Serie Geográfica
Número 12 - 2004-2005: 39 - 55

ISSN: 1136 - 5277

SPATIAL ALGORITHM FOR DETECTING DISEASE OUTBREAKS IN AUSTRALIA

S. Eagleson [1]
B. Veenendaal [1]
R. Watkins [2]
G. Wright [1]
A. Plant [2]

Fecha de Recepción: Agosto 2005
Fecha de Aceptación: Diciembre 2006

Australian Biosecurity Cooperative Research Centre for Emerging Infectious Disease
[1] Department of Spatial Sciences and
[2] Division of Health Sciences
Curtin University of Technology
GPO Box U1987 Perth WA 6845 Australia
s.eagleson@curtin.edu.au

RESUMEN

La detección temprana de brotes de enfermedades es esencial de cara a una intervención pronta en problemas de salud pública. Actualmente en Australia, las enfermedades notificables son recogidas y almacenadas, y referenciadas geográfica y temporalmente. Sin embargo, el proceso para la búsqueda de brotes de enfermedad sobre escalas espaciales distintas no está bien definido.

Los brotes son de detección difícil. Algunas enfermedades aparecen relativamente rápido, mientras otras requieren más tiempo para su incubación y sólo se hacen evidentes sobre largos intervalos temporales. En la práctica, los epidemiólogos combinan diferentes conjuntos de evidencias para determinar la probabilidad de la existencia de un brote. Gracias al progresivo incremento de disponibilidad de bases de datos electrónicas y de los Sistemas de Información Geográfica (SIG), el potencial para la utilización de técnicas de análisis espacial para la visualización, exploración y modelado de notificaciones de enfermedades para la detección temprana de brotes, es hoy mayor que en el pasado.

En este artículo, los autores presentan un algoritmo que emplea bases de datos de la administración, análisis espacial y SIG para la detección de clusters de enfermedades en el Estado de Australia Occidental. El algoritmo revisa los códigos postales de forma rutinaria hasta encontrar un número de casos que supera los valores que serían esperados en la región considerada. El algoritmo está diseñado para su uso por profesionales de la salud pública para asistir en la identificación y seguimiento de clusters en tiempo real.

Palabras Clave:

Cluster, SIG, análisis espacial, brotes de enfermedad

ABSTRACT

The early detection of disease outbreaks is essential for early intervention in potential public health problems. Currently in Australia, disease notifications are recorded, temporally and geographically referenced; however, the process of searching for outbreaks over different spatial scales is not well defined.

Disease outbreaks are difficult to detect. Some diseases appear relatively rapidly, while others take time to gestate and become apparent over long time intervals. In practice, epidemiologists combine different sets of evidence in different ways and apply reasoning to determine the likelihood of an outbreak. With an increase in the availability of electronic health-care data and geographic information systems (GIS), there is great potential to use spatial analysis techniques for the visualisation, exploration and modelling of disease notifications for the early detection of disease outbreaks.

In this paper, the authors present an algorithm that uses administrative databases, spatial analysis and GIS for the detection of disease clusters in Western Australia (WA). The algorithm routinely tests administrative areas (postcodes) and highlights the areas in which counts exceed the expected number for the particular region. This algorithm is intended to be used by public health officials to identify and track clusters in localised geographic areas in real-time.

Keywords:

Cluster, GIS, Spatial Analysis, Disease Outbreak

1. INTRODUCTION

A challenge for public health professionals is to identify an emerging outbreak of disease in its early stages so they can intervene and reduce its possible impact. In the past, a number of surveillance systems have been developed based on temporal analysis techniques. However, because the problem of disease outbreak is often dependent on the spatial diffusion of cases, many countries around the world are currently developing spatial surveillance systems.

Spatial surveillance is intended to identify deviations from the 'normal distribution of events in a region' (Lawson and Kleinman 2005). However, programming a computer to recognise if an abnormality has occurred is a difficult task. As outlined by Haggett (2000), the number of cases indicating the presence of an outbreak can vary according to the disease, the size and distribution of the population exposed, a lack of previous exposure and the diagnostic facilities available. Additionally, the cluster does not only depend on large numbers of cases or deaths. A single case of a disease long absent from the population, or the first invasion by a disease not previously recognised in that area, can require immediate investigation. Similarly, some surveillance systems adequately detect abrupt changes that might be caused, for example, by a serious bioterrorist attack. In many cases of public health surveillance, however, the change is gradual and difficult to detect (Haggett 2000).

The objective of this research is to develop a simple algorithm for the detection of clusters in routinely collected surveillance data. To be effective for public health professionals, the algorithm should be simple to understand and should produce unambiguous results. It should be able to search data hierarchically so that it can detect abnormalities at local, regional and national geographic scales. Most importantly, it must utilise existing surveillance data.

To achieve the objective, the algorithm presented within this paper applies a series of local area statistics to National Notifiable Disease Surveillance System (NNDSS) data to determine the regions in which a cluster is located. The paper has been structured into three sections. Section 1 provides background information about the data and the issues associated with detecting disease clusters in routinely collected surveillance data. Section 2 outlines the specifications and formalisation of an automated algorithm for the early detection of disease outbreaks in Australia. Section 3 illustrates the results based on the detection of Ross River virus (RRv) outbreaks in Western Australia.

