ELSEVIER

International Journal of Forecasting 11 (1995) 395–405

international journal of forecasting

Forecasting the final vintage of real personal disposable income: A state space approach

K.D. Patterson

Department of Economics, University of Reading, P O Box 218, Whiteknights, Reading RG6 2AA, England, UK

Abstract

This study reports a state space approach to integrating models of the data measurement process (DMP) and the data generation process (DGP) in order to reduce the root mean square error associated with using preliminary vintages of data on real personal disposal income as predictors of the final vintage. Our approach is more extensive than that taken previously with a DMP involving 13 vintages of data and a model of the DGP which exploits Johansen's cointegrating framework.

Keywords: Data measurement process; State space model; Kalman filter; Cointegration

1. Introduction

A common characteristic of data produced by national and international statistical agencies is that series which are published initially, to meet continual demands for timely data, are subsequently revised. These revisions, which are often quite substantial, imply that different vintages of data on a particular series become available sequentially. Thus, a very real problem for the users of such data is how to treat the preliminary vintages of revised series; a lengthy period of revision, as is often the case, raises the problem of how best to forecast the final vintage of data. If the preliminary vintages of data are efficient forecasts of the final vintage–see for example Mankiw et al. (1984) and Patterson and Heravi (1992)–then the most timely data is, in that sense, the best available. However, there may be information in, for example, the structure of the revisions which can be exploited to improve the preliminary vintages considered as forecasts of the final vintage. Previous studies to have addressed this problem include Howrey (1978, 1984), Conrad and Corrado (1979) and Harvey et al. (1983).[1] The consensus of these studies is that the best approach is through a state space framework with application of the Kalman filter providing an optimal combination of all the preliminary vintages of data with h-step ahead predictions. However, applications of this approach have been very limited, in terms of (i)

* Other studies on data revisions include Mankiw et al. (1984) on the rationality of preliminary data on the US money stock, Mankiw and Shapiro (1986) and Mork (1987) on revisions to US GNP, and Patterson and Heravi (1991a,b) on revisions to data in the UK and national income accounts.

0169-2070/95/$09.50 © 1995 Elsevier Science B.V. All rights reserved
SSDI 0169-2070(95)00602-8

the extent of the revisions, and hence the forecasting problem, (ii) the number of interrelated variables being considered, and (iii) the nature of the data generation process (DGP) which is used to provide the h-step ahead predictions. In this study we are able to extend previous work in all these areas whilst keeping a clear focus on the stimulus for this research, i.e. to reconsider the problem of forecasting the revisions to real personal disposable income (RPDI), following on from Howrey's (1978) seminal study.

We recognise that in practice the revisions process is extensive and consider a model with 13 vintages of data (i.e. 12 revisions to the initial data) on two variables and also allow a further variable which is not subject to revision. This introduces a number of important complications. With such a lengthy revisions process we cannot assume that the preliminary vintages are either unbiased or weakly efficient predictors of the final vintage. We also recognise that economic variables which are related in an economic context may also be revised in an interrelated way by the relevant statistical agency. Thus, we find evidence that revisions to income and consumption do have a systematic structure which can be exploited to reduce the unconditional variance of the measurement errors. The state space approach offers a convenient framework for incorporating this information into an optimal predictor of the final vintage of data.

The measurement equations of the state space model incorporate bias adjustments; the transition equations incorporate an interrelated model of measurement errors and a model of the data generation process for the final vintages of the economic variables. The measurement equations and the model of the measurement errors together constitute a model of the data measurement process or DMP. Models of the DMP and DGP together can be regarded as a complete or integrated model which explains what we actually observe. We model the DGP using recent developments in cointegration theory–see, for example, Johansen (1988). To briefly anticipate our results we find that it is possible, by integrating models of the DMP and DGP, to effect a substantial reduction in the root mean square error associated with using the preliminary vintages of data as predictors of the final vintage. This has the further implication that the preliminary vintages of data cannot be regarded as efficient forecasts of the final vintage.

This study is organised as follows. In Section 2 we briefly report on the scale of revisions to RPDI. In Section 3 we distinguish between the data measurement and data generation processes and outline the state space approach which integrates these processes. In Section 4 we report on the empirical components of the state space model and evaluate its success in providing a forecast of the final vintage of data on RPDI. Section 5 is devoted to some concluding remarks.

2. Notation and data characteristics

The notation used in this paper is as follows: a superscript v indicates the *vintage* of data, with $v = 0$ the initially published data and $v = m$ the final vintage of data, here $m = 12$. The first subscript indicates the variable and the second the period to which the observation refers; the vth vintage of RPDI for quarter t is Y^v_{1t}, with a lower case letter indicating the natural logarithm of the variable. Other variables will be defined in a similar manner as they are introduced. A variable without the v superscript is one which is *not* subject to revision; and our modelling framework is generally enough to incorporate both revised and unrevised data. All data on revised variables were obtained from successive issues of *Economic Trends* published by the UK Central Statistical Office (CSO). Our complete sample period for $v = 0$ data runs from 1970Q1 to 1992Q1, thus the last observation for $v = 12$ data is 1989Q1.

Whether or not researchers should be concerned about data revisions depends, in part, on how extensive such revisions are, and we summarise in Table 1 some relevant properties of revisions to UK RPDI. The revisions to the data span 3 years and are thus extensive both in time and, as is evident from Table 1, in nature.

