Spatial Data Analytics For The Pandemic

Peter Y. Wu
wu@rmu.edu

Diane A. Igoche
igoche@rmu.edu

Department of Computer & Information Systems
Robert Morris University
Moon Township, PA

Abstract

The recent COVID-19 pandemic season has presented several data analytics related problems especially for data that is spatially related. Often enough, we present the important information in maps for exploratory visualization of data. However, there is more that spatial data analytics can offer. Spatial data analytics calls for specific skill sets rather different from common practices in the commonly known approaches to analytics. Given the situation of a pandemic, we examine a few of these problems. We studied these problems and proposed approaches based on the nature of the problems. The approaches fall into three categories: spatial data aggregation for visualization, path finding for logistics and delivery planning, and site selection for resources allocation and distribution. These constitute a basic skill set for spatial data analytics in information system applications during a pandemic.

Keywords: Spatial Data Analytics, Geographic Information System, GIS.

1. INTRODUCTION

This paper refers to Data Analytics as applying the knowledge of data science techniques in problem solving. The proliferation of data science in various sectors coupled with the democratization of tools for applying techniques is reducing the barriers of use. Thus, making data analytics skill sets one of the most marketable for information system professionals. With the COVID-19 pandemic affecting almost every aspect of life, this paper reviews a common set of data related problems in dealing with pandemics, and analyzes the skill sets applicable to solving these problems. Focusing on COVID-19, the problems include visualizing the spread of the disease, monitoring the trend of changes related to the disease, and identifying the location of the disease epicenter. Issues with medical response also presented during the initial outbreak. Medical response issues include logistics and delivery of medicines, medical equipment, and personal protective equipment (PPE). There are also issues of contact tracing. Almost all these problems require the use of spatial data. The skill set associated with performing analytics using spatial data is primarily in the deeper understanding in the use of Geographic Information System (GIS).

In the next section, we will present a brief survey of the body of knowledge in data analytics, and focus on spatial data analysis. Section 3 introduces some of the selected problems that we often encounter during a pandemic. Section 4 will then describe our proposed solutions to these problems, highlighting the needed skills for spatial data analytics.

2. A BRIEF SURVEY

Information Systems (IS) can be roughly described as the discipline of practical problem solving with the combination of computer science, information technology and business management. Long favored as a field for statisticians, data analytics has become

integrated into the field of Information Systems. Over the last five (5) to ten (10) years, the focus has shifted to retrieving actionable insights from information systems applications; using data, statistical techniques, and computing to solve practical problems. With the broad availability of data through the internet and other communication means, the challenge of information technology is now going beyond efficiency and effectiveness in the processing of data, but the intelligent use of information for actionable real time business decisions.

General approaches to data analytics are broadly classified into the following categories: descriptive, predictive, and prescriptive. These categories often lend themselves to each other, especially descriptive and predictive. While these approaches are certainly closely related, they focus on different scopes in solving data related problems and applying results for strategic decision making. Descriptive analytics organizes and presents data using descriptive statistics and visual representations, this category explains the current state of a phenomena using data and assists in preparation for identifying trends, forecasting, and creating predictive models. Predictive analytics uses data, statistical techniques and algorithms to identify the possibility of future outcomes with the presence of historical data. Where descriptive analytics tells us what has happened, predictive analytics seeks to tell what will happen. Prescriptive analytics more comprehensively constructs a model of phenomenon. Applying simulation and other automation to act on the results of the analysis. Knowing what will happen is no longer enough, it is also important to take the best course of action. Prescriptive analytics involves the use of mathematical modeling and simulation to recommend the best actions to take.

