

SPATIAL ANALYSIS – A NEW CHALLENGE FOR VETERINARY EPIDEMIOLOGISTS

D U PFEIFFER *

The classic epidemiological triad includes space in addition to person (animal) and time as its components. Largely due to computational difficulties spatial analysis is an area that only recently has become more easily accessible for epidemiologists. Spatial epidemiological analysis includes hypothesis-driven as well as non-hypothesis driven investigations. The latter fits well within the new field of data mining, which has emerged as a result of the availability of large databases particularly in business. The complexity of spatial analysis is mainly the result of proximity inter-relationships between the observations. This causes a problem, because standard epidemiological analysis typically focuses on the attributes of observations, and makes the assumption that they are independent. Temporal analysis introduces consideration of one-dimensional dependence between repeated observations on the same subject. Spatial analysis extends this interdependence into two- or even three- dimensional space. The last 20 years have seen a significant development in the statistical methods that can deal with analysis of inter-dependent observations. Spatial analysis is a very challenging field as it introduces a whole new set of technical terms and techniques in the context of data storage as well as statistical methodology, which are different from what epidemiologists have used traditionally.

SPATIAL DATA

Any data with a spatial reference, either relative or absolute, should be considered spatial. In the simplest case, it can represent the geographic coordinates of a location where an outbreak of a disease has occurred. More general, the spatial reference associated with a single observation defines a spatial feature such as a point or a polygon. As an example, a point could represent the location where an animal was found dead, and a polygon might describe the boundaries of a farm. If the whole dataset represents spatially contiguous information it may describe either a continuous surface or a lattice structure reflecting for example the risk of disease outbreaks given specific environmental risk factors. In addition to the spatial reference, each observation can also have attribute information associated with it. For example, in the case of the location where a rabid animal had been found, attributes could be the animal species, its age as well as any diagnostic examination results. If the geographical data represents farm

* Dept. of Farm Animal and Equine Medicine and Surgery, Royal Veterinary College, Hawkshead Lane, North Mymms, AL9 7TA, United Kingdom

boundaries, its associated attribute data may for example record the name of the farmer and the tuberculin test history of the herd. Spatially referenced data can be obtained through direct data entry, digitising or remote sensing. Particularly the latter has over the last years become a more cost-effective source for example for digital vegetation data.

Coordinate point locations can be easily stored using standard databases and displayed using scatterplot graphs. But particularly in the case of lattice structures, geographical information systems (GIS) provide much more effective tools for input, storage, manipulation and presentation of spatial data.

One of the most important decisions, which has to be made during collection of spatial data, relates to the choice of the appropriate level of spatial aggregation. It has a strong impact on the cost of data collection, the power of the analysis and the level at which inferences can be drawn. Farms are often presented as point locations as this is the quickest and cheapest method for obtaining spatial reference data, and this may often be sufficient. But it becomes difficult to analyse neighbourhood relationships with this kind of data, as it does not take into account the shape and the size of the properties. Polygon (i.e. area) data could be presented at different levels of aggregation in that the number of infected animals is summarised per herd, district or country. Inappropriate aggregation can lead to effects such as the ecological fallacy and the modifiable areal unit problem (MAUP). The first relates to the difference between estimates at the aggregate and the individual level. The second, the MAUP, stems from the fact that areal units typically do not represent 'natural' but rather arbitrary constructs (Haining, 1998). This may result in a fixed number of areal units of the same spatial extent, but varying numbers of animals per unit. Alternatively, the number of areal units can be reduced, resulting in increased spatial extent of each unit. When analysing for presence of association, variability typically is underestimated, and therefore measures of association may increase. Ideally one would choose the scale revealing most detail, but this is likely to result in substantial costs, loss in data quality and in data quantities which cannot be processed. Therefore, a sensible compromise has to be found which still allows meaningful observations to be made and sensible inferential conclusions to be drawn.

The accuracy and precision of spatial data is an issue, which deserves special consideration. GIS often combines data from many different sources, which may well be of different accuracy and collected at different scales or precision. Often the error associated with the base maps is not documented. Combining such maps to generate new maps, one of the fundamental functions of GIS, can produce unpredictable results. This can result in error propagation particularly when these output maps are used as input for other operations.

SPATIAL DATA ANALYSIS

The analysis of spatial data can focus on the relationships between attribute variables, or on the spatial and space-time dimensions or a combination of attribute and space/space-time. The methods used in spatial data analysis can be broadly categorized into those concerned with visualizing data, those for exploratory data analysis and methods for the development of statistical models (Bailey and Gatrell 1995). Analyses can be hypothesis driven or they are used to trigger alarms if something unusual from a statistical perspective has been recorded. The first approach fits nicely into the classical framework of statistical analysis, whereas the second has resulted in controversy amongst scientists, as the risk of Type I errors can be high. Many studies will require post-hoc hypothesis formulation, and involve multiple comparison analyses. In spatial analysis, effects have to be distinguished which result in long-range spatial trends (i.e. first order effects) and those which produce localised dependence (i.e. second order effects). Often both effects will be present, which complicates spatial analysis as most procedures make the assumption that only one of the two effects is present. First-order effects can be modelled relatively easily using regression models, whereas the second-order effect has to be specifically incorporated in the error terms (Bailey and Gatrell 1995). During most analyses, a combination of techniques will be used with the data first being displayed visually, followed by exploration of possible patterns and possibly modelling.

Analytical approaches can be divided into visualisation, exploration and modelling. Visualisation of the data should be the first step in a spatial data analysis. It involves showing the actual data values as two-, three or more-dimensional maps. Data is presented as points, coloured points or continuous surfaces/ lattices. It allows detection of data errors, as well as generation of hypotheses. Exploration of spatial data is aimed at description and quantification of spatial structure. During this phase, hypothesis testing is limited to detection of clusters and spatial dependence. Methods for quantifying spatial autocorrelation, variograms and specific tests for detection of spatial clustering are applied. Modelling is used to explain and predict spatial structure. Testing for cause-effect relationships is the main purpose of these techniques, which include statistical and simulation modelling, and also multi-criteria/multi-objective decision modelling

