Spatio-temporal analysis of infectious disease outbreaks in veterinary medicine: clusters, hotspots and foci

Michael P. Ward

Summary

Analysis of disease data that has an implicit spatio-temporal component (such as disease outbreaks, data generated by surveillance systems and specific hypothesis-based veterinary field research) is a foundation of veterinary epidemiology and preventive medicine. Components of this process include exploratory spatial data analysis (finding interesting patterns), visualisation (showing interesting patterns) and spatial modelling (explaining interesting patterns). Spatio-temporal statistics and tests are valuable when adding precision to qualitative verbal descriptions, facilitating the comparison of distributions and drawing attention to characteristics unlikely to be noticed by visual inspection. Quantifying spatio-temporal patterns is important for understanding how disease phenomena behave. The application of a range of spatio-temporal statistics is illustrated by exploratory spatial data analysis and visualisation of the 2002 outbreak of West Nile virus encephalomyelitis in Texas equines. This large outbreak (1 698 reported cases) consisted of both point (latitude, longitude) and polygon (Texas counties) spatial data with a time component (reported date of onset of clinical disease) and case series and attack rate data. This example highlights the need to use a range of techniques to fully understand the spatio-temporal nature of disease occurrence. With knowledge of how disease occurs in time and space, appropriate and effective disease control, prevention and surveillance programmes can be implemented.

Keywords

Clustering, Disease, Epidemiology, Geographic information system, Space, Statistics, Texas, Time, United States of America, West Nile virus.

Analisi spazio-temporali di focolai di malattie infettive in medicina veterinaria: cluster, hotspot e foci

Riassunto

L'analisi dei dati di carattere sanitario che hanno un'implicita componente spazio-temporale (quali quelli che riguardano i focolai di malattie, i dati generati dai sistemi di sorveglianza e da ricerche di campo basate su specifiche ipotesi) rappresenta il punto di partenza fondamentale dell'epidemiologia veterinaria e della medicina preventiva. Le componenti di questo processo comprendono l'analisi spaziale esplorativa dei dati (ricerca di schemi fonte di interesse), visualizzazione (presentazione di schemi fonte di interesse) e modellazione spaziale (spiegazione di schemi che suscitino interesse). La statistica e i test spazio temporali sono preziosi quando aggiungono precisione alle descrizioni verbali qualitative, facilitando la comparazione di distribuzioni e attirando l'attenzione sulle caratteristiche improbabilmente rilevabili con un semplice esame visivo. La quantificazione dei modelli spazio-

Texas A&M University College of Veterinary Medicine and Biomedical Sciences, MS 4458, College Station, TX 77843-4458, United States of America
mward@cvm.tamu.edu

temporali è importante per comprendere il comportamento di alcuni fenomeni legati alle malattie. Viene illustrata l'applicazione di una gamma di statistiche spazio-temporali attraverso l'analisi esplorativa di dati spaziali nonché la visualizzazione del focolaio, di encefalomielite dal virus West Nile negli equini in Texas nell'anno 2002. La descrizione di questo esteso focolaio (i casi riportati sono 1.698) comprende sia dati spaziali puntuali (latitudine e longitudine) sia poligonali (contee del Texas) associati ad una componente temporale (data di insorgenza della malattia clinica) alla successione di casi e dati relativi al tasso di attacco. In questo studio si evidenzia la necessità di utilizzare una gamma di diverse tecniche di analisi per comprendere appieno la natura spazio-temporale del verificarsi della malattia. La conoscenza delle modalità con cui una malattia si verifica in termini di tempo e spazio, sta alla base dell'implementa-zione di specifici ed efficaci sistemi di controllo e prevenzione nonché di specifici programmi di sorveglianza.

Parole chiave

Clustering, Epidemiologia, Malattie, Sistema informativo geografico, Spazio, Statistica, Texas, Tempo, Stati Uniti d'America, Virus West Nile.

Introduction

Understanding the distribution of disease in time and space is a foundation of epidemiology and hence preventive medicine programmes. Knowledge of where and when a disease occurs enables the generation of disease causation hypotheses for diseases with unknown or poorly characterised aetiology, identification of disease risk factors and the design of efficient disease surveillance and control programmes in animal health. The drawing of epidemic curves and construction of maps are basic skills used by epidemiologists to investigate disease occurrence. Spatio-temporal statistics and tests are useful for adding precision to qualitative verbal descriptions, facilitating the comparison of distributions and drawing attention to characteristics that might not be identified upon visual inspection. Quantifying spatio-temporal patterns is important in the understanding of how spatio-temporal phenomena behave. Statistics quantify patterns.

Specific statistical tests and techniques have been available to analyse spatio-temporal disease data for at least 60 years. During the 1950s and 1960s, many techniques were developed and applied, including the nearest neighbour index (4), the autocorrelation statistic (18), Ederer-Myers-Mantel disease-clustering procedure (7), Knox test (13) and the temporal scan (19, 20). Within the last 20 years, additional techniques have been developed to meet specific needs, including the Cuzick and Edwards test for inhomogeneous populations (5), population density adjusted auto-correlation (21), spatial and spatio-temporal scan statistics (15, 16), local neighbourhood statistics (1, 9), empirical Bayes smoothing and kriging and other interpolation methods. Despite the development of a range of techniques that can be applied to analysing spatio-temporal phenomena, their application in the analysis of disease occurrence in animal health has been limited (31). We face a challenge in implementing these techniques as routine procedures within animal health control and prevention programmes. The aim of this paper is to show how a range of available spatio-temporal techniques and statistical tests can be applied to develop a better understanding of how disease occurs in time and space. This is illustrated by exploratory spatial data analysis and visualisation of the 2002 outbreak of West Nile virus encephalomyelitis in Texas equines. This example highlights the need to use a range of techniques to fully understand the spatio-temporal nature of disease occurrence.

