Graphical Modeling of Ecological Time Series Data

C.S. Meurk [1, 2], Brown, J.A. [1,] and M.Reale [1]

[1] Department of Mathematics and Statistics, University of Canterbury, Christchurch
[2] School of Social Sciences, University of Queensland, Brisbane
Email: Jennifer.brown@canterbury.ac.nz

Keywords: *statistical modelling, causality*

EXTENDED ABSTRACT

Graphical models offer a powerful tool for studying ecosystem function. Changes in relationships among extrinsic and intrinsic biological and environmental variables can be explored. We discuss the application of graphical modeling to ecological data and illustrate this with an example case study. Ecological datasets are characteristically small with few data points, covering only a short period of time, and with high seasonal variation. This high variation, along with the fact that the data sets are small, can present problems for graphical modeling. Despite this, in general, considerable insight into ecosystem function can be gained from the use of graphical modeling.

In our case study we modelled the ecosystem relationship between mice and food abundance. In New Zealand cyclical waves in the mice population size within beech forest roughly correspond to periods after a heavy beech tree seeding year. One explanation for this cycle is that years of heavy beech seeding causes an increase in mouse population. Understanding the ecosystem relationship will help understand possible causal links.

In this study we used data collected by the Orongorongo valley near Wellington between August 1971 and November 1996. We used three ecosystem measures: mouse population size, mouse breeding and seed fall and compared graphical models among seasons.

Mice population size was estimated from counts in mice traps adjusted for trapping-effort. Beech seed fall was measured using seed traps under mature trees. Mouse breeding was measured by the proportion of mice caught in traps that were pregnant females and the proportion of adult males in the population.

Direct assessment of seasonal effects on mice-beech forest ecosystem relationships was by comparison among seasonal-models. Separate graphical models were produced, one for each leading season: a model with spring as the most recent time, a model with autumn as the most recent time, and so on.

The seasonal-graphical models were helpful in understanding the relationship among variables. The winter observed mouse numbers are dependent on the numbers in the previous season, autumn, and on levels of seed fall. Similarly summer mouse numbers are dependent on the size of the population in the previous season, spring. There was no direct link with seed fall as there was with the size of the previous winter's mouse population. The number of mice in spring was related to mice numbers in the winter before. All these relationships are positive, i.e., with an increase in mouse numbers in winter, mouse numbers in spring will increase.

Graphical models for time series can be used for a wide range of environmental studies. The complexity of ecosystem interactions can be described by modelling the multivariate system with graphical links for the changing interactions through time. Comparison among models for different seasons, or for periods pre- and post-perturbation can be used for quantifying temporal change.

1. INTRODUCTION

Graphical modelling for time series is a method proposed for the visualisation and casual interpretation of time series data (Reale 2001, Reale 2002). Graphical modelling was first applied for analysis of medical data (Lauritzen and Spiegelhalter 1988).

There have been few applications of graphical modelling to ecological data yet the very nature of ecological study suggests that this is a useful approach. Ecological study is characterised by complex interactions among many biotic and abiotic factors. Ecology is the study of distribution and abundance of organisms and how these distributions are affected by their environment. Abiotic factors describe the physical environment of the habitat, including light, temperature, wind etc., and biotic factors describe the intrinsic population regulatory factors of the organism and of other living organisms.

The interest in ecological study is often in how these biotic and abiotic interactions change over time. Many ecological studies are interested in how a particular resource use or human activity could effect the environment. The environment is described by a model in some way and the impact of a perturbation or gradual change in resource availability is then estimated by changing model parameters. Modelling the environment is no simple task, given the multiplicity of environmental factors and complexity of their interactions.

Graphical models have potential as an analysis tool for ecologists for three reasons. The method requires process-complexity to be reduced to a visual graphical display thereby giving insight into the ecological system. The identification of causal links between events is the essence of ecological studies focussed on understanding the processes and mechanisms behind an observed data-pattern (Greig-Smith 1983). The third reason is that often seasonality in the ecological process adds complexity to a standard time series analysis, whereas in graphical modelling it can be dealt with directly.

