Creating High Frequency National Accounts with State Space Modelling: A Monte Carlo Experiment

By

H.Liu
(Imperial College)

and

S.G.Hall
(Imperial College)

ABSTRACT

This paper assesses a new technique for producing high frequency data from lower frequency measurements subject to the full set of identities within the data all holding. The technique is assessed through a set of monte carlo experiments. The example used here is gross domestic product (GDP) which is observed at quarterly intervals in the United States and it is a flow economic variable rather than a stock. The problem of constructing an unobserved monthly GDP variable can be handled using state space modelling. The solution of the problem lies in finding a suitable state space representation. A Monte Carlo experiment is conducted to illustrate this concept and to identify which variant of the model gives the best monthly estimates. The results demonstrate that the more simple models do almost as well as more complex ones and hence there may be little gain in return for the extra work of using a complex model

February 2000

JEL classification E21, E22, E23

Keywords; National accounts, Kalman Filter, Balancing, High Frequency

1.1 Introduction

The motivation of this study lies in the usefulness of high frequency data. The advantage of monthly GDP for example is apparent as it helps the government to examine the present state of the economy so as to respond more quickly to economic events. This has important impacts at the national level mainly because more timely and accurate policy can be decided and formulated according to the current economic circumstance. It also extends the government ability to assess the present economic environment and to conduct economy policy more efficiently.

Constructing high frequency data from lower frequency has many practical uses. A real world example is that GDP in Eastern Europe is only observed on an annual basis. It is much more sensible and useful for them to be constructed at a higher frequency, say, quarterly. Ideally this should be done not just for GDP but also for the complete national accounts. The benefit of having such national accounts immediately becomes apparent both for policy purposes and for making comparison with other nations.[1] The usefulness can also be extended into solving the problem of using mixed-frequency data. Efficient use of mixed-frequency data is a common problem in both financial and economic time series. It is clearly a loss of information to work only at the lowest common frequency.

Problems of using mixed-frequency data can be solved more effectively if the low frequency variable is constructed at the same interval as the high frequency variable. The state space formulation, proposed here, is a flexible way of handling complications involved in using mixed-frequency data and dealing with the issue of temporal aggregation of data. It allows unobserved components to be incorporated into a modelling process, and the Kalman filter and smoothing algorithms provide the means of estimating them.

The balancing method advocated by Stone and his colleague in the 1940s laid an important foundation in national accounting.[2] But this technique does not allow the construction of higher frequency data. The most widely used approach for deriving high frequency data is the conventional regression approach (see, for example, Chow and Lin (1971) (1976) and, more recently, Salazar, Smith and Weale (1997)), but this leaves the high frequency data unbalanced. The technique proposed here achieves both of these aims simultaneously by

[1] It is similar to the construction of the UN System of National Accounts (SNA) and a clear and concise description on the design and structure of the system of SNA 1993 can be found in Carson (1996).

means of Kalman filter and smoothing algorithms. It is also able to ensure that the high frequency data aggregates up to exactly match the low frequency data and uses any available information at the high frequency in constructing the new series.

This study begins with the introduction of state space modelling in the context of deriving unobserved high frequency estimates. Two different types of approaches are discussed in this study. A univariate approach is firstly considered and two different forms of the state space models are developed to estimate monthly GDP. A multivariate approach with the emphasis on linking available monthly data with quarterly GDP to derive monthly estimates of GDP is also discussed. The next section sets out the Monte Carlo experiment in this study as it is used to identify which form of the state space models gives the best monthly GDP estimates. Results of Monte Carlo experiments of the univariate and multivariate models are shown in the next section and conclusions are presented in the last.

1.2 The State Space Approach

The full state space form and Kalman filter prediction and updating equations are presented in appendix 1. The state space approach have been applied to economic data by many authors, including Conrad and Corrado (1979), Howrey (1984), de Jong (1987), Scadding (1988), Zadrozny (1990), Coomes (1992) and Patterson (1995). They have focused on the issue of data revisions and the existence of measurement errors in the provisional economic estimates and finding a better tool to improve the efficiency in handling provisional economic time series as well as to deriving optimal forecast for the initial estimates. These studies have used Kalman filters successfully to attack the analogous problem at the national level. The idea of using the state space framework to construct an optimal estimate of economic indicators has been adopted by Stock and Watson (1989), Stock and Watson (1991) and Garratt and Hall (1996).