During the COVID-19 pandemic, we have seen the application of the techniques including Taiwan's use of big data analytics to improve testing and medical response times (Wang, Ng, & Brook, 2020). The use of data from Ebola, and SARS outbreaks to determine how to tackle the COVID-19 crisis (Ting, Carin, Dzau, & Wong, 2020) among others. The scholarship on the use of analytics specifically these approaches is still growing due. There is a need to understand the approaches of analytics as related to spatial data. We much more often come across problems dealing with spatial data, as evidenced by information presented in maps, to be visualized and appropriately understood. Spatial data also presents a different challenge in analysis, in all three approaches broadly categorized above. Long before data analytics was a trend in Information System practices, geographic information system (GIS) was envisioned to be a more efficient and effective way to work with map data processing on the computer. The past two decades have seen that vision realized, bringing GIS applications in many different fields. The GIS was recognized as a viable tool for spatial data handling, particularly when that spatial relationship refers to places on the earth. The GIS is used not only for visualization of spatial data, as in descriptive analytics, reference mapping, but the GIS is also used in analytic mapping - with modeling and animation, as in predictive and prescriptive analytics.

Burrough (2001) saliently described the use of descriptive analytics for aiding geo-registration of data and facilitating spatial exploratory data analysis. The works related to spatial data and analytics have included the use of Geo-statistics. Geo-statistics is an exploratory analysis technique that is also used to predict values associated with spatial phenomena. Paez (2018) used spatial filtering and exploratory data analysis for the improvement of regression models of spatial data. Zhou et.al, (2020) expressed the concerns related to performing analytics using spatial data especially during the COVID-19 pandemic. The main issues identified are the challenges associated with data aggregation and using the right approaches for knowledge discovery and dynamic expression.

In the next section, we will highlight some problems that affect the application of spatial data analytics using GIS during a pandemic.

3. DATA RELATED PROBLEMS DURING A PANDEMIC

In this section, we describe some of the data analytics problems we commonly encounter during a pandemic. These may include visualization of data on a map, referred to as reference mapping with the traditional skill of cartography. Beyond visualization, we also want to analyze the data in the associated spatial relationships. Referred to as analytic mapping, we may apply methods of counting, apportionment and measurement on the map, or other more advanced geometric techniques. The GIS provides the viable tool to do these much more efficiently. We will first discuss the problems in further details, before we move on to our approaches to solutions in the next section.

Visualization

In a pandemic, we need to visualize on a map where the disease is occurring. We can collect the patient addresses which we will need to massively convert into map data for presentation. Figure 1 shows such a map of points of patient address locations.