In this paper we discuss the application of graphical modelling for time series to ecological data. We briefly describe graphical modelling for time series, and then introduce the case study. We discuss our results which were of mixed success, and conclude with suggestions for other applications.

2. GRAPHICAL MODELLING FOR TIME SERIES

It is useful to present a brief overview of the graphical modelling methods and to introduce some terms we use.

Time series analysis is concerned with modelling an observed temporal sequence of data. The extension to multivariate time series is natural where now the interest is in the individual series and in their interaction with each other. These interactions can be complex. The graphical modelling approach can be helpful in understanding the complexity.

The univariate autoregressive model (AR) extension to multivariate is the VAR, vector autoregressive model. A structural VAR, sVAR, has the form:

$$\alpha_0^* X_t = \alpha_1^* X_{t-1} + \alpha_2^* X_{t-2} \ldots \alpha_p^* X_{t-p} + Z_t$$

where X_t is the observed vector of variables in the series at time t, α are the model coefficients, and Z_t is an error term, mean 0 and variance σ^2.

Graphical modelling terminologies are nodes, edges, CIGs and DAGs. In the conditional independence graph (CIG) in figure 1 the circles are nodes which correspond to model-variables and the lines joining the nodes are edges, corresponding to the relationship between two variables.

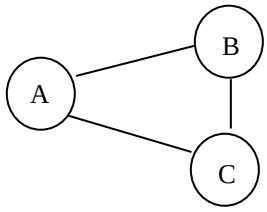

Figure 1. An undirected graph with three nodes and three edges.

When directions are added to the edges such that a cycle is not created the resultant graph is a directed acyclic graph (DAG), figure 2. The nodes with outgoing arrows are referred to as parent nodes, and child nodes are those with incoming arrows.

The DAG in figure 2 can be described by a system of equations relating one variable to others:

$$A = \alpha_1 B + \alpha_2 C$$
$$C = \alpha_3 B$$

.

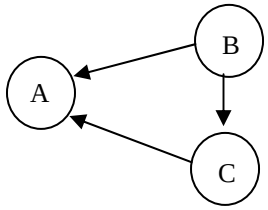

Figure 2. A directed graph with the node B being the parent of both nodes A and C. Node A has two parent nodes, B and C.

In the CIG derived from a DAG edges are drawn between nodes that are conditionally dependent. The CIG associated with the DAG in figure 2 has a link between B and C to describe the B-C relationship conditional on the three variables of the graph. The term 'moralisation' is used to describe the process of ensuring conditional dependence, by adding links in a CIG that has been created from a DAG. The DAG in figure 3 would result in the same figure 1 CIG as the DAG in figure 2, because there must be a moral link between B and C.

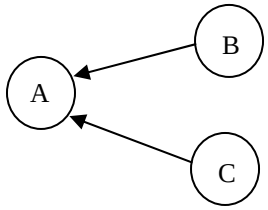

Figure 3. A directed graph with the node B being the parent of only node A. This DAG is different from figure 2, but both lead to the same CIG, figure 1.

Moving from a DAG to a CIG can be straightforward because there can be only one CIG for each DAG. Moving from a CIG to a DAG requires adding directional edges to indicate the movement of time, and is not straightforward. Both DAG's shown (figure 2 and 3) can produce the same CIG (figure 1). In fact there are more than two possible DAG's for the figure 1 CIG, the DAG in figure 4 would also produce the figure 1 CIG.

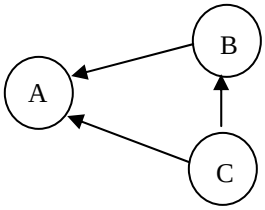

Figure 4. Yet another directed graph. This DAG is different from figure 2 (and figure 3), but would also lead to the same CIG, figure 1.

In creating a DAG from a CIG which direction to add the arrows can be obvious in some applications. When two nodes occur at the same time (contemporaneously) then it is not so obvious. Consider figure 2 and 4, B and C both occur before A but occur at the same time. Which direction should the arrow between B and C be, or should there even be an arrow between B and C (figure 3)?

