Department of Economics

A State Space Approach to Extracting the Signal from Uncertain Data

Alastair Cunningham, Jana Eklund, Chris Jeffery, George Kapetanios and Vincent Labhard

Working Paper No. 637 February 2009 ISSN 1473-0278

Queen Mary
University of London

A State Space Approach To Extracting The Signal From Uncertain Data [*]

Alastair Cunningham[†], Jana Eklund[†], Chris Jeffery[†],
George Kapetanios[†#] & Vincent Labhard[‡]

[†]Bank of England
[#]Queen Mary University of London
[‡]European Central Bank

January 30, 2009

Abstract

Most macroeconomic data are uncertain - they are estimates rather than perfect measures of underlying economic variables. One symptom of that uncertainty is the propensity of statistical agencies to revise their estimates in the light of new information or methodological advances. This paper sets out an approach for extracting the signal from uncertain data. It describes a two-step estimation procedure in which the history of past revisions are first used to estimate the parameters of a measurement equation describing the official published estimates. These parameters are then imposed in a maximum likelihood estimation of a state space model for the macroeconomic variable.

Keywords: Real-time data analysis; State space models; Data uncertainty; Data revisions
JEL codes: C32, C53

[*] This paper represents the views and analysis of the authors and should not be thought to represent those of the Bank of England, Monetary Policy Committee, or any other organisation to which the authors are affiliated. We would like to thank the associate editor for extremely helpful comments on an early version of the paper. We have further benefited from helpful comments from Andrew Blake, Spencer Dale, Kevin Lee, Lavan Mahadeva, Tony Yates and Shaun Vahey. We would also like to thank Simon Van Norden for his constructive and incisive comments especially on the relation between our model and comparable recent models in the literature.

1 Introduction

Most macroeconomic data are uncertain - they are estimates rather than perfect measures. Measurement errors may arise because data are based on incomplete samples or because many variables - for example, in-house software investment - are not easily observable. This necessitates the use of proxies. Without objective measures of data quality, it is difficult to gauge the potential for measurement errors. One symptom of data uncertainties is the propensity of statistical agencies to revise their estimates in the light of new information (larger samples) or methodological advances (better proxies). In the United Kingdom, the National Accounts are subject to a rich revisions process - staff at the Office for National Statistics (ONS) work through the implications of any changes to methodology for back data. As a result, past revisions give an indication of the likely incidence of revisions in the future and provide a measure of the potential for measurement errors surrounding the latest published estimates.

In practice, revisions have often appeared large relative to the variation observed in the published data. For example, the variance of revisions to the first *Quarterly National Accounts* estimates of real GDP growth was 0.08 percentage points over the period since 1993; compared with a variance of 0.07 percentage points in the latest estimates of quarterly GDP growth. This issue is by no means unique to the United Kingdom: see Mitchell (2004) for a review of work establishing the scale of historical revisions and Öller and Hansson (2002) for a cross-country comparison.

Uncertainty about the true profile of economic series now and in the past adds to the challenge of forming a forward-looking assessment of economic prospects and hence complicates policy formulation. Revisions to the recent profile of macroeconomic data may affect the forecasts generated by economic models. Taking published data at face value - ignoring the potential for future revisions - may result in avoidable forecast errors.

The data-user need not, however, treat uncertain data in such a naïve way. Indeed, there is some evidence that data-users have allowed for data uncertainties in interpreting macroeconomic data. In reviewing revisions to the United Kingdom's National Accounts, the Statistics Commission (2004) concluded that "the main users of the statistics knew that revisions should be expected, understood the reasons for them, and were able to make some allowance for them when taking important decisions." In other words, data-users appear to be aware that macroeconomic data provide a noisy signal of the current conjuncture.

One strategy that the data-user might adopt in the face of uncertainty in estimates of the past is to amend her model estimation strategy to recognise the imperfect signal in the published official data. For example, Harrison, Kapetanios, and Yates (2005) suggest that where measurement uncertainties are present in estimates of the recent past, models that downweigh recent 'experience' may have a superior forecasting performance to models in which all observations are weighted equally. In a similar vein, Jääskelä and Yates (2005) explore the implications of uncertain data for performance of competing simple policy rules. The intuition they develop is that the greater the uncertainty in current data compared to lagged data, the greater the weight on the lagged data should be.

However, integrating data uncertainty into model estimation strategies in this way adds to the complexity of model building and interpretation - the mapping from published official estimates to forecast economic variables conflates estimation of economic relationships with estimates of the signal contained in the published data. Such costs may be acute in a practical policy setting because of policymakers' preference for picking from a wide range of models appropriate for analyzing different economic developments; as described in Bank of England (1999). An alternative strategy is to unbundle the treatment of data uncertainty from estimation of specific forecasting models - first estimating the 'true' value of economic data and then using those estimates to inform economic modelling and forecasting.

This paper explores that signal extraction problem more formally. As long as revisions tend to improve data estimates - moving them towards the truth - the problem boils down to predicting the cumulative impact of revisions on the latest estimates of current and past activity. In addressing this problem, our paper contributes to a growing and long-standing literature on modelling revisions (or real-time analysis), of which Howrey (1978) was an early proponent.

1.1 An overview of the literature

One common approach to prediction of revisions is to estimate 'true' data using some form of state space model. One very simple possible setting would be to assume that: published data are unbiased; measurement errors are i.i.d; uncertainties are resolved after a single round of revisions; and that no alternative indicators are available. Then, the solution of the signal extraction problem is simply a matter of estimating the signal to noise ratio attaching to the preliminary estimates.

Early papers extended this basic story by allowing for any systematic biases apparent in previous preliminary estimates. Such biases appear to have been endemic in National Accounts data in the United Kingdom and elsewhere, as documented for example in Akritidis (2003), and Garratt and Vahey (2006). Early papers also allowed for serial correlation across releases - that is that errors in today's measure of activity in 1999 might be related to errors in yesterday's measure of growth in 1999. However, a number of features of real-time National Accounts data were left unexplored. Indeed, in a detailed review of the literature, Jacobs and Van Norden (2006) charge that the early papers "impose data revision properties that are at odds with reality". Recent papers have sought to enrich the representation on a number of fronts.