1.1 The Problem

The fundamental characteristic distinguishing spatial data from time series data is the spatial arrangement or pattern of observations. In spatial surveillance, the objective is to identify clusters or the occurrence of disease notifications in a community or region that are in excess of the number of cases normally expected (Lawson and Kleinman 2005).

Within spatial analysis theory, a number of cluster detection methods have been developed. In general, the cluster detection methods available can be grouped according to the structure of the data available; for example, point or area. The analysis of point data involves analysing the distance between points to determine if clustering is present. Although this method is very accurate, the method can only be used in circumstances where the address of the person with the disease is available. In Australia, to preserve individual confidentiality, notification data are aggregated to administrative boundaries. To effectively analyse this data, a second set of methods utilises disease counts aggregated to administrative areas; for example, postcodes.

Aggregated data is subject to two confounding problems that must be considered when analysing data in this form. The first well documented problem is the modifiable areal-unit problem (MAUP) and the second is the spatial hierarchy problem.

1.2 The Modifiable Areal-Unit Problem

The MAUP is classic problem associated with the design and display of boundaries. The MAUP is 'a form of ecological fallacy associated with the aggregation of data into areal units for geographical analysis'. This aggregated data is then treated as individuals in analysis (Openshaw and Taylor 1981). An example of this process is census data, which is collected from every household but released only at CD boundaries. When values are averaged through the process of aggregation, variability in the dataset is lost, and the values of statistics computed at various resolutions will be different. This is called the scale effect. Additionally, the data analyst gets different results depending on the placement of the areal boundaries and how the spatial aggregation occurs. This is called the zoning effect. Hence, the choice of areal units is important for determining how disease counts are aggregated and analysed.

1.3 The Spatial Hierarchy Problem

The spatial hierarchy problem has arisen due to the increasing amount of data being required and being integrated into a diverse range of applications. These data are often referenced to administrative and management areas whose boundaries are intended to align, but which do not because of historical and recording factors in the capture or digitisation of boundary lines. Due to the uncoordinated delineation of these boundaries, cross-analysis between them is restricted. Essentially, this problem has occurred because organisations historically hand drafted the majority of boundaries on hardcopy maps. With advances in technology, these hand-drafted maps have been digitised for incorporation into GIS, a technology for which they have not been adequately designed (Eagleson et al. 2003). It is recognised that, in the future, this problem may be overcome by the development of 'mesh blocks'. These are geographic units that have been specifically designed to meet the administrative needs of administrative agencies to protect individuals' confidentiality whilst enabling data integration (Australian Bureau of Statistics 2004).

1.4 Cluster Detection Techniques

One of the most important tools in determining clusters is visualisation. This allows the analyst to inspect the data and to quickly identify abnormalities, relationships and interactions between regions on a map. The choropleth map is the most common method for displaying and visualising disease rates (Boscoe et al. 2003). Although popular, this method of analysis can be deceiving for two reasons. First, the legend classifications chosen to display the data often affect the interpretation. Second, in sparsely populated regions data are aggregated to large geographic areas that dominate the map (Talbot et al. 2000) whilst smaller areas with large populations 'at risk' from infectious diseases are not identified. This can cause spatial patterns to be identified where none actually exist, and inferences can be made on invalid assumptions (Jacquez 1998).

As an alternative to mapping the number of cases, data used in health applications can be mapped according to a measure of relative risk where the maps are calculated by standardising the observed count in each postcode by the corresponding population. Alternatively, probability maps can be generated that illustrate the probability of obtaining a count that is more 'extreme' than that actually observed.

A further alternative involves Bayesian statistics, where maps are produced using prior knowledge or beliefs about parameters of interest. Bayesian statistics utilise three kinds of information:

1. the observed disease events in an area;
2. prior information on the variability of disease rates in the overall map; and
3. information on the disease rates in an area's neighbours since geographic areas in proximity tend to have similar rates of disease.

By combining this information about the rates in surrounding areas, estimates for postcodes with small populations can be effectively smoothed. Techniques for creating these maps are fully described by Bailey and Gatrell (1995). It should be noted that, to be effective, the map being analysed should be based on smoothed estimates, cleaned of noise and adjusted for variations in the 'at risk' population (Berke 2004).

As an alternative to mapping, spatial statistics can be used to measure the values for each geographic feature since the statistics are independent of how the map is displayed. These statistics are commonly divided into two categories: global and local. Global statistics indicate if the overall pattern is autocorrelated. In contrast to global statistics, local area statistics have been developed to identify significant clustering in a local neighbourhood. One of the major disadvantages of using local area indicators to determine clustering is the error associated with the design of the neighbourhood and how much each neighbour contributes to that local area. This information is recorded in what is referred to as a weights matrix. Such errors can be in the form of the design parameters of the weights matrix or the design of the administrative units used in the analysis. Where possible, techniques such as Moran's I autocorrelation statistics should only be used in regions that are of similar size and arranged in a regular pattern, as is the case for image analysis where these models originated (Wakefield 2003). Additionally, the scale at which the analysis is undertaken can affect the display of the results and, therefore, the perception of significance recorded by the analyst. To overcome these limitations, a number of researchers have developed methods that scan the map searching for abnormalities or regions in which the observations exceed those expected. These techniques are often referred to as 'moving window statistics' or 'scan statistics'.