Graphical modelling for time series (Reale 2001) is a three step process: creation of a CIG, creation of the equivalent DAG, and in step three, regression modelling and model selection. This third step is the process of selecting between competing DAG's.

In the graphs the time series of variables are lined up so that the most recent observation, *t*, is on the left (Figure 5). The directional arrows are used to specify a system of equations specified for each possible DAG. Model selection criteria, such as AIC (Burnham and Anderson 1998) are used to define the appropriate model.

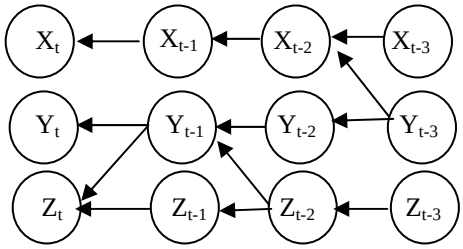

Figure 5. A graphical modelling for time series where all arrows are in the direct of time *t*.

The system of equations for the graphical model shown in figure 5 would be:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_4 Y_{t-3}$$
$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Z_{t-2} + \beta_4 Y_{t-3}$$
$$Z_t = \delta_1 Z_{t-1} + \delta_2 Y_{t-1} + \delta_3 Z_{t-2} + \delta_4 Z_{t-3}$$

Selecting the appropriate DAG with directional links places graphical models in the realm of addressing causality (Granger 1988, Pearl 2000). The system of equations used to describe the figure 2 DAG are appropriate for describing causality, even if the second equation is redundant mathematically.

3. CASE STUDY: MICE AND BEECH SEEDING

Cyclical waves in size of ecological populations have been observed in many habitats. Probably the best known is the lynx and snow shoe hare where peaks in one species' population size are followed by peaks for the other species. As the snow shoe hare population increases the predator species, the lynx, builds up in size. The hare population crashes followed by the lynx population crashing as their food declines. In the absence of predators the hare population rises, and so on.

A similar system has been observed in New Zealand with cyclical rises in the mice population size roughly corresponding to a period after a heavy beech tree seeding year. These years of heavy seeding are called beech-mast years. One explanation for this cycle is that beech seed masting causing an increase in mouse population, whereas an alternative view is that the increased flowering and seed fall cause an increase in invertebrate and it is this increase in invertebrate population size that causes the mice population size increases. The question is therefore related to causality – what causes what?

3.1. Data collection

Data on mice numbers, beech seed fall and mouse breeding were collected from the Orongorongo valley near Wellington between August 1971 and November 1996 by staff from the former DSIR and more recently, Landcare Research (M. Fitzgerald and B. Karl).

Mice population size was estimated from counts in mice traps adjusted for trapping-effort:

$$N_t = \frac{-1000 a_2}{(a_1 + a_2)} \log \frac{a_0}{T}$$

Where a_0 is the number of traps not sprung, a_1 is the number of traps sprung but empty, or containing another species and hence not available to catch mice, and a_2 is the number of traps with a mouse, and $T = a_0 + a_1 + a_2$ (Fitzgerald et al. 2004).

Beech seed fall was measured using seed traps under mature trees. The size of the traps was $0.28m^2$. Traps were cleared monthly but data was available only from annual records.

Mouse breeding was measured by the proportion of mice caught in traps that were pregnant females and the proportion of adult males in the population. Adulthood was assessed by toothwear and any mouse over 4 months old was considered an adult. All trapped mice were dissected and pregnancy assessed by the presence of either live or reabsorbing embryos in utero.

These data represented measurements from a system with strong seasonal differences both from intrinsic ecological reasons, and for other response variables, because data were only collected in some seasons. Mice were trapped four times a year (February, May, August and November) giving four estimates of mice numbers per year, roughly one estimate per season. Seed fall occurred in February through to May but only one estimate per year was available which was considered to be from the summer season. Mouse breeding occurs in both spring and summer, but is absent in the other two season. This strong seasonality can be a challenge for modelling.

Collecting ecological is very expensive and there are very few examples in New Zealand of long term studies were data has been collected with some consistency over more than 5 years. Even with longer term studies, the data collection techniques do evolve resulting in a subtle shift in what the sample estimate is measuring. In this study the there was a noticeable change in the data from mice traps after 1993 coinciding with a change in the field team personal. The difference was so large we used data from only up to 1993.