Most authors consider only the statistical agency's estimates as candidate measures. Ashley, Driver, Hayes, and Jeffery (2005) suggest weighting the signal extracted from alternative indicators in proportion to past performance in predicting revisions. Jacobs and Sturm (2006) model competing indicators more formally in a state space setting. Considering alternative measures in this way appears consistent with the wide array of indicators monitored by policymakers, see Lomax (2004), and is the approach pursued in this paper.

Following Howrey (1978), several papers restrict attention to revisions occurring in the first few quarters after the preliminary release. Assuming that estimates become ‘true’ after a few quarters is, however, violated by the presence of revisions to more mature estimates. Subsequent papers have explored a variety of approaches to dealing with the uncertainty surrounding more mature estimates. Some, such as Patterson (1994) and Garratt, Lee, Mise, and Shields (2005), increase the number of releases in the model so that estimates are not assumed to become ‘true’ for two or three years. In the case of the United Kingdom’s National Accounts, however, revisions have been applied to even more mature estimates. An alternative, followed by Jacobs and Van Norden (2006), is to restrict the model to a few maturities but allow that measurement errors may be non-zero for the most mature release modelled. Finally, Kapetanios and Yates (2004) impose an asymptotic structure on the data revision process - estimating a decay rate for measurement errors rather than separately identifying the signal to noise ratio for each maturity. The benefit of modelling the relationship between measurement errors of differing maturities in this way is that they can capture revisions to quite mature data relatively parsimoniously.

Many authors allow for serial correlation across releases, see, for example, Howrey (1984). Jacobs and Van Norden (2006) argue that spillovers in measurement errors within any release may be more important; in other words, that errors in today’s measure of growth in a given past period may be related to errors in today’s measure of growth in another past period.

Early models assumed measurement errors to be independent of the ‘true’ state. In an influential paper, Mankiw and Shapiro (1986) challenged whether early estimates should be viewed as ‘noisy’ in this way or whether we might expect some correlation with the level of activity, which they termed ‘news’. Ignoring such a correlation could lead models to underweight uncertain data. Jacobs and Van Norden (2006) propose a model that captures both ‘noise’ and ‘news’ elements.

The model developed in this paper extends the above literature with respect to a number of features. The set of available measures is expanded to include alternative indicators while the representation of measurement errors attaching to the latest official estimates allows for serial correlation, correlation with the true profile and for revisions to be made to quite mature estimates as well as the preliminary data releases. In allowing for mature data to be revised, we follow Kapetanios and Yates (2004) and assume the variance of measurement errors decays asymptotically.

The paper is structured as follows. Section 2 represents the signal extraction problem in state space. Section 3 describes the estimation strategy adopted; focusing on the use of the statistical properties of past revisions to estimate some parameters of the state space model. We also present the results of a small simulation exercise and an empirical illustration. Section 4 provides an illustrative example using United Kingdom investment data. Finally, Section 5 concludes.

2 A State Space Model of Uncertain Data

In this section, we present a state space representation of the signal extraction problem. Recognising that analysis of the latest official data may be complemented by business surveys and other indirect measures, we allow for an array of measures of each macroeconomic variable of interest. Then, for each variable of interest, the model comprises alternative indicators, a transition law and separate measurement equations describing the latest official estimates. The measurement equation is designed to be sufficiently general to capture the patterns in revisions observed historically for a variety of United Kingdom National Accounts aggregates.

The model is presented in a vector notation, assuming m variables of interest. However, we simplify estimation by assuming block diagonality throughout the model so that the model can be estimated on a variable-by-variable basis for each of the m elements in turn. One cost of this simplification is that estimates of the 'true' value of the various elements of National Accounting identities will not necessarily satisfy the accounting identities. In practical application of the model, it is relatively trivial to balance estimates as a post-model step - following Weale (1985) in allocating any accounting identity residual arising from estimation of the state space model across elements, to minimise some loss function.

2.1 The model for the true data

Let the m dimensional vector of variables of interest that are subject to data uncertainty at time t be denoted by $\mathbf{y}_t, t = 1, \ldots, T$. The vector $\mathbf{y}_t$ contains the unobserved true value of the economic concept of interest.

We assume that the model for the true data $\mathbf{y}_t$ is given by

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{i=1}^{q} \mathbf{A}_i \mathbf{y}_{t-i} + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\mathbf{A}_1, \ldots, \mathbf{A}_q$ are $m \times m$ matrices, $\mathbf{A}(L) = \mathbf{I}_m - \mathbf{A}_1 L - \ldots - \mathbf{A}_q L^q$ is a lag polynomial whose roots are outside the unit circle, $\boldsymbol{\mu}$ is a vector of constants, $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{mt})'$ and $\mathrm{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, where we denote the main diagonal of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ by $\boldsymbol{\sigma}_{\boldsymbol{\epsilon}}^2 = (\sigma_{\boldsymbol{\epsilon}_1}^2, \ldots, \sigma_{\boldsymbol{\epsilon}_m}^2)'$. We further assume that $\mathbf{A}_1, \ldots, \mathbf{A}_q$ are diagonal, so that the true value of each variable of interest is related only to its own historical values.

This representation has a number of limiting features in practical application. First, because we assume stationarity of $\mathbf{y}_t$, the model is more likely to be applicable to differenced or detrended macroeconomic data than to their levels. Second, we assume linearity for $\mathbf{y}_t$. Although this may be a restrictive assumption, it is unclear to what extent we can relax it as assuming one particular form of non-linearity is likely to be restrictive as well. Finally, because we assume $\mathbf{A}_1, \ldots, \mathbf{A}_q$ are diagonal, we do not consider transition laws that exploit prior views of any behavioural relationship between the variables of interest.

