Using graphical modelling in official statistics

Richard N. Penny[†] **and Marco Reale**[*]

[*] *Department of Mathematics & Statistics,*
University of Canterbury,
Private Bag 4800, Christchurch, New Zealand.

Report Number: UCDMS2004/10 May 2004

Keywords: Early estimates, Irregulars structures Moralization, Structural vector autoregressions

[†] *Fidelio Consultancy, Christchurch, New Zealand.*

Using graphical modelling in official statistics.

Richard N. Penny* and Marco Reale†

May 11, 2004

Abstract

People using economic time series would like them to be available as soon as possible after the end of the reference period. However there can be difficulties in getting all the responses required to produce a series of acceptable quality in a timely manner. The earlier the time series is released the more likely there will be tardy respondents, thus the series will have to be estimated without their responses. As QGDP is the aggregation of a large number of economic time series the difficulties are compounded.

An adequate preliminary estimate of QGDP may be made by using models parsimonious in the number of time series involved. Graphical models assist us in obtaining such parsimonious models by identifying the relevant components in a saturated structural VAR enabling us to eliminate unnecessary delays. Even if an earlier release is not possible we could target work to improve the timeliness of series identified in the parsimonious models.

KEY WORDS: Early estimates, Irregulars structures Moralization, Structural vector autoregressions

*Fidelio Consultancy, Christchurch, New Zealand.

†Department of Mathematics and Statistics, University of Canterbury, PB 4800 Christchurch, New Zealand

1. Introduction

A National Statistical Office (NSO) produces a large number of time series which are updated on regular basis. Some are estimates from surveys run by the NSO, some from data collected by another organization for their administrative purposes (e.g. customs records), and others are combinations of a range of time series (e.g. Consumer Price Index (CPI), Quarterly Gross Domestic Product (QGDP)).

It is required to produce these statistical outputs to published quality standards and in a timely manner. To some extent these criteria are related so that care must be taken that improving one aspect of the data (e.g. timeliness) does not impact on another (e.g. quality).

One key technical advance would be the ability to easily identify significant relationships between time series. Each time series has particular issues where a knowledge of the relation, or otherwise, between different time series would assist NSOs in assessing the benefits of changes in the way particular time series are compiled, particularly for those outputs, such as QGDP, that are combinations of a range of time series.

We have been investigating the feasibility of using graphical modelling to identify and model the relationships between time series, particularly to identify where improvements in timeliness could be made without materially affecting quality, or increasing cost.

2. Components of time series

As outlined above, the main concern of NSOs is to release data that reflect the social or economic concept that they are meant to represent, within the budget allocated for this work and, crucially, with little or no revisions after release. Much of the reporting on NSO outputs focusses on the movement represented by the new data point, that is the first difference, rather than its absolute value.

For any series that is seasonal a large part of any first difference is caused by movements in the seasonal component. Hence the desire for NSOs to seasonally adjust where appropriate. For this reason most statistical agencies provide the measured figure, along with the seasonally adjusted value (where appropriate) and, increasingly, the trend estimates, and direct users to the latter series rather than the unadjusted figures.

To provide seasonally adjusted and trend estimates Statistics New Zealand currently uses Census Method II Variant X-12, commonly called X-12 (Findley et al. 1998). For seasonally adjusted or non-seasonal series, work done by Statistics New Zealand (Kazakova 2002) shows that movements over short time spans will be dominated by the movement in the irregular component. As one of the key conditions for seasonal adjustment is that the seasonal pattern is stable we assume the seasonal factors are not changing significantly over a short time span. What is of interest to many uses are turning points in the trend estimate. It is well known that estimating trend components at the end of a series is difficult, with identification of turning points being particularly problematic. Often evidence of a turning point will appear first in atypical behavior of the irregular component. Therefore it is important that the estimate of the irregular component is not substantially revised. By definition estimating the the next value for all the components bar the irregular should be

done well. Therefore any attempt to find a more parsimonious model for any time series should ensure that the irregular component is consistent with that from the more complex model. For this reason we have focussed our attention on estimation of the irregular component. A byproduct of investigating the irregular component is that it is stationary.

