

Visualization and Simulation of Disease Outbreaks: Spatially-Explicit Applications Using Disease Surveillance Data

Joseph Burns¹, Charles Hatt¹, Colin Brooks¹, Eric Keefauver¹, Eden V. Wells², Robert Shuchman¹, and Mark L. Wilson³

¹ Environmental and Emerging Technologies Division, Altarum Institute, Ann Arbor, MI

² Bureau of Epidemiology, Michigan Department of Community Health, Lansing, MI

³ Department of Epidemiology, School of Public Health, The University of Michigan, Ann Arbor, MI

Abstract

Public health agencies are interested in making effective use of disease reporting data and developing a better understanding of disease outbreaks. Methods of using Geographic Information Systems (GIS) to accomplish these goals were investigated as part of a larger study of disease surveillance. Two GIS-related aspects of this work are presented. First, an ArcObjects-based tool was developed for visualizing and monitoring disease outbreaks using information that is now available in recently-developed electronic disease surveillance databases. This provides public health agencies with the ability to understand the geo-temporal nature of outbreaks in near real-time. The resulting tool has been demonstrated on both simulated and actual disease outbreak data. Second, a contact network-based simulation of disease outbreaks was developed to provide an environment for assessing new surveillance concepts. An important component of the model is a variable capturing the degree of connectivity for both individuals and areas. Concepts for generating contact networks from GIS data are presented. An example of generating an alternative index of connectivity based on travel times is discussed.

1. Introduction

Public health organizations regularly collect surveillance data which is increasingly being stored electronically. This information is potentially very useful to the public health practitioner to review endemic disease levels, recognize outbreaks, and determine the source and proper response of outbreaks. GIS can be used to present multidimensional space-time indexed data in a manner that enhances data interpretation and data exploitation efficiency. The simplest output from a disease surveillance database is a list of numbers, such as cases, indexed on location and

time. The interpretation of this information by a public health practitioner may be enhanced if the spatial locations are displayed on a geographic map, which can be updated over time intervals. In addition, alternate representations, where the data are aggregated over geographic area or time interval, may present occurrence information in a GIS framework that can potentially aid epidemiological investigations and outbreak response.

Methods of displaying disease outbreak data from an electronic database as a function of location and time were investigated using both simulated disease outbreak data and a historical disease dataset from the State of Michigan. A multidisciplinary team of researchers from the Altarum Institute has collaborated with epidemiologists from the Michigan Department of Community Health and the University of Michigan School of Public Health to study methods of simulating, detecting, analyzing, and visualizing disease outbreaks using computer simulations and geospatial analysis tools. The goal of this effort has been to develop enhanced methods of detecting and responding to emerging infectious diseases and potential bioterror events. This paper will discuss two aspects of the research that used ESRI software tools to investigate the potential of geospatial analysis to enhance the modeling and analysis of disease outbreaks. Specifically, the development of a GIS tool to visualize and analyze data in disease surveillance databases will be discussed, and then a novel concept for using GIS information to advance disease outbreak simulation will be presented.

2. GIS-based Visualization and Exploitation of Electronic Disease Databases

Desktop ArcGIS was used to display the data as a function of location and time. Results were also aggregated over space and time, e.g. county and month. The results were presented to public health practitioners to evaluate the utility of various GIS maps for displaying and understanding endemic disease levels, the variation of disease outbreaks with space and time, and the relationship of the disease data to demographic data.

The GIS products displayed were considered useful by the public health practitioners. However, the manual production of the initial examples was a labor-intensive process. Therefore, an ArcObject-based tool was developed to automate the production of GIS visualization from an electronic disease database, to compare the current disease data to the historical background level, and to allow the user to change evaluation parameters. This tool,

called the “disease outbreak GIS visualization tool” (or “GIS outbreak tool” for short), provides the capability to automate the production of GIS products from historical data available in electronic disease surveillance databases and to provide a means of comparing current disease data to the historical level of incidence. This GIS outbreak tool is designed to track how many cases have occurred in specific areas in a defined time period. The purpose is to be able to “ring the alarm”, or alert public health, if the number of cases exceeds a certain threshold in a defined geographic area. For example, if a user such as a public health official determines that a disease incidence occurring in greater than 1% of a county population over a two week time period is a matter of concern, then this threshold could be detected through our ArcObjects tool. In its current form, our GIS outbreak tool allows the user to select either counties or 2000 Census tracts as the evaluation area, the length of the time period to consider as “recent” cases, and what threshold in cases per 100 people should be considered as a level of concern. Figure 1 shows how a disease database has been visualized by cases within a one-week time window for Michigan counties, and how certain counties are above the user-selected threshold.