Visualisation

Visual analysis methods are extremely useful as they allow a process called visual thinking, and they eventually lead to visual communication through presentation of the data, for example, in map format. If the emphasis is purely on spatial occurrence such as whether a herd is infected or not, this can be shown using dot or polygon maps depending on whether point or polygon data has been used as the spatial reference. With dot maps any patterns become difficult to detect, as soon as the density of points increases. In this case, interpolation techniques can be used to generate continuous surfaces of the underlying point density. Kernel smoothing has become the standard method for these interpolations. The resulting surface represents the probability of the occurrence of a case, and is estimated using a bivariate probability density function (i.e. kernel), which is symmetric about the origin. The amount of smoothing is dependent on the chosen bandwidth, which can be fixed or adaptive. Fig. 1 shows kernel smoothed maps of the

density of all dairy herds and those infected with enzootic bovine leucosis in New Zealand (Teekayuwat, 1999). The images improve presentation of point data, but the effect is strongly influenced by the parameters of the kernel estimator and its algorithm. It is also possible to estimate the ratio of kernel estimates, and thereby adjusting for differences in the population at risk (Bithell, 1990; Kelsall et al. 1995). With polygon data, differences in denominators can be taken into account through cartogram presentations or density equalized map projection (Merrill et al. 1996), where the size and shape of the polygons is basically re-scaled to represent differences in denominator values, such for example population density. Standardisation used to be the main technique for mapping of disease rates and ratios. It had the disadvantage that it could not take into account the increased uncertainty about true estimates resulting from small counts and denominators. Empirical or fully Bayesian estimates are more appropriate as they use the overall or local risk as a prior, with the effect being that the local estimate is shrunken towards the overall or the neighbourhood mean risk particularly if it results from small numbers of observations in that particular location (Clayton et al. 1987; Langford, 1994).

Exploration

Exploratory analysis of spatial data is aimed at describing spatial patterns using inferential statistics, and it is used for the development of hypotheses. With the occurrence of disease it is mainly about whether diseases occur randomly in space or not. It is complicated by often having to take account of the spatially clustered distribution of the underlying population at risk. One effective method for dealing with this problem is the use of case-control data, where the cases of a particular disease are selected as usual and the controls are selected randomly from the non-diseased population, and therefore should represent the spatial distribution of the underlying population at risk. The controls could be matched to cases with respect to confounding factors other than spatial location (Lawson et al. 1996). This approach requires exact point locations for cases and controls to be available. If the data source is routine surveillance, it is likely to be aggregated at some administrative level such as veterinary district. As exploratory spatial data analysis involves statistical hypothesis testing, with the availability of fast computer technology bootstrap and permutation methods can be used to deal with the multiple testing problem (Kulldorff et al. 1995). In the statistical assessment of spatial clustering of point and polygon data global or local statistics can be generated. Global statistics will indicate if there is clustering somewhere in the area of investigation. Local statistics will also indicate where the likely clusters are and this will be particularly useful if the analysis is aimed at triggering an alarm.

Cuzick and Edwards (1990) developed a method which is based on nearest-neighbour distances. This aggregated test statistic compares the number of case-case pairs for a given number of nearest neighbors. Applying this technique to case-control data for New Zealand bovine tuberculosis breakdown herds, it suggests that there is significant clustering of cases compared with the control population according to the p-value for the Bonferroni statistic (see Fig. 2).

Other more recently developed techniques for analysing case-control point data for presence of spatial clustering are the K-function and the spatial scan statistic. The K-function produces an aggregated statistic. It expresses the mean number of cases with increasing distance from a given case scaled by the density of points in the area. The procedure is performed for cases and controls separately, and the difference between the two resulting K-functions can be plotted against distance. A random expectation statistic can then be generated by randomly permuting cases and controls at least 100 times and estimating K-functions and their differences for each permutation. If the observed curve extends beyond the simulation envelope, then significant clustering of cases relative to controls can be assumed (Bailey and Gatrell 1995; Jones et al. 1996). This technique assumes that only second-order effects are present and that these are isotropic (i.e. not dependent on direction). The example data presented in Fig. 3 represents locations of farms involved in the outbreak of an infectious animal disease (unpublished data). The shape of the difference of the kernel functions for case and control farms in relation to the simulation envelope indicates that there was significant clustering. The disadvantage of K-function is that it does not describe the location of the clustering. It is also quite common with animal diseases to have first-order effects present such as variation in climatic conditions as result of different levels of elevation.

In contrast, the spatial scan statistic is a local clustering statistic. The procedure involves generating ever-increasing circles around every point and calculation of relative risks based on disease risks within and outside the circle. A likelihood ratio test is calculated to assess for statistical significance, and the distribution of the test statistic is estimated using Monte-Carlo sampling (Kulldorff et al. 1995). Fig. 4 shows the map of trap locations used by tuberculous and non-tuberculous wild possums during a longitudinal field study in New Zealand (Pfeiffer, 1994). The spatial statistic identified a most likely cluster in one particular region of the study area. The associated relative risk was 2.1. A p value of 0.003 was estimated indicating that such a localised density of traps used by tuberculous possums only occurred 3 times in 1000 Monte-Carlo samples. The spatial scan statistic is one of the more robust techniques for exploratory analysis of spatial clustering.

For aggregate data such as disease rates, Moran's I, a global statistic, is used as an estimator of spatial autocorrelation. To allow a local description of spatial dependence for this type of data, the generic concept of Local Indicators of Spatial Association (LISA) has been developed recently (Anselin, 1995). It embraces Moran's I, Geary's C and the Getis-Ord G_i^* statistics under a single mathematical framework. This methodology can be complemented by visualisation of the resulting statistics as maps presenting spatial lag pies or bar charts (Anselin et al. 1993).

In dealing with infectious processes, testing for clustering in space and time may be of interest. This is usually done using the point locations of the cases and most statistics available will work on all possible pairs of time-space distances between the points. The

Knox test (Knox, 1964) and Mantel method (Mantel, 1967) have been the classical techniques used for these analyses. More recently, the space-time scan statistic (Kulldorff et al. 1998) and the K-nearest neighbour test (Jacquez, 1996) have been developed. All four techniques are using permutation statistics. They make the assumption that population sizes do not change in time, but the statistics are not affected by spatially heterogeneous populations. The data from a longitudinal study of *Mycobacterium bovis* infection in a wild possum population in New Zealand already mentioned above will be used to demonstrate the usage of the two newer techniques. The K-nearest neighbour test of space-time interaction allows an assessment of the statistical significance of a potential space-time interaction process. The test statistic indicates the number of case pairs which are K nearest neighbours in time and space. The statistic is based on an approximate randomisation of the Mantel product statistic. Fig. 5 presents the results from applying the K-nearest neighbour method to the possum tuberculosis data. The map shows the locations of the traps where tuberculous possums had been caught and the arrows indicate k=2 nearest neighbours. The test statistic produced on the basis of 1000 random permutations suggests that only the cumulative statistic J_k is statistically significant, whereas ΔJ_k is not. The latter parameter measures the statistical significance resulting from increasing K by 1. The test statistic supports the presence of space-time interaction, and suggests that the first 5 nearest neighbours are probably involved in space-time interaction, and thereby provides an indication of cluster size, but not of its actual geographical dimensions. The result has to be interpreted with caution though because none of the summary statistic for ΔJ_k is statistically significant.