3. Models for multivariate time series

The relation among several autoregressions can be modelled with the vector autoregression

$$x_t = c + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \ldots + \Phi_k x_{t-k} + e_t \tag{1}$$

of order k, VAR(k), where $x_t, x_{t-1}, \ldots, x_{t-k}$ are n-dimensional vectors, with the corresponding coefficient vectors $\Phi_1, \Phi_2, \ldots, \Phi_k$, c is the constant and e_t is the error vector (which is assumed iid). If the covariance matrix, H, of e_t is not diagonal, the set of linear equations (1) corresponds to a system of seemingly unrelated regressions (Zellner 1962) and in H are hidden the relations among the components of x_t. To highlight such relations we can represent the canonical VAR(k) in (1) in its structural form (SVAR) (Bernanke 1986, Blanchard and Watson 1986 and Sims 1986):

$$\Theta_0 x_t = d + \Theta_1 x_{t-1} + \Theta_2 x_{t-2} + \ldots + \Theta_k x_{t-k} + u_t \tag{2}$$

where $\Theta_i = \Theta_0 \Phi_i$ for $i = 1, \ldots, k$, $d = \Theta_0 c$ and $u_t = \Theta_0 e_t$ with covariance matrix $\Theta_0 H \Theta_0' = D$, which is diagonal.

If there are no zeros in the coefficient vectors the SVAR is saturated, but in many cases some lagged variables on the RHS in (2) do not play any role in explaining the current variables, x_t. In this case the value of the corresponding coefficient is zero and hence the SVAR is sparse.

The order p of the regression may be determined by various methods including inspection of a multivariate partial autocorrelation sequence, see (Reinsel 1993, pp 69-70), or minimization of an order selection criterion such as AIC (Akaike 1973), HIC (Hannan and Quinn 1979), SIC (Schwarz 1978), corrected AIC (Hurvitch and Tsai 1989). The latter is particularly advisable given its small sample properties and it is the one we use in this paper.

We require a further condition on Θ_0, that it represents a recursive dependence of each component of x_t on other contemporaneous components. This is equivalent to the existence of a re-ordering of the elements of x_t such that Θ_0 is triangular with unit diagonal. Each possible ordering of x_t therefore gives a potentially distinct form of (2), but these are all statistically equivalent, corresponding to different factorizations of

$$D = \Theta_0 H \Theta_0'. \tag{3}$$

This inverse problem contrasts with the unique form of (1), which is an attractive feature of that model from a time-series modelling viewpoint. The value of (2) therefore lies in the possibility that there is one particular form which, as a consequence of its representing the data generating process, is more parsimonious in its parameterization than either (1)

or the other forms of (2). This would be reflected in the ability to exclude many of the elements of Θ_0 and Θ_i from the model without penalizing the adequacy in comparison with the saturated or other forms of either (1) or (2).

4. Graphical models for time series

Neglecting, for the present treatment, any effects of time series model estimation, we suppose that we have observations on the vector Gaussian white noise innovations process e_t with the usual sample covariance matrix $\hat{H}$. We wish to determine *from the data* the form of possible sparse structural matrices Θ_0 which are compatible with $\hat{V}$. There may be no such unique form without imposing further constraints using insight from the modelling context.

Tunnicliffe Wilson (1992) and Swanson and Granger (1997) consider a similar problem. The latter authors focus more on testing for the constraints implied by a particular structural form of Φ_0 which has commonly occurred in practice. Their tests are expressed in terms of partial autocorrelations which, as they remark, are not directional and would therefore appear less appropriate for recursive models.

We follow the approach proposed by Reale and Tunnicliffe Wilson (2001) and use pairwise partial autocorrelations, conditioning on all remaining variables (i.e. components of e_t). With respect to backward stepwise regression this approach has the advantage of leaving the conditioning set unchanged. Nevertheless there is a problem of multiple testing to deal with and later we'll describe a strategy to tackle this issue.

