A novel approach to identifying space-time dependencies within data sets using the K-function and simulation.

G.T. Innocent[1*], I.J. McKendrick[2] & S.W.J. Reid [1,3]

[1] Comparative Epidemiology and Informatics, Department of Veterinary Clinical Studies, University of Glasgow Veterinary School, Bearsden Road, Glasgow G61 1QH, UK. [2]BioSS, James Clark Maxwell Building, Kings Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK. [3]Department of Statistics and Modelling Science, Univeristy of Strathclyde, Richmond Street, Glasgow G1 1XW, UK

Summary

Although a useful method when investigating stationary processes in space and time, analysis using the k-function can be difficult to interpret when it indicates that the process is over- or under-dispersed at certain distances in either space or time.This is particularly common in epidemiological investigations where, for example, the underlying population is not uniform.

The method proposed to investigate the space-time structure within the data is to simulate the data, assuming no space-time correlation in the infection process but incorporating any known space-time structure in the susceptible population, and then to evaluate the k-function for the data against those arising from simulated data

Introduction

The spatial k-function (Ripley, 1976) and its extension to time and space-time interactions (Diggle *et al*. 1996) are useful tools in examining the spatio-temporal structure of data collected with some positional attributes. This approach has been applied to diseases in the veterinary field (e.g., French, 2000). The method allows the visualisation of space-time interactions at a range of spatial distances and time lags via the D and D_0 plots. These can, however, lead to inappropriate inference. It is quite possible, for example, using totally random, uniformly distributed data in both time and space, to demonstrate apparent space-time interactions, due to random local under- or over-representation of Poisson-distributed events. Datasets are commonly simulated under the null hypothesis of no space-time structure for comparison with the k-functions arising from study data.

Objectives

To determine if simulation methods are approriate for extending the utility of the k-function to situations where there is an *a priori* model of dispersion, for example, where it is known that the population at risk of disease is not uniformly distributed.

Materials and methods

Data were simulated using the R statistical package (Ihaka and Gentleman, 1996), as a spatially-heterogeneous Poisson process defined as a seasonally variable (temporally heterogeneous) per-capita rate of infection applied to an underlying non-uniform population. The model was defined over a 100 by 100 unit square, for 100 units of time. A k-function analysis was conducted and then this was compared to

results arising from 999 simulations of a stationary random process, with an identical number of cases but with uniform population and prevalence, and from 999 simulations of a stationary random process with a uniform prevalence and non-uniform population. These simulated data were analysed using the same k-function method: the resulting statistics can be thought of as sample distributions of a test statistic under two models of the null hypothesis.

The results of the k-function analysis can be expressed in three parts; a spatial component, a temporal component, and a spatio-temporal component. These three components were calculated at integer points in both time and space from one to seventy. At all points the original data set was compared to the k-function calculated on the simulated uniform process, and its rank recorded.

Results

The ranks of the k-function of the data when compared to the stationary processes based on the uniform population were non-uniformly distributed, most frequently being ranked at the extremes (Figure 2). Figure 3 shows the distribution of the ranks of the k-function of the data compared to the Poisson process simulated subject to the underlying population distribution, and it is clear that the spatial component of the k-function is fairly uniform in distribution, whereas the temporal component still generates extreme rankings.

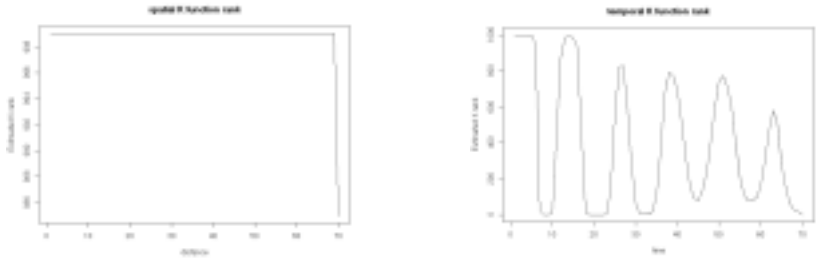

Figure 2 Plots of ranks of data-derived spatial and temporal k-functions relative to simulations based on a uniform population.

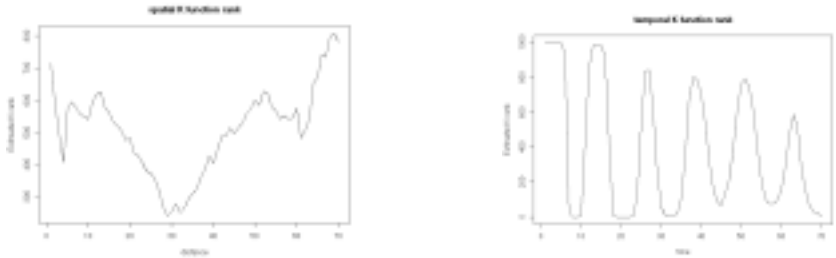

Figure 3 Plots of ranks of data-derived spatial and temporal k-functions relative to simulations based on a heterogeneous population.

Discussion

Many epidemiological investigations involve models of the null hypothesis that are non-uniform in either space or time. At its simplest this may merely reflect that cases of disease, for example, are associated with the centroid of population-census-based areas, and possibly are also aggregated in time, perhaps by week or month. If these

cases are sufficiently sparse compared to the resolution of space and time aggregation, then the k-function may not be unduely affected. However, it may then be difficult to determine if a structure arising in the k-function analysis represents a true spatio-temporal component of the (statistical) population, or merely represents an artefact of the method of data collection. By comparing the k-function of the data to those arising from data simulated from a null model which incorporates any assumptions inherent in the method of data collection used, it is possible to test for the presence of spatio-temporal structure in the model for prevalence, independent of any effects arising from heterogeneous population structures.

Although it is difficult, and arguably inappropriate, to interpret the results provided by this analysis in terms of a single hypothesis test, it does allow the examination of various models in exploring whether, in qualitative terms, they describe the spatio-temporal structure of observed data. Definition of just how extreme ranks must be in order to determine that spatio-temporal structure is present, which is not accounted for in the null model, is an on-going process.

References

Diggle, P., Chetwynd, A., Haggkvist, R. and Morris, S. (1995) Second-order analysis of space-time clustering. Statistical Methods in Medical Research, 4, 124-136

Ripley, B.D., (1976) The second-order analysis of stationary point processes. Journal of Applied Probability, 13, 255-266.

French, N.P., Clough, H.E., Berriatua, E, McCarthy, H.E., Proudman, C.J. & Hillyer, M.H., (2000) The use of k-function analysis to detect and describe space-time clustering of animal diseases, Proceedings of the 9th International Society for Veterinary Epidemiology and Economics Conference.

Ihaka, R. and Gentleman, R., (1996) R: A Language for Data Analysis and Graphics, Journal of Computational and Graphical Statistics, 5, 3, 299--314

Acknowledgements

This study is a part of the International Partnership Research Award in Veterinary Epidemiology (IPRAVE), *Epidemiology and Evolution of Enterobacteriaceae Infections in Humans and Domestic Animals*, funded by the Wellcome Trust. The contribution of BioSS to this study is funded as SEERAD project BSS/028/99.