2.2 The statistical agency's published estimate

Let $\mathbf{y}_t^{t+n}$ denote a noisy estimate of $\mathbf{y}_t$ published by the statistical agency at time $t + n$, where $n = 1, \ldots T - t$. The model for these published data is

$$\mathbf{y}_t^{t+n} = \mathbf{y}_t + \mathbf{c}^n + \mathbf{v}_t^{t+n} \tag{2}$$

where $\mathbf{c}^n$ is the bias in published data of maturity n and $\mathbf{v}_t^{t+n}$ the measurement error associated with the published estimate of $\mathbf{y}_t$ made at maturity n.

One of the main building blocks of the model we develop is the assumption that revisions improve estimates so that official published data become better as they become more mature. Reflecting this assumption, both the bias in the published estimates and the variance of measurement errors are allowed to vary with the maturity of the estimate - as denoted by the n superscript. Note also that the latest data release $(\mathbf{y}_{T-i}^{T-i+1}, \ldots, \mathbf{y}_{T-1}^{T})'$ includes data points of differing maturities ranging from preliminary estimates of the most recent past through more mature observations of data points that were first measured some years previously.

The constant term $\mathbf{c}^n$ is included in equation (2) to permit consideration of biases in the statistical agency's data set. Specifically, we model $\mathbf{c}^n$ as

$$\mathbf{c}^n = \mathbf{c}^1(1+\lambda)^{n-1}, \tag{3}$$

where $\mathbf{c}^1$ is the bias in published data of maturity $n = 1$ and λ describes the rate at which bias decays as estimates become more mature $(-1 < \lambda < 0)$. This representation assumes that the bias tends monotonically to zero as the estimates become more mature. It is possible that other specifications for the bias might fit the revisions history of specific variables better.

We assume that the measurement errors, $\mathbf{v}_t^{t+n}$, are distributed normally with finite variance. We allow serial correlation in $\mathbf{v}_t^{t+n}$. Specifically, we model serial correlation in the errors attaching to the data in any data release published at $t+n$, as

$$\mathbf{v}_t^{t+n} = \sum_{i=1}^{p} \mathbf{B}_i \mathbf{v}_{t-i}^{t+n} + \boldsymbol{\varepsilon}_t^{t+n}, \tag{4}$$

where $\mathbf{B}_1, \ldots, \mathbf{B}_p$ are $m \times m$ matrices, $\mathbf{B}(L) = \mathbf{I} - \mathbf{B}_1 L - \ldots - \mathbf{B}_p L^p$ is a matrix lag polynomial whose roots are outside the unit circle and $\boldsymbol{\varepsilon}_t^{t+n} = (\varepsilon_{1t}^{t+n}, \ldots, \varepsilon_{mt}^{t+n})'$ and $\mathrm{E}(\boldsymbol{\varepsilon}_t^{t+n}(\boldsymbol{\varepsilon}_t^{t+n})') = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^n$ as we are allowing for heteroscedasticity in measurement errors with respect to n. Equation (4) imposes some structure on $\mathbf{v}_t^{t+n}$ because we assume a finite AR model whose parameters do not depend on maturity. The representation picks up serial correlation between errors attaching to the various observations within each data release. We further assume that $\mathbf{B}_1, \ldots, \mathbf{B}_p$ are diagonal, so that the measurement errors attaching to published estimates of each of the m variables are treated independently from the measurement errors of the other variables.

Further, we allow that $\boldsymbol{\varepsilon}_t^{t+n}$ and therefore $\mathbf{v}_t^{t+n}$ has heteroscedasticity with respect to n. Specifically, we model the main diagonal of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^n$ as $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^n}^2 = (\sigma_{\varepsilon_1^n}^2, \ldots, \sigma_{\varepsilon_m^n}^2)'$, where $\sigma_{\varepsilon_i^n}^2 = \mathrm{E}(\varepsilon_{it}^{t+n})^2$. For future reference we also define $\sigma_{v_i^n}^2 = \mathrm{E}(v_{it}^{t+n})^2$. The model for $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^n}^2$ is given by

$$\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^n}^2 = \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^1}^2 (1+\delta)^{n-1}, \tag{5}$$

where $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^1}^2$ is the variance of measurement errors at maturity $n = 1$ and δ describes the rate at which variance decays as estimates become more mature $(-1 < \delta < 0)$. This representation imposes structure on the variance of measurement errors, because we assume that the variance declines monotonically to zero as the official published estimates become more mature. A monotonic decline in measurement error variances is consistent with models of the accretion of information by the statistical agency, such as that developed in Kapetanios and Yates (2004). We put forward three reasons for using this specification. Firstly, this model is parsimonious since it involves only two parameters. Secondly, δ has an appealing interpretation as a rate at which revision error variances decline over time. Thirdly, and perhaps most importantly Kapetanios and Yates (2008) provide empirical evidence in favour of this specification. In particular, tests of overidentifying restrictions implied by this specification cannot be rejected for any series in the United Kingdom National Accounts data.