Because of the expense in collecting ecological data surveys tend to have multiple objectives and this can compromise the quality of data. In this mouse study estimating mice population size was only one objective and the need for quality and consistent data on mice numbers, beech seeding and mice breeding would have been competing with the need to collect information on other ecological variables. Over time the survey objectives change and allocation of effort and data collection methods evolve. Comparison of data

trends over a long time span can be confounded by changes in data collection protocols. It will take many more years of study to address the competing hypothesis that mice numbers are directly relate to invertebrates rather than beech seeding because invertebrate data was not collected in the past – there so no compelling reason at the time to collect it.

Analysis of ecological data is also constrained in that ecological events can be infrequent. Beech masting occurs once every 4 – 7 years. In this study despite the considerable effort undertaken to collect this data over years beech masting, occurred only six times.

4. METHODS

Initial modelling used a simple GMTS approach with the seasonal variables X_t = mouse numbers, Y_t = seed fall and Z_t = mouse breeding . However the results were less than satisfactory. There were data from seasonal counts of mouse numbers, whereas for the seed fall there was only annual data. One approach to deal with the differences in temporal scale of the time series is to create synthetic values for seasonal series based on the annual responses. Data on annual seed fall for example could be used to model seasonal seed fall. However to be ecologically realistic the seasonal values for seed fall were set as 0 for seasons other than summer. An alternative would have been to model the data on an annual scale. The obvious problem here is that information on seasonal variation, where available, is lost.

For mouse breeding data were available only on an annual scale, but breeding occurs over two seasons. For this time series the annual breeding value was split between the two seasons, spring and summer. A random value from a uniform distribution with mean 0.5 and range (0.1) was used to estimate *p*, the proportion of the annual breeding that occurred in spring. The proportion of the breeding that occurred in summer was estimated by 1-*p*.

To ensure the modelling was ecologically realistic seasonality was preserves in the graphical modelling. Separate graphical models were produced, one for each leading season: a model with data from annual spring as the most recent time, a model with autumn as the most recent time, and so on.

5. RESULTS

The seasonal-graphical models are informative in understanding the relationship among variables (figure 5). In all three seasons modelled the size of the mouse population was related to the size of the population in the previous season. The effect of summer seed fall was only evident in the following winters' mouse numbers and there was no carryover effect to the following spring. The size of the winter mouse population had an effect on the summer's breeding population.

In summary, summer mouse numbers were dependent on the size of the population in the spring, while summer breeding was dependent on the previous winter's population size. And springs' mouse numbers were a function of the mice numbers in the winter before. The autumn model provided no useful information.

All these relationships were positive, i.e., with an increase in mouse numbers in winter, mouse numbers in spring will increase. With an increase in summer seed fall there is a corresponding increase in mouse numbers in winter.

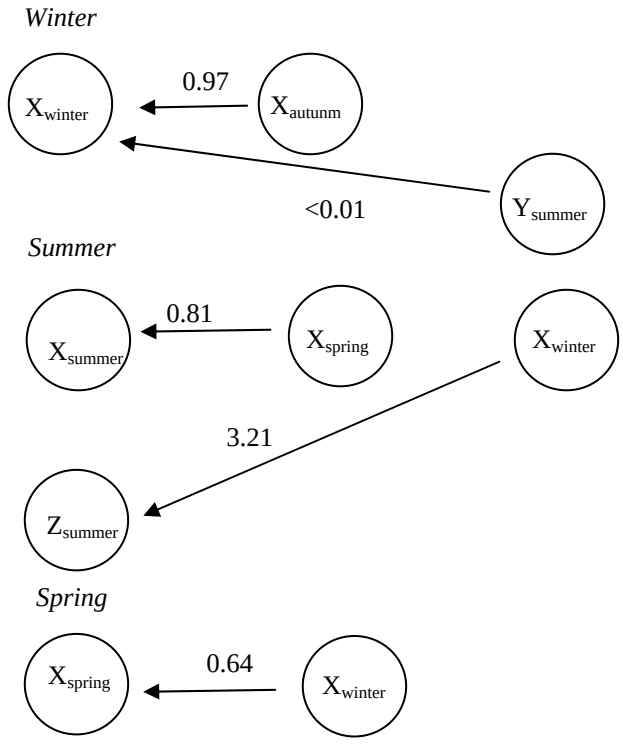

Figure 5 Seasonal graphical models, X_t = mouse numbers, Y_t = seed fall and Z_t = mouse breeding. The numbers on the arrows are parameter values in the linear model