An exploratory analysis of the spatial dependence of continuous type data involves the use of techniques such as the variogram method. This type of data is usually collected at sample point locations where some attribute such as density of disease vector populations is being measured. Spatial moving averages are used to represent first-order spatial effects, that is global trends in a particular geographical area, whereas variograms describe localised effects, i.e. second-order spatial effects. The presence of second order effects would result in positive covariance between observations a small distance apart and lower covariance or correlation if they are further apart. The covariogram describes the function of the covariance for varying distances h between sample points and the correlogram the corresponding correlation. The semi-variogram is a graphical representation of the variation between sampling points separated by a given distance and direction. For a stationary spatial process all three describe similar information. Estimates of the semi-variogram are considered to be robust to departures from stationarity represented as a general trend in the spatial process. A continuous process without spatial dependence will result in a horizontal line. A stationary process will reach an upper bound, referred to as the sill at a distance h called the range. Theoretically, the intercept with the y-axis should be at a value of 0 variation. In reality, sampling error and small scale variation will result in variability at small distances and the variogram will not meet the y-axis in the origin. This intercept with the y-axis is called the nugget effect. Variograms which do not reach an upper bound suggest non-stationarity in the data. Fig. 7 shows an isotropic sample semi-variogram for the proportion of tuberculous possums captured at trap sites during the above mentioned longitudinal study. The shape of the variogram suggests that the process is non-stationary, but given the relatively small nugget value there is also likely to be spatial dependence.

Modelling

Models derived from spatial data can be used to identify risk factors or they can remain within the spatial domain if these models are used for interpolation or smoothing. Particularly the latter are aimed at visual presentation. The most basic modelling techniques are based on map modelling (Bonham-Carter 1994). These involve overlaying different geographical layers of information using, for example, Boolean logic, whereas more advanced map modelling will apply fuzzy logic or Bayesian methods. A major area of continuing research development is the field of spatial risk factor modelling. The objective of these analyses in the context of epidemiology is to generate risk surfaces taking into account underlying geographical risk factor patterns. In epidemiology, parameters of interest are very often counts or proportions which should be modelled using generalised linear modelling techniques rather than ordinary least-squares regression. The most significant problem has been to determine mathematically sound techniques for taking into account the dependence in the spatial data structure. Bailey and Gatrell (1995) suggest introducing covariates into the regression model such as spatial coordinates or a variable representing broad regions to adjust for the effect of spatial dependence. Glass et al. (1995) developed a risk density map for Lyme disease based on a multiple logistic regression model, but they did not attempt to remove spatial dependence from the data. Williams et al. (1994) compared a number of different predictive modelling approaches for spatial data. They used linear and non-linear discriminant analysis, tree-based induction and neural networks to map tsetse distributions in Zimbabwe and concluded that while the simpler methods (linear discriminant analysis and tree-based induction) were less precise, they were easier to interpret. They also did not explicitly take account of spatial dependence. Fig. 6 presents results of a logistic regression analysis for prediction of *Theileria parva* presence in an African country (Pfeiffer et al. 1997). The regression model includes eight different environmental and land use variables and is based on information collected at random sample locations throughout the country. The model was used to generate a risk map representing the probability of *T.parva* presence at a particular location given a number of risk factors included in the model. This map is presented as a DTM and as a raster map. The receiver operating characteristic curve (ROC) characterizing the predictive accuracy of the model can be used to adjust the decision making cut-off for the prediction probability balancing sensitivity and specificity as required. In this analysis the possible presence of spatial dependence was taken into account using a categorical variable representing region as a random effect.

Three important developments have now provided more appropriate solutions to the problem of incorporating spatial dependence in regression models. Firstly, the advent of multi-level modelling provided a statistical framework for modelling spatial dependence in covariance structures (Langford et al. 1999). Secondly, the use of the prior in Bayesian statistics lends itself for representing the spatial dependence during the estimation of model parameters. Thirdly, the computation technique Markov chain Monte Carlo (MCMC) method using Gibbs sampling produces robust simulation- based estimates of

the likelihood or the posterior distribution in the case of Bayesian inference (Lawson et al. 1996). One particular approach for representing spatial dependence with binary outcome variables involves the use of an autologistic term which was originally described by Besag (1974). This term can be used as a covariate additional to other risk factors in a logistic regression model to reflect the dependence of the local risk of disease on the level of disease in the neighbours (Gumpertz et al. 1997); (Augustin et al. 1998). These authors applied Gibbs sampling for parameter estimation. A Bayesian hierarchical spatial modelling was described by Xia et al (1997) who used a mixed prior, one conditionally autoregressive and the other unstructured. This approach of including non-spatial as well as spatial random effects in the models which was first suggested by Besag et al (1991) is now commonly adopted when performing spatial regression modelling (Wakefield et al. 1999); (Langford et al. 1999). The difficulty with applying both, spatial logistic regression and MCMC estimation techniques, is that they are not available as standard procedures in statistical analysis packages. This is likely to change in the near future, though. For example, a module called GeoBUGS for MCMC estimation of spatial generalised linear models will be released for the BUGS software (MRC Biostatistics Unit, Institute of Public Health, Cambridge, United Kingdom) in the first half of 2000.

A number of approaches can be used to model or predict spatially continuous data. For the *first-order* processes *trend surfaces* can be generated with ordinary polynomial least squares regression. Results have to be treated with caution, because the standard regression assumptions of independent random errors and heteroscedasticity are likely to be violated. Lessard et al. (1990) applied this type of approach when using an inverse distance-weighted mathematical algorithm to interpolate climatic measurements between sample points. Most trend surface models may be able to describe an overall trend, but are not useful for local prediction. In the presence of weak first order, but strong second order effects it is more appropriate to use models fitted to variograms. Such models can be defined 'by eye' and are most commonly based on a spherical, exponential or gaussian model fitted to the variogram. The fit of a particular model can be assessed through cross-validation based on a comparison between the observed and interpolated values. Fig. 8 shows an isotropic exponential *variogram* model for the possum trap capture data from the longitudinal study discussed above. There is no upper bound to the variogram model indicating that the data is non-stationary. For this particular model to be a valid representation of the underlying data it would therefore be necessary to first remove the non-stationarity (i.e. first order spatial effect) through trend regression.