Corrado and Greene (1988) demonstrated the applicability of Kalman filtering to improve the predicative accuracy of a quarterly model. Corrado and Haltmaier (1988) described an overview of the underlying structure of the model linkage project at the Federal Reserve Board. The state space approach was utilised to make analogous interpolations so as to incorporate high frequency information on US economic activity into monthly GNP forecast.

[2] Stone, Champernowne and Meade (1942) "The precision of national income estimates".

Jones (1980), Kohn and Ansley (1983), Harvey and Pierse (1984), Ansley and Kohn (1985), Kohn and Ansley (1986), and G.∴mez and Maravall (1994) look at the issues of how the Kalman filter can be applied to deal with missing observations in economic time series.

2 State Space Forms for Generating High Frequency Data

This section discusses two different means of generating high frequency data - the univariate approach and the multivariate approach. Various state space forms associated with each approach are formulated.

2.1 Univariate Model

Quarterly GDP is an observable series and the unobservable monthly GDP variable can be filtered out giving the observable quarterly GDP. The concept for building a univariate model is to derive the estimate of unobservable high-frequency, monthly, estimates of GDP using its own low-frequency, quarterly, GDP.

2.1.1 Univariate Model I

This is the most straightforward of those considered. The construction of this model is based on the notion that observable quarterly GDP is the average of unobservable monthly GDP.[3] The measured variable is defined to be a monthly series of quarterly GDP figures such that Y_t is constant for each monthly of the quarter. S_t is the unobserved monthly GDP figure that varies continuously. This is constructed by having the *design matrix*, M, set to three so as to convert the monthly series into the given quarter. The problem with this construction is that monthly estimates do not add up to quarterly observations (the next model deals with this problem). The following state space form will then estimate S_t.

The measurement equation is constructed as follows:

$$\mathbf{Y_t = 3\ S_t + e_t} \qquad \mathbf{1.}$$

The state equation is constructed as follows:

[3] Salazar *et al*. (1995) assumed that there exists a (time-invariant) constraint linking the quarterly GDP and monthly GDP given by

$$Y_t = \sum_{t=1}^{K} MS_t$$

where the weights, M, are known a priori and are typically 1 if the low-frequency data (quarterly) are simply aggregates of the high-frequency data (monthly) or 1/k if the low-frequency data are the average of the high-frequency data.

$$S_t = 1\ S_{t-1} + \mu_t \qquad \textbf{2.}$$

The state equation has become a random walk process and this is mainly because GDP is a non-stationary series. Once the model has been cast into the standard state space form the prediction and updating formulas can be applied.

2.1.2 Univariate Model II

The construction of this model is based on the idea that quarterly GDP is the aggregate of three individual unobserved monthly GDP series as this is typical the case for a flow variable. This construction extends the previous model by forcing quarterly GDP to be the sum of the monthly series. This can be achieved using the following measurement equation and imposing restrictions on its errors. The idea of constructing the measurement equation in this way will be extended into the multivariate models. The measurement equation is constructed as follows:

$$Y_t = M_1\ S_{1t} + M_2\ S_{2t} + M_3\ S_{3t} + \varepsilon_t \qquad \textbf{3.}$$

The state equation is constructed as follows:

$$S_{1t} = T_1\ S_{1t-1} + u_{St} \qquad \textbf{4.}$$

$$S_{2t} = T_2\ S_{1t-1}$$
$$S_{3t} = T_3\ S_{2t-1}$$

The objective of making monthly estimates add up to the quarterly series is also achieved by having error terms in the measurement equation exhibit the following property;

$$\varepsilon_t \sim NID(0, H_t) \qquad \textbf{5.}$$

$$\text{where } H_t = (1\ 1\ 0, 1\ 1\ 0, \ldots) \qquad \textbf{6.}$$

The characteristic of the flow series is reflected in the error terms of the measurement equation. It follows the pattern of having estimation errors for the first month and second month. Whereas, there is no error for the third month as it adds to its quarter. This pattern can be viewed as imposing restrictions on the estimated monthly GDP, which the aggregation of three monthly GDP equals the observable quarterly GDP. Having first and second months fixed to unity is simply adopting the idea of the concentrated log likelihood function discussed in the apendix and is not a restriction on the model.

