A clustering environment for real-time tracking and analysis of Covid-19 case clusters

Jayakrishnan Ajayakumar[†]
Population and Quantitative Health Sciences
Case Western Reserve University
Cleveland Ohio US
jxa421@case.edu

Andrew Curtis
Population and Quantitative Health Sciences
Case Western Reserve University
Cleveland Ohio US
ajc321@case.edu

Jacqueline Curtis
Population and Quantitative Health Sciences
Case Western Reserve University
Cleveland Ohio US
jxc1546@case.edu

ABSTRACT

Spatial surveillance systems can be an effective and efficient way to provide early spatio-temporal warning signals of infectious diseases outbreaks. The development of spatial surveillance systems that can handle near-real time hospital spatial data have proven critical for directing intervention strategies and for risk mitigation during the ongoing monitoring and response to COVID-19. GeoMEDD, a geographic monitoring tool, was developed for such near-real time assessment of emergent localized disease. As the response to COVID-19 changed, moving through phases such as increased scientific understanding, testing variability, vaccine availability and uptake, and new variants, so GeoMEDD has also evolved. Here we present two advances for GeoMEDD; a fully automated cluster environment with a spatial database at its heart and a cluster tracking module to classify clusters based on the transition state of the cluster lifecycle. Our detailed use case analysis shows that these advances have improved local and global cluster analysis, contextual information dissemination, monitoring emergence based on underlying spatial structure and cluster evolution analysis. We believe that the addition of the fully automated cluster environment to GeoMEDD would be particularly beneficial for health institutions as well as governmental health organizations for disease outbreak detection due to the efficiency in data ingestion and analysis, , while the addition of the cluster tracking module will advance research into the mechanics behind disease diffusion in space and time.

CCS CONCEPTS

• Information systems• Information systems applications• Data mining• Clustering

KEYWORDS

Hotspot/Clustering, Syndromic Surveillance, Covid-19, Cluster Diffusion

ACM Reference format:

Jayakrishnan Ajayakumar, Andrew Curtis and Jacqueline Curtis. 2021. A clustering environment for real-time tracking and analysis of Covid-19 case clusters. In *Proceedings of 2nd ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology (SpatialEpi'21), November 2, 2021, Beijing, China. ACM, New York, NY, USA, 9 pages.* https://doi.org/10.1145/3486633.3491091

1 Introduction

The spread of Covid-19 involves multiple spatial processes [1] and therefore a geographic approach including geospatial analytics is vital in understanding and responding to the disease [2]. Typically, this is achieved using traditional spatial statistical analysis or hotspot/cluster detection of disease data [3]. Spatial statistical methods often incorporate socio-economic and demographic risk factors [4, 5]. For example, utilizing a multivariate model, Andree Ehlert [6], in his work on determining the socio-economic determinants of Covid-19 in Germany, determined deaths are positively associated with average age, population density, and the number of people working with the elderly. Similarly, a statistical modelling study in the United States by Mollalo et al. [7] identified income inequality, median household income, proportion of black females, and the proportion of nurse practitioners to be positively associated with Covid-19 incidence. While these types of spatial statistical analysis are important to unpack the reasons why Covid-19 spreads as it does, real-time monitoring and response analytics are also required to address that spread as it occurs, in as close to real-time as possible. To this end spatial and spatio-temporal hotspot/cluster, detection and surveillance are vital to identify emerging patterns and support intervention strategies [8, 9].

The high reproductive number (R0) associated with Covid-19 [10] suggests that cases will cluster in space and time, which in turn means methods should capture both those emerging concentrations [11], and their next step diffusion [12]. A frequently employed Covid-19 cluster detection approach is the space-time scan statistic (SaTScan) [13] or variants thereof [14, 15, 16, 17]. For example, Rosillo et al. [18] used SaTScan to develop a real-time surveillance system to detect active clusters of COVID-19 in Spain. They

SpatialEpi'21, November 2, 2021, Beijing, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9119-1/21/11..$15.00
https://doi.org/10.1145/3486633.3491091

suggested that SaTScan based surveillance of COVID-19 can be particularly useful during a low-incidence scenario to help tackle emerging outbreaks [18]. Desjardins et al. [15], using prospective SaTScan and county level case data from Johns Hopkins University, were able to identify 'active' and emerging clusters across the entire United States for a particular time period and they also posited that the same approach could be run continuously for the ongoing surveillance of Covid-19 clusters.

The majority of the published Covid-19 clustering work use coarse data due to data access and confidentiality constraints. A few exceptions include Greene et al. [16] who developed a "SARS-CoV-2 percent positivity cluster detection system" for census tract aggregations and a SaTScan prospective space-time scan statistic. Even more granular, Ladoy et al. [19] used SaTScan to identify clusters of positive cases by residential location in the state of Vaud, Switzerland. Using cluster attributes such as the median age of individuals and total number of cluster members, they were able to identify that cluster size was positively related to the presence of individuals with high Covid-19 viral load.

One fundamental challenge is that while more traditional approaches to disease cluster detection will always be important for spatial researchers and epidemiologists, the lack of a near-real time data-to-analysis mindset is a barrier for operationalization in a hospital setting [20]. When responding to an outbreak, it is vital to not only learn about the spatial structure and pattern of the continuously evolving pattern, but to do so in a time frame that is scalable, and operationally appropriate so that hospitals or health departments can use outputs to inform response teams. To be effective, data also needs to be as granular as possible, ideally at the residential level, and the analysis should be able to analyze continuously inflowing updates. To achieve this requires developing novel cluster methodologies that can work at various spatial scales as well as developing spatial data infrastructures (SDI) that can handle hospital "big" data in real-time [1]. GeoMEDD was developed to solve this problem as a near real-time assessment of emergent disease suitable to guide a local intervention strategy [20]. Through the integration of a spatial database and clustering algorithm, GeoMEDD provides multiple spatial and temporal perspectives on a highly dynamic disease landscape using various space and time thresholds [20]. Initial work has shown that GeoMEDD is effective in revealing clusters at various spatial scales as well as giving insights about why a cluster exists at a particular location and how it evolves through space and time [20].