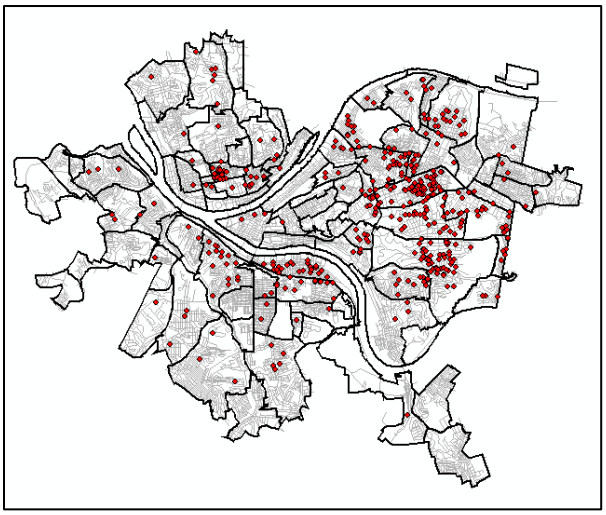

Fig.2 Point Map of Patient Address Locations, by Linear Geocoding

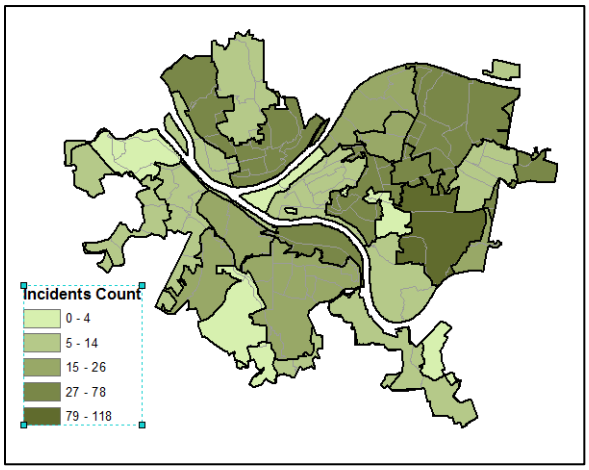

Fig.1 Counting Patients per Zip Code Area by Polygon Geocoding

There are, however, other issues. An important but simple one is that of privacy: when presenting the map, we cannot expose personal confidential information. Our point map of patient locations should not the viewer to zoom in to a level to identify an address. That generally may not be practicable. We should just present the patient locations in aggregate to protect privacy of personal information. An easy solution is to identify only the zip code, and present the number of patients in a zip code area. Figure 2 depicts the map of the same patient location data, but only showing the zip code areas.

Since the zip code areas may not be the way we want to visualize the data, we may want to derive our own regional area divisions to count the aggregate patient data there.

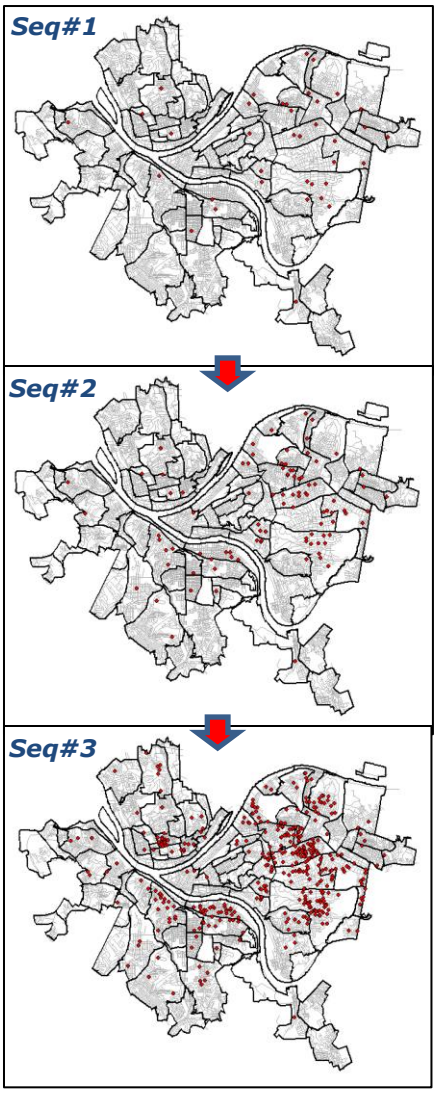

Fig.3 Map Sequence to show Changing Trends in the number of patient incidents

Trends and Temporal Changes

We also want to see how the situation changes in time, such as how the epicenter of the pandemic shifts from one location to another. The data to collected need to be appropriately time-stamped. When the spatial data is organized to present on a map, we may then follow the time setting to present the temporal changes in a sequence of maps accordingly, perhaps daily, weekly, or monthly. Figure 3 depicts a sequence of three

maps to illustrate the change of the situation from one to another. For a longer sequence, manual control of animated presentation may be desirable.

Logistics and Delivery

In response to a pandemic, we need to keep up the supply of medicine and medical equipment. The medical personnel and other first responders also need personal protective equipment (PPE) in large supply. In the case of covid-19 when neither viable treatment nor effective vaccine is presently available, medical supplies such as ventilators and accessorial materials can become suddenly in short supply when an outbreak occurs. PPE for service personnel will also be in the same situation. Logistics planning for quick delivery will involve with finding the best delivery path with estimate of how long that may take. In a simple case of delivery, determining an optimal delivery path is generally achievable, even when we may seek criteria such as cheapest or safest route different from the quickest. Complication may occur when we must deliver to multiple locations with different starting points for many delivery vehicles.

Resource Allocation and Distribution

Before responding to an emergency, it is better if we can design our plans for logistics and delivery. We need to identify the appropriate places to allocate our resources so that they may be distributed efficiently to the places of need. We want to prepare for where the needs will be and place important resources at the right places ahead of time.