The partial correlations, given the Gaussianity, are used to construct the conditional independence graph (CIG) of the variables, following procedures presented, for example, in Whittaker (1990). As Swanson and Granger (1997, p. 359) also remark, the structural form of dependence between the variables is naturally expressed by (and is equivalent to) a directed acyclic graph (DAG), in which nodes representing variables are linked with arrows indicating the direction of any dependence. A DAG implies a single CIG for the variables, but the possible DAGs which might explain a particular CIG may be several or none. The point is that, subject to sampling variability, the CIG is a constructible quantity and a useful one for expressing the data determined constraints on permissible DAG interpretations.

The CIG consists of nodes representing the variables, two nodes being *without* a link if and only if they are independent conditional upon *all* the remaining variables. In a Gaussian context this conditional independence is indicated by a zero partial autocorrelation:

$$\rho\left(e_{it}, e_{jt} | \{e_{kt}, k \neq i, j\}\right) = 0. \tag{4}$$

In the linear least squares context the linear partial autocorrelations in (4) still usefully indicate lack of linear predictability of one variable by another given the inclusion of all remaining variables. The link with Granger causality is quite evident. The set of all such partial correlations required to construct the CIG is conveniently calculated by making use of the *inverse variance lemma* (Whittaker, 1990) as

$$\rho\left(e_{it}, e_{jt} | \{e_{kt}, k \neq i, j\}\right) = -W_{ij}/\sqrt{(W_{ii}W_{jj})} \tag{5}$$

where $W = H^{-1}$. The sample values are obtained by substituting the sample value $\hat{H}$ for H.

We then test their significance using thresholds obtained by exploiting the relationship between a regression t value and the sample partial correlation $\hat{\rho}$ given by

$$\hat{\rho} = t/\sqrt{(t2 + \nu)} \tag{6}$$

(see Greene, 1993, p. 180), where ν is the residual degrees of freedom in the regression of one of the variables in the partial autocorrelation, upon all the other variables. The t value is that attached, in this regression, to the other variable in the partial autocorrelation. This is a relationship deriving from the linear algebra of least squares, and is not reliant upon statistical assumptions. Standard assumptions *are* needed to support the usual distribution of t under the null hypothesis that the true value of the relevant variable is zero, which is equivalent to $\rho = 0$. There are of course statistical pitfalls in applying the test simultaneously to all sample partial autocorrelations.

Our approach is to use these values to suggest possible models, and after fitting these, we apply more formal tests and diagnostic checks to converge on an acceptable model.

Central to the interpretation of a CIG is the separation theorem. The CIG is constructed by pairwise separation of variables which are independent conditional on the remainder. The separation theorem states that if two *blocks* of variables are separated, that is there is no link between any member of the first block and any member of the second, then the two blocks are completely independent conditional on the remaining variables. See, for example, Whittaker (1990, pp. 64-67) for a general proof and references to more straightforward proofs in the Gaussian case, where the result can be read directly from the joint density.

To illustrate the theorem, let us consider the conditional independence graph in Figure 1, where A, B and C are either random variables or groups of random variables. The graph implies that $A \perp\!\!\!\perp C|B$ or alternatively that $A|B, C = A|B$. While the CIG leaves

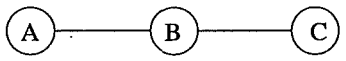

Figure 1: *Markov property for conditional independence graphs.*

room for several possible alternative factorizations of the joint density function, the DAG provides a more precise definition. As an example let us consider the DAG in Figure 2; it is very similar to the CIG in Figure 1 where the lines, also called *edges*, are replaced by arrows. Nevertheless the definition in terms of density is now very precise, $f(A, B, C) = f(A|B)f(B|C)f(A)$. Using the graphical modelling terminology we would say, in this case, that B is a *parent* of A and C is a parent of B. Although the DAG and the CIG represent a different definition of the joint probability, there is a correspondence between these two graphs which is embodied by the *moralization* rule (Lauritzen and Spiegelhalter 1988). Because of this result we can obtain the CIG from the DAG by transforming the arrows into lines and linking unlinked parents. These kinds of edges are defined as *moral.*