Even though GeoMEDD is currently being utilized as a syndromic surveillance tool by hospitals and public health organizations [21, 22], new strategies need to be developed as the disease situation changes. In this paper, we propose two advances for GeoMEDD, including a fully automated cluster environment and a cluster tracking module to classify clusters based on the transition state of the cluster lifecycle. We also provide the technical implementation details of setting up a GeoMEDD clustering environment in a hospital setting. A pseudocode implementation of the cluster tracking algorithm is provided and examples for real-world scenarios where GeoMEDD cluster surveillance and monitoring could be beneficial are provided as use cases.

2 Methods

2.1 GeoMEDD

GeoMEDD [20] conceptually utilizes a density-based spatial clustering of applications with noise (DBSCAN) [23] approach that groups together points, which are closely packed together. Similar to DBSCAN, GeoMEDD does not force any shape constraints (such as a circular buffer) on cluster growth thus providing a more realistic view of the underlying spatial process. GeoMEDD has three different parameters, the minimum neighbor parameter α_{min}, the maximum distance parameter β_{max}, and the interval parameter τ. While the α_{min} and β_{max} parameters determine the cluster core (connected to α_{min} neighbors which are within a distance of β_{max}) and fringe members (not a core member, but within a distance of β_{max} from a core member), the τ parameter is used to filter out old data points as well as to visualize the clusters at various time scales (Figure 1). At any time *t*, the dynamic dataset *D(t)* used for clustering in GeoMEDD can be represented using Equation 1, where p is the case occurring at time T(p). The neighborhood definition for a case *i* at time *t* is provided in Equation 2 where *j* is another case at a distance (*dist(i,j)*) less than β_{max}.

$$D(t) = \{p | t - \tau \leq \mathrm{T(p)} \leq \mathrm{t} \qquad (1)$$

$$NB(i,t) = \{j \in D(t) | dist(i,j) \leq \beta\mathrm{max}\,\} \qquad (2)$$

For initial Covid-19 syndromic surveillance, three types of clusters were utilized, *sentinel* (α_{min}=2 and β_{max}=100m), *micro* (α_{min}=5 and β_{max}=500m), and *neighborhood* (α_{min}=10 and β_{max}=1000m) [20]. *Sentinel* clusters acted as an early warning system of a potential outbreak, while *micro* clusters identified a growing outbreak or a higher concentration of disease cases and the *neighborhood* clusters signaled broader area growth – possibly as sentinel or micro clusters grew or merged. The common τ values utilized for Covid-19 surveillance were 21, 14, 7, and 3 days. The values for the different GeoMEDD parameters were set based on the suggestions from health experts and medical practitioners. These values are flexible, and it is entirely likely that new COVID-19 phases, or different geographies (such as rural settings), or even new disease combinations (such as adding in Flu), will result in parameter values being adjusted to capture the dynamic spatial patterns of the disease.

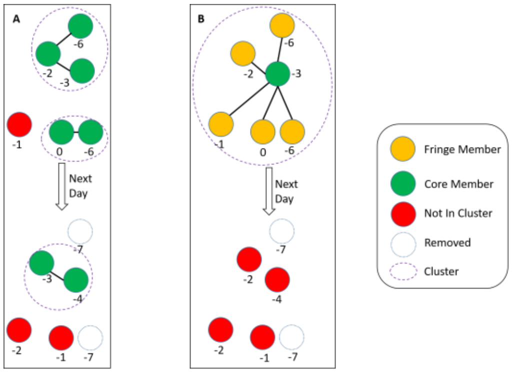

Figure 1. GeoMEDD transition for a single day for A) sentinel cluster B) micro cluster with τ = 7 days. The numbers at the bottom of each points indicate the days before the point was created.

2.2 GeoMEDD technical implementation

GeoMEDD was initially developed as a standalone software with positive COVID-19 case data being uploaded as comma separated value (CSV) files. The code repository along with software for the standalone implementation can be downloaded from GitHub[1]. While the standalone software was convenient for developing daily cluster reports, the manual intervention needed in uploading case data as well as copying cluster outputs was too laborious for the near real-time monitoring required in an ongoing pandemic response. The surveillance capability for GeoMEDD is defined as near real-time rather than real-time as manual correction of erroneous geocodes, and lag in test result data turnaround makes real-time monitoring impossible. There were also other impediments, such as saving historical clusters for retrogressive analysis to identify trend changes, and difficulties involved in combining cluster outputs with other spatial layers such as care home locations and socio-behavioral data. In response to these challenges, the GeoMEDD clustering environment was developed with a spatial database at its heart, acting as an interface between the collaborating hospital system and various data analytics.

2.2.1 GeoMEDD Clustering Environment

The GeoMEDD Clustering Environment (Figure 2) is a technological stack with each component performing a well-defined task to aid cluster monitoring, visualization, and analysis. The starting point and one of the key components of the environment is the geocoder module (Figure 2). Geocoding is a key aspect in cluster generation as the quality of clusters in-terms of cluster shape, size, and accuracy is directly dependent on the quality of the geocode. To improve the geocoding turnaround, three different geocoders are utilized. The addresses that are geocoded to a sufficiently accurate location (based on the geocoder quality codes) are pushed to the spatial database. The geocoder module is implemented in Python and works as an automated batch job process.

The spatial database (Figure 2) is the most important component in the GeoMEDD Clustering Environment, which stores the geocoded test data along with other contextual attributes. Along with the test data, a wide range of spatial socio-demographic data such as building parcels, census enumeration boundaries and data, and care homes are also ingested. The key motivation for utilizing a spatial database is the capability to retrieve the clusters which are saved back to the spatial database for retrospective analysis, query report generations, enhanced daily analysis and production cartography, and the generation of custom warnings based on cluster attributes. The spatial database also supports spatial querying (such as finding all COVID-19 cases within a set distance of a care home), some of which can be completely automated to again generate warnings and reports. PostgreSQL with PostGIS extension is used as the underlying spatial database infrastructure.