Over and above any serial correlation in revisions, we allow that measurement errors be correlated with the underlying true state of the economy, $\mathbf{y}_t$. This correlation relates to the degree of 'news' and 'noise' inherent in published estimates - addressing the challenge posed by Mankiw and Shapiro (1986). We specify that $\boldsymbol{\varepsilon}_t^{t+n}$ be correlated with shock $\boldsymbol{\epsilon}_t$ to the transition law in equation (1), so that, for any variable of interest

$$\mathrm{cov}(\epsilon_{it}, \varepsilon_{it}^{t+n}) = \rho_{\epsilon\varepsilon}\sigma_{\epsilon_i}\sigma_{\varepsilon_i^n}. \tag{6}$$

In principle, the model in equation (2) could be applied to previous releases as well as the latest estimates. One natural question is whether data-users should consider these previous releases as competing measures of the truth - that is, using $\mathbf{y}_t^{t+n-j}$ alongside $\mathbf{y}_t^{t+n}$ as measures of $\mathbf{y}_t$. In contrast with the treatment in much of the antecedent literature, we decide to exclude earlier releases from the set of measures used to estimate 'true' activity, see, for example, Garratt, Lee, Mise, and Shields (2005). The reason for using only the latest release is pragmatic. In principle, given that empirical work across a variety of data sets has found that revisions appear to be forecastable, using earlier releases should be useful. In practice, however, such a model would be complex. That complexity may be costly in various ways - the model would be more difficult to understand, more cumbersome to produce and potentially less robust when repeatedly reestimated. Given the importance of robustness in repeated reestimation, we feel this choice is justified. Further, by focusing on the latest release we are able to specify a model that is quite rich in its specification of other aspects of interest, such as heteroscedasticity, serial correlation and correlation with economic activity.

We note that there are circumstances where using only the latest release is theoretically optimal. An example of a set of such circumstances is provided in Appendix A. The model developed in the appendix makes a number of assumptions that imply a form of rational behaviour on the part of the statistical agency, which may well not hold in practice. Therefore, we must stress that such a model is restrictive. For example, it implies that revisions are not forecastable which contradicts the empirical evidence. Further, our modelling approach is obviously parametric and therefore has claims to efficiency only if, on top of rationality on the part of the statistical agency, the specification of the model for the unobserved true variable is correct. On the other hand, note that the use of such a parametric model for the unobserved variable can provide benefits as well. Even if the statistical agency is operating optimally in data collection, our state space model can provide further benefits by positing a model for $\mathbf{y}_t$, since that is not a part of the statistical agency's specification. A final point we should note is that previous releases are used to estimate bias and measurement error parameters as discussed in Section 3.2.

2.3 The alternative indicators

In addition to the statistical agency's published estimate, the data-user can observe a range of alternative indicators of the variable of interest; such as private sector business surveys. We denote the set of these indicators by $\mathbf{y}_t^s, t = 1, \ldots, T$. Unlike official published estimates, the alternative indicators need not be direct measures of the underlying variables. For example, private sector business surveys typically report the proportion of respondents answering in a particular category rather than providing a direct measure of growth. We assume the alternative indicators to be linearly related to the true data

$$\mathbf{y}_t^s = \mathbf{c}^s + \mathbf{Z}^s \mathbf{y}_t + \mathbf{v}_t^s. \tag{7}$$

The error term $\mathbf{v}_t^s$ is assumed to be i.i.d with variance $\mathbf{\Sigma}_{\mathbf{v}^s}$. This, of course, is more restrictive than the model for the official data. Simple measurement equations of this form may not be appropriate for all the alternative indicators used in routine conjunctural assessment of economic activity. One natural extension of the model presented would be to consider the potential for serial correlation in the measurement errors attaching to alternative indicators - recognising that business surveys often have a smoother profile than the related National Accounts variables. In particular, the model does not exploit any heteroscedasticity or serial correlation in measurement errors associated with the indicators; any correlation between the true state of the economy and the measurement errors surrounding the alternative indicators; or any correlation between the measurement errors attaching to the alternative indicators and those attaching to the published estimates.

2.4 The full model and further considerations

To summarise the model, we give its complete state space form for the latest available release. The model treats the most recent release of data published by the statistical agency and any alternative indicators as measures of the variable of interest. The state space representation of the model is

$$\begin{pmatrix} \mathbf{y}_t^T \\ \mathbf{y}_t^s \end{pmatrix} = \begin{pmatrix} \mathbf{c}^n \\ \mathbf{c}^s \end{pmatrix} + \begin{pmatrix} \mathbf{I} & \dots & \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} \\ \mathbf{Z}^s & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y}_t \\ \vdots \\ \mathbf{y}_{t-q+1} \\ \mathbf{v}_t^T \\ \vdots \\ \mathbf{v}_{t-p+1}^T \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{v}_t^s \end{pmatrix}, \tag{8}$$

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-q+1} \\ \mathbf{v}_t^T \\ \mathbf{v}_{t-1}^T \\ \vdots \\ \mathbf{v}_{t-p+1}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{A}_1 & \dots & \dots & \mathbf{A}_q & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{B}_1 & \dots & \dots & \mathbf{B}_p \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-q} \\ \mathbf{v}_{t-1}^T \\ \mathbf{v}_{t-2}^T \\ \vdots \\ \mathbf{v}_{t-p}^T \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \boldsymbol{\varepsilon}_t^T \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}. \tag{9}$$

Having completed the presentation of the model, it is worth linking our work to the literature that deals with the presence of measurement error in regression models. A useful summary of the literature can be found in Cameron and Trivedi (2005). This body of work is of interest as it can provide solutions to a number of problems caused by the presence of data revisions. In the context of the following simple regression model

$$z_t = \beta y_t + u_t \tag{10}$$

use of y_t^{t+1} as a proxy for y_t can lead to a bias in the OLS estimator of β. Then, the use of later vintages, y_t^{t+n}, $n = 2, ..., T - t$, as instruments in (10) can be of use for removing the bias in the estimation of β. One issue of relevance in this case is to choose if all available vintages should be used as instruments. The rapidly expanding literature on optimal selection of instruments, see, for example, Donald and Newey (2001), suggests useful tools for this purpose. Our analysis provides an alternative method of addressing this problem. In our modelling framework, equation (10) becomes a further measurement equation of the state space model and the overall estimation of the resulting model can provide unbiased estimates of β. However, our current state space formulation is of further interest since on top of giving estimates for relevant parameters it also gives an alternative and possibly superior proxy for the unobserved true series, in the form of an estimate for the state variable. This, can then be used for a variety of purposes including forecasting.

