A SPATIALLY PREDICTIVE LOGISTIC REGRESSION MODEL FOR OCCURRENCE OF THEILERIOSIS OUTBREAKS IN ZIMBABWE

Pfeiffer D.U.[1], Duchateau L.[2], Kruska R.L.[2], Ushewokunze-Obatolu U.[2], Perry B.D.[2]

Des bases de données dans l'espace fournissent des informations qui peuvent être utilisées en vue de développer des modèles de prédiction d'apparition de maladie en tenant compte de plusieurs facteurs de risque. Cependant, ce type de modèle doit tenir compte des auto-corrélations spatio-temporelles, car beaucoup de maladies infectieuses n'apparaissent pas de façon aléatoire dans l'espace. Le modèle de régression logistique peut être utilisé pour ces prédictions. Les méthodes statistiques pour tenir compte de ces auto-corrélations spatiales ne sont pas développées pour ce type de modèles tel qu'elles le sont pour les régressions par la méthode des moindres carrés.

Dans cette étude, des données sur la végétation et le climat au Zimbabwe ont été utilisées pour la prédiction de l'apparition de la theilériose dans certaines conditions spatiales.

Plusieurs méthodes d'estimation de l'auto-corrélation spatiale sont comparées. Elles sont basées sur l'inclusion des effets aléatoires représentant : le risque régional d'infection, la coordination géographique, les différences entre les tailles des régions ainsi que le terme de covariance spatiale. Le modèle avec la plus petite déviance était basé sur l'inclusion d'une variable indiquant la localité. Ce modèle montre une grande réduction dans l'auto-corrélation spatiale entre les résidus et le modèle prédictif. Une courbe ROC a été utilisée pour résumer l'exactitude de la prédiction du modèle en terme de sensibilité et de spécificité en fonction de différents seuils d'apparition de maladies. Des cartes géographiques ont été réalisées afin de faciliter l'interprétation visuelle des résultats du modèle. Ce type d'approche de modélisation prédictive peut être utilisé pour permettre une maîtrise locale efficace contre les maladies. Les décideurs pourront avoir recours aux cartes avec différents seuils pour faire leurs choix. La courbe ROC peut également être utilisée pour déterminer le seuil adéquat aux objectifs du programme.

INTRODUCTION

Multivariate logistic regression models are mostly used to identify risk factors associated with the occurrence of particular disease processes. Logistic regression models have also been used as tools for veterinary diagnosis by providing the probability of a particular disease in an individual animal given a set of characteristics such as diagnostic test results or other risk factors. They can also be applied to the prediction of the probability of the occurrence of future disease events. Decision making in animal disease control is constrained by cost-benefit considerations, which in turn should take into account the probability of the occurrence of particular disease events. The unit of interest in this context usually is an aggregate of spatial information such as an administrative district, province or state. With the advent of spatial databases and geographic information systems (GIS) the level of spatial aggregation can be easily controlled by the end user and is only limited by the spatial units at which the data has been collected. The relationships between various variables stored in a spatial database can be investigated and used to provide predictive tools allowing more cost-effective spatially optimised disease control.

In this study a logistic regression model was developed to estimate the probability of theileriosis occurrence in Zimbabwe, and the usefulness of measures of model goodness-of-fit for decision makers was investigated. Specific attention was given to the potential of effects of spatial autocorrelation on regression coefficient estimates.

MATERIALS AND METHODS

A spatial database from Zimbabwe with information on various climatic, vegetation, land use, topographic and animal demography was used to model the occurrence of theileriosis outbreaks recorded by the Zimbabwean disease control authorities between 1979 and 1989 (Kruska and Perry 1992). Each of the 4839 observations in the database represented a spatial grid-cell with a resolution of 5 arc-minutes. Rainfall was recorded using a millimeter and temperature using a tenth of a degree scale. In the case of climatic data there was strong collinearity between a number of variables and the approach described by Duchateau et al (in press) was used to reduce the database to its effective dimensionality. Spatial autocorrelation of theileriosis outbreaks was quantified using the Cuzick and Edwards' test (Cuzick and Edwards 1990). Stepwise logistic regression was then used to identify amongst the variables included in the database the most important risk factors for occurrence of theileriosis outbreaks. The impact of spatial autocorrelation on the regression model parameters was assessed using a number of different methods. Bailey and Gatrell (1995) suggest including location into the model as coordinates or geographical areas. In this analysis longitude and latitude or through local regions representing 16 and 25 aggregate grid cells were included as random effects into the regression model. In addition, a model was

developed where the dependent variable presence / absence of theileriosis outbreaks in a particular spatial grid cell was replaced by a binomial variable representing the probability of theileriosis outbreaks using the immediately neighbouring grid cells as sampling points (Eastman 1997). The same variable was used as a covariate representing autocovariance to generate an autologistic model as described in Augustin et al (1996). Goodness-of-fit of the model was assessed by producing spatial maps of the prediction residuals and the confidence limits of the prediction probabilities. The model residuals were also inspected for presence of autocorrelation based on the Moran's I statistic. Predictive accuracy of the model can also be quantified using sensitivity and specificity measures depending on cut-off values used to define occurrence and non-occurrence of theileriosis outbreaks. These results can be presented using a receiver-operating characteristic (ROC) curve.