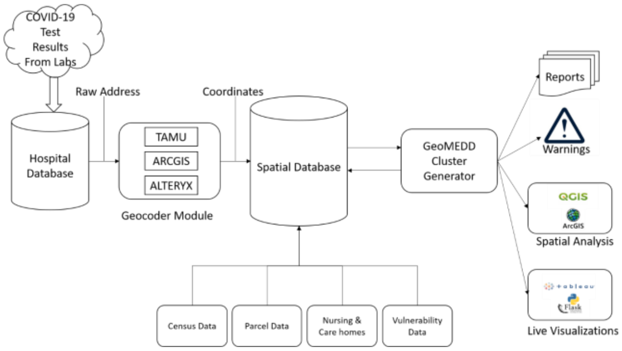

Figure 2. GeoMEDD Clustering Environment

2.2.2 GeoMEDD Cluster Generator

The cluster generator performs three main tasks including i) the generation of new clusters ii) saving generated clusters back to the spatial database and iii) tracking clusters and generating reports and warnings based on the cluster outputs. New clusters are generated from daily COVID 19 test results using a batch job. These data are generally represented as spatial points, but to facilitate better visualization, GeoMEDD clusters are converted to a polygon using the convex hull operation, which creates a boundary based on the outermost points assigned to any cluster. These polygons are converted to an ESRI shapefile (the most common GIS based map file) for additional spatial analytical or cartographic downstream work in a GIS or dashboard environment.

All output GeoMEDD clusters are saved to the spatial database. This strategy has a twofold advantage; firstly, it facilitates longitudinal analysis of clusters and secondly it helps in adding contextual information as the clusters can now be spatially joined with other data layers acquired from the census, or measures of social vulnerability, or even information about the built environment such as building outlines. The longitudinal analysis is vital as this can help show how the clusters (and therefore the epidemic) has diffused, or is diffusing at different geographic scales.

[1] GitHub url : https://github.com/JayakrishnanAjayakumar/SyndromicSurveillance

2.3 Cluster Tracking

In GeoMEDD, cluster tracking is implemented based on the MONIC (Modeling and Monitoring Cluster Transitions) framework [24]. According to the MONIC life cycle model, a cluster can be in any of the five transition states including *new*, *merge*, *split*, *survive*, and *dead*. When a *new* cluster emerges, it can transition to any of the other four states based on its interaction with other clusters over time. The interaction between clusters is assessed by first utilizing a spatial intersection test for identifying the matching candidates followed by cluster membership comparison for an accurate assessment. The membership comparison consists of two important methods, *overlap* and *matches*, from which the various transition stages are calculated.

DEFENITION 1 (Overlap). *Let X, Y be two clusters at time step t_i and t_j ($t_j > t_i$) respectively. Then "overlap of X to Y" at time t_j is defined in Equation 3 where X_j is the total number cases in X at time t_j.*

$$overlap(X,Y) = \frac{X_j \cap Y}{X_j} \tag{3}$$

DEFENITION 2 (Matches). *Let X, Y be two clusters at time step t_i and t_j ($t_j > t_i$) respectively. Further, let z such that $z_{match} \in [0.5,1]$ is a threshold value. Then "Y is a match for X at time t_j subject to z", only if Y is the cluster with maximum overlap for X and the overlap of X to Y is at least z.*

Based on the definitions for *overlap* and *matches* the external transition states for the clusters can be defined (Table 1). Apart from external transition, the clusters which have survived can undergo internal transitions that changes the size and orientation of clusters (Table 1). Figure 3 and Figure 4 show the various transition states for clusters. A detailed pseudocode for the tracking algorithm is provided in Appendix A. The technical workflow for cluster tracking in the GeoMEDD environment is shown in Figure 5. Unlike MONIC, the spatial intersection test used in this case provides a faster first pass for cluster similarity checking especially when the total number of clusters become high.

Table 1: Transitions states for a cluster

Transition Type	Transition	Indicator
External	X *survive* to Y	If Y after X and Y only *matches* to X
	X *splits* to $Y_1 \ldots Y_p$	X *overlaps* $Y_1 \ldots Y_p$
	X *merges* with Y	If Y after X and Y *matches* $X_1 \ldots X_p$
	X disappears	X does not *overlap* with any Y's
	X emerges	
Internal	X expands	X adds more members
	X contracts	X loses members
	No change in X	

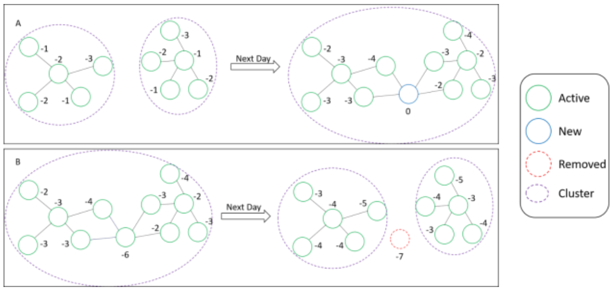

Figure 3. Cluster transition examples A) merge B) split for micro cluster with τ = 7 days

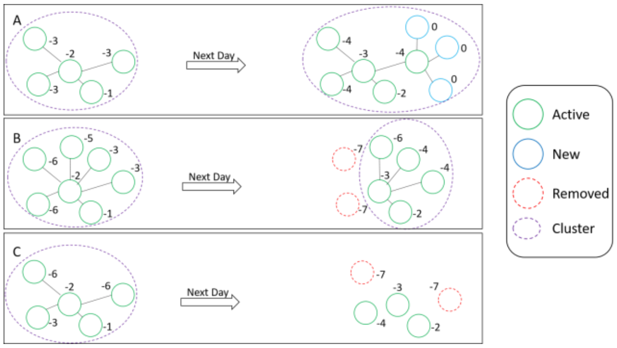

Figure 4. Cluster lifecycle A) expanding, B) contracting, and C) dead for micro cluster with τ = 7 days

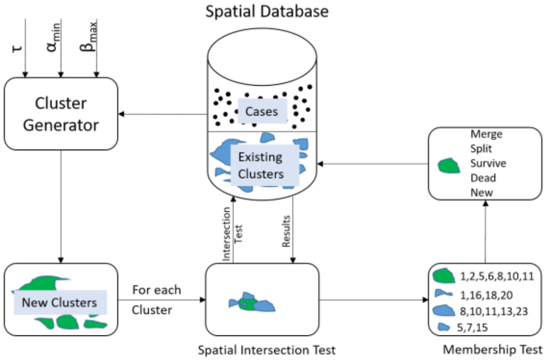

Figure 5. Cluster tracking workflow

3 Use cases

3.1 GeoMEDD to determine general epidemic trends

For the following section, we draw on our experiences of monitoring disease clusters in northeast Ohio. The results presented are typical of the type that were used to provide ongoing operational insights for local pandemic response.