Moving Window Statistics

Moving window statistics are based on the null hypothesis that the incidence rate is the same over the region. The alternative hypothesis is that the incidence rate is higher than that in the given neighbourhood. Three of the most recognised methods for implementing moving window statistics are Openshaw's geographical analysis machine (GAM), Besag and Newell's test for clusters, and Kulldorff's spatial scan statistics.

The geographical analysis machine was developed by Openshaw et al. (1987) for the exploratory analysis of points or data attached to small areal units. It works by using points on a regular grid spacing on a map as the centroids for generating a series of concentric, equal-sized circles. For each circle, the number of events falling within the circle are counted and compared with the expected number (assuming a random generating process). If the circle contains a significantly higher number of events than expected, a circle is drawn on the map (O'Sullivan and Unwin 2002).

The method of Besag and Newell (1991) works using a predefined cluster size k. For each postcode in turn, a circle is drawn, centred on the case with a radius such that the kth nearest neighbour is included in the analysis. Unlike the geographical analysis machine, the circles are comparable since they are all based on k number of cases. This can be a disadvantage as the method is highly dependent on the choice of k (Wakefield et al. 2001). Testing of this method indicates that it is sensitive to inputs and works well for low disease counts (Wakefield et al. 2001).

Moving window statistics aid the analyst in determining the regions in which clusters in the data may occur. In particular, they have the advantage of being less sensitive to the effects of scale than the visual and autocorrelation techniques. They are, however, limited in the way they search for clusters using discrete circles. In fact, disease outbreaks spread based on gradients following trajectories, such as common routes of transport or wind dispersion patterns.

In the search for emerging diseases, the primary interest is the changing pattern of the disease through space and time. This is particularly important because the spread of disease is a dynamic process, and the pattern at a fixed point in time is not very informative about the way the pattern has emerged through time (Bailey and Gatrell 1995). Within the literature, there is very little practical experience of space-time geographic analysis because, until recently, there were little relevant data available and very few statistical methods developed for detecting or measuring space-time patterns.

2. ALGORITHM DEVELOPMENT

This section of the paper details the development of the algorithm. It begins with a description of the routinely collected surveillance data collected in Australia. This is followed by the specifications for the algorithm and an outline of the algorithm's components.

2.1 Surveillance Data

In Australia, there are approximately sixty different notifiable diseases. Each time a person is diagnosed with one of these diseases, a notification is made to the appropriate health authority. Every fortnight these notifications are compiled into the National Notifiable Diseases Surveillance System (NNDSS). To preserve individual confidentiality, notification data are aggregated to the postcode of residence.

Due to the large quantity of data currently collected within the NNDSS, an effective algorithm is required to effectively search this data for clusters. To achieve this, the authors have developed an algorithm specific to the problem of spatially searching NNDSS data for abnormalities or clusters, for a time interval allocated by the user. To be certain that the algorithm meets the objectives of the users, the following specifications have been developed.

2.2 Algorithm Specifications and Requirements

The basis of the method involves matching the user requirements and specifications with the data and techniques available for analysis. The components outlined below are the specifications for the algorithm.

1. *Automated:* The automated approach to the detection of disease outbreaks has the advantage of being fast, repeatable and flexible. The flexibility of the system enables thresholds to be set and changed according to the disease and level of enquiry by the user.

2. *Simple:* The simplicity of the algorithm means that the process of searching the data can be easily understood and modified by the operators. The system should be able to be changed as needed, and changes should require minimal time, personnel or other resources.

3. *Hierarchical:* The ability of the system to search the data hierarchically means that abnormalities can be detected at local, regional and national scales.

4. *Portable:* The algorithm should be able to be duplicated to another setting without additional resources.

5. *Surveillance Data:* The system is being designed such that it uses routinely collected surveillance data. This means the data

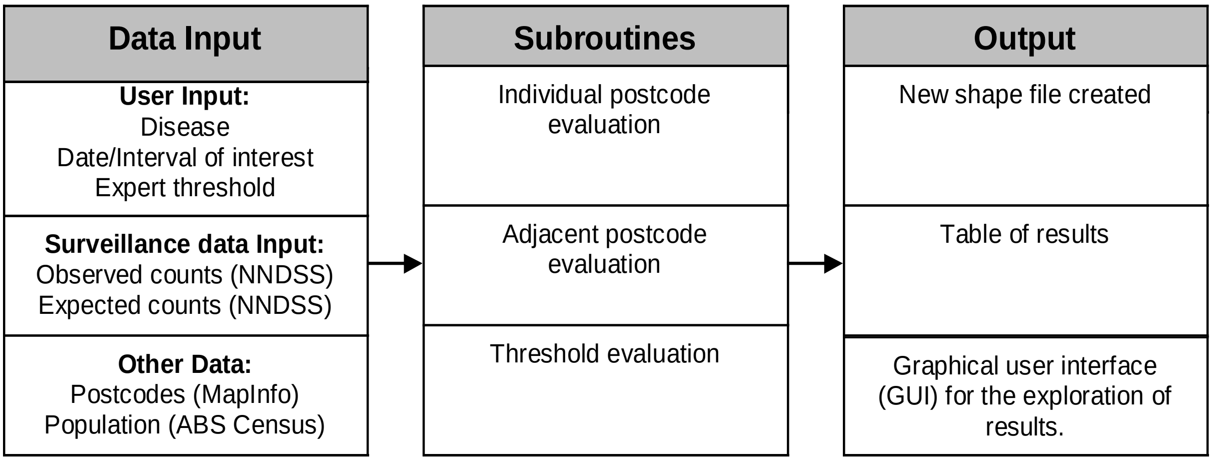

Figure 1.- The flow of inputs, algorithm subroutines and outputs.

processing standards are in place, and the data format and structure has been previously determined.