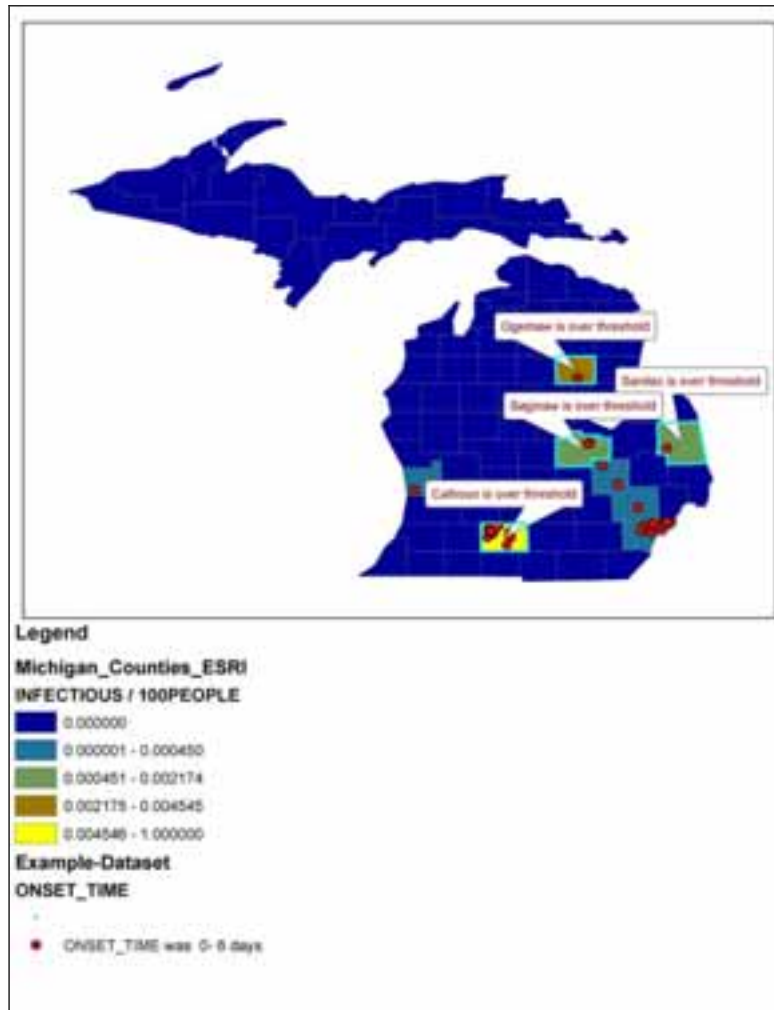


Figure 1: Example of output from outbreak visualization tool.

A description of the programming and geospatial layer parameters underlying the tool follows. The GIS outbreak tool was created as an ActiveX Dynamic Link Library (DLL) using Microsoft Visual Basic 6, and runs within ESRI ArcGIS version 9.0 and later. In order to use the tool, an input shapefile is needed that contains point data with a corresponding field that represents an event date (OnsetDate) that represents the time of disease onset (see Figure 2 for an example of the tool’s user interface). These points represent the geographic locations of people who at one time became infected with the disease in question. A layer is created from this shapefile and is assigned a “ClassBreaksRenderer” software object as its renderer. The user determines the infectious period with a dropdown box. When the user selects a date (SimulationDate) from the calendar control box, the ClassBreaksRenderer object is updated so that points will be displayed as a red dot if:

$$(\text{SimulationDate} - \text{InfectiousPeriod}) < \text{OnsetDate} \leq \text{SimulationDate}$$

That is, if the person is still considered to be infectious on the SimulationDate, display the person as a red dot on the map.

To use the alarm functionality, the user must also supply a geographic regions layer (such as counties or Census tracts), which must be a polygon shapefile. When the SimulationDate is set, the program code uses the ArcObjects parameter “ISpatialFilter” to count the number of people that are inside a region who are also considered infectious. This number (InfectiousLevel) is optionally normalized by user selectable field, and stored in another field within the geographic region shapefile. If the “InfectiousLevel” value is higher than a user defined threshold, a balloon style message (using the “IBalloonCallout” object available in ArcObjects) will originate from the centroid of the region alerting the user and the region will be automatically selected. Also, the user can then render the geographic region layer based upon the updated InfectiousLevel field using ArcMap’s native rendering options.