Materials and methods

Source data

West Nile virus (WNV) is a mosquito-borne flavivirus. It was first detected in the Americas in 1999 as a cause of neurological disease in humans, horses and birds in the vicinity of New York City. Horses are particularly susceptible to WNV infection and may present acute clinical signs of encephalomyelitis, such

as ataxia, rear limb paresis, muscle tremors and fasciculations, and recumbency (25, 32). Although 80% of affected horses recover in three to four weeks with supportive treatment, a small proportion may have persistent neurological disorders (24). Since the mid-1990s, the number of severe WNV disease outbreaks in equine populations has increased (2): recent outbreaks include Morocco (1996, 2003), Israel (1998-2000), Italy (1998) and France (2006). The North American outbreak of equine WNV encephalomyelitis exploded during 2002, with nearly 15 000 laboratory-confirmed cases in 44 states in the United States, 5 provinces in Canada and 3 states in Mexico.

Despite an active surveillance programme (sentinel flocks, mosquito-trapping, human and equine case reporting) and detection in neighbouring Louisiana and Oklahoma, WNV was not detected in the state of Texas until 2002. A total of 1 698 laboratory-confirmed (IgM-enzyme-linked immunosorbent assay [ELISA]) equine cases were reported during 2002 in 204 of 254 Texas counties (33). The first cases were reported on 27 June from eastern coastal Texas. The epidemic peak occurred on 5 October and 50% of cases were reported during a period of six weeks (3 September to 17 October). The epidemic lasted 25 weeks and appeared to consist of three phases, as follows:

- 27 June-25 July: 44 cases (2.6%)
- 26 July-27 September: 633 cases (37.3%)
- 28 September-17 December: 1 021 cases (60.1%).

The coordinates (latitude, longitude) were available for 1 334 of the 1 698 cases (78.6%). The county of origin was identified for all reported cases. In addition, estimates of the number of equines present in each Texas county in 2002 was available from the National Agricultural Statistics Service (NASS) website (www.nass.usda.gov/census/census02/profiles/tx/index.htm).

Spatio-temporal techniques and statistical tests

Investigations of spatio-temporal disease distribution generally focus on whether disease cases are likely to have occurred at random. If this null hypothesis is rejected, disease cases may be overdispersed or clustered. Our interest is usually in clustering, since this implies that common animal characteristics, a source of exposure, or common environmental characteristics have lead to foci of disease. Identifying these foci is the first step in elucidating aetiology and thus designing control, prevention and surveillance programmes. Thus, we generally search for evidence when and where disease events occur are correlated.

The first step in understanding empirical data (whether it has a spatial component or not) is the calculation of descriptive statistics. The central tendency of a set of points (or centroids of polygons) may be described by the arithmetic mean or a mean weighted by some attribute (for example, the date of onset of disease at each location or the estimated attack rate at each polygon centroid). The degree of dispersion of a set of points is measured by the standard distance deviation or standard deviational ellipse.

Three types of distributions may be observed when studying populations, diseases or other events: uniform (evenly distributed), clustered (aggregated) or random. In the clustered distribution, there is a definite, discernible aggregation of points. One of the most common methods of describing the distribution of a set of disease cases, whether measured as points or polygons, is Moran's autocorrelation statistic. This is an example of a global spatial test and is similar to the traditional Pearson correlation coefficient, except that the correlation of values of the same variable at different spatial locations is examined, with a weight matrix being included to define the spatial relationships between points or polygons. This weight matrix is often based on Euclidean distance, but may be modified to take into account neighbourhood relationships (for example, 1 for adjacent pairs and 0 otherwise). A positive autocorrelation implies clustering. A similar statistic is the nearest neighbour index (4), which is the ratio of mean Euclidean distance between nearest neighbour points in a given area and the mean distance expected

from a randomly distributed series of points, calculated based on the study area or Poisson probability density function (11). Moran's autocorrelation and the nearest neighbour index are sensitive to the spatial distribution of the underlying population at risk of disease. If the population at risk is clustered (for example, dairy herds in most countries are found in certain suitable ecoclimatic zones), then disease cases arising from that population are also expected to be clustered, even if the disease occurrence is not clustered *per se*. Adjustment for spatial variation in population density may be achieved using commonly available methods, such as standardisation. However, in many health studies and particularly in veterinary medicine, the information necessary to perform such adjustments and retain interpretability may be unavailable. To take into account population inhomogeneity, a modified autocorrelation (Ipop) was proposed by Oden (21) in which autocorrelation is adjusted to account for differences in population size across areas. A further development of the nearest neighbour index is the Cuzick and Edwards test for inhomogenous populations (15). This test compares the locations of case and control locations. Controls are drawn from the same underlying population as cases, thereby accounting for clustering that may occur in the population regardless of the clustering of cases. The test statistic is the number (summed over all cases and controls) of cases that are nearest neighbours to each individual case. The order of the analysis can range from 1 (nearest neighbour) to n (farthest neighbour within the study area).

