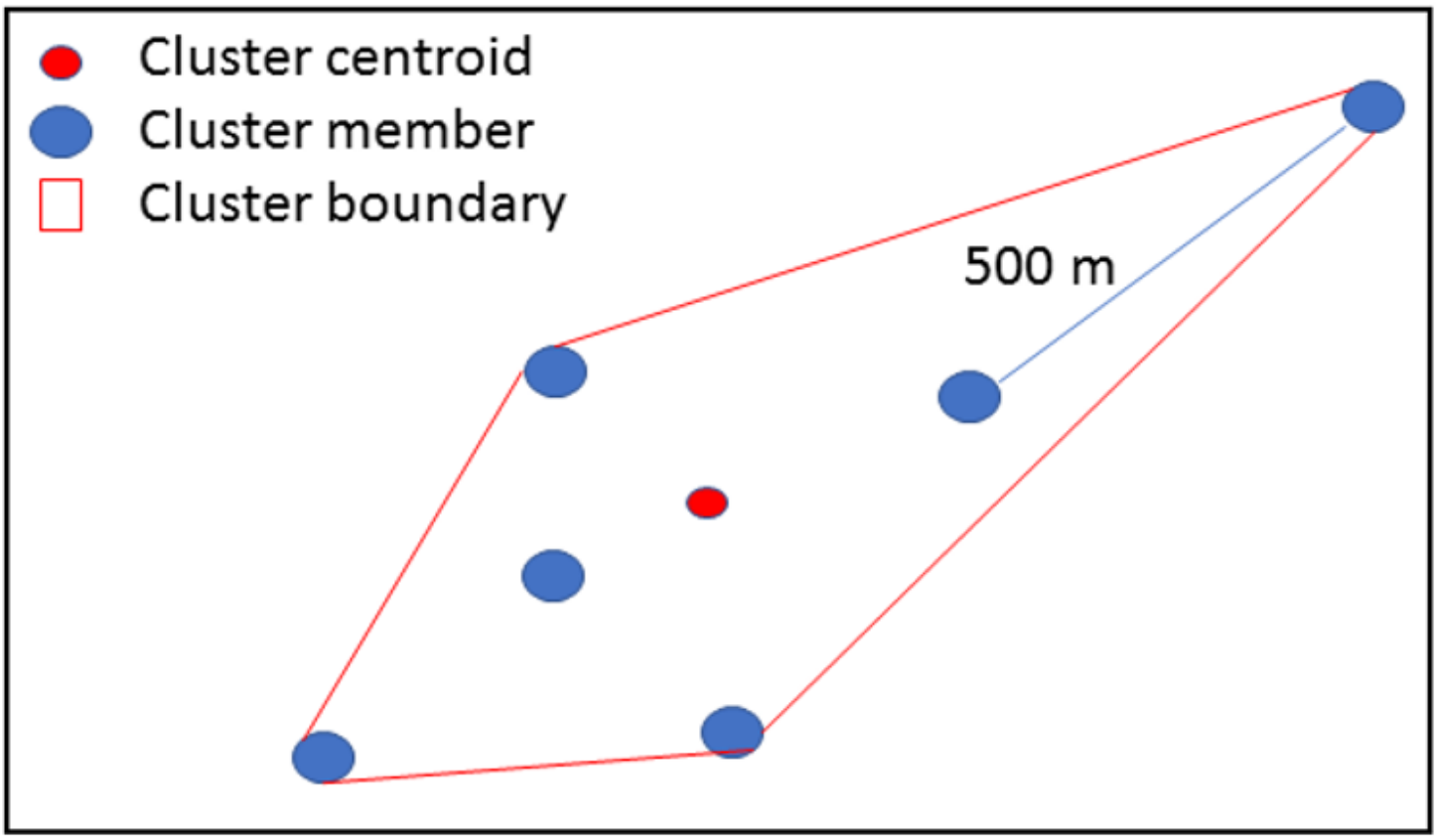

Figure 1

Example micro-cluster schematic

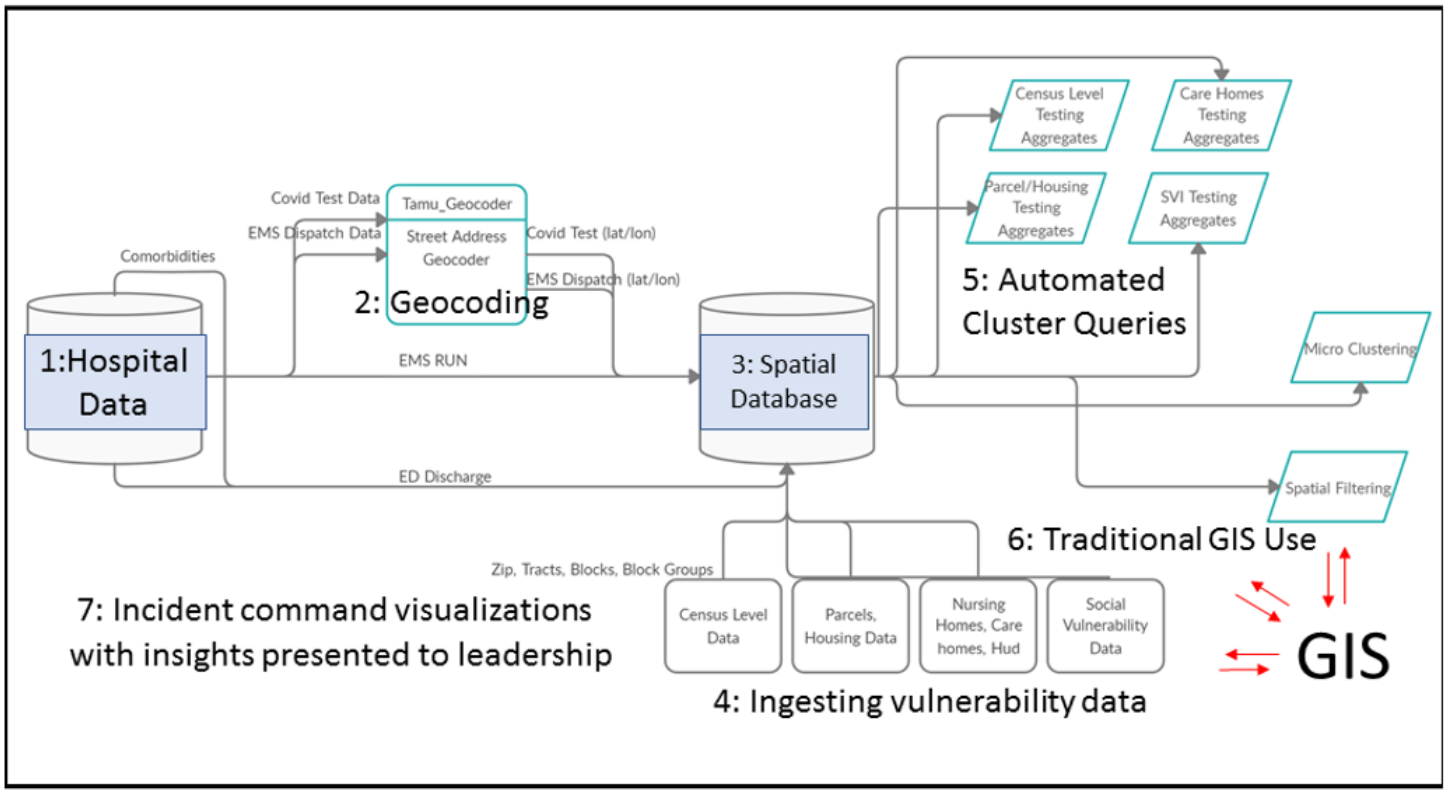

Figure 2

Schematic of the spatial database

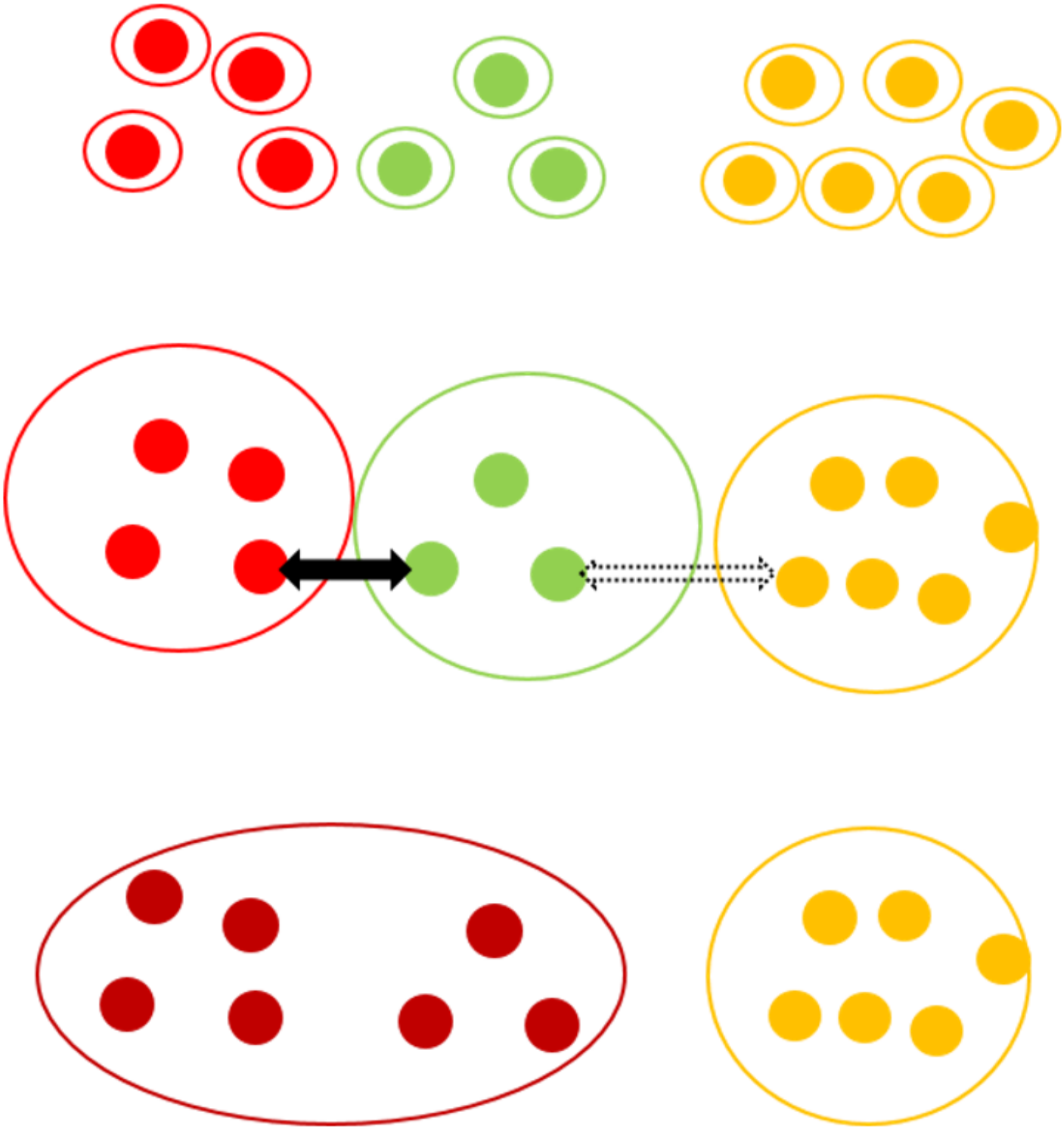

Figure 3