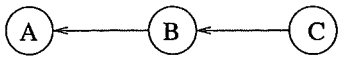

Figure 2: *Density factorization implied by a DAG.*

To better explain moralization let us consider a simplified example: obtaining the New Zealand residency. You can become NZ resident (C) for two reasons: general skills (A) or business reasons (B), which basically means having money to invest in New Zealand. We can effectively represent this system with the graph on the left hand side of Figure 3 where both A and B affect C: A and B are parents of C. Assuming no relationship between being rich and being skilful (many real cases would support this assumption) there is no link between A and B. The DAG clearly provides a precise description of the pairwise independence relations. The CIG on the other hand provides a description of more global independence relations, considering the effect of all the variables present in the graph. If the joint distribution of the variables in the graph is not Gaussian, relations should be interpreted in terms of partial correlation. In our example although we assumed no direct dependence between money and skills, the joint consideration of the third variable, the obtainment of the New Zealand residency, would change the situation. In fact, information, for a certain applicant, about the level of skills and the outcome of the application can be revealing about the business capability. The resulting CIG is shown on the right hand side of Figure 3.

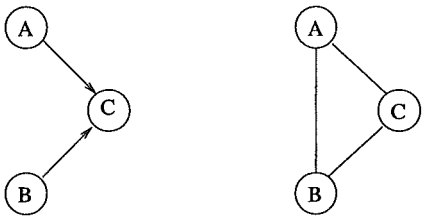

Figure 3: *Moralization of a directed acyclic graph.*

Moralization brings us to the next step which is to determine what DAG structures can explain a CIG. This is part of a much wider problem of the search for causal structure, covered for example by Spirtes, Glymour and Scheines (2000).

The DAG is very attractive because of its causal interpretation (Pearl 2000), but all we can observe in practice is the CIG obtained by the sample partial correlation. So actually we need to perform the inverse operation of the moralization, which we term *demoralization.* Unfortunately while the transformation of a DAG into a CIG is unique, there are several DAG's which can give the same CIG. As an example, consider the CIG on the right end side in Figure 3: it could result from the moralization of all the DAG's in Figure 4. So we need to identify the moral links and remove them. To do that we need to use all the knowledge we have about the relationships among the random variables in the system. As we shall see in the application in the next section, the search for the DAG is

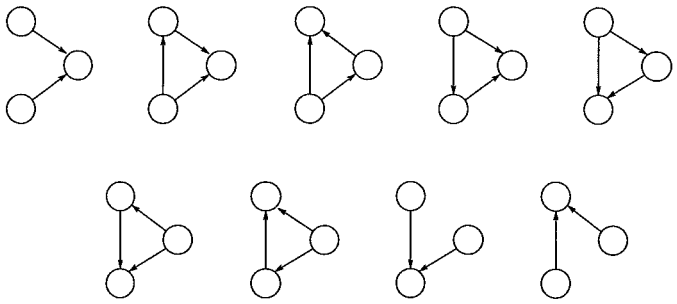

Figure 4: *Possible directed acyclic graph.*

simplified when we are in a structural VAR framework.

5. A graphical model for the quarterly gross national expenditure in New Zealand

QGDP is one of the most relevant time series for economic and social analyses and there is more and more pressure to release early reliable estimates for this aggregate. In our analysis we consider a subcomponent of QGDP but the same methodology could be extended to all the subcomponents and eventually we could possibly consider the main time series for each subcomponent jointly.

We have QGDP on an expenditure basis (QGDE) from the June quarter of 1988 to the December quarter of 2002, a series of 65 values. As QGDE is Quarterly Gross National Expenditure plus exports minus imports we have investigated the relationship between the top level components of QGNE only for our preliminary investigations.

QGNE is equal to the sum of Private Final Consumption Expenditure, Government Final Consumption Expenditure, Gross Fixed Capital Formation and Stock changes. That is

$$QGNE = PFCE + GFCE + GFKF + STOCK \tag{7}$$

Note that STOCK can have negative and positive values. We only model the irregular components produced by Statistics New Zealand's seasonal adjustment procedure. In order to provide earlier reliable estimates of the QGNE, our strategy is to focus on the most volatile component and rely on the stability of the others. The irregular components of PFCE, GFCE, GFKF and STOCK are plotted in Figures 5 and 6.

These time series are stationary by definition as confirmed by inspection. Nevertheless the methodology we are going to use can be applied to systems integrated of the first order without any concern for cointegration as proved by Tunnicliffe Wilson and Reale (2002).