Several statistical tests have been developed to assess global clustering of events jointly in time and space, including Barton's method, the Knox test (13), a nearest-neighbour test (12) and Mantel's time-space correlation statistic (17). The Mantel test statistic is the sum, across all pairs of events, of the time distance multiplied by the spatial distance. The standardised Mantel statistic (range –1 to 1) is a measure of matrix correlation and can be interpreted in a similar manner to Pearson's correlation statistic: a positive value implies that events at locations close (or far) in space tend to occur close (or far) in time (or that events occurring close [or far] in time tend to be located close to [or far from] each other). The null hypothesis tested is that the time and space distances are independent. The significance of Mantel's statistic can be tested using a randomisation process. Unlike some other methods (for example, the Knox test), Mantel's method does not require specifying critical or threshold distances for space-time association. However, a weakness in the Mantel approach is that it is based on a linear model and therefore is insensitive to non-linear dependence of time on space or space on time and it may be excessively influenced by larger distances in a data set (3). These issues may be addressed by various transformations of time and/or space distances.

Moran's autocorrelation, Oden's Ipop, the nearest neighbour index, the Cuzick and Edwards test, and Mantel's correlation statistic are examples of global spatial statistics, used to explore clustering without pre-determined hypotheses regarding cluster location or extent. If the interest in analysis is to identify local clusters, two recently developed techniques are commonly used: Anselin's local indicator of spatial autocorrelation (LISA) and the Getis-Ord Gi* statistic (1, 9). Given a set of weighted data points (for example, date of onset of disease cases, or disease rates for areas represented by centroids), the LISA identifies those clusters with values similar in magnitude, and those clusters with very heterogeneous values. In essence, the LISA decomposes Moran's autocorrelation statistic into contributions for each case location. Thus, the sum of LISAs for all observations is proportional to Moran's autocorrelation statistic.

When visually assessing choropleth maps of disease rates and proportions, an issue that must be addressed is the estimation of these metrics for areas with small denominator information. Given that areas, such as counties and other administrative units, may vary greatly in terms of animal populations, disease rates and proportions presented on a map can hide vastly different levels of confidence

implicit in these metrics. One solution to this problem is to use empirical Bayes smoothing. This procedure adjusts estimates for individual areas based on the overall (global) disease rate estimated for the entire study region (the 'prior' distribution). Those areas with small animal populations are adjusted more than those with large populations, reflecting the statistical reliability of the estimates. As a result, disease rates and proportions are made more stable and less variable.

A technique that allows detection of both global clustering and the identification of the location of specific clusters, and clustering in time, space and in time and space, is the scan statistic. The spatial scan statistic (15) uses a theoretical circular window placed on a map of all locations included in a study. This scanning window is sequentially centred around one of many possible centroids in the study area. For each centroid, the window radius may vary continuously from zero to some upper limit selected by the investigator. An upper limit of 50% of the study area is recommended (14). Thus, the procedure creates – in theory – an infinite number of distinct geographical circles, containing within them different sets of neighbouring locations. Each set of locations is a possible candidate for a cluster. However, since discrete locations (longitude, latitude) or the centroid of areas within a study are used in spatial analysis, the number of candidate circles that must be assessed is finite. The procedure is considered invalid if the choice of radius is made after examining the data and estimating the size of potential clusters, or if the procedure is used to identify the window that best fits the data (14).

The scan procedure is flexible in that data can be analysed using two different probabilistic models, based on the Bernoulli or Poisson distributions. For the Bernoulli model, the data has the form of cases and non-cases coded as '1' or '0'. Cases and non-cases may be selected from the study population, or may represent the entire study population. For the Poisson model, the number of cases at each location or within each area is assumed to be Poisson distributed. Under the null hypothesis, the expected number of cases at each location is proportional to the population size or population-time at risk at that location. For spatial analysis, results from using both models are generally similar (14). When there are few (<10%) cases compared to controls the Poisson model is a very good approximation of the Bernoulli model, although it may produce slightly conservative *P*-values. Calculations using the Poisson model typically take less computer time to complete than if the Bernoulli model is used (14).

The spatial distribution of grazing livestock is almost always heterogenous; some areas may be intensively grazed and other areas may not. Similarly, the distribution of livestock enterprises, such as dairies, feedlots or poultry houses, are commonly clustered away from human population centres. This has implications for selecting a spatial cluster statistic. The question of interest is usually 'does spatial clustering occur above and beyond the spatial clustering of cases that arises due to spatial variation in population density?' Adjustment for spatial variation in population density may be achieved using commonly available methods, such as standardisation. Regardless of the model (Bernoulli or Poisson) used in the spatial scan procedure, adjustment for lack of population homogeneity is achieved by conditioning on the total number of cases observed to calculate the expected number of cases for each location, a form of indirect adjustment.

The spatial scan statistic is a cluster detection test, able to both locate and test the significance of clusters (29). The scan procedure may be used to detect clusters with high, low or high and low rates of disease. The latter is equivalent to a two-sided test. The most common analysis is to scan for areas with high rates, that is, for clusters. For each location and size of scanning window used, the alternative hypothesis is that there is an elevated (or decreased or either elevated or decreased) rate within the window as compared to outside. The likelihood function is maximised over all windows, identifying the window that constitutes the most likely cluster – the cluster that is least likely to have occurred by chance. The likelihood ratio for

this window is the maximum likelihood ratio test statistic. Its distribution under the null-hypothesis and its corresponding *P*-value is obtained by repeating the same analytic exercise on a large number of randomly selected replications of the data set generated under the null hypothesis, in a Monte Carlo simulation.