While the major motivation for developing GeoMEDD is to understand the spatial spread, pattern and complexity of underlying phenomena at a local scale (at building, street or sub-neighborhood level), the clusters generated can be aggregated for insights into more regional trends. As an example, the seven-day rolling sum graph of the micro cluster count vs day (Figure 6A) clearly indicates a surge in cluster activity between November 2020 and January 2021 for NE Ohio. For this graph, only new clusters are accounted for and clusters that have been generated from a transition process such as splits or merges are not counted. The fourteen-day rolling average graph of the average cluster dispersion vs day (Figure 6B) clearly indicates that during the same period, the average cluster dispersion was also consistently higher. Cluster dispersion is defined as the median distance between all the members in a cluster and can be a crucial indicator for how clusters are dynamically changing with time. An increase in cluster dispersion generally indicates an increase in cluster merging activity, suggesting an increasing spatial spread of the underlying process. Finally, the bar chart between the average ages of members (in this case age of people who have been tested positive for Covid-19) and time in months (Figure 6C) indicate that clustering of Covid-19 cases were more prevalent in the elderly during the initial stages of the pandemic (April 2020 – May 2020). The chart also shows that there was a substantial dip in average age during September 2020 when education (especially universities) in Ohio were starting to re-open. From April 2021 onwards, there is a continuing decline in the average age of the cluster members indicating the disease was having more of any impact on younger populations, a situation locally addressed as being because of the Delta variant, and younger age cohorts remaining unvaccinated.

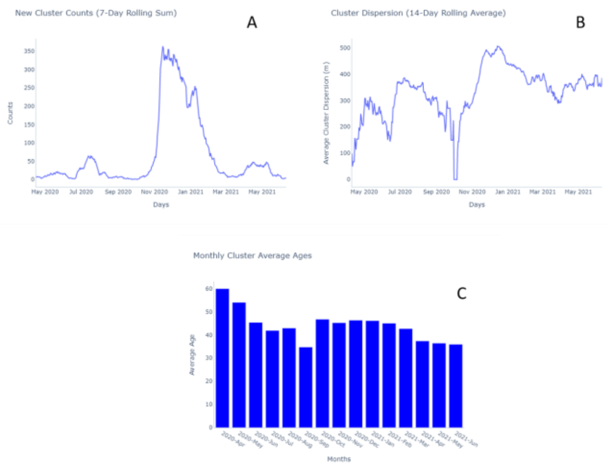

Figure 6. Cluster global analysis. A) Seven-day daily rolling sum of cluster counts, B) Fourteen-day daily rolling average of cluster dispersion in meters, and C) monthly cluster average ages.

3.2 *Local Scale Cluster Analysis*

The real strength of GeoMEDD cluster detection lies in its ability to identify local disease emergence so that intercept teams can get ahead of the pattern rather than just mapping / reporting it [20]. Here we provide a few use cases of how the GeoMEDD cluster analysis has been applied to monitor the pandemic.

3.2.1 Monitoring Congregated Facilities

In the early phases of the pandemic congregated facilities such as nursing homes, care homes, and assisted living facilities as well as correctional institutions are particularly vulnerable for Covid-19 due to both the advanced age and frequent chronic underlying health conditions of the residents and the movement of health care personnel among facilities in a region [25]. As a result, it was vital to monitor these for the first signs of any outbreaks. The spatial database can be leveraged for such monitoring by spatially joining cases and clusters to facilities based on different proximity risk distances. These automated daily reports about within and proximity disease presence around critical facilities can then be disseminated across local health organizations as well as hospitals. Such proximity-based analysis can be vital in taking precautionary measures before proximate disease spread causes an outbreak in the facility. As an example, the map on Figure 7 shows the cluster activity around a care home (indicated by the red star) as an outbreak develops. Figure 7A indicates that there was high cluster activity near the care home two weeks before the real outbreak (Figure 7B). These types of insights can provide an early warning to the care home facility administrators encouraging them to reevaluate the necessary preventive measures to reduce the likelihood of an outbreak occurring.

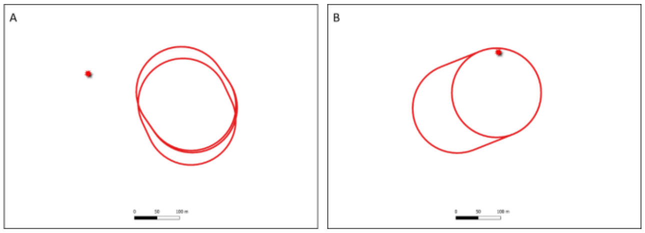

Figure 7. Sentinel cluster activity around a care home facility (indicated by red pin) for the time period A) T1 to T1+21 days and B) T1+30 days to T1+38 days

3.2.2 Generating custom warning messages based on cluster attributes

Additional cluster attributes such as the age of each cluster member can be used to enhance the contextual characteristic of a cluster as well as to filter cluster detection as an early warning system. For example, the clusters shown on Figure 8 have been filtered based on three age categories including an average age between 6 and 17 years (Figure 8A), 18 and 25 years (Figure 8B), and above 60 years (Figure 8C). Based on the different age categories automatic warnings can be generated within the GeoMEDD Clustering Environment and can be disseminated easily to the respective stakeholders (schools for 8A, and care homes for 8C).

3.2.3 Generating daily reports based on cluster tracking

While cluster tracking can be utilized for retrospective analysis to understand disease diffusion, a completely automated cluster tracking system backed by a spatial database offers the prospect for information dissemination in near real-time. Figure 9 shows the flow diagram for a message dissemination pipeline in GeoMEDD. Here each cluster is passed through three different filters including the transition, dispersion, age filters, and based on the filter criteria a new message is generated. Such filter-based messages are aggregated to warn any interested parties such as local health departments or hospital first responders. Figure 9 shows three possible scenarios where a local actor (health department / school board / university) could be apprised of the situation developing in their proximity. The benefit of including the generalized areas of risk also allows for these local experts to contextualize the outputs with their own local knowledge, for example are the clusters over student housing, or a mixed age congregate housing tower?