The spatial data was manipulated using the combination of the geographic analysis system IDRISI for Windows version 2.0 (The IDRISI Project, Clark University, Worcester, MA, U.S.A.) and Microsoft Access for Windows version 8.0 (Microsoft Corporation, Redmond, WA, U.S.A.). Spatial clustering was assessed using the software Stat! (BioMedware, Ann Arbor, MI, U.S.A.). The statistical analyses were conducted in SAS for Windows version 6.12 (SAS Institute, Cary, NC, U.S.A.) using PROC LOGISTIC for standard logistic regression and the macro GLIMMIX to implement a generalised linear mixed model allowing inclusion of spatial location as random effects.

RESULTS

For the analysis of spatial autocorrelation using Cuzick and Edwards' test all 387 grid cells with recorded occurrence of theileriosis outbreaks were treated as cases and a random sample of 888 grid cells without theileriosis outbreaks as controls. The results of this analysis indicate that the nearest neighbour of a theileriosis outbreak tend to be another theileriosis outbreak rather than a control. This suggests the presence of spatial clustering.

Interpretation of the loadings in a principal components analysis of rainfall data after varimax rotation suggested combining December to March rainfall into an average for the rainy season and May to September for the dry season (Duchateau et al in press).

The final logistic regression model for the binary dependent variable occurrence / non-occurrence of outbreaks ignoring any potential spatial autocorrelation included the main effect variables rainy season, dry season, their first –order interaction term, the main effects mean annual temperature, communal land use, commercial land use and maximum monthly normalised difference vegetation index (NDVI). The regression coefficients are presented in Table 1. The –2 Log Likelihood chisquare statistic comparing this model with an intercept only model was statistically significant ($_^2$=834.7, 7df, p=0.001). The residuals produced by this model showed strong evidence of spatial autocorrelation (Moran's I statistic =0.1467 with an expected value under independence of – 0.0002). Running the same analysis using the binomial dependent variable representing risk of infection including the immediate neighbourhood did not result in changes in the regression coefficients and in only very minor increases of their confidence limits. This dependent variable was therefore not considered further. Inclusion of different random effects terms produced models with the regression coefficients presented in Table 1. The models using a local region representing 16 adjacent grid cells and the grid-cell coordinates did not result in significant model parameter changes and were therefore not included in Table I.

Table I
Regression coefficients expressed as odds ratios with their 95% confidence limits for the four different logistic regression models

Parameter	No random effects	Including local region (25 grid cells)	Including autocovariance term	Including spatial covariance structure
Rainy season	1.046 (1.043-1.049)	1.04 (1.03-1.05)	1.03 (1.02-1.03)	1.048 (1.039-1.057)
Dry season	1.42 (1.316–1.533)	1.37 (1.04-1.79)	1.19 (0.99-1.42)	1.51 (1.2-1.9)
Interaction wet*dry	0.997 (0.9973-0.9975)	0.997 (0.996-0.998)	0.999 (0.997-0.999)	0.997 (0.995-0.998)
Average annual temperature	0.953 (0.949-0.956)	0.95 (0.94-0.96)	0.972 (0.96-0.98)	0.958 (0.95-0.97)
Communal land use	4.32 (2.97-6.30)	1.38 (0.66-2.88)	2.12 (0.96-4.7)	1.7 (0.83-3.5)
Commercial land use	13.3 (9.17-19.3)	4.11 (2.02-8.37)	6.25 (2.88-13.6)	6.05 (3.03-12.1)
Maximum NDVI	1.06 (1.05-1.07)	1.04 (1.02-1.07)	1.03 (1.001-1.053)	1.05 (1.02-1.08)
Deviance	1862.24	1453.6	1641.18	1872.02

The model with the smallest deviance term was selected and examined for spatial autocorrelation of the residuals. The Moran's I statistic of 0.0176 compared with an expected value of –0.0002 under independence indicated that there was still spatial autocorrelation present. Spatial maps were produced presenting the predicted probability of theileriosis outbreaks in Zimbabwe as well as their 95% confidence limits based on the estimates produced for the logistic regression model including large local region as random effect term. In addition, a ROC curve was produced showing sensitivity and specificity values for the same model (see Figure 1). The area under the ROC curve was 0.906 compared with 0.898 for the model without random effect.