Data analysis

The central location of the 1 334 WNV case locations was described using the mean centre statistic (8), and was compared with the mean centre weighted by date of onset (27 June to 17 December). The standard deviational ellipse was calculated for each mean centre estimated. Mean centres and standard distance deviational ellipses were also calculated for each of the three recognisable phases of the epidemic, phase I (44 cases; 27 June-25 July), phase II (633 cases; 26 July-27 September) and phase III (1 021 cases; 28 September-17 December).

The number of reported cases were summed by county (204 counties reporting equine WNV cases during 2002) and county attack rates (cases per 1 000 horses at risk) were estimated. The centroid (latitude, longitude) of each Texas county was identified (8). Mean centres and standard distance deviational ellipses were calculated, weighted by county number of cases, population at risk or attack rates (8).

The spatial distribution of all cases, and of cases in each of the three epidemic phases, was described using the nearest neighbour test and Moran's autocorrelation statistic (8). The distance between nearest neighbours expected under spatial randomness was calculated based on an estimation of the Poisson parameter (number of cases divided by study area). The spatial distribution of county attack rates was also described using Moran's autocorrelation statistic (8). Spatial clustering of counties reporting WNV cases was compared to those not reporting cases using the Cuzick and Edwards test, using county centroid as the indicator of spatial location and Euclidean distance to characterise the relationship between county centroids (28).

County-specific WNV attack rates were smoothed, using an empirical Bayes method (27). The rate in each county was smoothed using data from its nearest (estimated by inverse squared Euclidean distance) 10 neighbouring counties.

The location of case clusters was identified using spatial and spatio-temporal scan statistics. A Poisson model was used (since <10% of the study population – 327 563 horses in Texas in 2002 – were reported as cases). Case information consisted of the number, county and date of onset of reported cases. Population information consisted of the estimated number of horses in each Texas county in 2002. Location information consisted of the centroids (latitude, longitude) of each county in Texas. Data was scanned with a spatial window of 1% of study area (approximately 7 000 km^2) and a temporal window of 30 days. Only clusters with high (rather than low or low and high) WNV case attack rates were identified. The likelihood ratio test was used to test for statistical significance. Its distribution under the null hypothesis (that the rate of disease within a scanning window based on a certain location is not different from the rate of disease outside the window) and its corresponding *P*-value was obtained by repeating the likelihood calculations on a large number (999) of random replications of the data set generated under the null hypothesis using Monte Carlo simulation (14).

Results

The mean centre of Texas counties reporting cases of WNV weighted by the population at risk was approximately 148 km to the south-east of the case-weighted mean centre. The mean centre weighted by the estimated county attack rate was approximately 128 km to the north-west of the case-weighted mean centre (Fig. 1). The mean centres and standard deviational ellipses for each of the three phases of the WNV epidemic are shown in Figure 1. The centres were located in south-east, north-west and north-central Texas during these three phases, respectively. During phases I and

II, the distribution was ellipsoid (south-east to north-west), but during the final phase it was circular. The observed distance between cases was less than half that expected for all three phases of the epidemic. Moran's auto-correlation statistic indicated that both the distribution of date of onset of cases (I = 0.13; P<0.001) and the distribution of estimated county attack rates (I = 0.29; P<0.001) were clustered. The results of the Cuzick and Edwards test applied to case and control Texas counties are shown in Table I. Overall, no significant clustering of case counties was detected (combined P-value = 0.12), although significance (P<0.05) was present at nearest neighbour levels of 4, 5, 7 and 9. Mantel's correlation for all cases was 26.8% (P = 0.001). Mantel's correlation for each of the three epidemic phases was <1% (P = 0.492), 18.1% (P = 0.001) and 7.6% (P = 0.001).

Estimated crude WNV county attack rates and rates smoothed using an empirical Bayes algorithm are shown in Figure 2. Rates were smoothed using the 10 nearest neighbours weighted by inverse squared distance. Using this smoothing algorithm, the range of estimated attack rates was reduced from 65 cases per 1 000 horses at risk to 27 cases per 1 000 horses at risk and the standard deviation of the mean attack rate (8 cases per1 000 horses at risk) was reduced from 10.04 to 6.59.

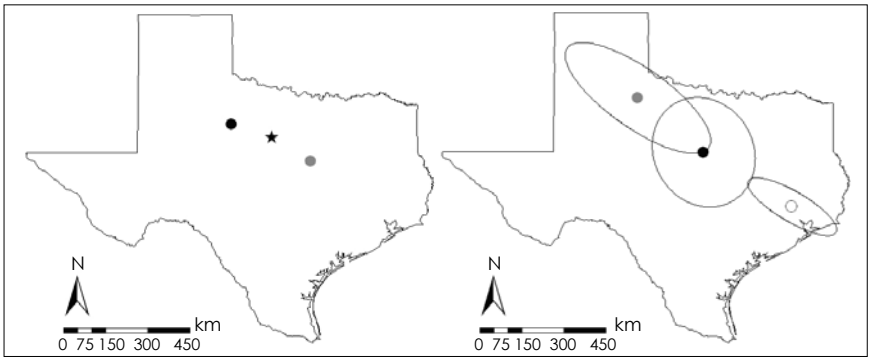

Figure 1
Equine West Nile virus encephalomyelitis in Texas countries in 2002
(Left) Mean centres of counties reporting cases, weighted by county case count (★) and estimated county attack rates (●; cases per 1 000 horses at risk)
The mean centre of Texas counties weighted by estimated county horse population is also shown (●)
(Right) Mean centres and standard deviational ellipses of reported cases during three phases of the epidemic:
27 June-25 July (O), 26 July-27 September (●) and 28 September-17 December (●)