Using the corrected AIC we identified a vector autoregression of order 4. We then proceeded with the methodology explained above to obtain the conditional independence graph in figure 7. We first calculated the sample partial correlation by using the inverse variance lemma (5) and then tested their significance by using (6).

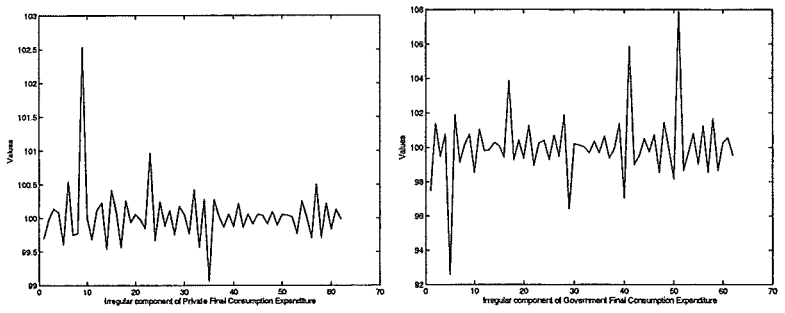

Figure 5: Irregular components of PFCE and GFCE.

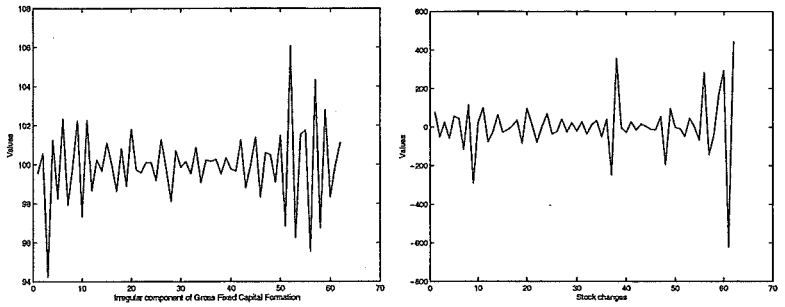

Figure 6: Irregular components of GFKF and STOCK.

In using this testing procedure we have to deal with the issue of multiple testing. A strategy to try to minimize type I and type II errors would be to use different levels of significance of partial correlations. This information combined with cross-correlations of residuals, prior information and moralization consistency will assist in selecting a specific DAG.

Looking at the independence graph (Figure 7) one of the advantages of graphical modelling is immediately obvious. It is considerably easier to see the intricacies of the relationship between different series at differing orders of lag.

Note that in the CIG we represent only the relations with current variables, excluding the relations between past variables. This is because it is the current relations we are interested in. Nevertheless relationships between past variables can sometimes be of help; their use and sampling properties have been studied by Reale and Tunnicliffe Wilson (2002).

From Figure 7 general higher level patterns can be clearly identified. It can be seen that there is a web of relationships between GFCE and GFKF at various lags, most at 0.99 significance. There are only two links connecting this group. Both are to PFCE, one from current GFKF and the other from lag 4 GFCE. It can also be seen that PFCE is linked to STOCKS. While this CIG is useful, for official statistics purposes we need also some indication of a causal structure.

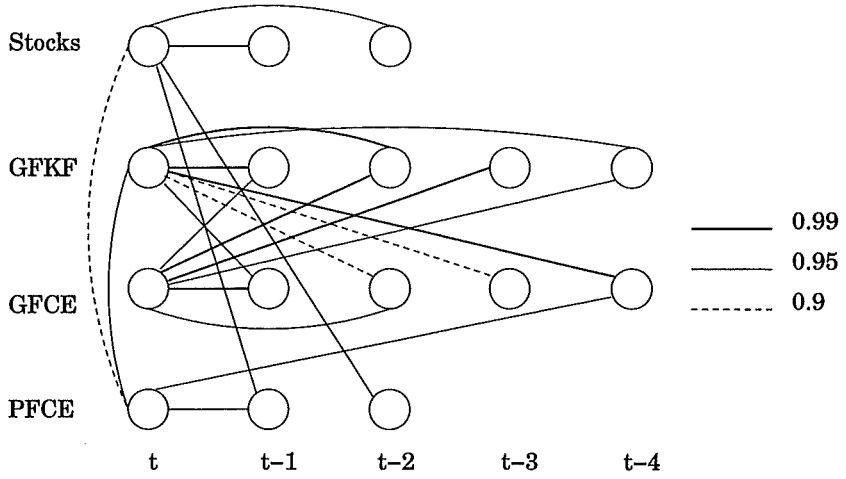

Figure 7: Conditional independence graph.