Table 1
Data characteristics: RPDI

Vintage v	0	1	2	3	4	5	6	7	8	9	10	11
Mean absolute revision (%)	1.96	1.66	1.49	1.37	1.30	1.23	1.00	0.88	0.76	0.66	0.50	0.30
Mean revision (%)	−1.70	−1.45	−1.25	−1.03	−0.92	−0.74	−0.54	−0.42	−0.33	−0.26	−0.17	−0.08
Standard deviation (%)	1.76	1.61	1.45	1.36	1.27	1.30	1./20	1.10	0.92	0.84	0.65	0.52
Test for bias												
α^v	0.018	0.016	0.013	0.012	0.011	0.009	0.006	0.005	0.004	0.003	0.002	0.0008
$t(\alpha^v)$	4.52	4.84	4.18	4.36	3.98	3.45	2.63	2.27	2.44	2.46	1.71	1.26
MSL	0.00	0.00	0.00	0.00	0.00	0.00	0.010	0.026	0.017	0.016	0.09	0.21
MA	4	4	4	4	4	4	4	4	4	2	1	0
DW	21.05	1.91	1.94	1.81	1.79	1.87	1.74	1.98	1.98	1.97	2.08	1.95
Biased	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no

There is a mean absolute revision (MAR) of just under 2% for $v = 0$ data with a standard deviation (defined about the mean) of similar magnitude; and it takes six revisions to halve the MAR–see $v = 6$. The bulk of the revisions are negative, as indicated by the mean revision, implying that there is a systematic upward revision to the preliminary vintage. This initial analysis suggests that the unadjusted preliminary vintages are unlikely to be good predictors of the final vintage, and in the next section we consider how to obtain an optimal estimator of the final vintage.

An interesting question relates to whether the mean revision is significantly different from zero or, equivalently, whether the preliminary vintage is an unbiased predictor of the final vintage. The statistical framework for this analysis is set out in Patterson (1992). The log revision (which is a good approximation to the percentage revision) for RPDI is modelled as

$$y_{1t}^m - y_{1t}^v = \alpha^v + \eta_{1t}^v \quad v = 0, \ldots, m-1$$

where η_{1t}^v is a random term which is likely to have a moving average structure–see Brown and Maital (1981). Following Holden and Peel (1990) y_{1t}^v is an unbiased predictor if $H_0: \alpha^v = 0$ is not rejected against the two-sided alternative–on the relationship between unbiased and weak efficiency in a predictor see also Patterson (1992). The results of this test for the different vintages are reported in the lower part of Table 1, where the point estimate of α^v is given followed by the (asymptotic) 't' statistic, the marginal significance level, MSL, of the test, the order of the MA process and the DW associated with the regression; and finally we indicate whether the mean revision is significantly different from zero, which carries the implication that the associated preliminary vintage is a biased predictor of the final vintage. The results are clearcut: for all but vintages 10 and 11 there is an indication that the mean revision is significantly different from zero at usual significance levels.

3. The data measurement and data generation processes

It is useful here to make a distinction between the data measurement process (DMP) and the data generation process (DGP). The usual emphasis in econometric work is on modelling the latter to the exclusion of the former. However, the DMP, which relates the final vintage of data to the preliminary vintages and summarises the evolution over time of the revisions, is an important part of the modelling process relating what is observed at time t (the preliminary vintages) to what is unobserved (the final vintage and, hence, revisions). We define an integrated econometric model as one which combines models of both the DMP and DGP. A natural framework for such a combination is the state space approach–see, Howrey (1978, 1984), Conrad and Corrado (1979), Harvey et al. (1983) and Patterson (1995). This combines a set of measurement equations, which relate the preliminary vintages available at time t to the final vintages plus measurement errors, with a transition equation which describes the evolution of the measurement errors and a model of the DGP.

In the empirical work summarised in Section 4 we report a joint model for the measurement errors arising from 13 vintages of data on RPDI and consumption. The latter variable in the natural logarithm is denoted by y_{2t}^{v}. The measurement errors model is combined with a cointegrating VAR (CVAR) to model the DGP. The specification of the CVAR is based on recent research–see Holmans (1991), Carruth and Henley (1992) and Patterson (1994)–which suggests that whilst income and consumption do not cointegrate the addition of a measure of housing equity withdrawal, HEW, ensures a cointegrating vector. The data on HEW is provided by the Bank of England and is not revised; for conformity with the previous notation we denote this variable as y_{3t}. Our empirical model gives rise to a state vector of dimension 128×1 and a 128×128 transition matrix; however, for illustrative purposes we adopt a model with all the essential ingredients of the empirical model but of lower dimensionality. In the following example there are two revised variables, y_{1t}^{v} and y_{2t}^{v}, with $v = 0, \ldots, 3$, so that there are three revisions and, thus, four vintages of data, and one unrevised variable, y_{3t}.

For convenience define the following notation:

$$\underline{y}_{it} = (y_{it}^{0}, y_{it-1}^{1}, y_{it-2}^{2}, y_{it-3}^{3})' ,$$

$$\underline{u}_{it} = (u_{it}^{0}, u_{it-1}^{1}, u_{it-2}^{2})' ,$$

$$\underline{y}_{it}^{m} = (y_{it}^{m}, y_{it-1}^{m}, y_{it-2}^{m}, y_{it-3}^{m})' ,$$

$$\text{for } i = 1, 2 \text{; and } m = 3$$

$$y_{it}^{3} = y_{it}^{m} \quad \text{for } i = 1, 2$$

$$\underline{y}_{3t} = (y_{3t}, y_{3t-1}, y_{3t-2}, y_{3t-3})'$$

We adopt the convention that underlining and double underlining indicate vectors and matrices, respectively. Note that the vector $\underline{y}_{it}$ comprises data which are observable for time t. This is the data made available by the CSO in a particular issue of *Economic Trends*–that is, they publish the 0th vintage for time t, the first revision of $t-1$ data, the second revision of $t-2$ data and so on. Then the measurement equations of the state space model are,

$$\underline{y}_t = \begin{bmatrix} \underline{y}_{1t} \\ \underline{y}_{2t} \\ \underline{y}_{3t} \end{bmatrix}$$