Table I
Analysis of spatial clustering of Texas counties that did (cases) or did not (controls) report equine West Nile virus encephalomyelitis in 2002
(using the Cuzick and Edwards test)
Combined P-values (based on Monte Carlo randomisation) for all k were 0.12 (Bonferroni correction method) and 0.06 (Simes correction method)

k	Test statistic, T	Expected (T)	Variance (T)	Z statistic	Monte Carlo P-value
1	168	164	26.6	0.84	0.23
2	336	327	43.7	1.31	0.27
3	502	491	58.3	1.43	0.38
4	677	655	73.4	2.60	0.02
5	853	818	102	3.42	0.01
6	1 020	982	120	3.46	0.31
7	1 196	1 146	150	4.09	0.02
8	1 368	1 309	180	4.37	0.09
9	1 542	1 473	214	4.70	0.04
10	1 711	1 637	261	4.59	0.22

Using the scan statistic, the most likely (log likelihood ratio = 90.3, $P = 0.001$) spatial cluster of WNV cases occurred in two counties in north-central Texas (Fig. 3). A total of 84 cases were reported from a population at risk of 2 402 (35 cases per 1 000 horses at risk). Using the Poisson model, 12.45 cases would be expected to be reported from this size population. Thus, 6.7 times as many cases were observed as expected. An additional 12 significant ($P<0.01$) clusters were identified, located in north-west (24 counties), north-central (10 counties) and south (one county) Texas (Fig. 3). Attack rates in these clusters ranged from to <1 to 57 cases per 1 000 horses at risk, and included from one to six counties. The most likely spatio-temporal cluster identified (log likelihood ratio = 205, $P = 0.001$) occurred in 5 counties in north-west Texas between 15 August and 10 September (Fig. 3). A total of 67 cases were reported from a population of 3 117 equines (21 cases per 1 000 horses at risk). Using the Poisson model, 1.2 cases would be expected to be reported from this size population. Thus, 56 times as many cases were observed as expected. An additional 28 significant ($P<0.01$) clusters were identified, located in north-west (42 counties), north-central (24 counties), east (3 counties) and south (9 counties) Texas (Fig. 3). Attack rates in these clusters ranged from 3 to 27 cases per 1 000 horses at risk, and included from one to six counties. Clusters occurred between 7 July and 11 November (Fig. 4). Most clusters began during the week of 29 September.

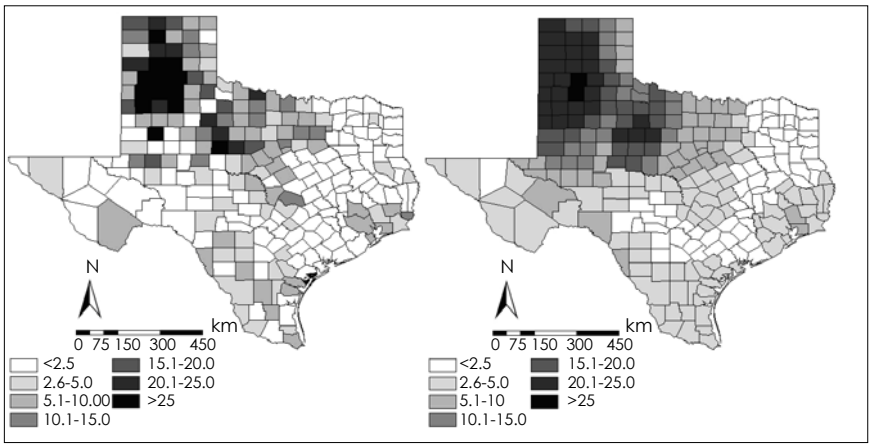

Figure 2
County attack rates of equine West Nile virus encephalomyelitis in Texas in 2002
Left: Estimated crude attack rates
Right: Attack rates smoothed using an empirical Bayes method based on the 10 nearest neighbouring counties determined by inverse distance squared

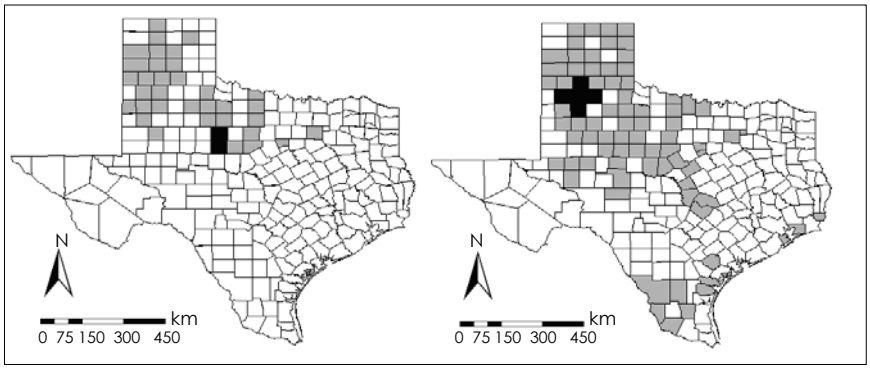

Figure 3
Location of clusters of equine West Nile virus encephalomyelitis in Texas during 2002
Left: Significant ($P<0.05$) spatial
Right: Spatio-temporal
The most likely clusters are shown (■)