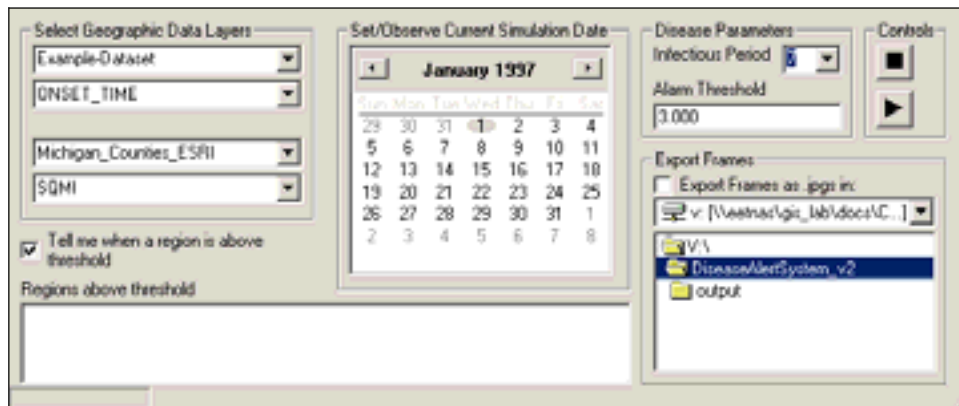


Figure 2: A screenshot of the user interface for the GIS outbreak visualization tool.

3. Enhanced Disease Propagation Simulations Using Contact Networks From GIS Information for Bio-Surveillance

We have also developed and demonstrated a computer simulation of disease propagation through contact networks. This statistical simulation models disease propagation through a population via person-to-person contact. The simulation allows the detection performance of bio-surveillance approaches to be studied as a function of spatial resolution (such as the size of geographic areas used in reporting) and the temporal resolution of reporting. Temporal resolution reflects both the rate at which time samples are collected, as well as the time delay in

reporting. Further, the simulation framework provided a virtual environment for assessing new surveillance concepts and algorithms, such as syndromic surveillance.

The computer simulation, which is based on disease propagation via contact, tracks the spread of disease over time through individuals displaced over a geographic area. The probability of transmission between individuals was based on a transmission rate, which was deduced from historical data, and a contact probability, which described how likely it was for two individuals to come into contact. The contact model used in the baseline simulation simply considers the distance between two individuals to determine the probability of contact. The baseline simulation appears to do a reasonable job of modeling temporal evolution of the disease, but has a rather simple contact model.

The computer simulation code is split into two parts. The first part is a Mathworks MATLAB script that takes simulation parameters such as transmission rate and contact probability, as well as a table that contains geographic locations for each individual that is part of the simulation, and outputs a text-formatted table that contains rows of data describing simulated individual case data such as unique ID, generation of infection (how many degrees of separation from the source infection), transmission time, symptom onset time, and doctor visit time. This table is then imported into two programs. For statistical analysis, it is processed with additional MATLAB scripts. A desktop application created using Visual Basic 6 is used for spatial viewing. The application used a customized ArcObjects-based MapControl to convert the ASCII text file into a shapefile and display the data. The user provides the number of days after the initial onset event and the software code calculates the infection status for all cases at that point in time, including the stage of infection, presence of symptoms, and level of medical attention for each individual. These three parameters are determined by comparing the user selected time with time fields stored in the shapefile that represent onset reporting events. A three character string describing the current state of the disease for each individual is stored in a field and rendered according to a specific scheme. Because the individual's stage of infection, the presence of symptoms, and the level of medical attention received may be independent from each other, a rendering scheme is used that describes infection stage with color, symptom presence with either hollow or filled points, and medical attention level with shape, as shown in Figure 3. Figure 4 displays the results of the disease propagation simulation with two different transmission rates, with the results being generated through the ArcObjects-based code.