$$= \begin{bmatrix}
\gamma_1^0 & \beta_1^0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\gamma_1^1 & 0 & \beta_1^1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\gamma_1^2 & 0 & 0 & \beta_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\gamma_2^0 & 0 & 0 & 0 & 0 & \beta_2^0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\gamma_2^1 & 0 & 0 & 0 & 0 & 0 & \beta_2^1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\gamma_2^2 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_2^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \underline{\alpha}_t$$

with $\underline{\alpha}_t = (1; \underline{y}_{1t}^{12}; \underline{y}_{2t}^{12}; \underline{y}_{3t}; \underline{u}_{1t}; \underline{u}_{1t-1}; \underline{u}_{2t}; \underline{u}_{2t-1})$; or in an obvious notation,

$$\underline{y}_t = Z\underline{\alpha}_t \tag{1}$$

The measurement equations embody bias adjustment coefficients γ_i^v and β_i^v–see Howrey (1978); if these are 0 and 1, respectively then the u_{it}^v are simply the (log) revisions, otherwise they define the measurement errors. Note that provided preliminary vintages and the final vintage cointegrate on a pairwise basis 'super-consistent' estimators of γ_i^v and β_i^v are obtained by OLS–see Stock (1987). $\underline{\alpha}_t$ is the state vector, the definition of which, for illustrative purposes, assumes a second order AR process for modelling the measurement errors considered further below.

The transition equations are given by

$$\underline{\alpha}_t = T\underline{\alpha}_{t-1} + \xi_t \tag{2}$$

with $E(\xi_t\xi_t') = \sigma^2 Q$. The specification of T depends upon the 'structural' model relating y_{1t}^m, y_{2t}^m and y_{3t}, and the model governing the evolution of the measurement error vector $\underline{u}_t = (\underline{u}_{1t}'; \underline{u}_{2t}')$. We deal with the measurement error model first taking into account two important factors in the modelling strategy. We allow measurement errors on closely related variables, such as consumption and income, to be interrelated; and allow for AR processes not only through lags on the same vintage of data but also through lags on the vintage. To illustrate these factors consider the AR(2) interrelated model of measurement errors given by

$$\begin{pmatrix} u_{1t}^0 \\ u_{1t-1}^1 \\ u_{1t-2}^2 \\ u_{2t}^0 \\ u_{2t-1}^1 \\ u_{2t-2}^2 \end{pmatrix} = \left[\begin{array}{ccc:ccc} \psi_{11}^1 & 0 & 0 & \psi_{14}^1 & 0 & 0 \\ \psi_{21}^1 & \psi_{22}^1 & 0 & \psi_{24}^1 & \psi_{25}^1 & 0 \\ 0 & \psi_{32}^1 & \psi_{33}^1 & 0 & \psi_{35}^1 & \psi_{36} \\ - & - & - & - & - & - \\ \psi_{41}^1 & 0 & 0 & \psi_{44}^1 & 0 & 0 \\ \psi_{51}^1 & \psi_{52}^1 & 0 & \psi_{54}^1 & \psi_{55}^1 & 0 \\ 0 & \psi_{62}^1 & \psi_{63}^1 & 0 & \psi_{65}^1 & \psi_{66}^1 \end{array}\right] \begin{pmatrix} u_{1t-1}^0 \\ u_{1t-2}^1 \\ u_{1t-3}^2 \\ u_{2t-1}^0 \\ u_{2t-2}^1 \\ u_{2t-3}^2 \end{pmatrix} + \left[\begin{array}{ccc:ccc} \psi_{11}^2 & 0 & 0 & \psi_{14}^2 & 0 & 0 \\ 0 & \psi_{22}^2 & 0 & 0 & \psi_{25}^2 & 0 \\ \psi_{31}^2 & 0 & \psi_{33}^2 & \psi_{34}^2 & 0 & \psi_{36}^2 \\ - & - & - & - & - & - \\ \psi_{41}^2 & 0 & 0 & \psi_{44}^2 & 0 & 0 \\ 0 & \psi_{52}^2 & 0 & 0 & \psi_{55}^2 & 0 \\ \psi_{61}^2 & 0 & \psi_{63}^2 & \psi_{64}^2 & 0 & \psi_{66}^2 \end{array}\right] \begin{pmatrix} u_{1t-2}^0 \\ u_{1t-3}^1 \\ u_{1t-4}^2 \\ u_{2t-2}^0 \\ u_{2t-3}^1 \\ u_{2t-4}^2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t}^0 \\ \varepsilon_{1t-1}^1 \\ \varepsilon_{1t-2}^2 \\ \varepsilon_{2t}^0 \\ \varepsilon_{2t-1}^1 \\ \varepsilon_{2t-2}^2 \end{pmatrix}$$

In a more compact notation, following the above partitioning, we can rewrite this as

$$\begin{pmatrix} \underline{u}_{1t} \\ \underline{u}_{2t} \end{pmatrix} = \begin{bmatrix} \underline{\underline{\psi}}_{11}^1 & \underline{\underline{\psi}}_{12}^1 \\ \underline{\underline{\psi}}_{21}^1 & \underline{\underline{\psi}}_{22}^1 \end{bmatrix} \begin{pmatrix} \underline{u}_{1t-1} \\ \underline{u}_{2t-1} \end{pmatrix} + \begin{bmatrix} \underline{\underline{\psi}}_{11}^2 & \underline{\underline{\psi}}_{12}^2 \\ \underline{\underline{\psi}}_{21}^2 & \underline{\underline{\psi}}_{22}^2 \end{bmatrix} \begin{pmatrix} \underline{u}_{1t-2} \\ \underline{u}_{2t-2} \end{pmatrix} + \begin{pmatrix} \underline{\varepsilon}_{1t} \\ \underline{\varepsilon}_{2t} \end{pmatrix} \tag{3}$$

with

$$E\left[\begin{pmatrix} \underline{\varepsilon}_{1t} \\ \underline{\varepsilon}_{2t} \end{pmatrix} (\underline{\varepsilon}_{1t}' \quad \underline{\varepsilon}_{2t}')\right] = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$$