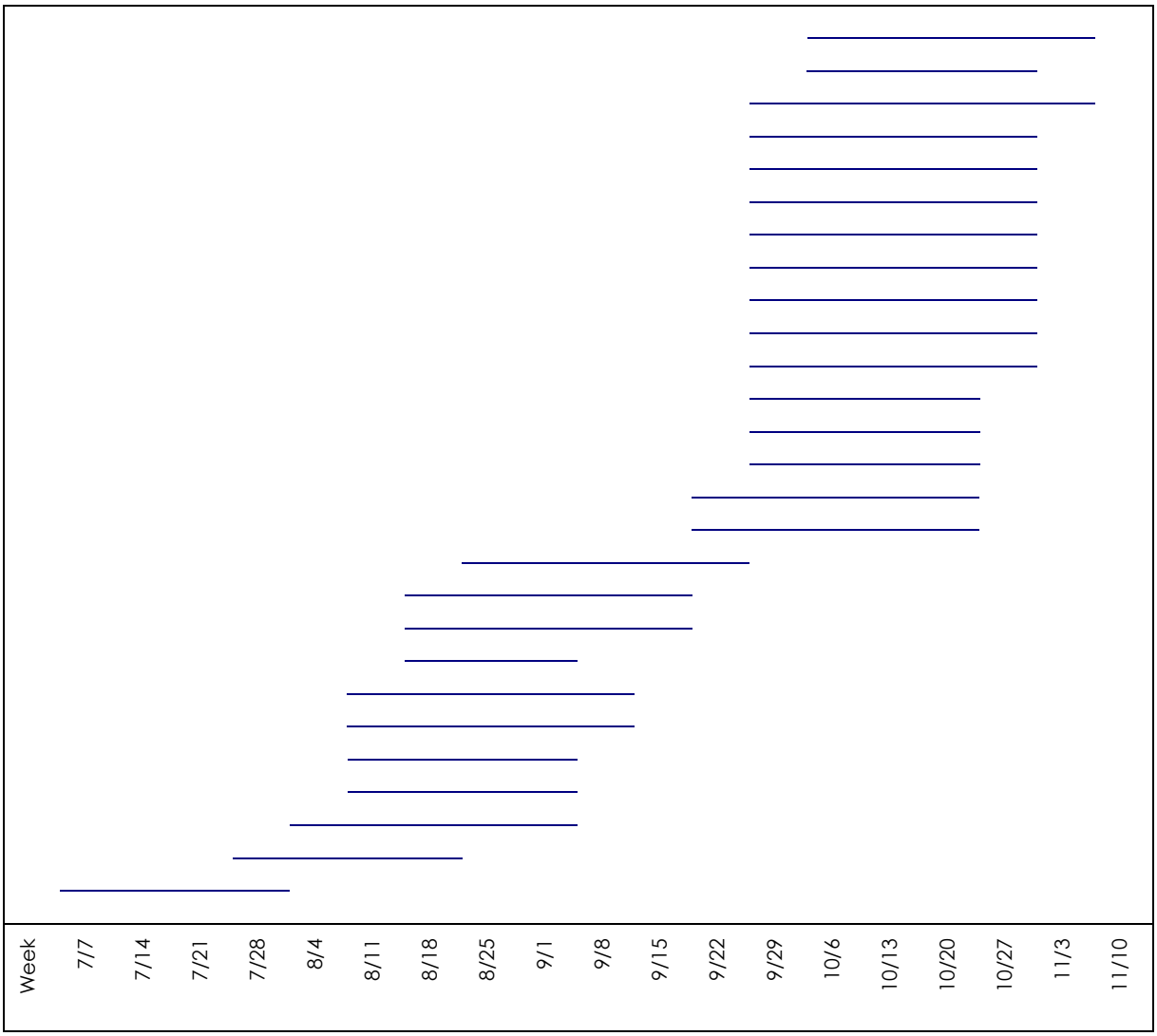

Figure 4
The timing of occurrence 27 significant ($P<0.05$) spatio-temporal clusters of equine West Nile virus encephalomyelitis in Texas in 2002

Discussion

Clustering of disease can be subtle and quite complex – for example, when populations change substantially over time and are not uniformly distributed in space. As with all epidemiological investigations, statistical techniques are helpful, and sometimes essential, in understanding the disease process. Spatio-temporal statistics have three special attributes in these circumstances, namely:

- they add precision to qualitative verbal description
- they facilitate the comparison of distributions by offering objective, quantitative criteria
- they may draw attention to characteristics unlikely to be noticed by visual inspection (10).

Clearly, the use of spatio-temporal statistics can enhance our understanding of how disease occurs in animal populations.