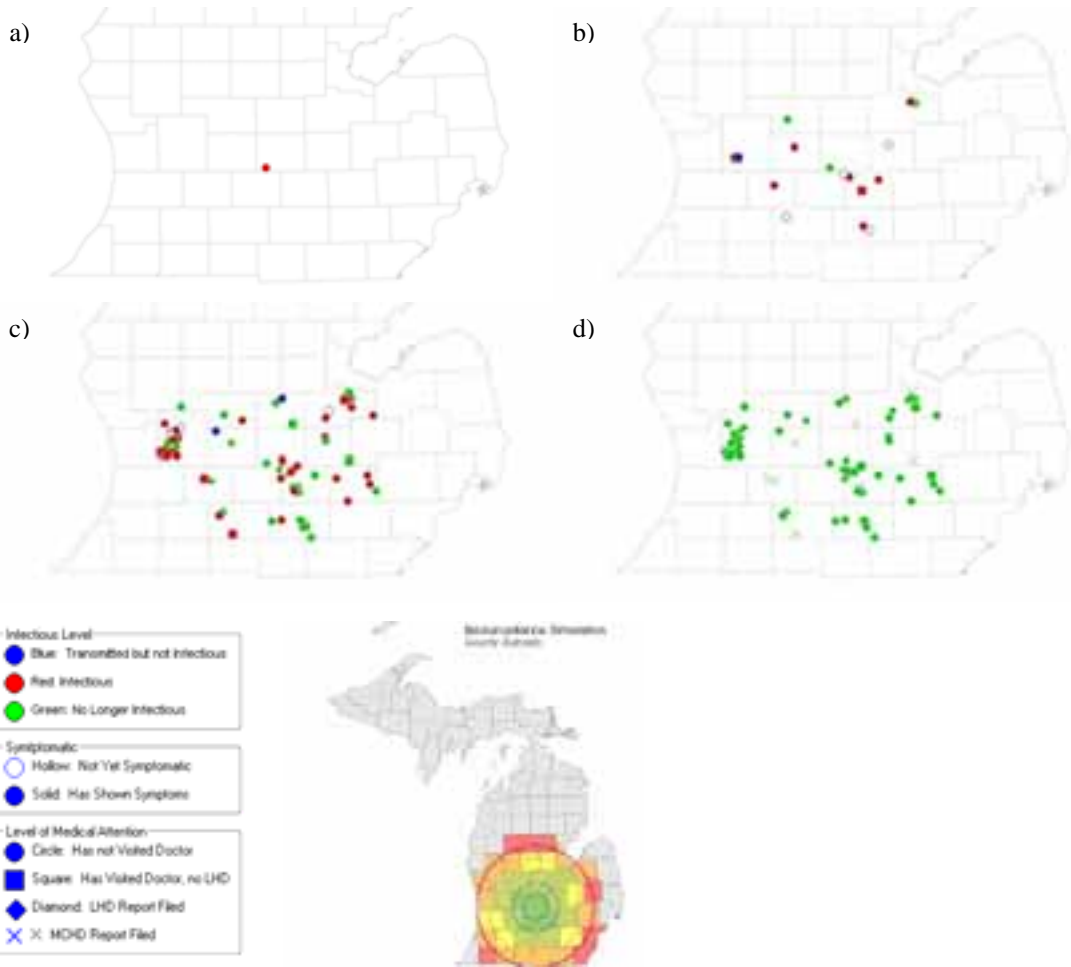


Figure 3: Example GIS output from the bio-surveillance simulation. The results for an outbreak of an influenza-like virus in a 13 county area are shown. The images are from days 2 (a), 12 (b), 20 (c) and 30 (d) of the outbreak. The transmission rate is 0.2.

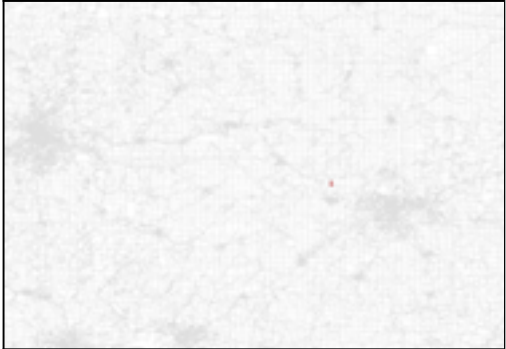
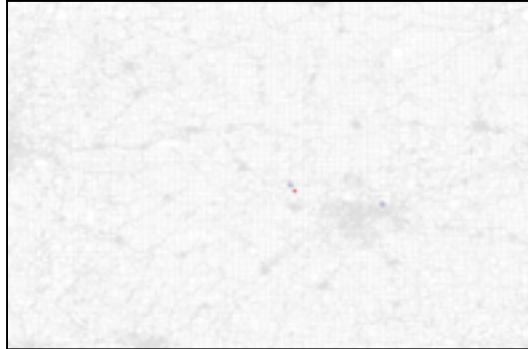




Day	Transmission Rate	
	0.2	0.6
3		
12		
25		

Figure 4: Example output from the viewer for the disease propagation simulation, shown for a four county area in Michigan with two different transmission rates.