6. *Sensitivity:* The sensitivity of the system is the ability of the system to detect unusual events. If the main objective of the system is to monitor trends, a constant sensitivity with a reasonably low number of false alarms may be acceptable. However, if the objective is to detect outbreaks before they become widespread, the system must be highly sensitive to small changes in the number of notifications.

2.3 Algorithm Components

The user invokes a script that implements the algorithm, either directly through the Python interface or through the ArcGIS interface. The algorithm can be best described using the flow of events shown in Figure 1. Each of these events, data inputs, algorithm components and outputs are further detailed below.

2.4 Data Input

User Input

The user inputs three of the six inputs required by the algorithm. These are the disease being investigated, the date or time interval for investigation, and the threshold. The threshold is based on the experts' opinion of the number of notifications beyond the expected number of observations. An expert must determine these inputs, as they will affect the sensitivity of the system.

Surveillance System Input

The observed and expected counts are retrieved from the National Notifiable Disease Surveillance System (NNDSS) data and processed for input into the algorithm as follows

- *Observed:* This is the number of notifications recorded for the period of interest. Depending on the disease, the user can indicate if this is to be a daily or weekly count. Alternately, they can specify a date range.

- *Expected:* This is the number of notifications that would be expected. There are

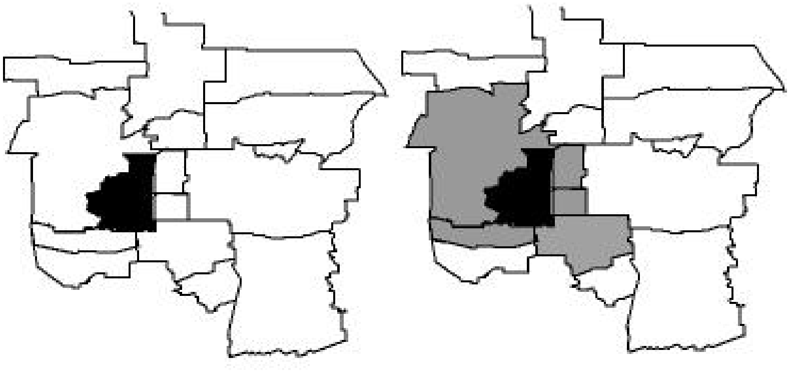

Figure 2.- (a) Initial polygon selection and comparison. (b) Intersection of adjacent polygons and comparison.

many different ways to calculate the expected number of notifications. In this instance, the value is an average of the number of counts received over the past three years within the same time interval as set for the observed number of notifications.

Other data sets that can be used by the algorithm include the population counts for each postcode, which is available from the Australian Bureau of Statistics, and the postcode boundaries obtained from MapInfo Australia.

2.5 Subroutines

The algorithm consists of three subroutines. Each of these has been developed to meet the user requirements.

1. *Postcode evaluation:* For each postcode, the number of diseases observed for a given time interval is subtracted from the expected number of disease notifications recorded. If the difference between the observed and expected exceeds the allowable threshold then an alert is sent.

2. *Adjacent postcode evaluation:* This routine involves comparing notifications aggregated for adjacent postcodes to determine if the neighbourhood is exceeding the expected values. In some instances, postcode values may be abnormal and this may not be cause for alarm. However, if the individual postcode and the surrounding postcodes are exceeding expectations then it could be significant. This configuration can also shed light on if the disease has begun to diffuse out from its central source of contamination.

3. *Threshold evaluation:* Given the variable size of postcodes across WA, it is important that the user is not just alerted to a particular area because of its large size. One way to overcome this problem is to alert the user to a specific area that exceeds the normal expectation. This routine is computed by using the expert-derived threshold per person and multiplying this value by the population in each postcode. If the observed number of notifications within the postcodes exceeds the threshold then the regions exceeding the threshold are displayed using a natural breaks classification.

3. CASE STUDY: ROSS ROVER VIRUS IN WESTERN AUSTRALIA

This case study investigates clustering of one disease: Ross River virus (RRv). As outlined in the introduction, this study is focused on the detection of outbreaks based on routinely collected disease notification data. RRv has been chosen for this study because of the relatively large number of notifications. Even with under-reporting, major peaks in activity can be identified.