3 Estimation of the State Space Model

In this section, we discuss the strategy adopted in estimating the model. Section 3.1 outlines the creation of a real-time database and Section 3.2 discusses the use of real-time data for estimating bias and measurement error parameters. Section 3.3 summarises the results of a Monte Carlo simulation exercise aimed at establishing the model's performance relative to taking published estimates at face value.

The estimation is performed in two steps: first using the revisions history to estimate equations (2) through (6); and then, as a second step, estimating the remaining parameters via maximum likelihood using the Kalman filter. Approaching estimation in two steps simplifies greatly the estimation of the model and has the additional benefit of ensuring that the model is identified. Were all parameters to be estimated in one step, the state space problem represented by equations (8) and (9) would not always satisfy the identification conditions described in Harvey (1989).

3.1 Extracting revisions form real-time data

In recent years, a number of real-time data sets have been developed - describing the evolution of estimates through successive data releases. Using this real-time data to estimate the parameters in (2) to (6) requires us to first manipulate the real-time data set to derive a matrix of revisions to published data of differing maturities.

The real-time data set for each variable of interest is an upper-triangular data matrix with publication dates ordered horizontally and reference dates vertically down. Each column represents a new release of data published by the statistical agency, and each release includes observations of differing maturities. By way of illustration, Table 1 shows an extract of the real-time database for whole economy investment used in the illustrative example presented in Section 4; and Table 2 shows the maturity of the various observations.

Table 1 Extract from the real-time database for quarterly growth of whole economy investment

		Release date				
		2003 Q1	2003 Q2	...	2006 Q3	2006 Q4
Reference date	2002 Q4	-0.15	0.16	...	3.51	3.51
	2003 Q1		-1.13	...	-3.18	-3.18
	⋮			⋱	⋮	⋮
	2006 Q2				1.31	1.21
	2006 Q3					1.32

Table 2 Stylised real-time database - maturity of observations

		Release date				
		2003 Q1	2003 Q2	...	2006 Q3	2006 Q4
Reference date	2002 Q4	1	2	...	15	16
	2003 Q1		1	...	14	15
	⋮			⋱	⋮	⋮
	2006 Q2				1	2
	2006 Q3					1

Define the revisions to published estimates of an individual variable between maturities n and $n + j$ as

$$w_t^{n,j} = y_t^{t+n+j} - y_t^{t+n}. \tag{11}$$

For estimation purposes, we take revisions over the J quarters subsequent to each observation to be representative of the uncertainty surrounding that measure of activity. For example, with $J = 24$, we evaluate uncertainties surrounding data of maturity 1 by considering revisions between the 1st and 25th release; and we evaluate uncertainties surrounding data of maturity 12 by considering revisions between the 12th and 36th release.

If the real-time data set contains W releases of data, and we are interested in the properties of N maturities, we can construct an $N \times (W - J)$ matrix of revisions $\mathbf{W}^J$, over which to estimate the parameters of equations (2) through (6). Each column of the matrix $\mathbf{W}^J$ contains observations of revisions to data within a single data release. Each row describes revisions to data of a specific maturity n. N and J are both choice variables and should be selected to maximise the efficiency of estimation of the parameters driving equations (2) to (6). There is a trade-off between setting J sufficiently large to pick up all measurement uncertainties and retaining sufficient observations for the estimated mean, variance, and serial correlation of revisions and their correlation with mature data to be representative. In the remainder of the paper we arbitrarily set $N = J = 20$.

3.2 Estimation of bias and measurement error parameters

We use the sample of historical revisions in matrix $\mathbf{W}^J$ to estimate c^1 and λ quite trivially. Recall that we assume $\mathbf{B}_1, \ldots, \mathbf{B}_p$ to be diagonal. As a result, the functions can be estimated for individual variables rather than for the system of all variables of interest. In the remainder of this section, we therefore consider estimation for a single variable and discard vector notation. The sample means of revisions of each maturity $n = 1$ to N are simply the average of observations in each row of $\mathbf{W}^J$. Denoting the average revision to data of maturity n by mean$(w^{n,J})$, the parameters c^1 and λ are then estimated from the moment conditions mean$(w^{n,J}) = c^1(1+\lambda)^{n-1}$ via GMM, where $-1 < \lambda < 0$.

We cannot use historical revisions to estimate $\rho_{\epsilon\varepsilon}$ directly, because neither ϵ_t nor ε_t^{t+n} are observable. But we can use the historical revisions to form an approximation of ρ_{yv} - denoted ρ_{yv}^*. The manipulation in obtaining $\rho_{\epsilon\varepsilon}$ from ρ_{yv}^* is summarised in Appendix B. We start by estimating ρ_{yv}^*. We can readily calculate the correlation between revisions to data of maturity n and published estimates of maturity $J+n$, denoted by $\rho_{yv}^{n} = \operatorname{corr}(y_t^{t+J+n}, w_t^{n,J})$. Averaging across the N maturities in $\mathbf{W}^J$ gives an average maturity-invariant estimate of ρ_{yv}^{*}. Where the variance of measurement errors decays sufficiently rapidly, we do not introduce much approximation error by taking this correlation with mature published data as a proxy for the correlation with the true outcome, y_t. We do not apply any correction for this approximation because derivation of any correction would require untested assumptions about the relationship between measurement errors across successive releases (such as those described in Appendix A) which we do not wish to impose on the model.

The variance-covariance matrix of historical revisions may be used to jointly estimate both the heteroscedasticity in measurement errors and their serial correlation. This requires us to first express the variance-covariance matrix of measurement errors as a function of the parameters in equations (4) and (5) and then to estimate the parameters consistent with the observed variance-covariance matrix of revisions.