Note that each coefficient matrix in (3) is partitioned into four blocks. The diagonal elements of the diagonal blocks relate to the order of the AR process on *own* measurement errors; whereas the lower off-diagonals on these blocks relate to AR lags on the vintage. The zeroes in the lower triangle of this sub-matrix reflect the order of the AR process; for example, ψ_{32}^1 is the coefficient relating u_{1t-2}^2 to u_{1t-2}^1 which is a first order lag on the *vintage*; ψ_{31}^2 is the coefficient relating u_{1t-2}^2 to u_{1t-2}^0, which is a second order lag on the *vintage*. Each of these submatrices has an upper triangle of zeroes otherwise *leads* rather than lags on the vintages would be present. The off-diagonal blocks follow the same scheme but relate measurement errors with different variable subscripts so that non-zero coefficients here capture the interrelated nature of data revisions.

Note that (3), in this example, is a set of six equations which looks similar to a VAR (for which OLS would be an efficient estimation procedure). The set is not, however, a VAR because the explanatory variables differ between equations; for example, no first order lag on the vintage is possible for u_{1t-2}^0, and no second order lag on the vintage is possible for u_{1t}^1. The equations given by (3) are a set of seemingly unrelated equations (SURE) which can be estimated efficiently by maximum likelihood or iterated SURE. The model of the DMP comprises the measurement equations, (1), and the measurement error model, (3); in an integrated model they are combined with a model of the DGP, with the latter referred to here for convenience as the structural model.

The structural model could be as simple as a univariate time series model, and it turns out that this performs moderately well in combination with the DMP in reducing the mean square error of the preliminary vintages as predictors of the final vintage, or more complex as in a BVAR or CVAR. We illustrate our approach with the CVAR framework due to Johansen (1988). It is well known that a VAR should be modelled in stationary variables, however with variables which are I(1), as reported below, a VAR in I(0) variables implies the absence of a long run specification. Following Johansen (1988) if there is a stationary linear combination of the I(1) variables–that is a cointegrating vector–then this can be incorporated in the VAR to tie down the long run relationship between the variables. Whilst multiple cointegrating vectors may exist we find evidence, reported below, for a single cointegrating vector which can be interpreted as defining the consumption equilibrium and we illustrate the structural model in this context. Consider the CVAR in y_1^{12}, y_2^{12} and y_3, with an error correction mechanism, ecm, given by

$$
\begin{aligned}
\Delta y_{1t}^{12} &= \Pi_{11}\Delta y_{1t-1}^{12} + \Pi_{12}\Delta y_{2t-1}^{12} + \Pi_{13}\Delta y_{3t-1} \\
&\quad + \Pi_{14}\mathrm{ecm}_{t-1} + v_{1t} \\
\Delta y_{2t}^{12} &= \Pi_{21}\Delta y_{1t-1}^{12} + \Pi_{22}\Delta y_{2t-1}^{12} + \Pi_{23}\Delta y_{3t-1} \\
&\quad + \Pi_{24}\mathrm{ecm}_{t-1} + v_{2t} \\
\Delta y_{3t} &= \Pi_{31}\Delta y_{1t-1}^{12} + \Pi_{32}\Delta y_{2t-1}^{12} + \Pi_{33}\Delta y_{3t-1} \\
&\quad + \Pi_{34}\mathrm{ecm}_{t-1} + v_{3t} \\
\mathrm{ecm}_t &= y_{2t}^{12} - \theta_1 y_{1t}^{12} - \theta_2 y_{3t}
\end{aligned}
\qquad (4)
$$

with $E(vv') = \sigma^2\Omega$, where $v = (v_{1t}, v_{2t}, v_{3t})'$. To embody the equation system (4) into the transition equation it needs to be written in companion form in the levels of the variables. In this form and bringing together the measurement error model and the structural model we have:

$$
\underline{\underline{T}} = \left[\begin{array}{c|cccccccccccc|cccc}
1 & 0 & \cdots & & & & & & & & & & \cdots\ 0 & & & & \\
\hline
0 & \phi_{11} & \phi_{12} & 0 & 0 & \phi_{13} & \phi_{14} & 0 & 0 & \phi_{15} & \phi_{16} & 0 & 0 & & & & \\
\vdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\
 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\
 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\
\hline
 & \phi_{21} & \phi_{22} & 0 & 0 & \phi_{23} & \phi_{24} & 0 & 0 & \phi_{25} & \phi_{26} & 0 & 0 & & & & \\
\vdots & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\
 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & & & & \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & & \underline{\underline{0}}^{+} & & \\
\hline
 & \phi_{31} & \phi_{32} & 0 & 0 & \phi_{33} & \phi_{34} & 0 & 0 & \phi_{35} & \phi_{36} & 0 & 0 & & & & \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & & & & \\
\vdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & & & & \\
\hline
0 & & & & & & & & & & & & & \underline{\underline{\psi}}_{11}^{1} & \underline{\underline{\psi}}_{11}^{2} & \underline{\underline{\psi}}_{12}^{1} & \underline{\underline{\psi}}_{12}^{2} \\
\vdots & & & & & & & & & & & & & I & \underline{\underline{0}}_3 & \underline{\underline{0}}_3 & \underline{\underline{0}}_3 \\
 & & & & & & \underline{\underline{0}}^{+\prime} & & & & & & & & & & \\
 & & & & & & & & & & & & & \underline{\underline{\psi}}_{21}^{1} & \underline{\underline{\psi}}_{21}^{2} & \underline{\underline{\psi}}_{22}^{1} & \underline{\underline{\psi}}_{22}^{2} \\
0 & & & & & & & & & & & & & \underline{\underline{0}}_3 & \underline{\underline{0}}_3 & I & \underline{\underline{0}}_3
\end{array}\right]
$$