The variogram model itself does not allow prediction of values. This can be achieved with kriging. This is a weighted moving average technique for estimating the value of a spatially distributed variable from adjacent values while considering interdependence expressed in a variogram. It allows the interpolation error to be mapped and from a statistical viewpoint is considered to be an effective technique for interpolation of continuous type spatial data according to Oliver and Webster (1990). Webster et al (1994) used kriging to describe the risk of cancer in children for the West Midlands of England. The resulting maps showed that child cancer in this region clearly had a patchy distribution, in that areas of high risk were near to each other as was the case amongst those of low risk. The authors emphasise that with low incidence disease such as in their

analysis it is important to have large amounts of observations available. They expressed concern about the validity of the confidence limits they had estimated for their binomial data, as the technique is more appropriate for prediction of continuous type variables. Carrat and Valleron (1992) applied kriging to generate a map of weekly influenza cases for France. They concluded that kriging had the advantage that it was not constrained by geographical boundaries and that it can be used to satisfactorily replace missing values. Pfeiffer (1994) used ordinary kriging to produce a surface of possum population density based on possum capture data at sample points (see Fig. 9). As discussed above, the isotropic variogram (see Fig. 8) suggests that this data is not stationary, but it also suggests strong spatial dependence. An exponential model was fitted and used as the basis for kriging. The crossvalidation scatterplot indicates that the model substantially under-estimates high trap catch. For regression modelling of continuous type dependent variables, spatial covariance structures can be specified as part of mixed models to estimate and adjust for spatial dependence (Littell et al. 1996).

A number of multivariate methods can be used for modelling of spatially continuous data. Principal components combine the information from multiple variables into a small number of components, each of them representing a particular combination of variables and explaining a particular proportion of the variation in the data. Eastman and Fulk (1993) used the technique to analyse the information contained in a time series of NDVI maps for Africa, thereby conducting a space-time analysis of continuous type variables. This technique could be used to assess the relative importance of spatial in comparison with temporal variation, for example the pattern of tick-borne disease incidence across a country could be separated into spatial and seasonal/cyclical variation. Cliff et al. (1995) discuss the application of multidimensional scaling (MDS) to spatial epidemiological data. They use the technique to map geographical information about measles mortality in Australia and New Zealand as a disease space where points with similar disease risks are closer to each other on the MDS map even though they are far removed geographically. Bailey and Gatrell (1995) discuss a range of other multivariate analysis techniques for spatially continuous data.

Recently, the use of spatial data for optimisation of resource allocation has been explored. Methods include multi-criteria and multi-objective evaluation techniques which have been adapted for spatial problems. They can take account of the uncertainty in the underlying input data as well as of the risk of making the wrong decision (Eastman et al. 1995). Systems are being developed which will take data from various spatial input data sets to, for example, define optimal types of animal production for specific geographical regions. With this methodology, it will be possible to define optimal disease control strategies given various spatially defined constraint variables.

CONCLUSION

Spatial data analysis provides a range of new techniques for descriptive epidemiological analysis and also explanatory or predictive investigations. In the context of the increasing number of geographically referenced databases, beyond the use of purely descriptive methods spatial analysis will allow hypothesis generation for example if used for triggering alarms for unusual disease occurrence. In the field of predictive modelling, the production of risk maps generated from observed disease risks as well as from predictions based on risk factor patterns provide a means for more effective resource allocation. New developments still have to be expected in the field of generalised linear modelling of spatial data, and the recent developments do point towards Bayesian techniques being the appropriate methodology for taking account of spatial dependence. In the new Millennium, spatial methods will become a standard component of the epidemiologist's tool chest of analysis techniques.

REFERENCES

- Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis* 27, 93-115.
- Anselin, L., Dodson, R.F. and Hudak, S. (1993). Linking GIS and spatial data analysis in practice. *Geographical Analysis* 1, 3-23.
- Augustin, N.H., Mugglestone, M.A. and Buckland, S.T. (1998). The role of simulation in modelling spatially correlated data. *Environmetrics* 9, 175-196.
- Bailey, T.C. and Gatrell, A.C. (1995). *Interactive spatial data analysis*. Longman Group, Harlow, Essex, England, 413p.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B.* 36, 192-236.
- Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics* 43, 1-59.
- Bithell, J.F. (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine* 9, 691-701.

- Bonham-Carter, G. F. (1994). Geographic information systems for geoscientists: Modelling with GIS. Elsevier Science Ltd, Kidlington, United Kingdom. 398p.
- Carrat, F. and Valleron, A.-J. (1992). Epidemiologic mapping using the "Kriging" method: application to an influenza-like illness epidemic in France. *American Journal of Epidemiology* *135*:1293-1300.
- Clayton, D. and Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* *43*, 671-681.
- Cliff, A.D., Haggett, P., Smallman-Raynor, M.R., Stroup, D.F. and Williamson, G.D. (1995). The application of multidimensional scaling methods to epidemiological data. *Statistical Methods in Medical Research* *4*, 102-123.
- Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society Series B* *52*, 73-104.
- Eastman, J.R. and Fulk, M. (1993). Long sequence time series evaluation using standardized principal components. *Photogrammetric Engineering and Remote Sensing* *59*, 991-996.
- Eastman, J.R., Jin, W., Kyem, P.A.K. and Toledano, J. (1995). Raster Procedures for multi-criteria/multi-objective Decisions. *Photogrammetric Engineering and Remote Sensing* *61*, 539-547.
- Glass, G.E., Schwartz, B.S., Morgan, J.M., Johnson, D.T., Noy, P.M. and Israel, E. (1995). Environmental Risk Factors for Lyme Disease identified with Geographic Information Systems. *American Journal of Public Health* *85*, 944-948.
- Gumpertz, M.L., Graham, J.M. and Ristaino, J.B. (1997). Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological and Environmental Statistics* *2*, 131-156.
- Haining, R. (1998). Spatial statistics and the analysis of health data. In: GIS and Health. A.C. Gatrell and M. Löytönen, eds. Taylor & Francis, London, pp. 29-47.

- Jacquez, G.M. (1996) A k nearest neighbour test for space-time interaction. *Statistics in Medicine* 15, 1935-1949.
- Jones, A.P., Langford, I.H. and Bentham, G. (1996). The application of K-function analysis to the geographical distribution of road traffic accident outcomes in Norfolk, England. *Social Science Medicine* 42, 879-885.
- Kelsall, J. E. and Diggle, P. J. (1995). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine* 14, 2335-2342.
- Knox, E.G. (1964). The detection of space-time interaction. *Applied Statistics* 13, 25-29.
- Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine* 14, 799-810.
- Kulldorff, M., Athas, W.F., Feuer, E.J., Miller, B.A. and Key, C.R. (1998). Evaluating cluster alarms: A Space-Time Scan Statistic and Brain Cancer in Los Alamos, New Mexico. *American Journal of Public Health* 88, 1377-1380.
- Langford, I.H. (1994). Using empirical Bayes estimates in the geographical analysis of disease risk. *Area* 26, 142-149.
- Langford, I.H., Leyland, A.H., Rasbash, J. and Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. *Applied Statistics* 48, 253-268.
- Lawson, A.B. and Waller, L.A. (1996). A review of point pattern methods for spatial modelling of events around sources of pollution. *Environmetrics* 7, 471-487.
- Lessard, P., L'Eplattenier, R., Norval, R.A.I., Kundert, K., Dolan, T.T., Croze, H. et al (1990). Geographical information systems for studying the epidemiology of cattle diseases caused by *Theileria parva*. *Veterinary Record* 126, 255-262.
- Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996). SAS System[®] for Mixed Models. SAS Institute, Cary, North Carolina, 633p.

- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* 27, 209-220.
- Merrill, D.W., Selvin, S., Close, E.R. and Holmes, H.H. (1996). Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. *Statistics in Medicine* 15, 1837-1848.
- Oliver, M.A. and Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems* 4 (3), 313-332.
- Pfeiffer, D.U. (1994). The role of a wildlife reservoir in the epidemiology of bovine tuberculosis. Unpublished PhD thesis, Massey University, Palmerston North, New Zealand. 496p.
- Pfeiffer, D.U., Duchateau, L., Kruska, R.L., Ushewokunze-Obatolu, U. and Perry, B.D. (1997). A spatially predictive logistic regression model for occurrence of theileriosis outbreaks in Zimbabwe. *Epidemiologie et santé animale* 31-32, 12.12.1-3.
- Teekayuwat, T. (1999). Geographical reporting and analysis of infectious animal disease occurrence in Thailand and New Zealand. Unpublished MVSc thesis, Massey University, Palmerston North, New Zealand, 187p.
- Wakefield, J. and Elliott, P. (1999). Issues in the statistical analysis of small area health data. *Statistics in Medicine* 18, 2377-2399.
- Webster, R., Oliver, M.A., Muir, K.R. and Mann, J.R. (1994). Kriging the local risk of a rare disease from a register of diagnoses. *Geographical Analysis* 26, 168-185.
- Williams, B., Rogers, D., Staton, G., Ripley, B., and Booth, T. (1994). Statistical modelling of georeferenced data: Mapping tsetse distributions in Zimbabwe using climate and vegetation data. In: Modelling vector-borne and other parasitic diseases. Perry, B. D. and Hansen, J. W. (eds) The International Laboratory for Research on Animal Diseases, Nairobi, Kenya. 267-280.
- Xia, H., Carlin, B.P. and Waller, L.A. (1997). Hierarchical models for mapping Ohio lung cancer rates. *Environmetrics* 8, 107-120.

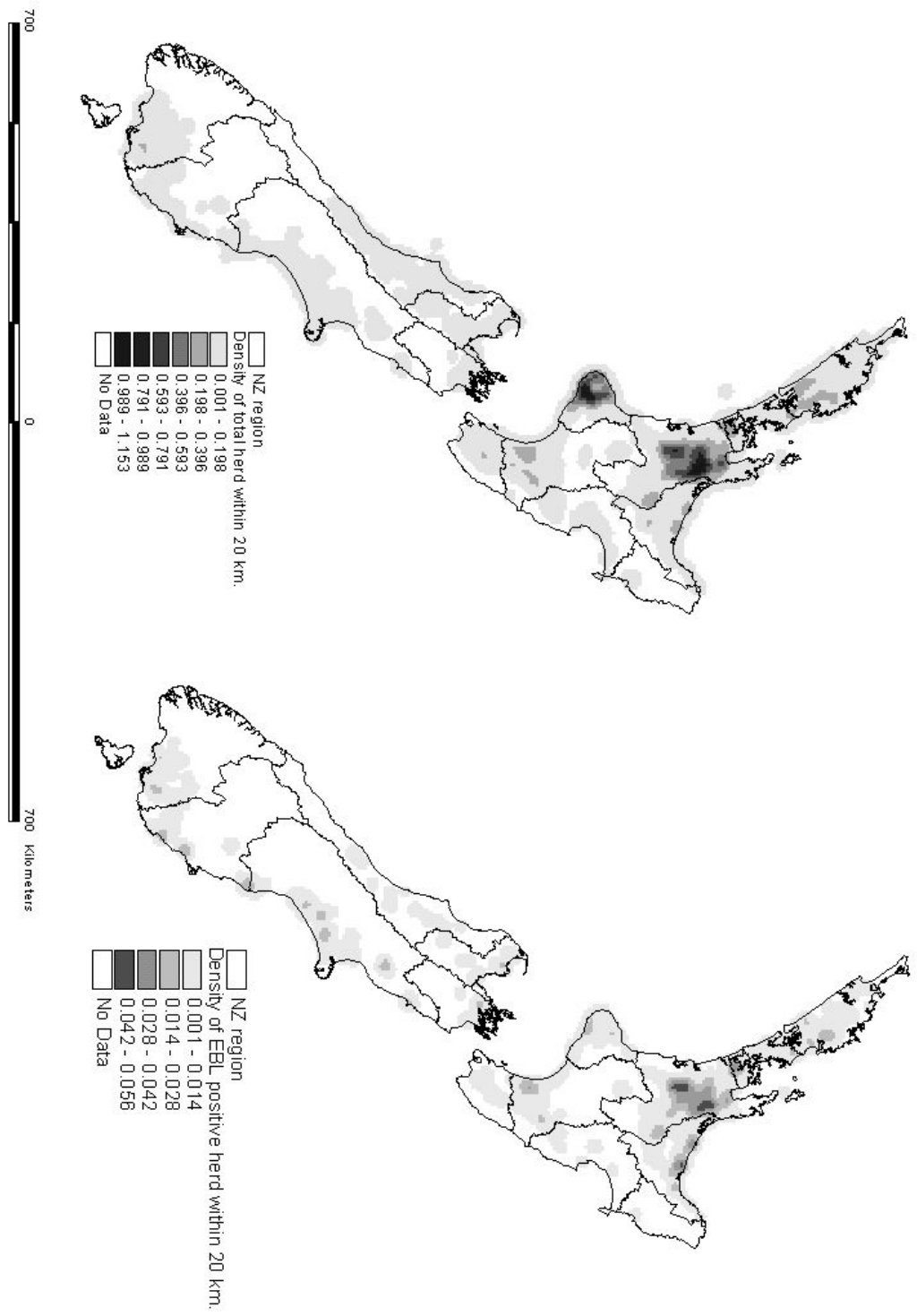


Fig. 1 Kernel density maps of all dairy herds and those infected with EBL in New Zealand

```

Swap drive : e:
Cases file : tbcases.txt
N : 96

Virtual drive : e:
Controls file : tbcontro.txt
N : 91