The study area for this investigation is Western Australia (WA). WA is the largest state in Australia. It is comprised of over 380 postcode code units, which contrast in shape from very compact postcodes in urban areas to highly irregular postcodes along the coast and rural regions. They also vary in size from very small

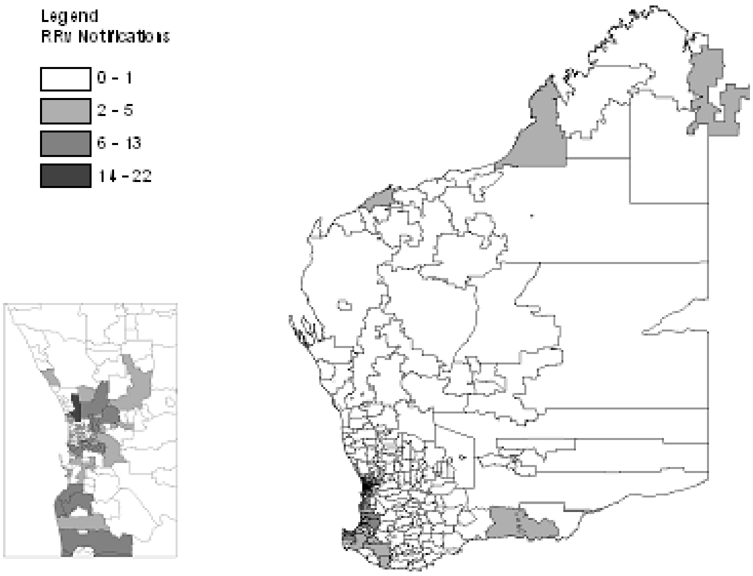

Figure 3.- Raw counts for RRv, WA January 2004

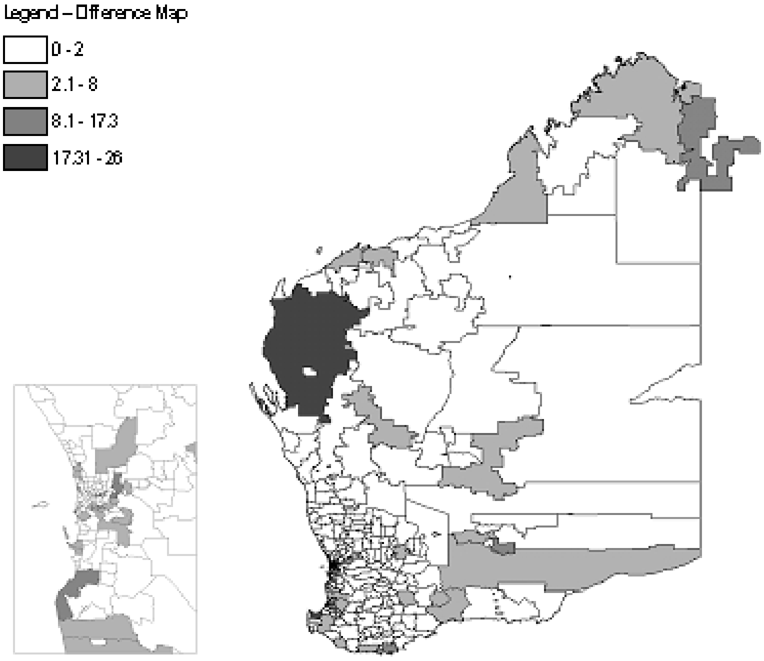

Figure 4.- Disease count comparison (Expected - Observed). The expected data is the average from 1999 to 2003, and the observed data is the total number of notifications received in 2004.

postcodes of approximately 200 m^2 to very large postcodes approximately 300 km^2.

Figure 3 illustrates an example choropleth map of RRv disease rates aggregated to postcodes for WA. The inset map identifies the Perth metropolitan region and surrounding areas. Although popular, this method of analysis can be deceiving for three reasons. First, the legend classifications chosen to display the data often affect the interpretation. Second, in sparsely populated regions data are aggregated to large geographic regions that dominate the map (Talbot et al. 2000). Third, the raw counts do not take into account the number of cases expected in a non-outbreak period. Within Figure 3, it appears that the outbreak is dispersed primarily within the metropolitan region.

Implementing the algorithm for RRv notifications has yielded the following results.

First, the results of the postcode evaluation (Figure 4) illustrate the difference between the observed and expected number of notifications for RRv in the month of January 2004 in Western Australia. It can be seen that the majority of regions are 0-2 notifications from the expected values. It can also be seen that, within the insert of the Perth metropolitan area, there appears to be more cases observed than expected. This is typical of the outbreak period selected for analysis. What this map does is inform analysts where the intensity of notifications is greatest. This intensity, whist still present in the metropolitan region, appears to be significant in some of the rural postcodes to the north and south of the state. Some of these areas have recorded more than 17 additional cases over that calculated in the baseline.