0^+ is a conformal zero matrix. The ϕ_{ij} coefficients are the levels coefficients obtained from the underlying Π_{ij} coefficients in the CVAR. The covariance matrix of ξ_t–see (2) above–is specified by a suitable ordering of the elements of the covariance matrices of the CVAR and the measurement error model, with zero covariance between the two sets of disturbance terms. Univariate and VAR models are easy to accommodate in the transition matrix. The former, for example, would reduce the dimensionality considerably by setting all the ϕ_{ij} coefficients equal to zero except for ϕ_{11} and ϕ_{12} for a second order AR model. Application of the Kalman filter produces a minimum mean square linear estimator, mmsle, of the state vector or mmse if the disturbances and initial state vector are normally distributed.

It may be useful here to briefly summarise the steps in setting up the model in state space form:

(i) estimate the measurement equation coefficients γ_i^v and β_i^v, these serve to define the (log) measurement errors u_{it}^v; super-consistent estimates of γ_i^v and β_i^v can be obtained by OLS;

(ii) having defined u_{it}^v from step (i), estimate an autoregressive model linking the measurement errors to own vintage past measurement errors and to own variable lags on the vintage; and follow the same scheme to relate measurement errors on different variables–here income and consumption;

(iii) estimate the dynamic equations for the final vintage of income (y_{1t}^m) and consumption (y_{2t}^m), using the unrevised data for HEW (y_{3t});

(iv) construct the measurement equations of the state space model using the coefficients estimated in step (i), and construct the transition matrix using the coefficients estimated from steps (ii) and (iii);

(v) finally apply the Kalman filter to the complete state space model to obtain optimal estimates of all vintages of data.

4. Empirical results

4.1. The measurement equations

The coefficients of the measurement equations, that is γ_i^v and β_i^v, for $v = 0, 1, \ldots, 11$ and $i = 1, 2$, serve to bias adjust the preliminary vintages of data. They can be super-consistently estimated by OLS provided y_{it}^v and y_{it}^{12} cointegrate–see Stock (1987). Estimation details are reported in Table 2 with point estimates of γ_i^v and β_i^v; in all cases cointegration is achieved (the AIC model selection criteria favoured the DF test).

4.2. Estimation of the structural model

Tests for the order of integration of y_{1t}^m, y_{2t}^m and y_{3t} confirmed that they were I(1) whereas their first differences were consistent with I(0) processes–see Table 3. Johansen's (trace) test statistic for the number, r, of cointegrating vectors suggested $r = 1$, with a sample value of 36.18 against the 5% critical value of 34.91, which was readily interpretable as the long run consumption function–see Table 3. Incorporating the stationary combination given by the

Table 2
Measurement equation coefficients

$v=$	0	1	2	3	4	5	6	7	8	9	10	11
RPDI												
γ_1^v	0.113	0.917	0.211	0.214	0.283	0.274	0.254	0.214	0.173	0.170	0.162	0.115
β_1^v	0.986	0.977	0.976	0.976	0.969	0.970	0.972	0.977	0.981	0.982	0.982	0.987
DF	−4.07	−4.21	−3.90	−3.65	−3.56	−4.51	−4.82	−4.92	−4.64	−4.89	−5.11	−8.16
Consumption												
γ_2^v	0.415	0.364	0.245	0.170	0.114	0.065	0.031	0.003	0.000	0.001	−0.006	−0.002
β_2^v	0.954	0.959	0.972	0.981	0.987	0.992	0.996	0.993	0.999	0.999	1.000	1.000
DF	−3.60	−3.52	−3.85	−3.87	−4.20	−3.84	−4.49	−4.24	−4.41	−5.00	−6.05	−8.41

Table 3
Structural model: empirical results

	Unit root tests						Cointegration (Johansen) test		
	y^{12}_{1t}	Δy^{12}_{1t}	y^{12}_{2t}	Δy^{12}_{2t}	y_{3t}	Δy_{3t}	Null	Alternative	Test statistic
DF	−2.48	−9.28	−0.23	−8.49	−2.73	−9.36	$r=0$	$r\geq 1$	36.18(34.91)
ADF()	−2.73	−3.35	−1.09	−3.28	−1.83	5.66	$r\leq 1$	$r\geq 2$	16.00(19.96)
							$r\leq 2$	$r=3$	4.78(9.24)

Cointegrating vector normalised on consumption: $\mathrm{ecm}_t = y^{12}_{2t} - 1.41 - 0.832y^{12}_{1t} - 1.693y_{3t}$.
Notes: The 5% critical values for the DF and ADF tests are −3.47 and −2.90 for the I(1) and I(0) tests, respectively; the I(1) auxiliary regression contained a time trend; without a trend the test statistics were even further in favour of the series being I(1). Phillips–Perron versons of these tests (not reported) gave the same conclusion. 5% critical value in parentheses for the Johansen test statistic which was based on a 4th order VAR with intercepts.