```

(C) Biomedware, 1994

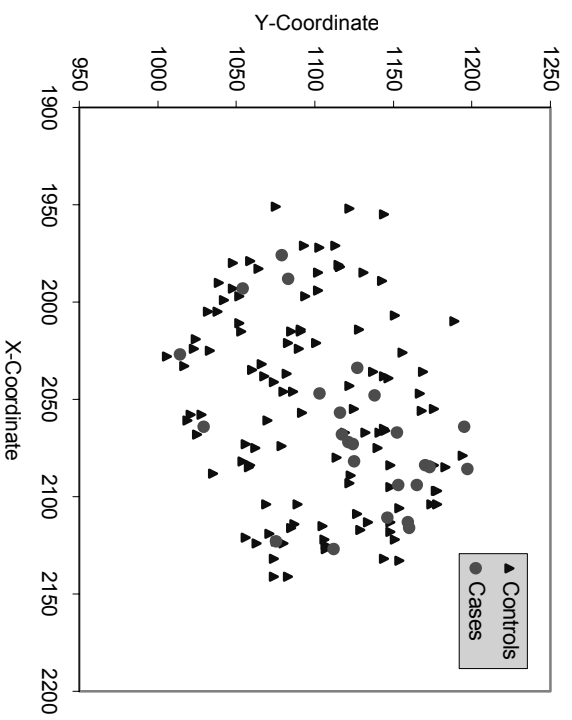


There were no case-control ties

k	TK[]	ETT[]	VARRT[]	z	P
1	53	49.03	27.05	1.15	0.1256
2	117	98.06	55.60	2.54	0.0056
3	172	147.10	84.00	2.72	0.0033
4	231	196.13	113.07	3.28	0.0005
5	285	245.16	141.42	3.35	0.0004
6	337	294.19	176.82	3.22	0.0006
7	388	343.23	210.37	3.09	0.0010
8	445	392.26	253.73	3.31	0.0005
9	507	441.29	286.74	3.88	0.0001
10	563	490.32	327.69	4.01	0.0000

Bonferonni P : 0.0003
Simes P : 0.0000

Fig. 2 Cuzick and Edwards' method applied to tuberculosis breakdown case control study data (+ = cases, □ = controls, arrows identify nearest neighbours)

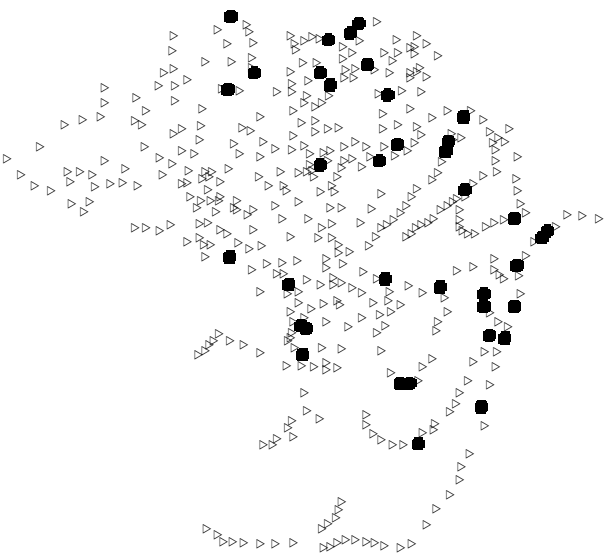


a



b

Fig. 3 Map of case-control farm locations and difference K-function based on case-control data for animal disease outbreak



a



b

Fig. 4 Locations of trap sites where possums with (full circles) and without (empty triangles) tuberculosis had been caught and the location of the most likely cluster (full circles) according to the spatial scan statistic

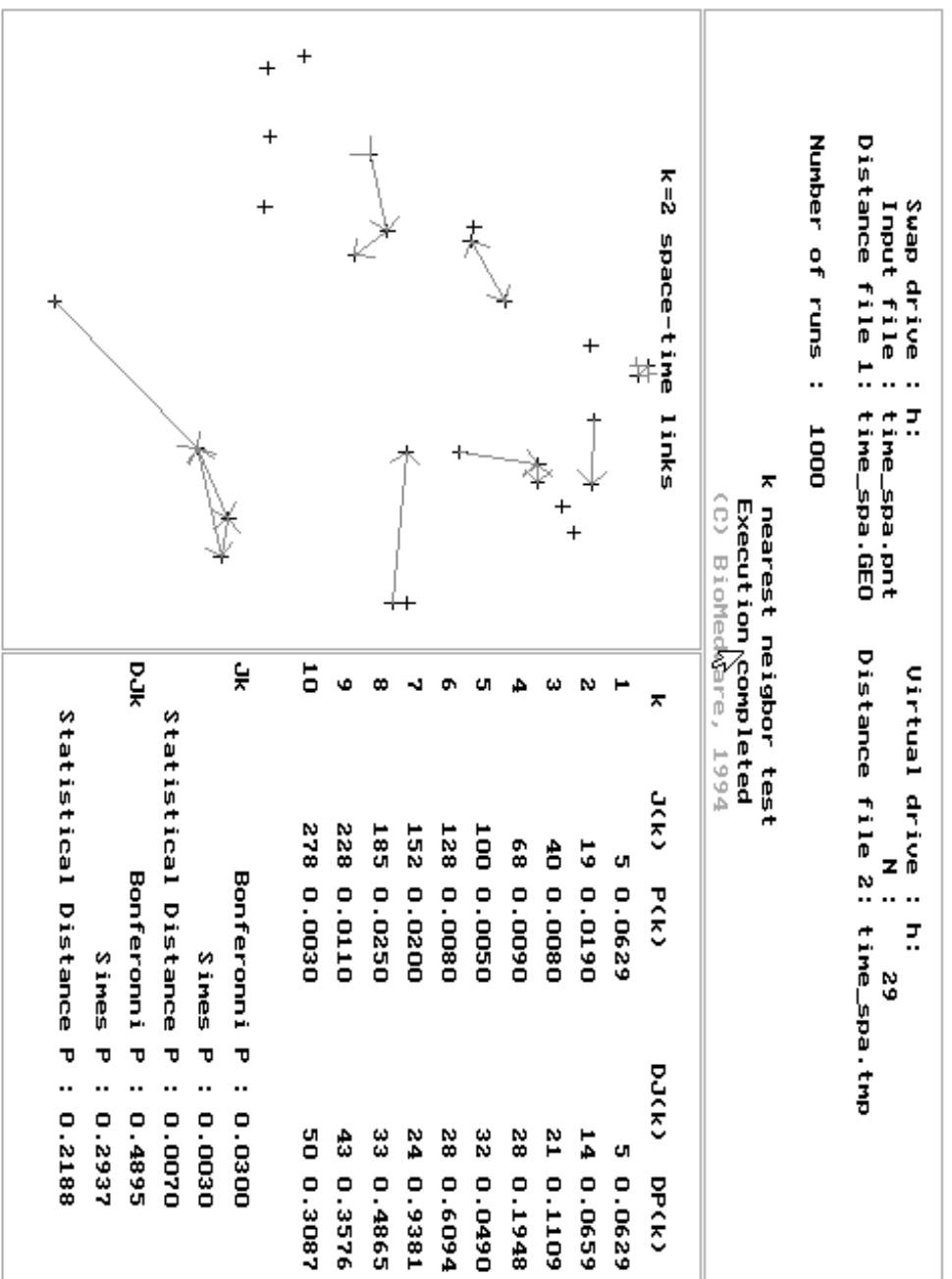
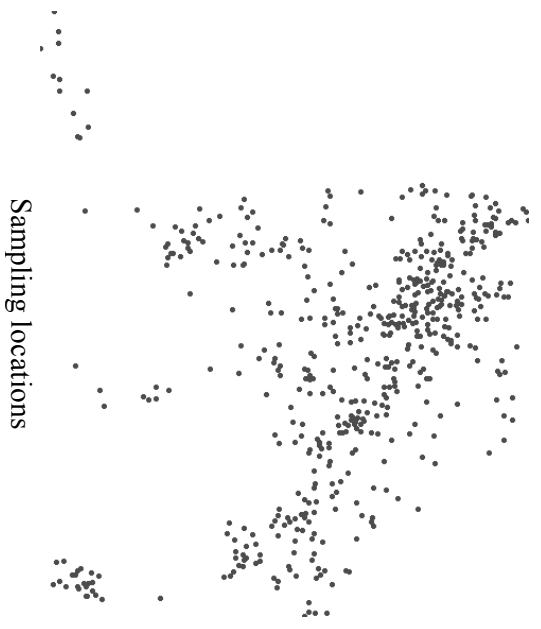
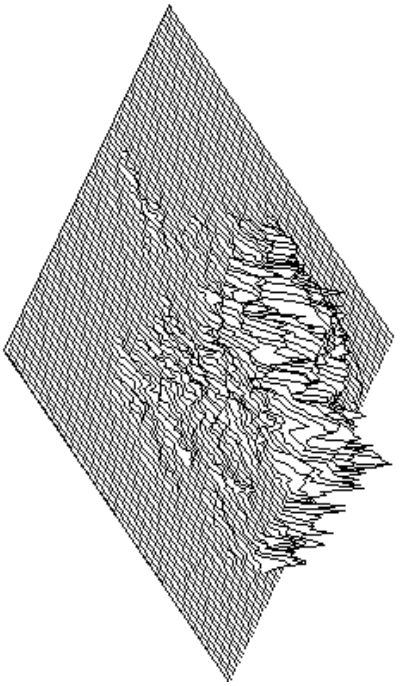