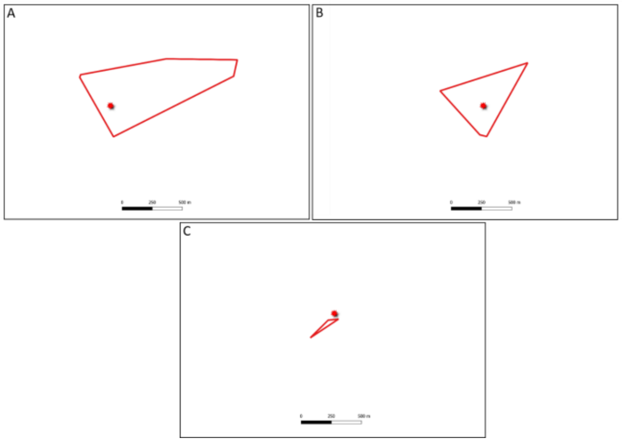

Figure 8. Cluster filtering based on age categories A) 6 to 17 years (school students), the red marker represents a school, B) 17 to 25 years (university students), the red marker represents a University, and C) above 60 years (senior citizens), the red marker represents a care home facility.

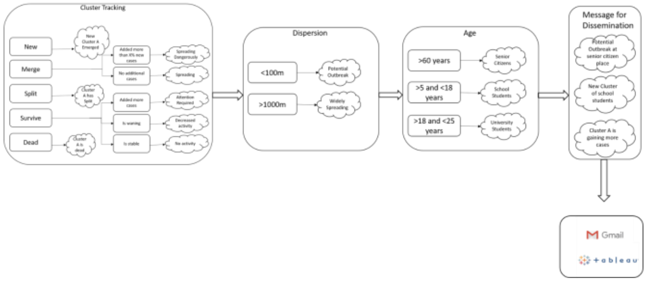

Figure 9. Generating daily reports based on cluster tracking and cluster attributes. Messages for dissemination are generated based on the results from the different filters.

3.2.4 Identifying cluster drivers at various spatial scales

The GeoMEDD cluster parameters α_{min} and β_{max} can be varied to generate clusters at various levels of the hierarchy. The sentinel clusters ($\alpha_{min} = 2$ and $\beta_{max} = 100$) might act as an early warning signal for a potential outbreak, and micro clusters ($\alpha_{min} = 5$ and $\beta_{max} = 500$) might help identify surges spanning across multiple streets and buildings. However, when viewed together, the sentinel clusters nested within the micro clusters can help identify the key drivers, meaning where the majority of the larger cluster activity is found. This might be, for example, the hearth areas of this cluster. This same logic applies for all levels of the hierarchy; sentinel clusters identify drivers in the micro cluster layer, which in turn show the drivers for the neighborhood cluster, which in turn can be viewed as driver areas for a super cluster event. Figure 10 show an example of this multi cluster view with a combination of micro and sentinel clusters. The three sentinel clusters represent the key drivers for the larger micro cluster. Contextual attributes such as age can again be used to gain insights into why these are cluster drivers.

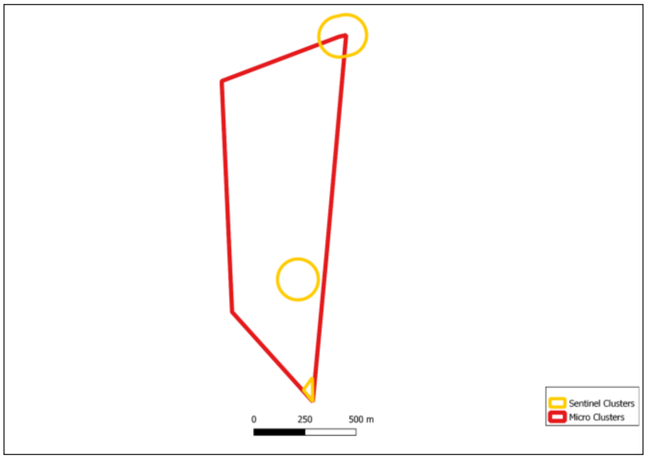

Figure 10. Cluster analysis and monitoring at various spatial scales. Micro clusters (red polygon) and sentinel clusters (yellow polygons) when viewed together displays a hierarchical relationship with the sentinel cluster indicating the major micro cluster drivers.

4 Discussion

Disease surveillance analytics is an important tool to inform both the public health and health care systems about what is happening during an outbreak. While such analytics are important irrespective of the size of the event, what Covid-19 has shown us is that during a pandemic, with multiple waves and variants, and with the added complexity of vaccine status, having excellent and appropriate spatial analytics is vital [12, 15, 16, 20]. Clustering at coarser spatial scales such as counties [15], Zip Codes (or by ZCTA) [14], and even census tracts [16] are important to gauge overall disease trends and inform the public as to what is happening. For an incident command setting however, where intercept teams are mobilized, or hospital resources managed, these data and the associated analytics need to work in near real-time and at as granular a spatial scale as possible. GeoMEDD [20] has been developed as part of north east Ohio's Covid-19 response to achieve those goals, support local hospitals, health departments and other aligned health groups such as federal qualified health centers. However, as the pandemic changes in space and time, it is imperative that GeoMEDD should be conceptually and operationally flexible enough to adapt and evolve. In this paper, we have developed two new components for GeoMEDD including a fully automated clustering environment and a cluster tracking module.

For large hospital systems and governmental health agencies, timely and actionable spatial insights are required. While clustering strategies such as GeoMEDD offer actionable insights, the various steps involved in data curation, geocoding, cluster generation, and finally information dissemination tend to be laborious and time consuming when performed manually. The clustering environment we have developed plays a major role in efficiently automating these manual tasks so that organizational resources can be focused on other critical tasks. We have given particular focus to developing a completely automated and thorough geocoding system, as the quality of any fine scale (at sub neighborhood, street or even building level) clustering methodology is heavily dependent on the quality of the underlying spatial data. For example, many current standard geocoders, when not able to geocode an address to point level, default to the next granular spatial scale such as street centroid, locality, or even to a postal (zip) code. Such artificial aggregation can generate spurious clusters, which can lead to the erroneous deployment of resources. Such geocoding nuances are carefully handled, while the three-layer geocoding pipeline also improves the overall turnaround by caching in on the strengths of each geocoder. The advantage of the spatial database, which is the core component in the clustering environment, is manifold. Based on new surveillance requirements that arise due to the dynamic nature of the disease, new spatial queries can be developed and completely automated. The use case describing monitoring around a care home (Figure 7) provides an example of a completely automated spatial query. The spatial database is also an excellent tool for performing layer based analysis, such as ingesting different Social Vulnerability Indexes [26] so that cluster output can be contextualized by the underlying neighborhood risks.