The probability of contact between two individuals, either within a population group, or between population groups is influenced by more factors than their simple separation. Social, demographic, and geographic factors may influence the potential for interaction. Methods to incorporate this information into the contact model, especially tailored to local geographic regions, would provide an enhanced simulation capability.

In the context of the simulation, the contact model basically reflects the interaction between individuals or groups of individuals. These interactions generally do not occur uniformly in all directions. Some groups of individuals are more likely to interact with other groups of individuals. Researchers in several fields have begun to use network representations to model processes that propagate among groups of people along specific paths connecting the groups (for instance, see Newman [2002] and Meyers, et al., [2003]). In this representation, individuals or groups of individuals are represented as nodes. Interacting nodes are connected by paths, which usually have an associated probability that information is transferred or contacts occur along the path.

In a network-based approach, paths are defined to connect groups, and probabilities of movement along the paths are assigned. Some groups may be connected, while others may not. Some paths may have high movement, and others low. The network connections and probabilities are defined using some source of information. For instance, a network model for disease transmission in a hospital could potentially be deduced from a survey of the staff that described their movements in the hospital. For a population spread over a large geographic area, a method of defining the network model is not obvious.

Many researchers have studied the mathematics of processes operating on networks, but there does not seem to be a lot of work related to deriving the network. A possible solution is to use social, demographic, and geographic data contained in a GIS to define network connections between population centers. In our existing simulation, this approach, in principle, could weight interactions between groups based on GIS information. Specifically, the distance-based contact probability would be replaced by a function that reflects factors calculated using GIS databases.

The current simulation, even with a limited contact model, can simulate an outbreak of different contact-related diseases by changing the transmission parameters. Network information can be used to modify the probability of contact function. A network approach could provide a method of modeling other disease vectors. For waterborne disease, hydrologic information contained in a GIS could be used to develop a network model that connects communities along a watercourse. For an animal borne disease, natural resources and wildlife data contained in a GIS could be used to develop a network model that connects populations of animals and/or people associated with the animal group.

A similar approach may be applicable to developing a network model for our simulation. A road network could be used to provide an alternative estimate of the connectivity between individuals, and that estimate could be used in the contact model instead of the straight-line distance between points. For example, two individuals could be one mile apart in terms of straight-line distance, but could be connected by a long and curved road with a 35 mph speed limit. Another set of individuals might be 1.5-miles apart in terms of straight line distance, but could have a four-lane highway with a 65 mph speed limit directly connecting them. The second set of individuals, despite being further apart “as the crow flies,” would be considered closer together when estimating connectivity using a road network.

To evaluate this potential for developing an alternative index of connectivity between individuals, we used the ESRI StreetMap 2000 GIS data collection of U.S. roads included in the Desktop ArcGIS 9.1 data collection along with ArcGIS Network Analyst 9.1 as our primary GIS tools. We created 15 example point locations in Ann Arbor, Michigan, to serve as our point locations for simulated disease cases, and calculated the straight-line distance and travel times (with Network Analyst) for each of those points to a central location, which could be considered the disease origin. The 15 points were assigned a unique ID, the straight-line distance in kilometers and the travel time in minutes were calculated in ArcGIS for each point to the origin point, and the rank orders were calculated for distance versus travel time. To assess difference that using road-based connectivity might have, we were looking for cases where points were close to the origin point in distance and far away in travel time, or vice versa. Table 1 shows the 15 points with their calculated attribute values ordered by length of travel time, and Figure 5 shows the 15 points and the road networks used to calculate the travel time with Network Analyst. Points 2 and 7, along with points 6 and 14, form two sets of interesting examples for considering connectivity. Points 2 and 7, highlighted in yellow in Table 1 and Figure 5, are in eighth and tenth place in terms of straight-line distance, but are fifth and sixth in terms of travel time, while points 6 and 14 (highlighted in green) are fourth and fifth in straight-line distance but are ninth and tenth in travel time. The yellow points are “far away” in straight-line distance but close in travel time, while the green points are “nearby” in straight-line distance but “far” in travel time. A closer inspection of the underlying StreetMap road network reveals that the yellow points have a direct, major road connecting them to the central location, while the green points have meandering neighborhood streets with slower speed limits connecting them to the