Note that the resources include not just materials but also services. Normally we keep needed medical supplies at the hospitals and clinics where the patients may also be treated. In the presence of an outbreak during a pandemic, the need may suddenly overwhelm the planned capacity, both material and service resources. Even though the brick-and-mortar hospitals and clinics are not easily relocated, it is not at all uncommon we need to use a temporary clinical site or even an impromptu hospital to be built as quickly as possible. These will involve major resources planning, and site selection is a spatial data analytics problem.

4. SPATIAL DATA ANALYSIS

We will examine each of the problems indicated above and discuss our approaches to solution. In each case, we may also need to study the variations that may be there, and how we may also adjust our approach to solution to handle that.

Spatial Aggregation for Visualization

To show where the disease is occurring, we may have collected relevant patient addresses. Address geocoding is a common GIS featured function to process a massive collection of addresses into locations for display on a map. It is often an expert system to cope with common human mistakes in ill-formed addresses: the system can properly handle "dirty" data and also support manual data correction with interactive computer-aid (Wu & Rathswohl 2010).

A related issue is the confidentiality of personal patient data. To protect privacy, we do not want to allow viewer zooming in on the map to the extent that specific location of a patient address can be identified. A simple solution is aggregation of patient data by location. Instead of linear geocoding to match an address to a point location on the street map, we apply polygon geocoding so that an address is matched to a region such as the zip code area or the state (as illustrated in Figure 2). However, when we want to organize and present the patient data aggregated in some other way which may not be coded in the address information, geocoding will not be sufficient.

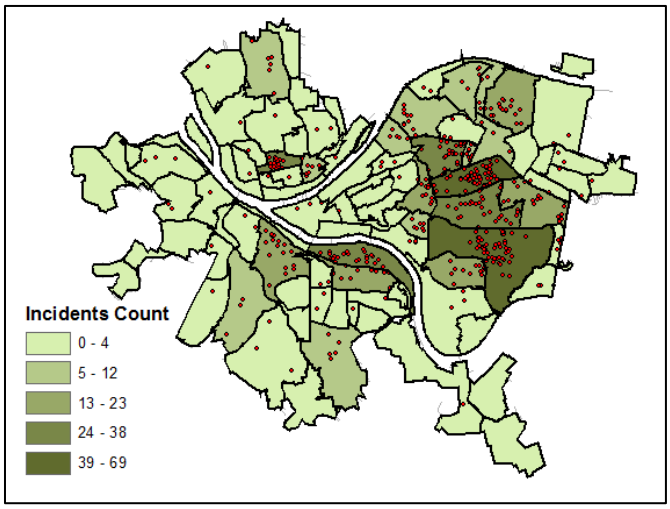

Fig.4 Using Spatial Join in Order to Count The Patient Incidents in each Municipality

If we know the regions to sub-divide the patient information into aggregate areas, we must have the polygon map for these areas, or be able to create the map using GIS. Then with the point map of the patient locations of the geocoded addresses, we can apply spatial join in which we can associate each point to the area it falls inside, thus creating the aggregate of the points inside

each area (Wu 2013). Figure 4 illustrates the spatial join allows us to identify and count the points in each area.

In the case of a pandemic, we may aggregate and count the number of cases by municipalities, counties, states, or any regional areas we may want to create.

Map Animation for Temporal Changes

To visualize how the disease is spreading in time during a pandemic, we will time-stamp the data as we collect so that we can present time sequence of maps showing data at different day or time intervals.

Modern GIS such as ArcGIS 10.8 (Gorr & Kurland 2020) supports the use of map data with embedded time sequencing. With the feature, it becomes convenient to animate the map display as a timed sequence of maps and provide manual control of the display for viewing. For example, we may visualize the changes in the concentration of disease cases with the shifting of the epicenter on a map for each day in a week, or every week in a number of weeks.