The cluster attributes, which can be explicitly (age) or implicitly (dispersion) derived, can also be used to classify cluster output or for filter-based cluster monitoring (Figure 8). For example, dispersion can be a good indicator of the characteristics of the underlying spatial structure of the cluster. Typically, a small dispersion value and large cluster count would indicate a potential outbreak at a congregated living facility such as a care home or an assisted living facility, while a large dispersion value indicates potential community spread. Along with supporting these types of near real-time surveillance and monitoring, the spatial database is also an indispensable resource for longitudinal and retrospective analysis such as tracking how the cluster landscape changes, output from which could eventually provide previously unavailable insights for a new family of diffusion models. For example, tracking clusters helps to understand the disease diffusion process across geographic scales (both global (Figure 6) and local). Cluster tracking can also be used to classify geographical areas based on cluster emergence (new clusters) and cluster merging. For example, an area could be classified as having severe community spread if there is frequent cluster merging in a short time frame. While currently we determine these patterns through exploratory analysis in near real-time, eventually we would need to theorize exactly what is happening so that effective trigger thresholds can be developed.

Similarly, it might be that a combination of cluster deaths, cluster splits and reduction in cluster size might indicate a reduction in cluster intensity, or even a shortage in Covid-19 testing. Cluster tracking when used in conjunction with attribute based filtering, could be used to generate daily disease reports (Figure 9), the content of which could be tailored based on the interested parties (hospitals, county and city health organizations). For example, a hospital will be interested in knowing about cluster prevalence in their catchment area, while a county health department will be interested in knowing about how often cluster merges occur (indicating community spread). Of course, introducing such analytics would also require an associated training as what is being suggested here is the next step in operational health care analytics.

As a clustering technique, GeoMEDD is similar to DBSCAN with an additional τ parameter for filtering data based on time. As with DBSCAN, the lack of statistical rigor is also an issue in GeoMEDD [27]. As GeoMEDD does not rely on denominator values (underlying population), there is no inherent normalization of the case data as compared to other techniques such as SaTScan. We emphasize that the purpose of GeoMEDD is not to replace traditional forms of disease cluster detection, but to enhance investigation in near real-time and in the context of health system operations to a dynamic situation. In this setting it is important to know when any new disease cluster is emerging rather than if it has reached a normalized threshold [20]. With the addition of cluster tracking module, GeoMEDD is comparable to the Modified Space–Time DBSCAN (MST-DBSCAN) [28]. Compared to MST-DBSCAN, GeoMEDD is implemented within a spatial database ecosystem, which is particularly helpful in identifying the cluster interaction patterns (external transitions) efficiently. Even though not implemented here, adding a time-based weight parameter along with the time filtering parameter would be beneficial if knowledge about a disease life cycle (incubation period, those infected, those recovered) is clear and accurate.

Even though there are many benefits in developing a fully automated clustering environment and cluster tracking system, there are some inherent limitations too. Firstly, the cluster output is heavily dependent on the quality of the underlying geocoder. Even though we utilize a three-step geocoding process, the underlying geocoder is still vulnerable to geocoding nuances. For example, based on the underlying spatial structure, a care home could have more than one address (commonly seen with multiple complexes) which can affect that particular local cluster generation. Secondly, setting up an operational cluster environment and deploying a cluster tracking module are also a challenging task for many resource depleted health organizations, especially those lagging in IT support and infrastructure. To tackle this, we plan to develop a single application package containing all the required modules for building the cluster environment and the tracking module. While packaging a database such as PostgreSQL along with PostGIS is not a trivial task, we plan to develop a friendly standalone version of the software with SpatiaLite (a spatial extension for the lightweight SQLite database) as the core spatial database system. Finally, clustering is also prone to dynamic structural changes, which can be problematic for cluster tracking. For example, a cluster that has died due to the recovery of a single member can be reinstated as a new cluster by the addition of a new member by the next day. Such fast transitions are difficult to capture, and we are planning to address this issue by devising various member weighting schemes. While we acknowledge these challenges, we are still confident in the current version of GeoMEDD due to its eighteen month (and ongoing) evaluation by those who are most in need of these data analytical insights.

5 Conclusions

As Covid-19 evolves in space and time, new automated spatial syndromic surveillance tools are required to keep apace of the changes. New approaches should improve on existing data analytic methods, but also be scalable for deployment in multiple settings, and be conceptually and operationally flexible enough to morph in to any new challenge. In this study, we have presented a method which has been at the heart of the COVID-19 geospatial response in northeast Ohio. More specifically we have developed a completely automated clustering environment, along with a cluster tracking module to capture the space time changes in Covid-19 spread. In a unique twist, all cluster and architecture development, and all cluster and query outputs, have been evaluated in near real-time by those who will use them on the front lines; health practitioners and managers.

ACKNOWLEDGMENTS

The team would like to thank all at UH Ventures and the Cleveland Clinic Foundation who have been involved in discussions surrounding this work, especially Eric Beck, Sam Brown, David Sylvan, Andrew Proctor, Eric Hixson, Tim Sullivan, Sarah Mihalik, Maulik Purohit, Zachary Scott, James Muisyo, James Labadorf, Sorapat Vijitakula, Justin Yax, and James Simon. We like to thank Daniel Goldberg for helping us with geocoder setup. We would also like to thank all the reviewers for the insightful comments and suggestions.