central location. In terms of the original contact network, the green points would have been considered more closely connected to the origin point and more likely to receive a disease transmission. With our new travel-time based index of connectivity, the yellow points would now be more connected than the green points to the origin point, and would have an increased probability for disease transmission.

Table 1: Rank order list of 15 example points for straight-line distance versus travel time.

POINT UNIQUE ID	STRAIGHT- LINE DISTANCE in KM	ORDER BASED ON DISTANCE	TRAVEL TIME IN MINUTES	ORDER BASED ON TRAVEL TIME	NOTE
11	2.27	1	3.66	1	
8	2.56	2	3.96	2	
5	3.56	6	5.02	3	
3	3.07	3	5.24	4	
2	3.62	8	5.32	5	"Far" away in straight-line distance but close in travel time
7	3.93	10	5.48	6	"Far" away in straight-line distance but close in travel time
1	3.59	7	5.74	7	
10	3.81	9	6.09	8	
14	3.20	4	6.28	9	"Nearby" in straight-line distance but far in travel time
6	3.33	5	6.66	10	"Nearby" in straight-line distance but far in travel time
12	3.96	11	6.68	11	
13	4.18	13	7.20	12	
15	3.99	12	7.27	13	
9	4.90	15	7.84	14	
4	4.84	14	8.14	15	

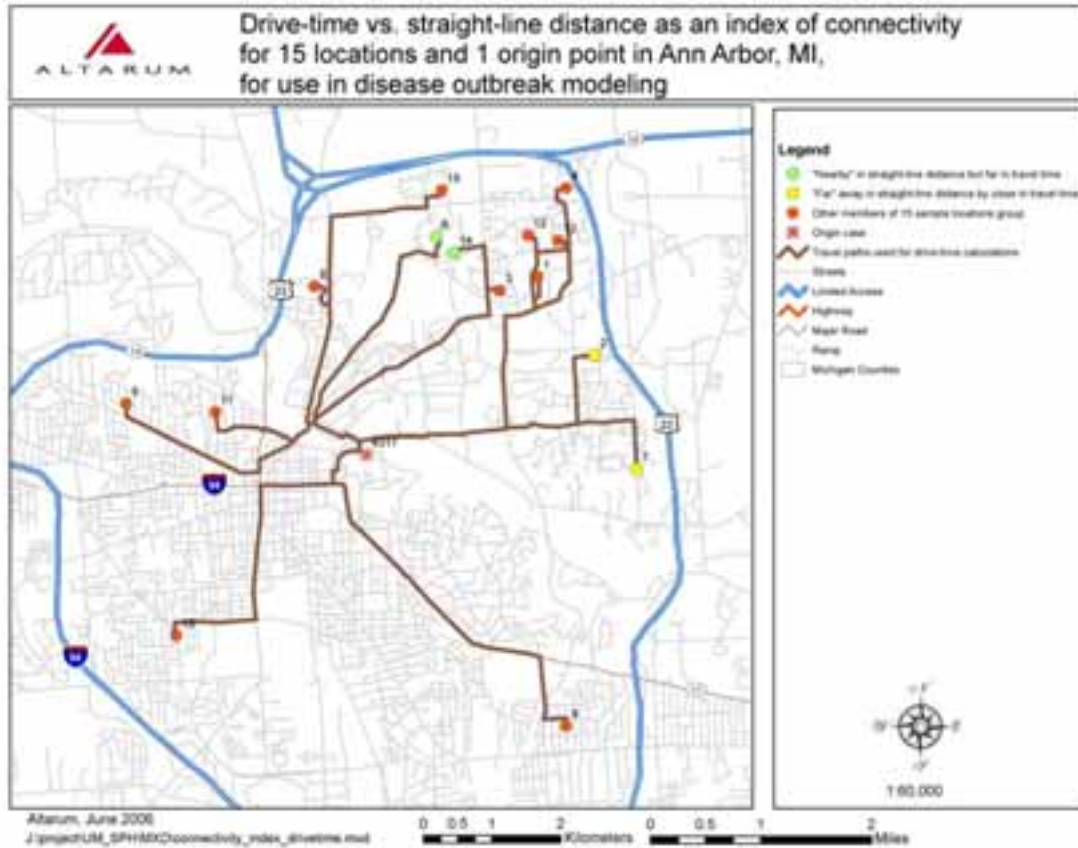


Figure 5: Comparison of drive time to straight-line distance for 15 example points.