Path Planning for Delivery

In a pandemic, medicine and medical supplies as well as PPE for medical personnel and other responders are in constant needs. Shipment of materials or personnel is basically a problem of spatial data analytics. For the simple problem of delivery from one place to another, an algorithm for shortest path will serve well for a solution. The criteria for best delivery path may sometimes be different. We often want a shortest path for quickest delivery, but sometimes we may want a delivery path that costs the least. When we can model the problem in a directed graph of weighted arcs, and the generic shortest path algorithm will still work right seeking for the optimized solution (Lanning, Harell, Wang 2014). Hence modern GIS is often augmented with network analysis package option to provide the routing function. Thanks to the elegance and efficiency of Dijkstra's algorithm (Knuth 1977), the routine function works well even with the dynamic changes to re-route delivery in cases of emergency. An evident example is in the popular use of Google Map for routing.

However, there is also the complication at times with the need to deliver to multiple locations at different starting points. Arrivals at different destination points may need to follow certain sequence, or there may be options in the required sequence. There may not be an easy algorithm to find an optimized path. The data analyst in such a situation will need to have a good understanding of the practicable algorithms (such as Dijkstra's algorithm) and adapt them to different segments of the problem. Integration of partial solutions may need to do programming in the augmented packages such as Python Script (Zanbergen 2013).

Site Selection for Resources Distribution

Noting the needs for equipment and supplies, where should we place them in order to best handle deliveries at the time of need including emergencies? That is the problem of selecting appropriate sites of storage. It is basically the same problem when we need a place for a clinic to serve patients in the vicinity and allow service personnel to come conveniently. The spatial analytics problem is that of site selection so that resources or services can be efficiently and effectively distributed. While there may not be a theoretical general solution, there are many applicable tools for analysis.

To analyze the efficacy of service at a single site, we often consider a circular buffer around the site to assess the apportionment of clients that the site may serve. Figure 5 illustrates assessing the potential population a testing site may serve, by counting the population within multiple ring buffers around the site. GIS is commonly built-in with this kind of spatial analysis functionality.
When we want to analyze the service efficacy of

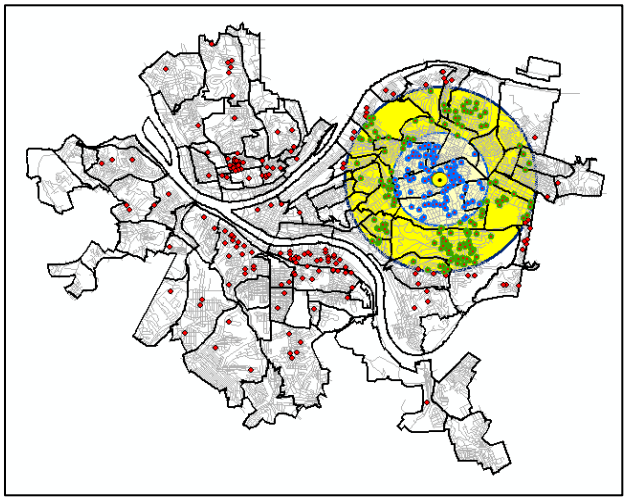

Fig.5 Use of Ring Buffers to Evaluate The Site's Service to The Clients in The Vicinity

multiple sites in an area, we will need a different approach since the ring buffers of the different sites overlap with intersections. One common approach is to partition the area of service into polygons around each site so that for every point within a polygon, the closest service site is the

site within that polygon. It is known as the Thiessen Polygons, also known as the Voronoi Diagram (Yamada 2017). The algorithms to construct the solution were developed in the field of computational geometry (Preparata & Shamos 1985). Figure 6 illustrates the Thiessen Polygons constructed over 5 service sites on a map. GIS today may or may not provide such function, and the spatial data analyst will have to more technically trained, such as in programming extensions to build customized tools for special case solutions (Bivand, Pebesma & Gomez-Rubio 2013).