2.2 Multivariate Model

In this section we outline a model which allows us to exploit any high frequency data which may contain information about the unobserved high frequency variable we are modelling. This is an important real world extension as often many of the components of GDP are available at higher frequencies than the total index. For example we often have monthly industrial production but only quarterly GDP, but as industrial production is a major sector of GDP it obviously contains important information about the way GDP may move..

In this case we need to extend the standard state equations to include exogenous variables. The prediction formula given in the appendix is then altered in an obvious way. The state equations then become;

$$\sqrt{}_{t|t-1} = T \sqrt{}_{t-1} + \alpha_{it} \lambda_t \qquad \textbf{7.}$$

λ_t represents the changes of all monthly variables being considered in the model and α_{it} is the corresponding coefficients.

The state equation of the multivariate model is formulated in two different ways – one with a trend growth term and the other one without it. The former construction enables the model to pick up the upward trend if λ_t fails to capture the growth rate of monthly estimates of GDP.

3.1 Monte Carlo Experiment

The main purposes for conducting a Monte Carlo experiment is to investigate the relative performance of the four methods. It is also used to test whether the estimations of monthly GDP can be improved with more monthly economic variables being considered in the multivariate case. Quarterly GDP is constructed from three monthly series so as to mimic the economic environment of the real world. The three series used for DGP are designed to be similar to industrial production (IP), total business sales (RS) and employment for non-farm industries (EM) to represent production (supply side), consumption (demand side) and employment (labour side) markets of the economy.

3.2 Deriving the Data Generation Process

The data generation process (DGP) plays a pivotal part in the experiment. It is important that the experimentation tries to ensure the sample to be representative and to design an artificial world which is similar to the real world. Each monthly series (MS) has a general AR(n) representation as shown below:

$$MS_t^{\ m} = C_{MS} + \sum_{i=1}^{k} \beta_{MSi}\, MS^{m}{}_{t-i} + R_{MS}\varepsilon_t \qquad \textbf{8.}$$

$$\varepsilon_t \sim \text{IID}(0, 1)$$

where k is the order of the AR process. These models are estimated on real world data both to choose k and to select the parameters. C_{MS} is the constant term. Standard normal *pseudo-random* numbers generation are simulated to generate random errors, ε_t, and then multiplied by the standard error, R_{MS}, from the regression analysis of each monthly series. Error terms are generated using standard normal *pseudo-random* numbers generation and a GAUSS program is written to carry out the experiment. One thousand replications are conducted in this study.

Having set out the DGP, monthly GDP and consequently quarterly GDP can be derived accordingly. Filtering and smoothing techniques in the state space modelling can then be applied to the quarterly series to derive the monthly estimates. Four different forms of the state space model will be tested.

3.3 Estimation of Monthly Models

Table 3.1, 3.2 and 3.3 give the estimation results and diagnostics for the three AR models.

Table 3.1: Estimation Results for IP, RS and EM

Variables	Coeff	T-Ratio	Variables	Coeff	T-Ratio	Variables	Coeff	T-Ratio
Const	0.80	0.426	*Const*	100.48	0.824	*Const*	35.780	0.596
IP (-1)	1.309	23.325	*RS (-1)*	0.621	11.375	*EM (-1)*	1.267	22.818
IP (-2)	-0.166	-1.792	*RS (-2)*	0.114	1.756	*EM (-2)*	0.044	0.480
IP (-3)	-0.143	-2.539	*RS (-3)*	0.273	4.959	*EM (-3)*	-0.112	-1.233
						EM (-4)	-0.199	-3.565

Table 3.2: Autocorrelation Function, Box Pierce and Ljung Box Statistic for IP, RS & EM