CVAR model

$$\Delta y^{12}_{1t} = 0.002 - 0.147\,\Delta y^{12}_{1t-1} + 0.181\,\Delta y^{12}_{2t-1} + 0.291\,\Delta y^{12}_{2t-2} + 0.825\,\Delta y_{3t-1} + 0.479\,\Delta y^{3t-2} + 0.43\mathrm{ecm}_{t-1}$$
(1.10)(−1.61) (1.45) (2.27) (3.25) (1.71) (3.86)

$\%\hat{\sigma} = 1.39$ LM (4) = 0.96(0.44) HS(1) = 0.14(0.71) FF(1) = 0.19(0.66) NORM(2) = 0.13(0.64)

$$\Delta y^{12}_{2t} = 0.005 + 0.175\,\Delta y^{12}_{2t-3} + 0.169\,\Delta y^{12}_{2t-1} - 0.122\mathrm{ecm}_{t-1}$$
(2.27) (2.26) (1.95) (−2.25)

$\hat{\sigma} = 1.20$ LM(4) = 0.58(0.68) HS(1) = 0.00(0.99) FF(1) = 0.96(0.33) NORM(2) = 3.43(0.18)

$$\Delta y_{3t} = 0.02 + 0.083\,\Delta y^{12}_{1t-3} - 0.097\,\Delta y^{12}_{2t-2} - 0.140\,\Delta y^{12}_{2t-3} - 0.156\,\Delta y_{3t-3} - 0.171\,\Delta y_{3t-4}$$
(2.02) (1.66) (−1.46) (−1.94) (−1.81) (−1.93)

$\%\hat{\sigma} = 0.70$ LM(4) = 1.19(0.32) HS(1) = 1.74(0.19) FF(1) = 2.48(0.12) NORM(2) = 3.12(0.21)

LM(4) is the Lagrange-Multiplier test for 4th order serial correlation; HS(1) is the MICROFIT test for heteroscedasticity; FF(1) is the RESET functional form test; NORM(2) is the Bera–Jaque normality test; degrees of freedom indicated in parentheses; marginal significance level of the test in parentheses following the test statistics; 't' statistics in parentheses beneath estimated coefficients.

cointegrating vector, lagged once, as an error correction mechanism, ecm, in a CVAR gave the results reported in the lower part of Table 3. All the equations of the CVAR passed a range of diagnostic statistics, and these coefficients and the covariance matrix of the residuals form the necessary input to the transition equation in the state space model.

4.3. Measurement error model

Recall that the two distinct sets of elements in the transition equation are the structural model and the measurement error model. We report the empirical version of the latter here. We found it necessary to consider a fourth order autoregressive interrelated model of measurement errors on income and consumption. The empirical model was obtained by sequential simplification ensuring that at each stage the equations passed a range of diagnostic statistics. The estimated equations are reported in Table 4.

The results show how important it is to consider an interrelated model of measurement errors on income and consumption with a number of significant coefficients on the 'cross' effects. The dominant effect is the AR process on the vintage (except where this is not possible as for $v=0$)–for example, u^3_{1t-3} depends on u^2_{1t-3} with a coefficient of 0.872 and 't' statistic of 28.6, and this is a pattern reflected for $v=2,\dots,11$. Also note that there is a considerable reduction in the unconditional standard deviation of the measurement errors, $\hat{\sigma}_{uc}(\%)$; for example, for income and $v=2$, $\hat{\sigma}_{uc}$ is reduced from 1.45% to 0.69% by the estimated model. The estimated coefficients from Table 4 correspond to the ψ^p_{ij} coefficients in (3) and T, and the variance–covariance matrix is partitioned into estimates of Σ_{11}, Σ_{22} and Σ_{12}.

4.4. Combining and evaluating the models

The measurement equations and the measurement error model which comprise the model of the DMP are combined with the structural