In order to identify a causal structure among the irregulars we need to identify the DAG and hence the direction of the edges among contemporaneous variables, the direction of the other edges being obvious given the time framework. Therefore we need to determine the causal structure for the contemporaneous relationship between STOCKS, PFCE and GFKF. Using moralization there are three possible directed structures among contemporaneous variables. They are presented in figure 8.

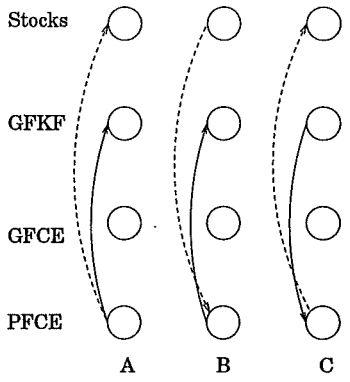

Figure 8: Possible structures among contemporaneous variables.

Because of the knowledge we have of the system we can exclude structure B. We then proceed with subset selection and use information criteria, in particular the one proposed by Schwarz (1978), to select the best models for contemporaneous structures A and C (figures 9 and 10). According to both the Schwarz and Hannan and Quinn criteria the model in figure 9 provides a better representation of the data.

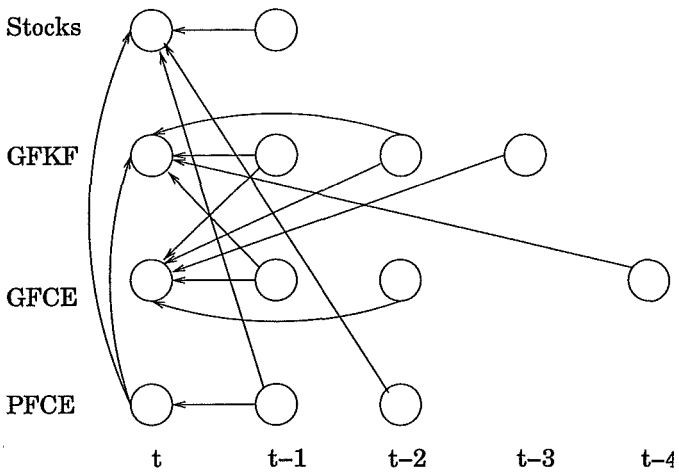

Figure 9: Best model.

The following table provides the number of parameters, deviance, AIC, HIC and SIC for the saturated model, best model and alternative model.

Model	k	DEV	AIC	HIC	SIC
Saturated	70	416.47	556.47	612.65	700.70
Best	14	541.36	569.36	580.59	598.20
Alternative	20	555.12	595.12	611.17	636.33

At this point we now have a model that may be useful for official statistical purposes. We see that GFKF and GFCE are linked to past values of both, so both would be required for total QGNE. Current STOCKS and GFKF are related to current PFCE, which is an AR(1) process. So there is some evidence that a good current value for PFCE is not necessary to produce current QGNE, but rather its information is already contained in the other variables in the graph. We would still need to eventually have a good value for PFCE in order to use it in the PFCE AR(1) model for the next quarter's estimate of QGNE.

However any decision as to when to release QGNE would require further investigation of the subcomponents of the series that we have used, plus using information on the time the various series are available for use. Also we would need to analyse the early estimates from any proposed model with the final estimates as given by Statistics NZ.