Assuming for simplicity first-order serial correlation in the measurement errors, we can easily build-up a full variance-covariance matrix at any point in time. The variance-covariance matrix of the measurement errors in the most recent N maturities, will be invariant with respect to t and is given by

$$\mathbf{V} = \frac{\sigma_{\varepsilon^1}^2}{1-(1+\delta)\beta_1^2} \times \begin{pmatrix} 1 & (1+\delta)\beta_1 & \cdots & (1+\delta)^{N-1}\beta_1^{N-1} \\ (1+\delta)\beta_1 & (1+\delta) & \cdots & (1+\delta)^{N-1}\beta_1^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ (1+\delta)^{N-1}\beta_1^{N-1} & (1+\delta)^{N-1}\beta_1^{N-2} & \cdots & (1+\delta)^{N-1} \end{pmatrix} \tag{12}$$

A sample estimate of the variance-covariance matrix $\hat{\mathbf{V}}$ can be calculated trivially from the matrix of historical revisions $\mathbf{W}^J$. Taking the variance-covariance matrix to the data, we can estimate $\beta_1, \sigma_{\varepsilon^1}^2$ and δ via GMM by minimising

$$(\operatorname{vec}(\mathbf{V}) - \operatorname{vec}(\hat{\mathbf{V}}))'(\operatorname{vec}(\mathbf{V}) - \operatorname{vec}(\hat{\mathbf{V}})). \tag{13}$$

The derivation of the variance-covariance matrix for higher lag-orders requires some further manipulation, as outlined in Appendix B. It is worth noting here that there exist a interesting special case where the first step estimation does not affect the second step ML estimation via the Kalman filter. This is the case where the number of available vintages, N, tends to infinity. In this case, the GMM estimation outlined above, results in parameter estimates that are $\sqrt{NT}$ consistent whereas the second step ML estimation is only $\sqrt{T}$ consistent implying that the parameters that are estimated in the first step can be treated as known for the second step and the resulting approximation error associated with the first step estimation is asymptotically negligible.

More generally, the fact that more data are used in the first step implies that the variability of the first step estimates is likely to be lower than that of the second step estimates. However, the use of a two step estimation procedure implies that, in practice, the variability of the first step estimates is not taken into account when the likelihood based second step variance estimates are obtained. As pointed out above, the advantages of the two step estimation, in our view, outweigh this disadvantage.

Of course, if the variances of the parameter estimates are of particular interest, a parametric bootstrap can provide a standard avenue for obtaining variance estimates that implicitly take into account the variability arising out of both estimation steps. The parametric bootstrap would have to replicate both steps of the two-step estimation procedure to capture appropriately the parameter uncertainty associated with the first step estimation. However, note that the validity of the bootstrap in this two-step estimation context has not been formally shown, to the best of our knowledge, in the relevant literature. Further, use of the bootstrap requires the specification of a model for all vintages used in the first step GMM estimation, which may be problematic in practice. For these reasons, we provide standard errors for the estimated parameters obtained from the second estimation step, using standard likelihood based inference.

3.3 Monte Carlo simulations

As a check on the small-sample performance of our estimator, we run a simple Monte Carlo simulation exercise. The focus of the exercise is on the performance of the model in fitting the true state, $\mathbf{y}_t$, rather than on the estimation of specific parameters.

The data are generated according to the model described by equations (8) and (9). It is assumed that the model is of quarterly growth, with only one release per quarter. We assume only one variable of interest, y_t, that evolves as an AR(1) process. The constant in the true model is set to $\mu = 0$. For further simplicity we assume $c^n = c^1 = 0$. This reduces the complexity of the model. For the measurement errors we also assume an AR(1) process. Further, we assume no additional indicators are available. The output of the model is an estimate of the true state prevailing in each period. The model is estimated over a sample of length $T = 100$; corresponding to 25 years of data. We run 1000 replications in total for each parametrisation and the results presented are averages over the replicates.

We evaluate the properties of the model across differing assumptions about the degree of persistence in the transition law and the measurement errors for the official estimates - assigning the AR coefficients α and β, values 0.1 and 0.6. We also consider different assumptions about the degree of correlation between transition shocks and measurement errors - setting $\rho_{\epsilon\varepsilon} = -0.5, 0$ and 0.5. We set the heteroscedasticity decay parameter to $\delta = -0.05$; broadly in line with the decay rates found in the revisions history to United Kingdom National Accounts data since 1993. We have not explored alternative values. The transition error, ϵ_t, and the error of the measurement error, ε_t, are assumed to be $i.i.d.N(0,1)$. The variance of the measurement error at maturity one is $\sigma^2_{v_T^{T+1}} = 1$ implying that the signal to noise ratio is also one at maturity one.

We use the simulation results to gauge the degree to which using the model is superior relative to taking the latest published estimate, $\mathbf{y}_t^{t+n}$, at face value. The metric used is the standard deviation of the difference between the smoothed estimates of the truth and the unobserved truth across replications and relative to the standard deviations of the difference between the latest published estimate and the unobserved truth. We evaluate this metric separately for each maturity of the latest data to check whether any performance gain is restricted to recent maturities.

Figure 1 compares the performance of estimated and published data for $\alpha = 0.6, \beta = 0.1$ and $\rho_{\epsilon\varepsilon} = 0$. The model has a smaller standard deviation of prediction errors than the published data for all maturities up to 58 quarters. Thereafter, the measurement errors attaching to the published estimates have become small enough so that any gains from filtering are more than offset by parameter uncertainties. Table 3 contains Monte Carlo results for various combinations of parameters α, β and $\rho_{\epsilon\varepsilon}$. The results show that the model performs always better than taking published data at face value for the first 18 maturities.