				IP				
Lag	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Coeff	-0.0128	-0.0270	0.0934	0.0079	-0.0298	-0.0157	-0.0531	0.0543
Box P.	0.05	0.28	3.03	3.05	3.33	3.40	4.29	5.22
Ljung B.	0.05	0.28	3.08	3.10	3.38	3.46	4.37	5.33
Lag	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Coeff	0.0226	0.0300	0.1082	-0.0282	-0.0456	-0.0668	-0.0647	-0.0384
Box P.	5.38	5.66	9.35	9.61	10.26	11.66	12.98	13.45
Ljung B.	5.50	5.79	9.64	9.90	10.59	12.07	13.46	13.96
				RS				
Lag	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Coeff	0.0023	-0.0305	-0.0323	-0.1218	0.0252	0.0095	0.0668	0.0543
Box P.	0.00	0.29	0.62	5.30	5.50	5.53	6.93	9.92
Ljung B.	0.00	0.30	0.63	5.40	5.60	5.63	7.07	10.16
Lag	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Coeff	0.0287	0.0500	0.0040	-0.0130	-0.0602	0.0140	0.1812	-0.0486
Box P.	10.18	10.96	10.97	11.02	12.16	12.22	22.56	23.31
Ljung B.	10.43	11.24	11.25	11.31	12.51	12.57	23.50	24.28
				EM				
Lag	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Coeff	-0.0181	-0.0069	0.0087	0.0461	-0.0023	-0.0544	-0.0303	0.0384
Box P.	0.10	0.12	0.14	0.81	0.81	1.74	2.03	2.49
Ljung B.	0.10	0.128	0.14	0.82	0.83	1.78	2.07	2.55
Lag	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Coeff	0.0663	-0.0074	0.0194	0.0200	-0.0253	-0.0306	-0.0529	-0.0147
Box P.	3.87	3.89	4.01	4.13	4.33	4.63	5.50	5.57
Ljung B.	3.98	4.00	4.12	4.26	4.47	4.77	5.70	5.77

Table 3.3: Lagrange Multiplier Test for IP, RS and EM

IP		**RS**		**EM**	
LM (1)	2.40520	LM (1)	0.0221588	LM (1)	0.458591
F Form (1,309)	*2.38517*	*F Form (1,309)*	*0.0218075*	*F Form (1,307)*	*0.450460*
LM (2)	2.67488	LM (2)	4.61844	LM (2)	0.739949
F Form (2,307)	*1.31880*	*F Form (2,307)*	*2.29139*	*F Form (2,305)*	*0.361345*
LM (4)	3.26165	LM (4)	5.54345	LM (4)	1.00658
F Form (4,303)	*0.794908*	*F Form (4,303)*	*1.36111*	*F Form (4,301)*	0.244320
LM (8)	6.73044	LM (8)	9.26225	LM (8)	4.49817
F Form (8,295)	*0.806928*	*F Form (8,295)*	*1.11992*	*F Form (8,293)*	0.537092

4.1 Monte Carlo Experiment of Univariate Model I

It did not generally prove possible to choose the size of the measurement error through estimation in this very simple model. We therefore propose a grid search of this parameter and to select the best parameter through the Monte Carlo exercise itself. That is we can select the measurement error which makes our constructed series as close as possible to the actual series for monthly GDP.

After 1000 replications, the error value of 0.5 yields the minim root mean squared error (RMSE) and it is, therefore, to be used to make comparison with other models. RMSEs with 10 different errors in the state equation for univariate model I are shown in Table 4.1.[4]

[4] One thousand replications for this experiment is conducted separately to the one used in 4.3.

Table 4.1: RMSEs with Different State Errors in the Univariate Model I

Errors in the State Equation	RMSEs
0.1	877.65831
0.2	753.78142
0.3	716.17738
0.4	703.61205
0.5	***701.78876***
0.6	704.78832
0.7	709.70494
0.8	715.09908
0.9	720.29863
1.0	725.02750

4.2 Monte Carlo Experiment of Univariate Models and Multivariate Models

We now compare the 4 models. ML estimation is used for the univariate model II and for the multivariate models as being the appropriate estimation technique to be applied. The concentrated log likelihood function is adopted for the Monte Carlo experimentation.

In the results we also investigate how much improvement can be achieved by drawing on extra high frequency information. In our model true GDP is constructed from three series. We then consider two variantes of the multivariate model. Model (a) is the model with only one, IP, monthly variable in the state equation and multivariate models (b) is the model with two, IP and EM, monthly variables in the state equation.

We do not allow the constant terms of the state equation to be stochastic. This is mainly because GDP is a series that exhibits a constant upward trend over the years. However if the underlying growth rate of the series did seem to vary systematically it would be an easy extension of the model to make the constant a stochastic state variable.