Perhaps the most interesting features are the graphs are the missing links indicating lack of evidence of a relationship. Mouse breeding was not related to previous seed fall, although mouse numbers were. Only in winter were mouse

numbers related to breeding (in summer). The only relationship with summer seed fall was with mouse numbers in winter. Increased mouse numbers in winter was associated with increased breeding in the next summer, but the link between summer seed fall and summer breeding, was not direct.

6. DISCUSSION

Graphical modelling of time series data is a useful tool for biologists. The complexity of ecosystem interactions can be described by modelling the multivariate system with graphical links for the changing interactions through time.

In this case study example, results were constrained by the size of the dataset and consistency of the data collection procedure (see earlier discussion in 3.1). Our seasonal models were derived from data there were strictly not collected seasonally and we had to derive seasonal mouse breeding estimates. The analysis was limited to pre-1993 data because of changes in survey protocol. While we did create three seasonal models, the results were not outstanding. Perhaps what is most relevant is that these data were from one of New Zealand's longest term ecological studies and it is a reminder of the high cost of environmental data collection. Rather than dismissing the dataset as being too small, or inferior in some way, appropriate analysis methods need to be developed.

The primary advantage of using graphical models for ecology is the visual display of the ecosystem relationships. The identification of the temporal variables, and the resultant graph can be a very informative data-display. The process of creating the CIG and the DAG and the resultant graphical model quantifies the variable relationships. Inference to causal relationships is a productive step for conservation management. For example, quantifying the time lag between an observed increase in seed fall and in mouse numbers would allow managers to target the optimal timing for a mouse-control operation. Identifying whether the link between seed fall and mouse numbers is direct and causal would allow managers to measure appropriate environmental indicators of potential mouse population size increases, well before the increase occurred.

In this case study there were insufficient data to directly model the effect of inter-annual variation in seed fall and the specific effect of beech mast years. With more data, variation in graphical models in the periods before, and after a perturbation (such as beech mast year) could be used to measure the effect of the change. Questions on the temporal "effect" of mast years can be addressed by measuring how long it takes for the system to return to pre-perturbation state. The graphical modelling for time series framework is adaptable to assist in analysis of a wide range of environmental research studies.

7. ACKNOWLEDGEMENTS

We are very grateful for the use of the data kindly supplied by Mike Fitzgerald, who alone with Brian Karl was responsible for collection for over 20 years.

8. REFERENCES

Burnham, K and D. Anderson (1998), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition, Springer-Verlag.

Fitzgerald, B., M. Efford, M. and B. Karl (2004), Breeding of house mice and the mast seeding of Southern beeches in the Orangorongo Valley, New Zealand, *New Zealand Journal of Zoology* 31(2), 167-184.

Granger, C.W.J. (1988), Some recent developments in a concept of causality *Journal of Econometrics* 39, 199-211.

Greig-Smith, P., (1983), *Quantitative plant Ecology*, 3rd edition, Studies in Ecology vol 9, Blackwell Scientific Publications, Oxford.

Lauritzen, S.L. and D. J. Spiegelhalter, (1988), Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal Royal Statistical Society B* 50, 157-224.

Pearl, J., (2000), *Causality*. Cambridge University Press, New York.

Reale, M. and G. Tunnicliffe Wilson (2002), The sampling properties of conditional independence graphs for structural vector autoregressions, *Biometrika*, 89, 457-61.

Reale, M. and G. Tunnicliffe Wilson (2001), Identification of vector AR models with recursive structural errors using conditional independence graphs, *Statistical Methods and Applications* 10, 49–65.