Figure 1 Standard deviation of errors in predicting y_t

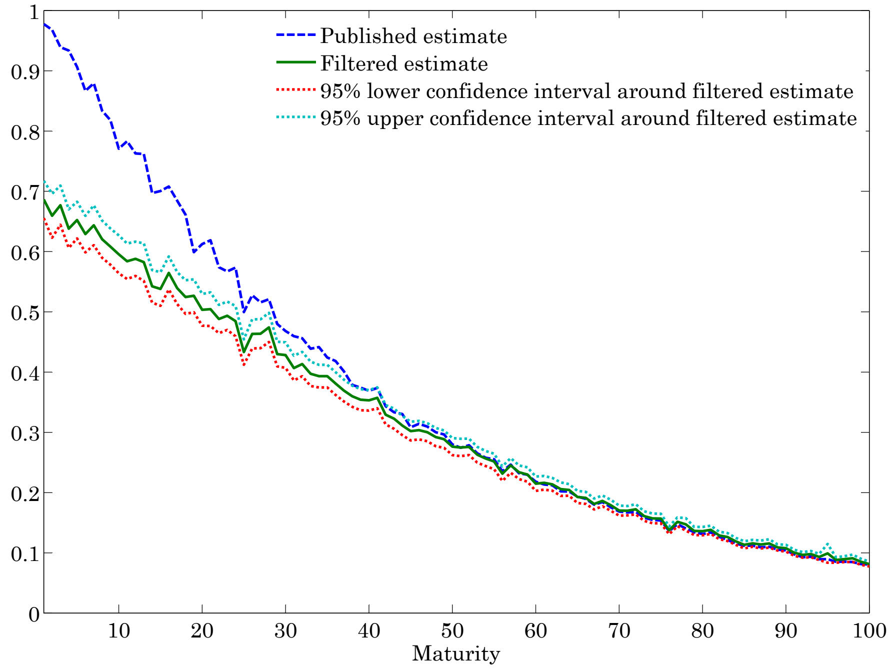


Table 3 Gains from filtering (in %) and the earliest maturity at which published data outperform filter estimates

			Gain at maturity		Earliest
$\rho_{\epsilon\varepsilon}$	α	β	1	9	maturity
0.5	0.1	0.1	47.7	43.6	-[a]
0.5	0.1	0.6	47.4	41.2	80
0.5	0.6	0.1	51.2	46.4	-[a]
0.5	0.6	0.6	46.0	39.0	70
0	0.1	0.1	30.3	19.9	52
0	0.1	0.6	31.2	26.1	41
0	0.6	0.1	29.8	25.7	58
0	0.6	0.1	29.2	18.8	42
-0.5	0.1	0.1	12.4	6.0	18
-0.5	0.1	0.6	17.0	10.3	23
-0.5	0.6	0.1	16.5	11.1	26
-0.5	0.6	0.6	9.7	6.3	18

[a] The filter outperforms the published data at all evaluated maturities.

4 An Illustrative Example

As an illustrative example, we apply the state space model to quarterly growth of whole-economy investment. The Bank of England's real-time data set was described in Castle and Ellis (2002) and includes published estimates of investment from 1961. The Bank of England's real-time data set is available at www.bankofengland.co.uk/statistics/gdpdatabase. As an indicator, we consider the British Chambers of Commerce's Quarterly Survey. Specifically, the balance of service sector respondents reporting an upward change to investment plans over the past three months. This is an arbitrary choice made to explore the functioning of the model rather than following from any assessment of competing indicators. We do not provide such an assessment as part of this example. We restrict estimation to the period 1993 to 2006 because an earlier study of the characteristics of revisions to the United Kingdom's National Accounts (Garratt and Vahey (2006)) found evidence of structural breaks in the variance of revisions to National Accounts aggregates in the years following the *Pickford Report*.

4.1 Estimation results

Table 4 sets out some summary statistics describing the revisions history of published data of differing maturities - evaluating revisions over a 20-quarter window as discussed in Section 3.1. Table 5 reports estimated heteroscedasticity, bias, serial correlation and correlation parameters.

Table 4 Quarterly growth of whole economy investment - revisions summary statistics, 1993Q1 to 2006Q4

	Maturity					
	1	4	8	12	16	20
Mean	0.49	0.32	0.22	0.31	0.03	0.11
p-value[a]	*0.41*	*0.23*	*0.37*	*0.14*	*0.76*	*0.44*
Variance	3.09	3.28	2.26	1.65	1.35	1.57
p-value[b]	-	*0.18*	*0.03*	*0.00*	*0.00*	*0.00*
Mean > 0	1.70	1.49	1.25	1.13	0.85	0.96
Mean < 0	-1.21	-1.51	-1.07	-0.85	-0.88	-0.96
Skewness	-0.08	-0.55	-0.16	-0.05	-0.74	-0.22
Excess kurtosis	-0.67	0.06	-0.06	0.60	1.24	0.77

[a] p-value of a test that mean revision are zero at each maturity.

[b] p-value of a test that revisions variance at each maturity is smaller than revisions variance at maturity one.

The summary statistics suggest that, on average, upward revisions have been larger magnitude than downward revisions. However, the null hypothesis that mean revisions are zero cannot be rejected at the 5% level for any maturity. The variance of revisions is 3.09 percentage points for estimates with a maturity of one quarter. That is similar to the variance of whole-economy investment growth (3.12 percentage points). For immature data there is little evidence of heteroscedasticity, but the variance of revisions does decline quite markedly once data have reached a maturity of 8 quarters. The null hypothesis that the variance of revisions is equal to that at maturity 1 is rejected at the 5% level for maturities beyond 8 quarters.

The bias was not found to be significant and hence was excluded from the model. This is not surprising given that Table 4 shows bias to be insignificant at all maturities. The measurement error variance parameters also map fairly easily from the summary statistics quoted in Table 4. The variance decay pa-

Table 5 Quarterly growth of whole economy investment - Estimated parameters

		Parameter	Standard error
Initial variance	σ^2_{v1}	3.584	0.296
Variance decay	δ	-0.058	0.013
Serial correlation	β_1	-0.220	0.055
Correlation with data	ρ^*_{yv}	-0.315	0.162

rameter, δ, suggests a half-life for measurement errors of 12 quarters. There is significant first order negative serial correlation across revisions: successive quarters of upward/downward revision are therefore unusual. Revisions appear to have been negatively correlated with mature estimates, although the parameter is only significant at the 10% level.