The description of the 15 points serves as an example of alternative ways of calculating distance, and would need to be scaled up and modified to be more directly applicable to disease modeling. For example, the thousands of points evaluated in the GIS outbreak tool could have their travel times calculated to a suspected origin point to see which ones were most closely connected. Instead of specific locations representing individuals cases, the disease locations could be cities connected by highways, and the degree of connectivity of uninfected cities to a city with an outbreak could be calculated to see how likely the cities are to become infected. Cities could have their index of connectivity based on additional parameters, such as the number of people commuting from one city to another. U.S. Census tract to census tract commuting data are available in the Census 2000 Special Tabulation document STP-64, “Census Tract of Work by Census Tract of Residence.” The GIS layer providing the thread of connectivity would not necessarily have to be comprised of roads. For the waterborne disease idea mentioned earlier, the connection layer could be a river system, and the points could be cities along the river, with

the connectivity being estimated for various cities downstream. Another waterborne example could be for a water distribution system for a local water utility, and the points could be houses connected by the water lines. The index of connectivity could be calculated by water line length and pipe capacity.

4. Conclusion

Public health agencies are increasingly interested in making effective use of reported disease data and developing improved surveillance for detection and a better understanding of disease outbreak patterns. Two GIS-related methods of achieving these goals were presented. First, an ArcObjects-based tool was developed for visualizing and monitoring disease outbreaks using information that is available in newly-developed electronic disease surveillance databases within public health departments. This provides the ability for public health agencies to understand the geo-temporal nature of outbreaks in near real-time. The resulting tool has been demonstrated using both simulated and actual disease outbreak data. Second, a disease propagation simulation based on contact networks was developed using ArcObjects and MATLAB-based programming. One of the more time-consuming tasks in developing a contact-based simulation is to define the underlying contact network. A method of inferring contact networks from GIS data would make the development of these simulations more efficient and allow them to be rapidly adjusted to specific geographic areas. An example of generating alternative connectivity indices based on travel times was analyzed and discussed.

Acknowledgements

We would like to acknowledge Sean Savage for help with the initial development of the underlying GIS. Funding was provided by the Michigan Department of Community Health through a grant to the University of Michigan School of Public Health and a cooperative research agreement with the Altarum Institute.

References

Meyers, L.A., M.E.J. Newman, M. Martin and S. Schrag (2003), Applying Network Theory to Epidemics: Control Measures for Mycoplasma pneumoniae Outbreaks, *Emerging Infectious Diseases*, Vol. 9, No. 2, pp. 204-210.

Newman, MEJ (2002), Spread of epidemic disease on networks, Phys Rev E Stat Nonlin Soft Matter Phys; 66:016128.

Author Information

Joseph Burns

Altarum Institute
3520 Green Court, Suite 300
Ann Arbor, MI 48105
Phone 734-302-4725
Fax 734-302-4992
joseph.burns@altarum.org

Charles Hatt

Altarum Institute
3520 Green Court, Suite 300
Ann Arbor, MI MI
734-302-4726
charles.hatt@altarum.org

Colin N. Brooks

Altarum Institute
P.O. Box 134001
Ann Arbor, MI 48113
734-302-5633
colin.brooks@altarum.org

Eric Keefauver

Altarum Institute
P.O. Box 134001
Ann Arbor, MI 48113
734-302-4724
eric.keefauver@altarum.org

Dr. Eden V Wells

Michigan Department of Community Health
201 Townsend
P.O. Box 30195
Lansing, MI 48913
517-335-8165
wellsed@michigan.gov

Dr. Robert Shuchman

Altarum Institute

3520 Green Court, Suite 300
Ann Arbor, MI 48105
734-302-4600
robert.shuchman@altarum.org

Dr. Mark L. Wilson
University of Michigan
109 Observatory Street
Room 2006
Ann Arbor, MI 48109-2029
734-936-0152
wilsonml@umich.edu