REFERENCES

[1] Charlotte D. Smith and Jeremy Mennis. 2020. Incorporating Geographic Information Science and Technology in Response to the COVID-19 Pandemic. Preventing Chronic Disease 17, (July 2020), E58. DOI:https://doi.org/10.5888/pcd17.200246

[2] Maged N. Kamel Boulos and Estella M. Geraghty. 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. International Journal of Health Geographics 19, 1 (March 2020), 8. DOI:https://doi.org/10.1186/s12942-020-00202-8

[3] Ivan Franch-Pardo, Michael R Desjardins, Isabel Barea-Navarro, and Artemi Cerdà. 2021. A review of GIS methodologies to analyze the dynamics of COVID-19 in the second half of 2020. Transactions in GIS (2021). DOI: https://doi.org/10.1111/tgis.12792

[4] Chaowei Yang, Dexuan Sha, Qian Liu, Yun Li, Hai Lan, Weihe Wendy Guan, Tao Hu, Zhenlong Li, Zhiran Zhang, John Hoot Thompson, Zifu Wang, David Wong, Shiyang Ruan, Manzhu Yu, Douglas Richardson, Luyao Zhang, Ruizhi Hou, You Zhou, Cheng Zhong, Yifei Tian, Fayez Beaini, Kyla Carte, Colin Flynn, Wei Liu, Dieter Pfoser, Shuming Bao, Mei Li, Haoyuan Zhang, Chunbo Liu, Jie Jiang, Shihong Du, Liang Zhao, Mingyue Lu, Lin Li, Huan Zhou, and Andrew Ding. 2020. Taking the pulse of COVID-19: a spatiotemporal perspective. International Journal of Digital Earth 13, 10 (October 2020), 1186–1211. DOI:https://doi.org/10.1080/17538947.2020.1809723

[5] Ayodeji Emmanuel Iyanda, Richard Adeleke, Yongmei Lu, Tolulope Osayomi, Adeleye Adaralegbe, Mayowa Lasode, Ngozi J Chima-Adaralegbe, and Adedoyin M Osundina. 2020. A retrospective cross-national examination of COVID-19 outbreak in 175 countries: a multiscale geographically weighted regression analysis (January 11-June 28, 2020). Journal of infection and public health 13, 10 (2020), 1438–1445. DOI:https://doi.org/10.1016/j.jiph.2020.07.006

[6] Andree Ehlert. 2021. The socio-economic determinants of COVID-19: A spatial analysis of German county level data. Socio-Economic Planning Sciences (May 2021), 101083. DOI:https://doi.org/10.1016/j.seps.2021.101083

[7] Abolfazl Mollalo, Behzad Vahedi, and Kiara M. Rivera. 2020. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Science of The Total Environment 728, (August 2020), 138884. DOI:https://doi.org/10.1016/j.scitotenv.2020.138884

[8] Osvaldo Fonseca-Rodríguez, Per E. Gustafsson, Miguel San Sebastián, and Anne-Marie Fors Connolly. 2021. Spatial clustering and contextual factors associated with hospitalisation and deaths due to COVID-19 in Sweden: a geospatial nationwide ecological study. BMJ Glob Health 6, 7 (July 2021), e006247. DOI:https://doi.org/10.1136/bmjgh-2021-006247

[9] Fuyu Xu and Kate Beard. 2021. A comparison of prospective space-time scan statistics and spatiotemporal event sequence based clustering for COVID-19 surveillance. PLOS ONE 16, 6 (June 2021), e0252990. DOI:https://doi.org/10.1371/journal.pone.0252990

[10] Sheng Zhang, MengYuan Diao, Wenbo Yu, Lei Pei, Zhaofen Lin, and Dechang Chen. 2020. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. International Journal of Infectious Diseases 93, (April 2020), 201–204. DOI:https://doi.org/10.1016/j.ijid.2020.02.033

[11] David De Ridder, José Sandoval, Nicolas Vuilleumier, Silvia Stringhini, Hervé Spechbach, Stéphane Joost, Laurent Kaiser, and Idris Guessous. 2020. Geospatial digital monitoring of COVID-19 cases at high spatiotemporal resolution. The Lancet Digital Health 2, 8 (August 2020), e393–e394. DOI:https://doi.org/10.1016/S2589-7500(20)30139-4

[12] Alexander Hohl, Eric M. Delmelle, Michael R. Desjardins, and Yu Lan. 2020. Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. Spatial and Spatio-temporal Epidemiology 34, (August 2020), 100354. DOI:https://doi.org/10.1016/j.sste.2020.100354

[13] Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assunção, and Farzad Mostashari. 2005. A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. PLOS Medicine 2, 3 (February 2005), e59. DOI:https://doi.org/10.1371/journal.pmed.0020059

[14] Jack Cordes and Marcia C. Castro. 2020. Spatial analysis of COVID-19 clusters and contextual factors in New York City. Spatial and Spatio-temporal Epidemiology 34, (August 2020), 100355. DOI:https://doi.org/10.1016/j.sste.2020.100355

[15] Michael R. Desjardins, Alexander Hohl, and Eric M. Delmelle. 2020. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. Appl Geogr 118, (May 2020), 102202. DOI:https://doi.org/10.1016/j.apgeog.2020.102202

[16] Sharon K. Greene, Eric R. Peterson, Dominique Balan, Lucretia Jones, Gretchen M. Culp, Annie D. Fine, and Martin Kulldorff. 2021. Detecting COVID-19 Clusters at High Spatiotemporal Resolution, New York City, New York, USA, June–July 2020. Emerging Infectious Disease 27, 5 (May 2021), 1500–1504. DOI:https://doi.org/10.3201/eid2705.203583

[17] Sarah L. Jackson, Sahar Derakhshan, Leah Blackwood, Logan Lee, Qian Huang, Margot Habets, and Susan L. Cutter. 2021. Spatial Disparities of COVID-19 Cases and Fatalities in United States Counties. International Journal of Environmental Research and Public Health 18, 16 (August 2021), 8259. DOI:https://doi.org/10.3390/ijerph18168259

[18] Nicolás Rosillo, Javier Del-Águila-Mejía, Ayelén Rojas-Benedicto, María Guerrero-Vadillo, Marina Peñuelas, Clara Mazagatos, Jordi Segú-Tell, Rebeca Ramis, and Diana Gómez-Barroso. 2021. Real time surveillance of COVID-19 space and time clusters during the summer 2020 in Spain. BMC Public Health 21, 1 (May 2021), 961. DOI:https://doi.org/10.1186/s12889-021-10961-z

[19] Anaïs Ladoy, Onya Opota, Pierre-Nicolas Carron, Idris Guessous, Séverine Vuilleumier, Stéphane Joost, and Gilbert Greub. 2021. Size and duration of COVID-19 clusters go along with a high SARS-CoV-2 viral load: A spatio-temporal investigation in Vaud state, Switzerland. Science of The Total Environment 787, (September 2021), 147483. DOI:https://doi.org/10.1016/j.scitotenv.2021.147483