Figure 1
Receiver operating characteristic curve for logistic regression model with large
local region as a random effect
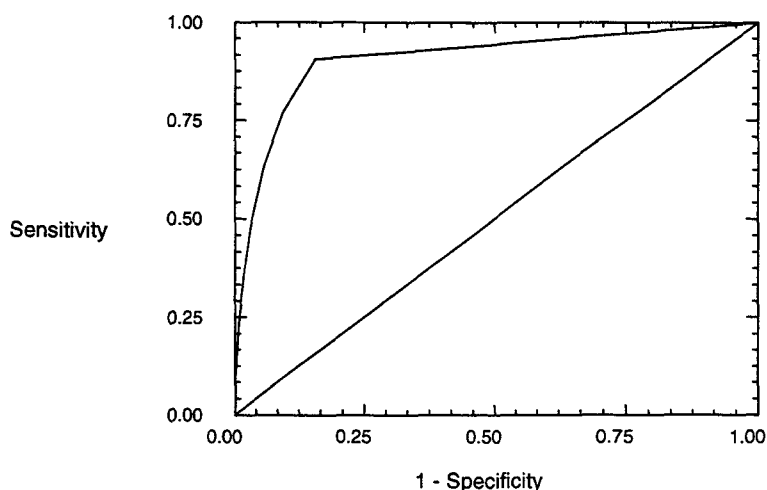
DISCUSSION

Prediction of events on a spatial scale has been the subject of a number of investigations. Glass et al (1995) used logistic regression to generate a lyme disease-risk density map. Williams et al (1994) compare the use of non-linear discriminant analysis, neural networks, decision tree induction methods and k-nearest neighbour analyse for prediction of tsetse fly distribution in Zimbabwe. These authors did not take spatial autocorrelation into account.

Comparison of the deviances between the four models described in Table I suggests that the model without random effect term did severely overestimate the odds ratios for the variables representing land use. Visual comparison of the deviances between the four models suggests that the model including the term representing the large local region did fit the data better than the other models. The residuals of this model do indicate that there is still some autocorrelation but much less than what was found for the model without random effects. The model with the spatial covariance structure probably did not perform as well because a covariance term suitable for modelling continuous dependent variables had been used. There is still further research required specifically with regard to appropriate spatial covariance structures when dealing with binomial or binary dependent variables.

The spatial maps of model predictions provide a visually effective basis for making disease control decisions taking into account local outbreak risks. Knowledge of the uncertainty of these estimates as revealed by their 95% confidence limits will allow the decision maker to choose between risk averse and –prone approaches. The decision can then be based on which level of disease risk is sufficient to justify implementing certain control measures such as for example preventive vaccination. The ROC curve very effectively summarises the probability of missing potential outbreaks or unnecessarily applying the control measures to individual spatial units depending on which cut-off point has been selected.

BIBLIOGRAPHY

Augustin N.H., Mugglestone M.A., Buckland S.T., 1996. An autologistic model for the spatial distribution of wildlife. Journal of Applied Ecology 33, 339-347.

Bailey T.C., Gatrell A.C., 1995. Interactive spatial data analysis. Longman Scientific & Technical, Burnt Mill, Harlow, Essex, England, p 313.

Cuzick J., Edwards R., 1990. Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society B 52, no. 1: 73-104.

Duchateau L., Kruska R.L., Perry B.D., in press. Reducing a spatial database to its effective dimensionality for logistic-regression analysis of incidence of livestock disease. Preventive Veterinary Medicine.

Eastman R.J., 1997. IDRISI for Windows Version 2.0 - Tutorial Exercises. IDRISI Productions, Clark University, Worcester, MA, p. 70-71.

Glass G.E., Schwartz B.S., Morgan J.M., Johnson D.T., Noy P.M., Israel E., 1995. Environmental risk factors for lyme disease identified with geographic information systems. American Journal of Public Health 85, no. 7: 944-48.

Kruska R.L., Perry B.D., 1992. Development of spatial databases for analysis of tick-borne diseases of cattle in Zimbabwe. SADDC Regional Workshop on Geographic Information Systems for Natural Resource Management, Harare, Zimbabwe, April 1992.

Williams B., Rogers D., Staton G., Ripley B., Booth T., 1994. Statistical modelling of georeferenced data: Mapping tsetse distributions in Zimbabwe using climate and vegetation data. Perry B.D., Hansen J.W., editors. Modelling vector-borne and other parasitic diseases. Nairobi, Kenya: The International Laboratory for Research on Animal Diseases. p267-80.