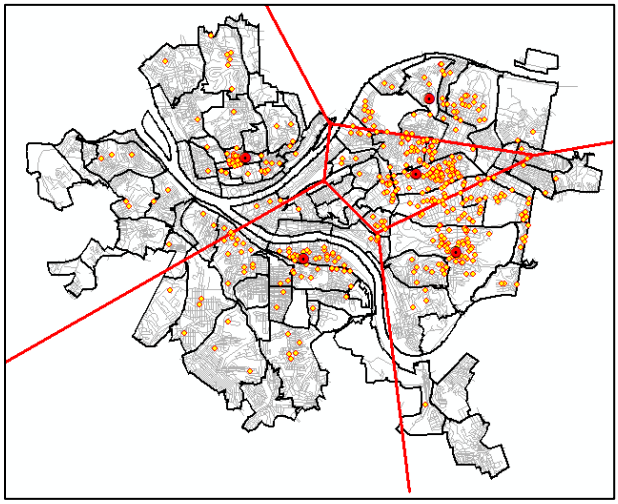

Fig.6 Using Thiessen Polygons to Analyze Service Around Five Service Centers

5. CONCLUSION

We examined the data analytics problems that are common during a pandemic and sketched out our approaches to solution. Different from more traditional Information Systems problem solving, these problems are not only data analytics problems, but more often, they involve spatially related data. We examined three categories of problems: visualization of data on a map, logistics and delivery path finding, and site selection for resources allocation and distribution. From our sketched approach to solution, these skills build on the use of GIS but also require deeper thinking about the geometry and geographical nature of the problems. These therefore constitute a skill set for spatial data analytics during a pandemic.

6. REFERENCES

Bivand, R.S., E. Pebesma and V. Gomez-Rubio. (2013) *Applied Spatial Data Analysis with R*. Springer Science.

Burrough, P.A. (2001) GIS and geostatistics: Essential partners for spatial analysis. *Environmental and ecological statistics*, *8*(4), 361-377.

Gorr, W.L., K.L. Kurland. (2020) *GIS Tutorial for ArcGIS Desktop 10.8*. ESRI Press. Redlands, CA.

Haining, R. (2003) *Spatial Data Analysis: Theory and Practice*. Cambridge University Press.

Knuth, D.E. (1977) *A Generalization of Dijkstra's Algorithm*. Information Processing Letters 6(1): 1-5. [doi:10.1016/0020-0190(77)90002-3]

Lanning, D.R., G.K. Harell, J. Wang. (2014) *Dijkstra's Algorithm and Google Maps*. Proceedings of the ACM Southeast Regional Conference. [doi: 10.1145/2638404.2638494]

Paez, A. (2019) Using spatial filters and exploratory data analysis to enhance regression models of spatial data. *Geographical Analysis*, *51*(3), 314-338.

Preparata, F.P. and M.I. Shamos. (1985) *Computational Geometry – An Introduction*. Springer-Verlag. ISBN 0-387-96131-3.

Ting, D.S.W., L. Carin, V. Dzau, and T.Y. Wong. (2020) Digital Technology and COVID-19. *Nature medicine*, *26*(4), 459-461.

Wang, C.J., C.Y. Ng, and R.H. Brook. (2020) Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. *Jama*, *323*(14), 1341-1342.

Wu, P.Y. (2013) *Aggregation in Spatial Data Management: Prerequisite Database Concepts for GIS Skills in Retail Marketing*. Proceedings of Information Systems Education Conference (ISECON 2013). ISSN: 2567-1435, v.30, #2567, San Antonio, TX.

Wu, P.Y., E.J. Rathswohl. (2010) *Address Matching: An Expert System and Decision Support Application for GIS*. Proceedings of Information Systems Education Conference (ISECON 2010). ISSN: 1542-7382, #1339, Nashville, TN.

Yamada, I. (2017) *Thiessen Polygons*. The International Encyclopedia of Geography. Richardson, et al, eds. John Wiley & Sons, Ltd. [DOI: 10.1002/9781118786352.wbieg0157]

Zanbergen, P.A. (2013). *Python Scripting for ArcGIS*. ESRI Press. Redlands, CA.

Zhou, C., F. Su, T. Pei, A. Zhang, Y. Du, B. Luo, and C. Song. (2020) COVID-19: Challenges to GIS with big data. *Geography and Sustainability*.