Table 4
The empirical model for measurement errors

	Income					$\hat{\sigma}_{uc}$(%)	$\hat{\sigma}$(%)	DW	Consumption						$\hat{\sigma}_{uc}$(%)	$\hat{\sigma}$(%)	DW
$v=0$	u^0_{1t}	$=0.582u^0_{1t-1}$ (7.17)				1.80	1.45	2.12	u^0_{2t}	$=0.608u^0_{2t-1}$ (8.92)	$+0.246u^0_{2t-3}$ (3.63)	$-0.213u^0_{1t-2}$ (−4.17)	$+0.239u^0_{1t-3}$ (4.39)		1.32	0.85	1.97
$v=1$	u^1_{1t-1}	$=0.784u^0_{1t-1}$ (2.71)				1.63	0.78	1.64	u^1_{2t-1}	$=0.790u^1_{2t-2}$ (11.76)		$-0.102u^1_{1t-3}$ (−1.86)			1.23	0.81	2.11
$v=2$	u^2_{1t-2}	$=-0.089u^1_{2t-2}$ (−1.99)	$+0.783u^1_{1t-2}$ (20.91)			1.45	0.69	1.71	u^2_{2t-2}	$=0.867u^1_{2t-2}$ (31.45)		$-0.060u^2_{1t-4}$ (−2.17)	$+0.041u^2_{1t-6}$ (1.74)		1.20	0.51	1.89
$v=3$	u^3_{1t-3}	$=-0.107u^3_{2t-5}$ (−2.73)	$+0.117u^3_{2t-7}$ (2.93)	$+0.872u^2_{1t-3}$ (28.62)		1.37	0.61	1.81	u^3_{2t-3}	$=0.870u^2_{2t-3}$ (31.45)	$+0.072u^0_{2t-3}$ (1.81)	$+0.087u^0_{2t-6}$ (−2.88)	$+0.095u^2_{1t-3}$ (3.86)	$-0.153u^3_{1t-4}$ (−5.00)	1.15	0.43	1.93
$v=4$	u^4_{1t-4}	$=0.41u^3_{1t-4}$ (14.67)	$+0.115u^1_{1t-4}$ (2.40)			1.27	0.58	1.60	u^4_{2t-4}	$=0.827u^3_{2t-4}$ (38.40)			$-0.049u^1_{1t-4}$ (−2.80)		1.01	0.38	1.74
$v=5$	u^5_{1t-5}	$=-0.221u^3_{2t-5}$ (−2.51)	$+0.196u^2_{2t-5}$ (2.30)	$+0.920u^4_{1t-5}$ (23.43)	$-0.078u^5_{1t-7}$ (−1.94)	1.32	0.59	2.17	u^5_{2t-5}	$=0.561u^4_{2t-5}$ (13.88)	$+0.190u^5_{2t-7}$ (4.82)	$+0.082u^1_{2t-5}$ (2.58) $-0.24u^1_{1t-5}$ (−6.07)	$-0.097u^4_{1t-5}$ (−2.69) $+0.151u^5_{1t-9}$ (5.45)	$+0.234u^2_{1t-5}$ (5.05)	0.87	0.38	1.87
$v=6$	u^6_{1t-6}	$=0.125u^4_{2t-6}$ (−2.63)	$+0.789u^5_{1t-6}$ (18.49)	$-0.121u^5_{1t-6}$ (−2.82)		1.20	0.47	1.70	u^6_{2t-6}	$=0.961u^5_{2t-6}$ (35.70)	$-0.130u^6_{2t-7}$ (−4.23)				0.82	0.37	2.12
$v=7$	u^7_{1t-7}	$=0.068u^4_{2t-7}$ (2.34)	$+0.884u^6_{1t-7}$ (21.84)			1.10	0.45	1.68	u^7_{2t-7}	$=0.791u^6_{2t-7}$ (31.09)	$-0.073u^7_{1t-8}$ (−3.85)	$+0.091u^5_{1t-7}$ (4.43)	$-0.110u^3_{1t-7}$ (−5.15)	$+0.128u^7_{1t-11}$ (6.05)	0.72	0.29	2.08
$v=8$	u^8_{1t-8}	$=0.274u^8_{2t-11}$ (5.06)	$+0.779u^7_{1t-8}$ (29.62)	$+0.073u^8_{1t-11}$ (2.22)		0.93	0.45	2.10	u^8_{2t-8}	$=0.617u^7_{2t-8}$ (18.06)	$+0.097u^4_{2t-8}$ (3.68)	$-0.243u^8_{2t-12}$ (−6.08)	$+0.072u^8_{1t-7}$ (3.31)	$-0.067u^4_{1t-8}$ (−3.81)	0.57	0.24	2.05
$v=9$	u^9_{1t-9}	$+0.275u^9_{2t-12}$ (4.62)	$+0.76u^8_{1t-9}$ (25.72)			0.84	0.45	2.07	u^9_{2t-9}	$=0.669u^8_{2t-9}$ (14.94)	$+0.128u^6_{2t-9}$ (3.78)	$-0.143u^9_{2t-12}$ (−3.34)	$-0.092u^7_{2t-9}$ (−2.85)	$+0.117u^6_{1t-9}$ (3.84)	0.46	0.19	2.11
$v=10$	u^{10}_{1t-10}	$=0.204u^{10}_{2t-13}$ (2.81)	$+0.561u^9_{1t-10}$ (15.77)			0.63	0.40	1.81	u^{10}_{2t-10}	$=0.560u^9_{2t-10}$ (8.82)	$-0.232u^{10}_{2t-12}$ (4.41)	$+0.155u^6_{2t-10}$ (3.36)	$-0.216u^{10}_{2t-14}$ (−2.98)		0.37	0.22	2.29
$v=11$	u^{11}_{1t-11}	$=0.864u^{10}_{1t-11}$ (25.77)	$-0.403u^{11}_{1t-12}$ (−10.02)	$+0.076u^{11}_{1t-13}$ (−2.08)		0.50	0.21	2.50	u^{11}_{2t-11}	$=0.763u^{10}_{2t-11}$ (21.44)	$-0.320u^{11}_{2t-12}$ (−6.44)	$-0.050u^9_{1t-13}$ (−3.40)	$+0.033u^7_{1t-11}$ (2.70)		0.29	0.14	2.07

Notes: $\hat{\sigma}_{uc}$ is the unconditional standard deviation of the measurement errors; $\hat{\sigma}$ is the equation standard error; effective sample period 1973Q1–1989Q1; 't' statistics in parentheses; DW is the Durbin-Watson statistic.

Table 5
Root mean squared error for alternative predictors of RPDI

v	0	1	2	3	4	5	6	7	8	9	10	11
[1]												
Preliminary vintage[a]	302.1	270.3	239.6	212.4	195.9	186.5	162.8	143.7	117.4	106.8	81.4	63.8
[2]												
Bias adjusted	224.1	203.7	181.6	168.0	152.3	153.3	142.0	129.5	106.0	98.1	74.3	60.8
	0.74	0.75	0.76	0.79	0.77	0.82	0.87	0.90	0.90	0.92	0.91	0.95
[3]												
Univariate RPDI	189.9	161.0	140.6	140.7	136.4	132.9	123.8	119.9	100.2	89.2	72.1	57.2
	0.63	0.59	0.59	0.66	0.70	0.71	0.76	0.83	0.85	0.84	0.88	0.90
[4] CVAR	132.9	126.9	116.7	113.7	115.9	116.0	106.0	105.0	83.3	76.5	71.9	57.0
	0.44	0.47	0.49	0.53	0.59	0.62	0.65	0.73	0.71	0.72	0.88	0.89

Note: [a] In 1970 £mn. In rows [2]–[4] the rmse is reported and then expressed as a proportion of row [1].

CVAR to provide an optimal estimator of the state vector. Using the estimates reported in Tables 2–4 we have a feasible state space model to which the Kalman filter can be applied. To gauge the extent to which reductions in uncertainty are achieved with this approach we report in Table 5 the root mean squared error, rmse, of a number of alternative estimators of the final vintage of RPDI. The first, see row [1], is the preliminary vintage for $\nu = 0, \ldots, 11$; this corresponds to taking the latest available vintage as an estimator of the final vintage. It is clear from a comparison of the rmse for the early vintages with those for the later vintages that the former are very noisy predictors of the final vintage; the rmse for $\nu = 0$ is 4.74 times that for $\nu = 11$. There is scope, therefore, for the state space approach to effect a reduction in the rmse compared to using just the latest available vintage.