Second, the results of the postcode intersection evaluation (Figure 5) illustrate the smoothed number of cases; that is, the combined value of the postcode when added with all intersecting postcodes. This process smoothes

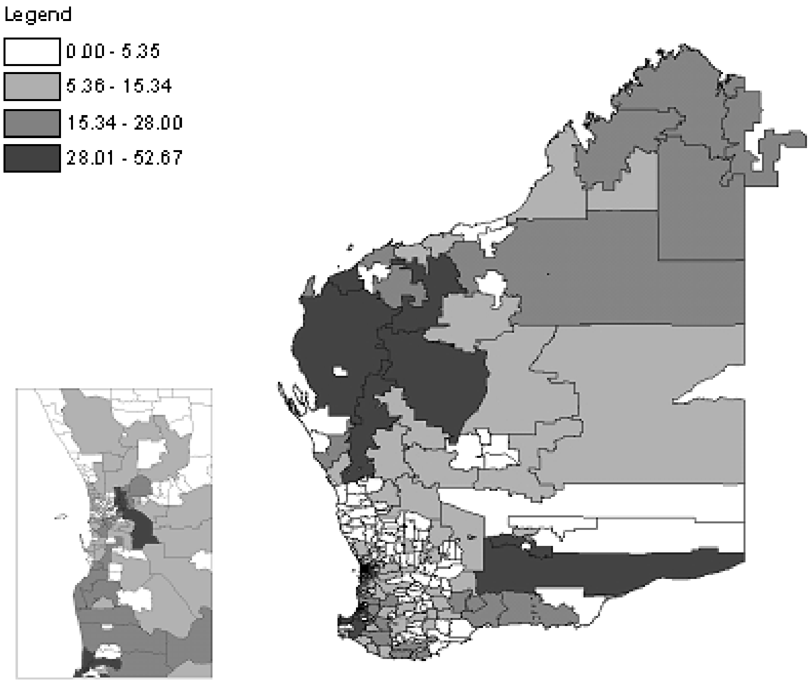

Figure 5.- Smoothed difference map. This is calculated using the adjacent postcode subroutine in which each postcode and its adjoining postcodes are compared as a group to expected counts. (This is calculated as the average for the previous 4 years.)

the data and provides information on the group of postcodes. This method is not appropriate in large rural areas where it appears to be giving a distorted interpolation of the data. This is a result of the large and irregular shape of postcodes in these areas. Results for the urban area are, however, consistent with research from the Department of Health and Ageing's Communicable Diseases Surveillance Report (2004). This states that the majority of cases from the 2004 outbreak of RRv were '... reported from the south-west region of the state in residents of, or visitors to, coastal areas stretching from Mandurah to Busselton. Transmission has also occurred across the Perth metropolitan area, particularly around the fringes.'

Third, the results of the threshold evaluation were computed based on the expert knowledge of two experienced epidemiologists. To determine an appropriate threshold, the epidemiologists reviewed the mean of non-outbreak days versus the mean of outbreak days within the NNDSS database over the past ten years. For the individual outbreaks identified, the mean number of cases reported per day for the whole outbreak period was as low as 0.9 cases per day for the smaller outbreaks in 1993 and 1995. It was as high as 8.4 cases per day for a large outbreak in 1996. The mean number of cases reported for all outbreak days was 3.1 cases per day, and the variance was 2.94 cases per day.

Using these values, the smallest outbreak figure of 0.9 notifications per day was chosen for the 1,962,100 people in WA. This threshold was then adjusted to reflect the number of residents in each postcode area to determine if the postcode was exhibiting values above or below the threshold. The results illustrate that

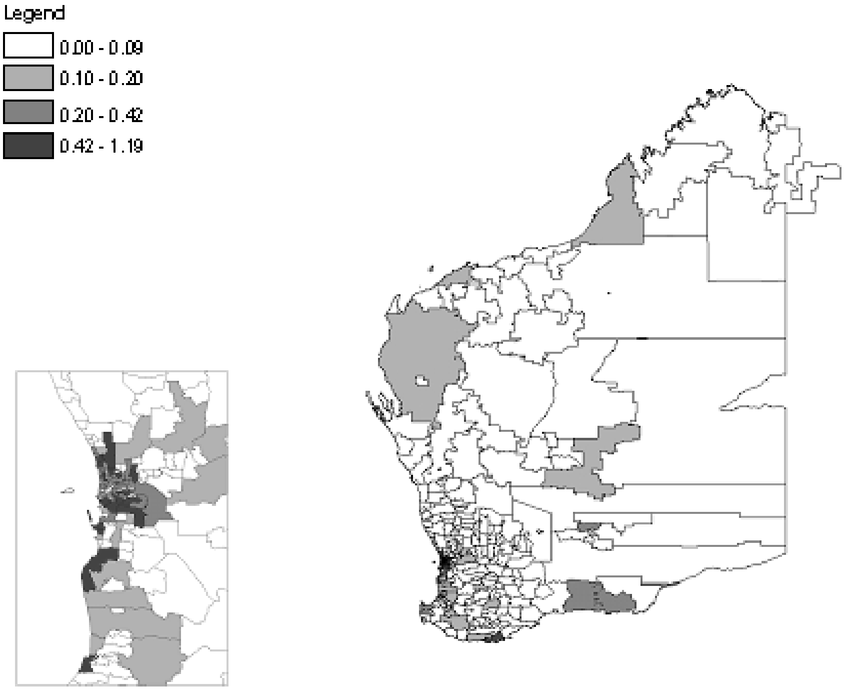

Figure 6.- Threshold evaluation

the majority of cases exceeding the threshold are in the southwest and metropolitan areas of the state.

It appears that second process, that of group calculation, is clearer for urban regions in which the postcodes are smaller and more compact. This set of analyses is likely to be affected by the MAUP because the data are aggregated to large postcode areas that cover a large area on the map.