4.3 Results for the Univariate and Multivariate Models

Results of the RMSEs for the univariate and multivariate models are presented in Table 4.2. It can be identified that the multivariate model with a constant term has the smallest RMSEs and it is, therefore, the best model among those being considered. As we would expect multivariate models (b) is better than multivariate models (a) as it contains a wider information set. However, the striking thing about the results is that the improvement is not large and the simpler model works almost as well. Indeed the very simple univariate model II,

by simply imposing aggregation of months to the quarterly figure seems to do remarkably well.

Table 4.2: RMSEs Results for the Univariate and Multivariate Models

Models	RMSEs
Univariate Model I	697.8913
Univariate Model II	690.5032
Multivariate Model (a)	687.6932
Multivariate Model (b)	686.7581
Multivariate Model with constant term (a)	682.4036
Multivariate Model with constant term (b)	682.3256

5.1 Conclusions

This paper has proposed a range of state space forms for simultaneously balancing and generating high frequency data. Those models have been compared by means of a Monte Carlo study. This has demonstrated the practical feasibility of the proposed technique and also that although the more complex models always give more accurate answer the improvement is fairly small. This suggests that for many purposes the more simple technique may be preferable.

Appendix 1 The State Space Form

Let

$$\mathbf{Y_t} = \mathbf{M}\,\mathbf{S_t} + \mathbf{e_t} \qquad (1)$$

be the measurement equation (observation equation), where Y_t is a measured variable. S_t is the state vector of unobserved variables. M is a *design matrix*. The state equation (transition equation) is then given as:

$$\mathbf{S_t} = \mathbf{T}\,\mathbf{S_{t-1}} + \boldsymbol{\mu_t} \qquad (2)$$

where T are parameters. The time subscripts of M and T are ignored here, as they both are time-invariant in this study. Disturbances, ε_t and μ_t, are assumed to be normally distributed and they are mutually and serially uncorrelated with each other.

$$\varepsilon_t \sim \mathrm{NID}(0, H_t),\ \mu_t \sim \mathrm{NID}(0, Q_t)$$

The Kalman Filter

The Kalman filter is a recursive procedure and it works in two parts, prediction and updating. Prediction is simply the attempt to make a best guess at the state variables given the knowledge of the system and estimates. The Kalman filter prediction equations are as follows;

$$\sqrt{}_{t|t-1} = T\,\sqrt{}_{t-1} \qquad (3)$$

and

$$P_{t|t-1} = T\,P_{t-1}\,T' + Q_t \qquad (4)$$

Updating is the process of combining the initial estimate of the state variable with the information contained in the current observed variables. Once the current observation on Y_t becomes available, these estimates can be updated using the following equations:

$$\sqrt{}_t = \sqrt{}_{t|t-1} + P_{t|t-1}\,M'\,(M\,P_{t|t-1}M' + H_t)^{-1}\,(Y_t - M\,\sqrt{}_{t|t-1}) \qquad (5)$$

and

$$P_t = P_{t|t-1} - P_{t|t-1}\,M'\,(M\,P_{t|t-1}M' + H_t)^{-1}\,M\,P_{t|t-1} \qquad (6)$$

Prediction Error Decomposition and Maximum Likelihood Estimation

The likelihood function can be written in terms of the one-step ahead prediction errors, v_t, and their variances, F_t. The one-step ahead prediction error (innovation) is defined as

$$v_t = Y_t - M\,\sqrt{}_{t|t-1} = M\,(S_t - \sqrt{}_{t|t-1}) + \varepsilon_t \qquad (7)$$

and

$$F_t = M\,P_{t|t-1}\,M' + H_t \qquad (8)$$

With the Gaussian condition, the Kalman filter provides the means of constructing the likelihood function by the prediction error decomposition (see Harvey (1989) p.126) and the log likelihood function can then be written as

$$\log L(y) = -\frac{NT}{2}\log 2p - \frac{1}{2}\sum_{t=k}^{T}\log|F_t| - \frac{1}{2}\sum_{t=k}^{T} v_t' F_t^{-1} v_t \qquad (9)$$

where N=T-k and k is the number of periods needed to derive estimates of the state vector. The above expression is the prediction error decomposition form of the likelihood function. The concentrated log likelihood function is having the error variance fixed, either in the measurement equation, H_t, or in the state equation, Q_t, and then estimates the other. This function can be shown to be proportional to

$$\log(L) = \sum_{t=1}^{T} \log(F_t) + N \log\left(\sum_{t=1}^{T} \frac{v_t^2}{NF_t}\right) \qquad (10)$$