Demonstration of agglomerative clustering. The black arrow indicates a distance that is within the threshold distance (α), and the dashed arrow indicates a distance that is above the threshold.

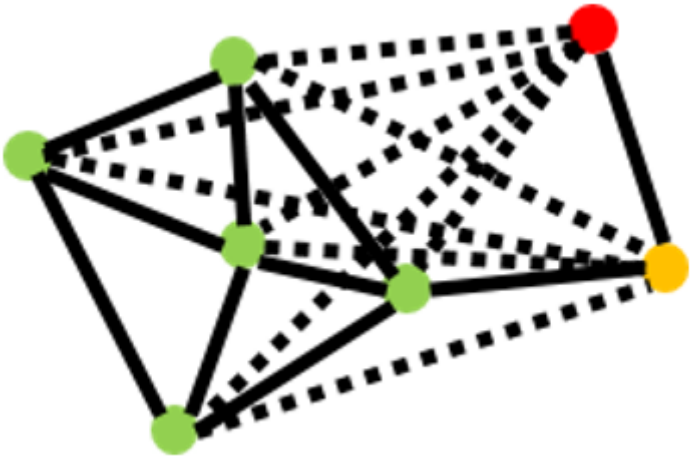

Figure 4

DBSCAN components. The green points indicate core-members, yellow indicates non-core members, and red indicates outlier. The dashed lines indicate distances that are greater than threshold distance (dmax) and the thick lines indicate distances that are with in dmax.