3.1 Discussion

The results from the algorithm clearly indicate the regions in which elevated levels of disease activity have occurred. This is consistent with the results from the research from the Department of Health and Ageing's Communicable Diseases Surveillance Report (2004). Although the overall trends obtained from using the methods are similar, the sensitivity of the methods varies. It appears that in urban areas, where the postcodes are more regular in shape, the method of postcode intersection evaluation is appropriate. In contrast, in regions in which the postcodes are irregular in their design and size methods, threshold evaluation and individual postcode evaluation are more appropriate.

One of the fundamental impediments to be overcome in the development of an effective algorithm for the spatial analysis of disease outbreaks in Australia is the design of the administrative boundaries to which the data is aggregated. Currently, the data used in this analysis are aggregated to postcodes. Originally, postcodes were designed to aid the distribution of mail. This meant they were publicly recognisable units from which it was easy for organisations to collect and attribute information. However, Australia Post's boundary design does not adequately facilitate spatial analysis. It is recognised that this problem may, in the future, be overcome by the development of 'mesh blocks' (Australian Bureau of Statistics 2004). Alternatively, within each state or health district where address level data is stored, point-pattern analysis techniques may provide more accurate results.

3.2 Limitations

This investigation has a number of limitations. Retrospective review is not an absolute standard. Additionally, the algorithm is based on a number of assumptions. These assumptions are as follows:

Calculation of Expected

Major outbreaks of RRv occurred approximately every four years between 1992 and 2004. The multiple-year disease cycle is associated with large variation in baseline statistics when baseline periods of less than four years are used due to the timing of the epidemic (1992-1996) and inter-epidemic years.

Consequently, the calculation of the expected number of counts and the sensitivity of the system is highly dependent upon the baseline chosen in the analysis and major outbreaks during this time. One of the further developments of the system will involve removing past outbreaks from the baseline calculation as well as integrating the results between the three subroutines. For instance, because the threshold evaluation is independent of baseline calculations these results could be used to weight the values within the postcode adjacency evaluation to provide a more robust results.

Stability of the Population

Typically, the occurrences of a disease are expected to vary with the at-risk population. In the simplest case, everybody is equally at risk. More than likely, the at-risk population varies with the demographics of the region; for example, the percentage of children (O'Sullivan and Unwin 2002). When direct comparisons are made between observed and expected counts within a region, the assumption made is that the population and the demographics within the population have remained constant between the periods of analysis. This assumption may not be true, especially with the increasing number of coastal developments to which people of retirement age are migrating. Further research will involve the integration of population statistics into the algorithm.

NNDSS Data

The analysis in this paper is based on routinely collected surveillance data. In using surveillance data, the following assumptions have been made:

1. People acquire infection with RRv near their homes. However, in some circumstances people acquire diseases whilst on holiday in areas of increased activity.

2. The analysis is based on the registration of diagnosed cases. In reality, this may have its limitations as it is reliant on access to consultation by the patient, the correct diagnosis by the doctor, appropriate and accurate laboratory testing, and the case being reported to authorities.

Real-time detection of outbreaks requires that real-time data are available. Although processing systems are constantly being improved, it is recognised that diseases take time to gestate and to be diagnosed.

3.3 Further Research

Testing and Evaluation

Patterns of outbreaks can vary and have a significant impact on the performance of different detection methods. To improve the sensitivity of the system for the early detection of disease outbreaks, the algorithm thresholds will need to be calibrated to the specific disease. One way of doing this will be to run a temporal model in parallel to the algorithm. The temporal model will improve the sensitivity of the system to know when an outbreak is occurring, whilst the spatial algorithm developed above will provide information on where.

The model derived in this research is highly dependent upon accurate and current expected data sets as the input for the data analysis; however, it is recognised that this level of data sophistication may not always be available for surveillance. With further research, it is anticipated that indicators such as syndromic and sentinel indicators maybe incorporated into the algorithm.

The selection of baseline parameters to which the algorithm compares observed counts is based on available historic data. Changes in surveillance methods over time, including case notification methods and case definitions, can produce apparent changes in disease incidence when no real change in incidence has occurred. There is little specific information available to guide the selection of baselines in public health surveillance applications. Further research being undertaken by the authors will explore the influence of the selection of different baseline periods and, thus, the influence of variations in estimated baseline parameters on the accurate detection of disease outbreaks.

Temporal Analysis

The problem with detecting emerging disease outbreaks using only spatial tests is the low power to detect recently emerging clusters (Kulldorff 2001). This problem can be partially resolved by assigning the appropriate period in which to undertake the analysis. However, the appropriate number of years or cases to include is often unknown. If too few years are included, this might decrease the power to detect a low-to-moderate excess risk that has been present for considerable time. If too many years are included, the power to detect a very recent, high-excess cluster is reduced. A more rigorous solution to solving the problem is by integrating space-time analysis techniques.

The approaches outlined in this paper are spatial and largely ignore valuable attribute and temporal information that is often available. It is recognised that detection procedures need to incorporate many types of evidence to be effective for the early detection of outbreaks (Wagner et al. 2000). In practice, an epidemiologist combines the evidence in different ways to reason about potential outbreaks. Example datasets that aid the analyst include demographics, historical information, current notifications and expert knowledge.