Despite the rapid increase in the application of geographic information system (GIS) technology, the use of statistical tests to investigate clustering of disease in veterinary medicine remains relatively uncommon (30, 31). Analysis of data in GIS does not routinely employ statistical tests of spatial clustering. Rather, GISs have generally been used to analyse (through visual interpretation) the relationships between potential risk factors

and the occurrence of disease (incidence or prevalence) on a geographical basis. The lack of availability of and user familiarity with statistical software has restricted the spatial and temporal analyses of data sets for disease clustering. One of the barriers to the use of advanced spatial analytical techniques has been the lack of compatibility between GIS and specialist statistical software. Newer GIS software, such as ArcGIS™ version 9.0 and GeoDa™ 0.9.5-i5, include limited statistical functionality (in the case of ArcGIS™ version 9.0, this includes descriptive statistics [mean centre, standard deviational ellipse], global spatial statistics [Moran's autocorrelation, nearest neighbour] and statistics for identifying local clusters [Anselin's local indication of spatial autocorrelation, Getis-Ord Gi*]). Some more recent specialised software for spatial analysis, such as ClusterSeer® version 2.0, now has the option of importing shapefiles produced within GISs. Similarly, software capable of performing empirical Bayes smoothing, for example, STIS™ version 1.0.6 and GeoDa™ 0.9.5-i5, can also make use of GIS shapefiles. The statistical analysis of spatial distributions remains a weak point in the application of GIS technology. If GIS technology is to fulfil its potential as a general-purpose tool for handling spatial data, it needs stronger analytical capabilities (23). Development of statistical software to investigate disease clustering and integration of these routines into GISs, will improve the ability of epidemiologists to identify and describe determinants of disease.

No omnibus test exists for assessing spatial and temporal clusters of disease. Thus, investigators have been advised to 'perform several related tests and to report the results that are most consistent with validated assumptions' (3). As part of an overall approach to investigating clusters, the information provided by these tests is useful for developing a better understanding of disease causation. Autocorrelation is an easily understood technique. The LISA approach and Mantel's correlation for spatio-temporal clustering are useful for variables measured on a continuous scale, such as disease prevalence and incidence, and dates of onset of disease. In situations in which disease case-control data exist, the nearest neighbour test and Cuzick and Edwards test are appropriate to test for global clustering. The most flexible technique for investigating spatial and spatio-temporal disease clusters is the scan statistic. It can be used for both disease rate and case-control data, point or area data, it accounts for heterogeneity among the population at risk and potential confounders can be controlled. It can both test for global clustering and can identify the location of specific clusters. A disadvantage of the scan statistic is that subjectivity is introduced into the testing procedure by the need to select the size of the spatial and temporal scanning windows. It is recommended that the scanning window should be based on biological characteristics of the disease being studied. For example, Paré *et al.* (22) used a window of 2 to 14 days in length to investigate temporal clustering of *Salmonella krefeld* infection in horses admitted to an intensive care unit of a veterinary hospital. This was based on the average duration of hospitalisation, the known lag period between infection and shedding of *Salmonella* species, and the need to perform multiple cultures to detect *Salmonella* organisms. In contrast, Singer *et al.* (26) and Doherr *et al.* (6) used a range of window lengths in their studies, apparently in an attempt to generate disease causation hypotheses. Obviously, the more analyses performed using a wide range of scanning windows the more likely it is that significant clustering will be detected. Thus, the investigator needs to carefully consider the aim of analysis and to review available literature and expert opinion on the disease of interest prior to selecting one or more scanning windows to use. The selection of a scanning window *a priori* will provide more robust results.

The power of techniques to detect time-space clustering is not well-characterised, but is probably only low to moderate. For example, the Mantel test has been found to be insensitive to clusters characterised by a gradual change in the risk of event occurrence ('clinal'), but moderately sensitive (up to 40%)

in detecting 'hot spot' clusters (a situation in which one sub-region of the study area has a uniform and greater risk than the rest of the study area), using a sample size of 50 events (34). The false-positive rate was found to be close to the nominal type-I error (0.05) used. The increasing sophistication of database management and GIS may increase the number of studies investigating the interaction between the temporal and spatial occurrence of disease in veterinary epidemiology. Several techniques should be used when attempting to identify and describe whether events are clustered in time and space in order to maximise the power of the analysis. In addition, it is important to consider the spatial and temporal model implicit in techniques chosen when interpreting results of analysis.

Recent developments in livestock production are likely to increase the need for epidemiologists to undertake time-space analyses. For example, better identification of livestock and routine recording of their location and health and production status offered by modern animal health monitoring systems incorporating novel technology will provide data which readily can be used to detect unusual disease clusters and to generate and test hypotheses regarding causes of sub-optimal health and productivity. Without appropriate techniques to analyse and interpret these data, the costs of database construction may exceed the benefits realised.

Conclusion

Investigations of disease clustering in animal health can be greatly enhanced through the use of a variety of analytical techniques. These techniques add considerable information to disease investigations and provide the veterinary epidemiologist with a firm foundation on which to build causal hypotheses and implement control strategies. A challenge is to implement these techniques as routine procedures within animal disease control and prevention programmes. Increased data quality and availability through the development of modern animal disease and production monitoring and surveillance systems, new techniques such as remote sensing, the ability to sort and recombine data using GIS, and the increasing availability of software packages over the past three decades have created an ideal environment for epidemiologists to apply spatial and temporal analytical techniques to disease problems.

Grant support

Support for the collection of data used in this study was provided by the Texas Equine Research Advisory Committee Grant Number 203314.