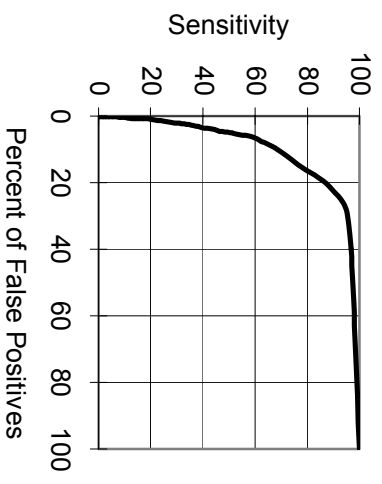
Fig. 5 Results from applying the K-nearest neighbour method to test for time-space interaction between cases of tuberculosis infection in wild possums



Sampling locations



DTM of predicted probability of *Theileria parva* presence



ROC curve for logistic regression model



Raster map of predicted probability of *T. parva* presence

Fig. 6 Results of a multiple logistic regression analysis for prediction of *Theileria parva* presence in Zimbabwe

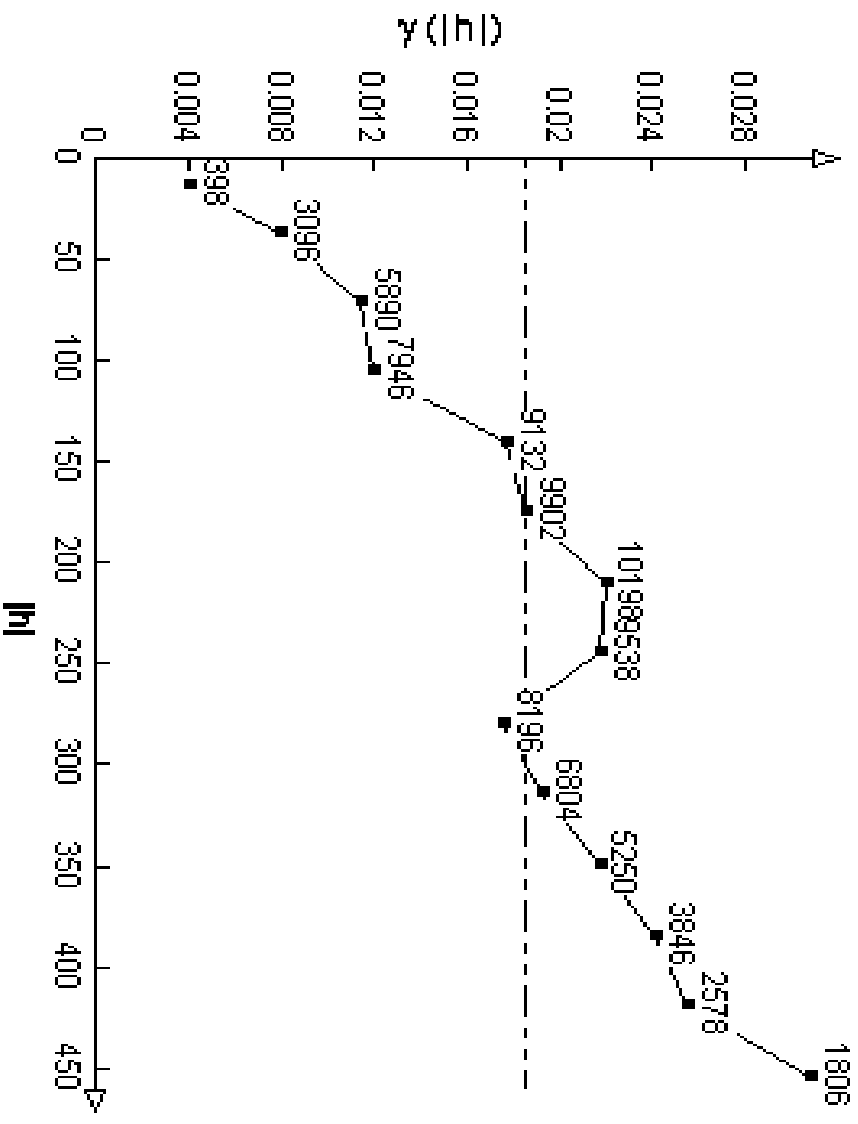
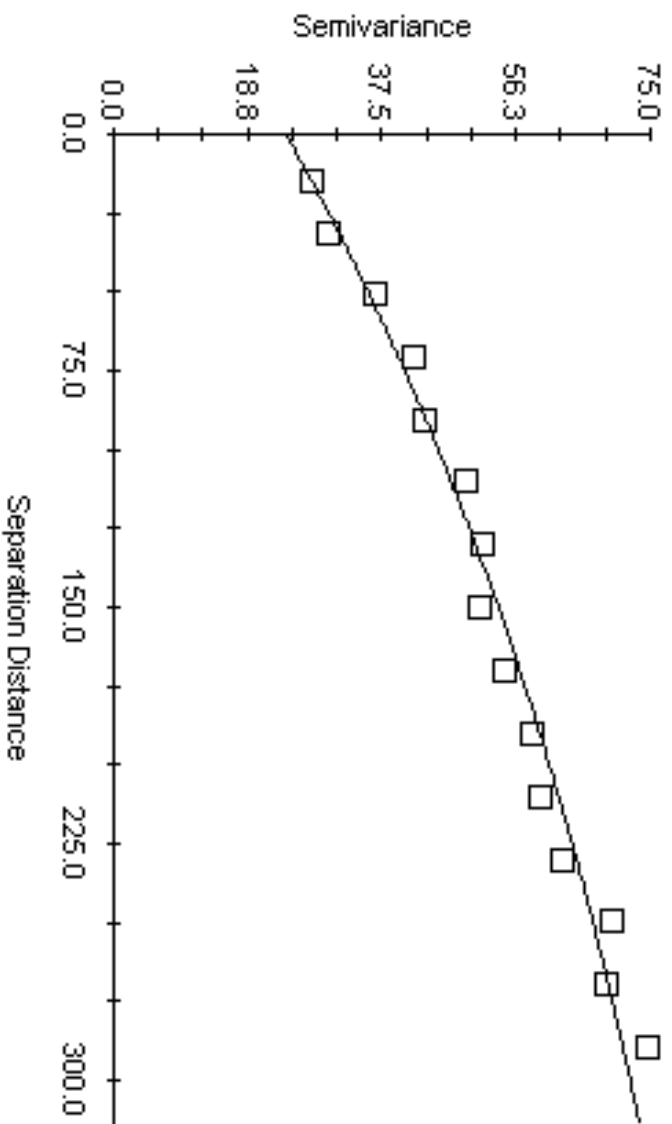