Predictive Modelling

Once we are able to provide a descriptive illustration of the patterns for a given point in time, the next question of interest will be 'What will happen in the future?'. Directional bias and associated physical distance in disease transmission are two important attributes that can be modelled and monitored using spatial data. One technique used to make predictions into the future is Bayesian modelling. A Bayesian model defines prior and conditional probability distributions for each node and then uses combination rules to propagate conditional probability distributions through the model. The probability distributions can be derived from a combination of data and expert opinion. This process of combining probabilities produces conditional probabilities for each possible outcome (Bonham-Carter 1994). This approach will be further investigated in subsequent stages of the research project.

Implementation

Once complete, the algorithm will be incorporated with other spatial information - for example, data from hospitals and schools - in a GIS and made available to health officials. There are a number of options for the provision of this tool. These include distributing the Python script and providing the output files through a secured Internet site.

It is expected that the way to offer the analyst the most satisfactory solution will be to provide the results of the algorithm along with tools that will allow analysts to view and review their data to select appropriate representations. This information coupled with their training and expertise will then lead to the early detection of disease outbreaks.

4. CONCLUSION

GIS can offer quantitative and statistical measures along with visualisation tools to examine patterns of disease spread with respect to disease clusters (Lai et al. 2004). When applied to surveillance data in real-time, it can also aid in monitoring and enhancing the understanding of the transmission dynamics of an infectious agent. This facilitates the design, implementation and evaluation of potential intervention strategies.

This paper identifies current problems associated with disease outbreak detection in Australia. In response to these problems, the primary objective of this research has been to develop a new method through which spatial analysis and surveillance data can be used for the early detection of disease outbreaks.

The proposed solution involves the routine calculation of local area statistics using GIS. Initial results indicate that the process is promising, and future research will involve testing this algorithm with trained epidemiologists to assess the value of using the algorithm for the detection of disease outbreaks. At the very least, the tool developed will extend the capability of public health officials to analyse the spatial distribution of routinely collected surveillance data in Australia.

REFERENCES

Australian Bureau of Statistics (2004). 1209.0 Information Paper: Mesh Blocks. Canberra.

Bailey, T. C. and A. Gatrell (1995). *Interactive Spatial Data Analysis*. Essex, Prentice Hall.

Berke, O. (2004). "Exploratory Disease Mapping: kriging the Spatial Risk Function from Regional Count Data." *International Journal of Health Geographics* 3(18): 11.

Besag, J. and J. Newell (1991). "The Detection of Clusters in Rare Diseases." *Journal of the Royal Statistical Society* Series A(154): 143-155.

Bonham-Carter, G. F. (1994). *Geographic Information Systems for Geoscientists.* Ottowa.

Boscoe, F. P., C. McLaughlin, M. J. Schymura and C. L. Kielb (2003). "Visualisation of the Spatial Scan Statistic Using Nested Circles." *Health & Place* 9(3): 273-277.

Department of Health and Ageing (2004). *Communicable Diseases Surveillance Report* http://www.health.gov.au/ Date of Access: 3/08/2005

Eagleson, S., F. Escobar and I. P. Williamson (2003). "Automating the Administration Boundary Design Process Using Hierarchical Spatial Reasoning Theory and Geographical Information Systems." *International Journal of Geographical Information Science* 17(2): 99-118.

Haggett, P. (2000). *The Geographical Structure of Epidemics*. London, Clarendon Press.

Jacquez, G. M. (1998). GIS as an Enabling Technology. *GIS and Health*. A. Gatrell and M. Loytonen. Philadelphia, Taylor and Francis: 17-28.

Kulldorff, M. (2001). "Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic." *J.R Statistical Society A* 164, Part 1: 61-72.

Lai, P., C. Wong, A. Hedley, S. Lo, P. Leung, J. Kong and G. Leung (2004). "Understanding the Spatial Clustering of Severe Acute Respiratory Syndrome (SARS) in Hong Kong." *Environment Health Perspectives* 112.(15): 1550-6.

Lawson, A. and K. Kleinman (2005). *Spatial & Syndromic Surveillance*. Chichester, John Wiley & Sons, Ltd.

O'Sullivan, D. and D. J. Unwin (2002). *Geographic Information Analysis*. New Jersey, John Wiley & Sons, Inc.

Openshaw, S. and Taylor, P. (1981), 'The Modifiable Areal Unit Problem', in *Quantitative Geography: A British View, eds.* Wrigley, N. & Bennett, R.J., Routledge, London.

Talbot, T. O., M. Kulldorff, S. P. Forand and V. Haley (2000). "Evaluation of Spatial Filters to Create Smoothed Maps of Health Data." *Statistics in Medicine* 19: 2399-2408.

Wagner, M., J. Espino, F. C. Tsui, L. Harrison and W. Pasculle (2000). *Real-time Detection of Disease Outbreaks.* Pittsburgh, University of Pittsburgh.

Wakefield, J. (2003). *Spatial Epidemiology: Methods and Applications.* Seattle, Department of Statistics and Biostatistics, University of Washington: 282.

Wakefield, J., J. E. Kelsall and S. E. Morris (2001). Clustering, Cluster detection and Spatial Variation in Risk. *Spatial Epidemiology Methods and Applications*. P. Elliot, J. Wakefield, N. Best and D. Briggs. Oxford, Oxford Medical Publications: 128-150.