[20] Andrew Curtis, Jayakrishnan Ajayakumar, Jacqueline Curtis, Sarah Mihalik, Maulik Purohit, Zachary Scott, James Muisyo, James Labadorf, Sorapat Vijitakula, Justin Yax, and Daniel W. Goldberg. 2020. Geographic monitoring for early disease detection (GeoMEDD). Scientific Reports 10, 1 (December 2020), 21753. DOI:https://doi.org/10.1038/s41598-020-78704-5

[21] University Hospitals. 2020. Stronger Together. Stronger Together. Retrieved October 16, 2021 from https://www.uhhospitals.org/about-uh/publications/corporate-publications/stronger-together

[22] Brightsurf. 2021. New paper describes use of geographic monitoring for early COVID cluster detection. Brightsurf. Retrieved October 16, 2021 from https://www.brightsurf.com/news/article/010621528530/new-paper-describes-use-of-geographic-monitoring-for-early-covid-cluster-detection.html

[23] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and others. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd, 226–231. DOI: https://dl.acm.org/doi/10.5555/3001460.3001507

[24] Myra Spiliopoulou, Irene Ntoutsi, Yannis Theodoridis, and Rene Schult. 2006. MONIC: modeling and monitoring cluster transitions. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06), Association for Computing Machinery, New York, NY, USA, 706–711. DOI:https://doi.org/10.1145/1150402.1150491

[25] Michael L. Barnett and David C. Grabowski. 2020. Nursing Homes Are Ground Zero for COVID-19 Pandemic. JAMA Health Forum 1, 3 (March 2020), e200369–e200369. DOI:https://doi.org/10.1001/jamahealthforum.2020.0369

[26] Barry E. Flanagan, Edward W. Gregory, Elaine J. Hallisey, Janet L. Heitgerd, and Brian Lewis. 2011. A Social Vulnerability Index for Disaster Management. Journal of Homeland Security and Emergency Management 8, 1 (January 2011). DOI:https://doi.org/10.2202/1547-7355.1792

[27] Yiqun Xie, Shashi Shekhar, and Yan Li. 2021. Statistically-Robust Clustering Techniques for Mapping Spatial Hotspots: A Survey. arXiv:2103.12019 [cs, stat] (October 2021). DOI:https://doi.org/10.1145/3487893

[28] Fei-Ying Kuo, Tzai-Hung Wen, and Clive E. Sabel. 2018. Characterizing Diffusion Dynamics of Disease Clustering: A Modified Space–Time DBSCAN (MST-DBSCAN) Algorithm. Annals of the American Association of Geographers 108, 4 (July 2018), 1168–1186. DOI:https://doi.org/10.1080/24694452.2017.1407630

Appendix

A. Cluster Tracking Algorithm

Algorithm 1 Cluster Tracking

```
/* Run Clustering for every day and compares with existing clusters to determine the current state of a cluster*/
existingClusters = {} /* Data structure for storing clusters*/
clustId = 0 /*Counter for generating cluster ids*/
clustContinuing={}/*New Clusters that are continuations*/
probablyLiving = {}/*Clusters that might continue to exit*/
newToOldMapping = {}/*New cluster to old cluster relation */
minMatchThreshold = 50 /*Minimum percentage of members that should be similar for a new cluster to be a continuation of an existing cluster*/
minSplitThreshold = 10 /* Minimum percentage of members that should be similar for an existing cluster to be a part of a new cluster formed by splitting process*/
currentCases = getCurrentCases(previous=21) /*Get recent cases for previous n days*/
/*Run GeoMEDD clustering algorithm with currentCases as input*/
newClusterSet = runGeoMEDD(currentCases,minNeighb,maxDist)
/*Store each cluster bounds to an rtree for fast intersection test*/
clustBounds=rtree(newClusterSet)
for each existingClust in existingClusters
    /*if the cluster status is not active continue*/
    if existingClust.status != ‘active’
        Continue
    /*get new clusters that intersects with existing clusters*/
    matchClusts = clustBounds.intersects(existingClust.bounds)
    /*if there is no intersecting cluster, then the existing cluster has died*/
    if matchClusts.size == 0
        existingClust.status = ‘Dead’
    /*if there is a single match*/
    else if matchClusts.size == 1
        /*Check if the members are similar up to threshold */
        sim= similarity(existingClust.members,matchClusts[0].members)
        if sim < minMatchThreshold
            /*since the clusters are not strongly similar we consider  the existing cluster to be dead*/
            existingClust.status = ‘Dead’
            Continue
        else
            /*The existing cluster is continuing to live or have merged*/
            probablyLiving[existingClust.id]= matchClusts[0].id
            newToOldMapping[matchClusts[0].id]= existingClust.id
        else
            /*There are multiple matches. This could be a split*/
            localT = 0/*Local variable to accumulate thresholds*/
            localMatchIds=[]/*for matching clusters*/
            for each newClusts in matchClusts
                sim=similarity(existingClust.members, newClusts.members)
                if sim > minSplitThreshold
                    localT = localT+sim
                    localMatchIds.add(newClusts.id)
                /*If the accrued total is less than threshold, cluster is dead*/
                if localT < minMatchThreshold
                    existingClust.status = ‘Dead’
                    Continue
                else
                    /*If there is only one valid match the cluster can be added to probably living set*/
                if localMatchIds.size==1
                    probablyLiving[existingClust.id]= localMatchIds[0]
                    newToOldMapping[localMatchIds[0]]= existingClust.id
                else /*This is a pure split*/
                    existingClust.status = ‘Split’
                    for each ids in localMatchIds
                        newToOldMapping[ids]= existingClust.id
/*Now for each probablyLiving cluster check if its continuation or a merge*/
for each pClusters in probablyLiving
    /*If there is one to one relation between old and new, then it’s a continuation, and otherwise it’s a merge */
    if newToOldMapping[probablyLiving[pClusters]]==1
        existingClust.status = ‘Active’
        clustContinuing[probablyLiving[pClusters]]=0
    else
        /*This is a merge. The old cluster has merged to a new cluster*/
        existingClust.status = ‘Merge’
/*Now add each new member to the existingCluster set*/
for each newClust in newClusterSet
    /*If the cluster is continuation, don’t add to existing*/
    if newClust.id is not in clustContinuing
        existingClusters[newClust.id]= newClust
```