6. Conclusions

Graphical modelling has been developed to help draw population inferences. While NSOs produce models as part of their outputs the primary task of an NSO is to produce a broad range of timely quality data for use by society. To this end it would be useful to identify the relationships between time series to identify those that are crucial to the release of an

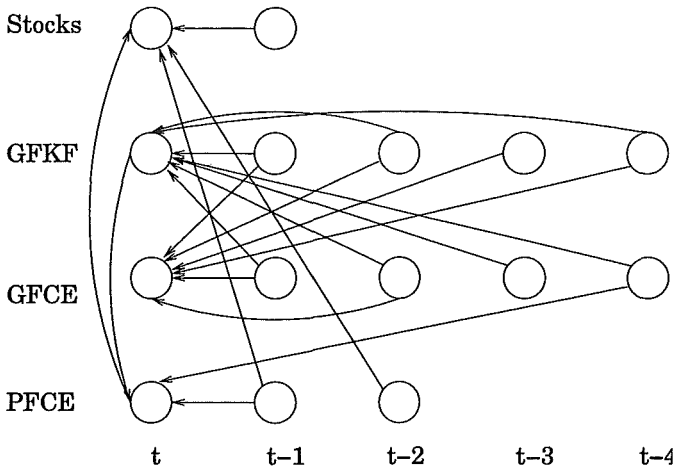

Figure 10: Alternative model.

acceptable first estimate. These crucial series could merit work to improve their timeliness, whereas less important series could be either not collected or have less resources applied to their collection.

Our preliminary work shows graphical modelling has potential, but as a NSO often approaches time series analysis with different purposes than straight prediction more work will be required to identify under what conditions and in what areas it will be most useful.

An extension of this approach for instance could be devised by including all the components of the time series. Graphical modelling could also be succesfully applied in reducing the number of time series collected by eliminating time series giving information already given by others.

The large quantity of data available to a NSO offers applications to data mining, a field where graphical modelling is successful (Borgelt and Kruse 2002).

References

Akaike H. (1973), A new look at Statistical Model Identification, *IEEE Transactions on Automatic Control*, AC-19, 716-723..

Anderson T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, New York.

Bernanke B.S. (1986), Alternative explanations of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy* 25, 49-99.

Blanchard O.J., Watson M.W. (1986), Are business cycles all alike?, *The American Business Cycle: Continuity and Change*, (Gordon R.J. ed.), University of Chicago Press, Chicago.

Borgelt C., Kruse R. (2002), *Graphical Models: Methods for Data Analysis and Mining*, Wiley, New York.

Findley D.F., Monsell B.C., Bell W.R., Otto M.C., Chen B.-C. (1998), New capabilities and methods of the X-12 ARIMA seasonal adjustment program, *Journal of Business and Economic Statistics*, 16, 127-177.

Greene W.H. (1993), *Econometric Analysis*, Prentice-Hall, Englewood Cliffs.

Hannan E.J., Quinn B.G. (1979), The determination of the order of an autoregression, *Journal of the Royal Statistical Society Series B*, 41, 190-195.

Hurvitch C.M, Tsai C-L. (1989), Regression and time series model selection in small samples, *Biometrika*, 76, 297-307.

Kazakova V. (2001), What dominates movements in SNZ seasonally adjusted time series, Internal Report, Statistics New Zealand.

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

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

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

Reale M., Tunnicliffe Wilson G. (2002), The sampling properties of conditional graphs for structural vector autoregressions, *Biometrika*, 89, 457-461.

Reinsel G.C. (1993), *Elements of Multivariate Time Series Analysis*, Springer-Verlag, New York.

Schwarz G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464.

Sims C.A. (1986), Are forecasting models usable for policy analysis?, *Federal Reserve Bank of Minneapolis Quarterly Review*, 10, 2-16.

Spirtes P., Glymour C., Scheines R. (2000), *Causation, Prediction and Search*, MIT University Press, Cambridge.

Swanson N.R., Granger C.W.J. (1997), Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions, *Journal of the American Statistical Association*, 92, 357-367.

Tunnicliffe Wilson G. (1992), Structural models for structural change, *Quaderni di Statistica e Econometria*, 14, 63-77.

Tunnicliffe Wilson G., Reale M. (2002), Causal diagrams for I(1) structural VAR models, *University of Canterbury Department of Mathematics and Statistics Research Reports*, 2002/6.

Whittaker J.C. (1990), *Graphical Models in Applied Multivariate Statistics*, Chichester, Wiley.

Zellner A. (1962), An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association*, 57, 348-368.