References

1. Anselin L. 1995. Local indicators of spatial association-LISA. *Geogr Anal*, **27**, 93-115.
2. Castillo-Olivares J. & Wood J. 2004. West Nile virus infection of horses. *Vet Res*, **35**, 467-483.
3. Centers for Disease Control 1990. Guidelines for investigating clusters of health events. *MMWR Recomm Rep*, **39** (RR-11), 23 pp.
4. Clark P.J. & Evans F.C. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, **35**, 445-453.
5. Cuzick J. & Edwards R. 1990. Spatial clustering for inhomogenous populations. *J Roy Stat Soc Series B*, **52**, 73-104.
6. Doherr M.G., Carpenter T.E., Wilson W.D. & Gardner I.A. 1999. Evaluation of temporal and spatial clustering of horses with *Corynebacterium pseudotuberculosis* infection. *Am J Vet Res*, **60**, 284-291.
7. Ederer F., Myers M.H. & Mantel N. 1964. A statistical problem in space and time: do leukemia cases come in clusters? *Biometrics*, **20**, 626-638.
8. Environmental Systems Research Institute, Inc. (ESRI) 2007. Spatial statistics: ArcGIS™ 9.0. ESRI, Inc., Redlands, California.
9. Getis A. & Ord J.K. 1992. The analysis of spatial association by use of distance statistics. *Geogr Anal*, **24**, 189-206.

10. Hammond R. & McCullagh P.S. 1978. Quantitative techniques in geography: an introduction, 2nd Ed. Clarendon Press, Oxford, 364 pp.
11. Jacquez G.M. Stat! Software for the clustering of health events. BioMedware Press, Ann Arbor, 168 pp.
12. Jacquez G.M. 1996. A k nearest neighbour test for space-time interaction. *Stat Med*, **15**, 1935-1949.
13. Knox G. 1964. The detection of space-time interactions. *Applied Stat*, **13**, 25-29.
14. Kulldorff M. 2006. SaTScan™ version 6.1.2. Software for the spatial and space-time scan statistics (www.satscan.org/ accessed on 16 June 2007).
15. Kulldorff M. & Nagarwalla N. 1995. Spatial disease clusters: detection and inference. *Stat Med*, **14**, 799-810.
16. Kulldorff M., Athas W.F., Feuer E.J., Miller B.A. & Key C.R. 1998. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos. *Am J Public Health*, **88**, 1377-1380.
17. Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res*, **27**, 209-220.
18. Moran P.A.P. 1950. Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17-23.
19. Naus J.I. 1965. The distribution of the size of the maximum cluster of points on a line. *J Am Stat Assoc*, **60**, 532-538.
20. Naus J.I. 1966. A power comparison of two tests of non-random clustering. *Technometric*, **8,** 493-517.
21. Oden N. 1995. Adjusting Moran's I for population density. *Stat Med*, **14**, 17-26.
22. Paré J., Carpenter T.E. & Thurmond M.C. 1996. Analysis of spatial and temporal clustering of horses with *Salmonella krefeld* in an intensive care unit of a veterinary hospital. *Am J Vet Med Assoc*, **209**, 626-628.
23. Paterson A.D. 1995. Problems encountered in the practical implementation of geographical information systems (GIS) in veterinary epidemiology. *In* Proc. Society for Veterinary Epidemiology and Preventive Medicine (SVEPM meeting), 29-31 March, Reading. SVEPM, University of Edinburgh, Roslin, Midlothian, 162.
24. Porter M.B., Long M.T., Getman L.M., Giguere S., MacKay R.J., Lester G.D., Alleman A.R., Wamsley H.L., Franklin R.P., Jacks S., Buergelt C.D. & Detrisac C.J. 2003. West Nile virus encephalomyelitis in horses: 46 cases (2001). *J Am Vet Med Assoc*, **222**, 1241-1247.
25. Salazar P., Traub-Dargatz J.L., Morley P.S., Wilmot D.D., Steffen D.J., Cunningham W.E. & Salman M.D. 2004. Outcome of equids with clinical signs of West Nile virus infection and factors associated with death. *J Am Vet Med Assoc*, **225**, 267-274.
26. Singer R.S., Case J.T., Carpenter T.E., Walker R.L. & Hirsh, D.C. 1998. Assessment of spatial and temporal clustering of ampicillin- and tetracycline-resistant strains of *Pasteurella multocida* and *P. haemolytica* isolated from cattle in California. *J Am Vet Med Assoc*, **212**, 1001-1005.
27. TerraSeer 2004. Space-time intelligence system® version 1.0.6. TerraSeer, Ann Arbor, Michigan.
28. TerraSeer 2004. ClusterSeer® 2.0. TerraSeer, Ann Arbor, Michigan.
29. Turnbull B.W., Iwano E.J., Burnett W.S., Howe H.L. & Clark L.C. 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *Am J Epidemiol*, **132**, S136-S143.
30. Ward M.P. & Carpenter T.E. 2000. Analysis of time-space clustering in veterinary epidemiology. *Prev Vet Med*, **43**, 225-237.
31. Ward M.P. & Carpenter T.E. 2000. Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. *Prev Vet Med*, **45**, 257-284.
32. Ward M.P., Levy M., Thacker H.L., Ash M., Norman S.K.L., Moore G.E. & Webb P.W. 2004. An outbreak of West Nile virus encephalomyelitis in a population of Indiana horses: 136 cases. *J Am Vet Med Assoc*, **225**, 84-89.
33. Ward M.P., Schuermann J.A., Highfield L.D. & Murray K.O. 2006. Characteristics of an outbreak of West Nile virus encephalomyelitis in a previously uninfected population of horses. *Vet Microbiol*, **118** (3-4), 255-259.
34. Wartenberg D. & Greenberg M. 1990. Detecting disease clusters: the importance of statistical power. *Am J Epidemiol*, **132**, S156-166.