Fig. 7 Isotropic semi-variogram for the proportion of tuberculous possums captured at individual trap sites in the longitudinal study



Exponential model ($C_0 = 24.0000$; $C_0 + C = 102.2500$; $A_0 = 314.10$; $r^2 = 0.981$;
 RSS = 54.48)

Fig. 8 Isotropic exponential variogram model for possum trap capture data

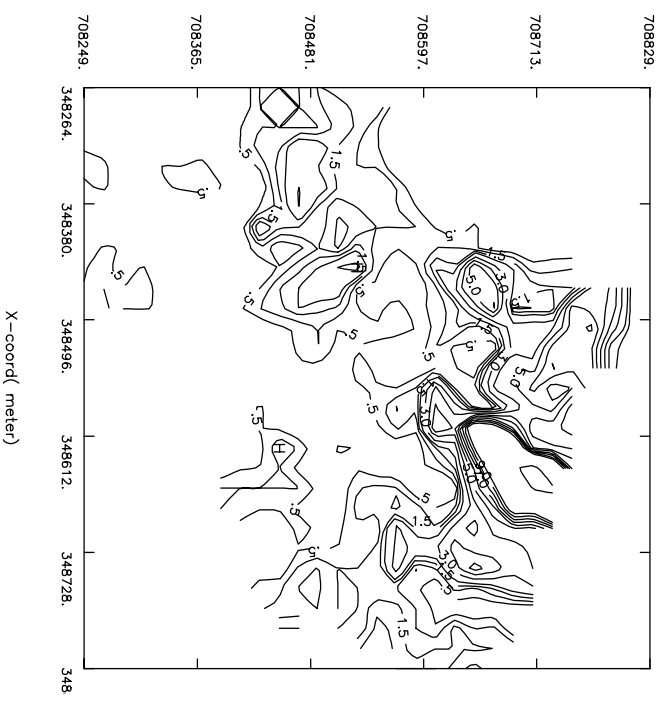
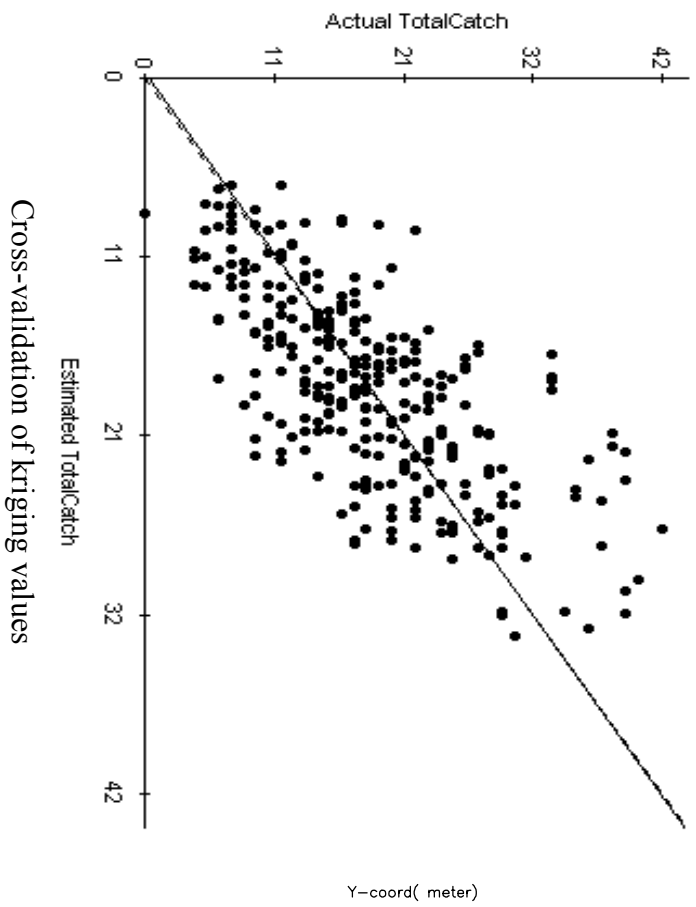


Fig. 9 Crossvalidation and contour map estimates for interpolated possum density in the longitudinal study on possum tuberculosis epidemiology

Contour map based on kriging estimates

Cross-validation of kriging values