The first improvement on this can be achieved by the simple process of bias adjustment; that is invert the measurement equations for RPDI and anti-log the result for each vintage of data–the results are reported in row [2]. Next, an estimator based on perhaps the simplest modelling approach is obtained by treating both the measurement error model and the structural model for income as univariate processes. This results in a further quite substantial reduction in rmse–see row [3]. Finally, we report in row [4] the results using the interrelated measurement error model and the CVAR outlined earlier with notable reductions in the rmse. For example, at the same point in time that $\nu = 0$ data are available by optimally combining the elements of the state space model there is a reduction of 56% in rmse; and about half-way through the revisions process the reduction is of the order of 38%. As is the case with all the model-based predictors the relative decline in rmse, as we approach the final data, is attenuated, but even at $\nu = 11$ a reduction of 10% in rmse is possible.

5. Conclusions

It is clear from our analysis that revisions to important macroeconomic time series such as real personal disposable income are extensive both in time and magnitude, and that preliminary vintages of data are not optimal predictors of the final vintage despite the frequent use made of such data for this purpose. The optimal predictor is obtained by an application of the Kalman filter to the problem set up as a state space model. At an empirical level there are considerable gains from: (a) adjusting the preliminary vintages for bias; (b) modelling revisions on related variables as a joint process; and (c) using a cointegration approach to ensure that an equilibrium relationship ties down the levels of the variables for the longer prediction horizons. For example, compared to the initial vintage, that is the most timely data, the state space approach embodying an interrelated measurement errors model, a CVAR structural model and bias adjustments resulted in a reduction of 66% in the root mean squared error. Alternatively, the state space model in this case effects a reduction in uncertainty as to the final vintage equivalent to about 8 quarters of vintages. A further consequence of these results is that the preliminary data cannot be viewed as an efficient forecast of the final data–see Mankiw et al. (1984)–as it is possible to reduce the rmse by combining the preliminary vintages with information available at the time these were compiled.

Acknowledgements

The author gratefully acknowledges financial support from the Nuffield Foundation. Saeed Heravi provided able research assistance and constructive comments on an earlier version of this paper. I am particularly grateful to two referees for their comments on an earlier version of this paper.

References

Brown, B.W. and S. Maital, 1981, What do economists know? An empirical study of experts' expectations, Econometrica 49, 491–504.

Carruth, A. and A. Henley, 1992, Consumer durables spending and housing market activity, Scottish Journal of Political Economy 39, 261–271.

Conrad, W. and C. Corrado, 1979, Application of the Kalman filter to revisions in monthly sales estimates, Journal of Economic Dynamics and Control 1, 177–198.

Harvey, A.C., C.R. McKenzie, D.P.C. Blake and M.J. Desai, 1983, Irregular data revisions, in: A. Zellner, ed., Applied time series analysis of economic data (US Bureau of the Census, Washington, DC).

Holden, K. and D.A. Peel, 1990, On testing for unbiasedness and efficiency of forecasts, Manchester School of Economic and Social Studies 58, 120–127.

Holmans, A.E., 1991, Estimates of housing equity withdrawal by owner-occupiers in the United Kingdom, 1970 to 1990, Government Economic Service Working Paper No. 116.

Howrey, E.P., 1978, The use of preliminary data in econometric forecasting, Review of Economics and Statistics 60, 193–200.

Howrey, E.P., 1984, Data revision, reconstruction and prediction: An application to inventory investment, Review of Economics and Statistics 66, 386–393.

Johansen, S., 1988, Statistical analysis of cointegration vectors, Journal of Economic Dynamics and Control 12, 231–254.

Mankiw, N.G. and M.D. Shapiro, 1986, News or noise: An analysis of GNP revisions, Survey of Current Business 66, 20–25.

Mankiw, N.G., D.E. Runkle and M.D. Shapiro, 1984, Are preliminary announcements of the money stock rational forecasts?, Journal of Monetary Economics 14, 15–27.

Mork, K.A., 1987, Ain't behavin: Forecast errors and measurement errors in early GNP estimates, Journal of Business and Economic Statistics 5, 165–175.

Patterson, K.D., 1992, Revisions to the components of the trade balance for the United Kingdom, Oxford Bulletin of Economics and Statistics 54, 103–120.

Patterson, K.D., 1994, Consumers' expenditure on nondurables and services and housing equity withdrawal in the United Kingdom, Manchester School of Economic and Social Studies 62, 251–274.

Patterson, K.D., 1995, An integrated model of the data measurement and data generation processes with an application to consumers' expenditure, Economic Journal 105, 54–76.

Patterson, K.D. and S.M. Heravi, 1991a, Data revisions and the expenditure components of GDP, Economic Journal 101, 887–901.

Patterson, K.D. and S.M. Heravi, 1991b, Are different vintages of data on the components of GDP cointegrated?, Economics Letters 35, 409–413.

Patterson, K.D. and S.M. Heravi, 1992, Efficient forecasts or measurement errors? Some evidence for revisions to the United Kingdom GDP growth rates, Manchester School of Economic and Social Studies 60, 249–263.

Stock, J.H., 1987, Asymptotic properties of least squares estimates of cointegrating vectors, Econometrica 55, 1035–1056.

Biography: Kerry PATTERSON, Professor of econometrics at the University of Reading, has published widely in leading journals with, for example, recent articles in The Economic Journal and the Review of Economics and Statistics. He advised the Bank of England for 15 years on areas of macroeconomics and forecasting.