Table 6 reports the parameters estimates from the Kalman filter, while Table 7 sets out some standard diagnostic tests of the various residuals of the Kalman filter to give an indication of the degree to which modelling assumptions are violated in the data set. Higher orders of q were not found to be statistically significant, therefore the transition equation does not include an autoregressive component.

Both the prediction errors for the published ONS data and the smoothed estimates of the errors on the transition equations pass standard tests for stationarity, homoscedasticity and absence of serial correlation at the 5% level. Prediction errors are the ‘surprise’ in the observable variables (i.e. official published data and alternative indicators) given the information available about previous time periods. These errors enter into the prediction error decomposition of the likelihood function. Standard maximum likelihood estimation therefore assumes that these errors are zero-mean, independent through time, and normally distributed. If this is not the case, then the Kalman filter does not provide an optimal estimator of the unobserved states. The errors surrounding predictions for the indicator variable are less well-behaved. In particular, there is evidence of significant serial correlation in these residuals. We have assumed that residuals associated with the indicator variables are i.i.d. This assumption could be relaxed in future work.

We next turn to the estimate of quarterly growth of whole economy in-

Table 6 Estimated Kalman filter parameters

		Parameter	Standard error
True data parameters			
Constant	μ	1.177	0.238
Error variance	σ_ϵ^2	3.217	0.673
Indicator parameters			
Constant	c^s	1.177	0.219
Slope	Z^s	0.369	0.138
Error variance	$\sigma_{v^s}^2$	2.629	0.567

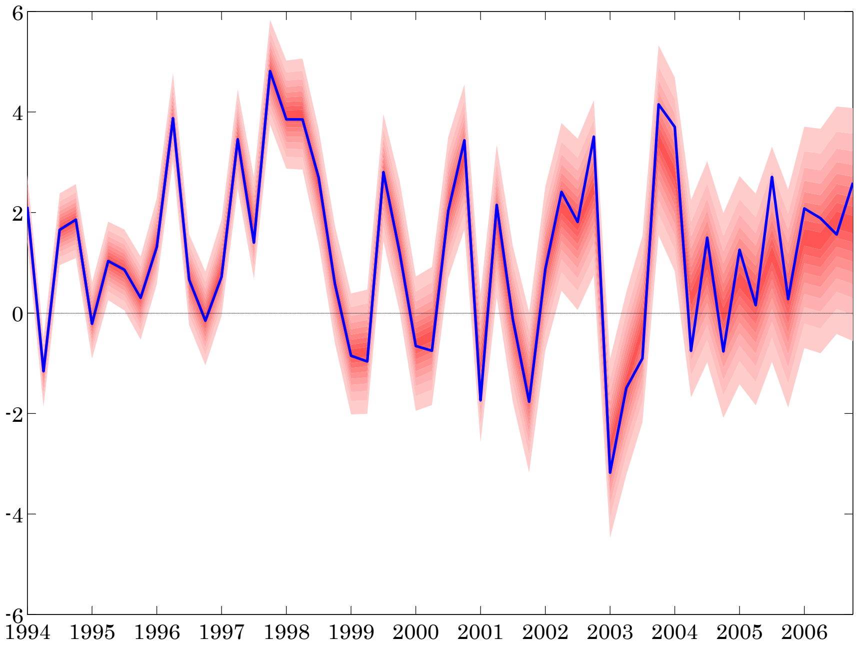
Table 7 Model residual diagnostics
Table reports p-values for all tests except for the ADF tests, where t-statistics is reported. Entries in bold indicate rejection of the null hypothesis at 5% significance level.

	$\hat{\epsilon}_t$	$\hat{v}_t^s$	$\hat{\varepsilon}_t^T$
ADF test: no constant or trend	**-6.114**	**-2.795**	**-5.405**
ADF test: constant, but no trend	**-6.054**	-2.781	**-5.346**
ADF test: constant and trend	**-5.984**	-3.439	**-5.401**
Normality test	0.598	0.921	0.891
Serial correlation test: 1 lag	0.313	**0**	0.061
Serial correlation test: 4 lags	0.538	**0**	0.294
ARCH test: 1 lag	0.069	**0.006**	0.166
ARCH test: 4 lags	0.401	0.064	0.646

vestment - that is, the smoothed backcast. Figure 2 reports the estimates of quarterly growth of whole economy investment. Following the presentational convention of the GDP and inflation probability forecasts (more commonly known as fan charts) presented in the Bank of England's *Inflation Report* each band contains 10% of the distribution of possible outcomes. In this application, because the normality assumption is not rejected by the data, the outer (90%) band is equivalent to a ± 1.6 standard error bound.

The central point of the fan chart tracks the statistical agency's published estimates quite closely once those estimates are mature. This follows from the fact that the heteroscedasticity and bias in measurement errors decline reasonably rapidly. Over the most recent past, the central point differs more materially. This mainly reflects the higher measurement error variance attaching to earlier releases.

Figure 2 Fan chart for quarterly growth of investment and the official estimate (solid line). Each band contains 10% of the distribution of possible outcomes.



4.2 Real-Time Evaluation of the State Space Model

In this subsection we provide an evaluation of the real-time performance of the model. For this experiment, the evaluation period starts at $s_0 = 1998Q1$ and terminates at $s_1 = 2002Q4$. That is the model is estimated and outputs are produced based on samples from $1993Q1$ to $1998Q1$. The estimation period is then extended to include observed data for the following time period, i.e. $1998Q2$. This is repeated until $2002Q4$, which gives 20 evaluation observations. For each run, we compare the performance of the smoothed backcast with that of the official published estimates available at the time the smoothed backcast was formed. Because each official data release includes data points of differing maturities, we evaluate backcasting performance for each maturity from 1 to 24.

In standard forecasting applications, real-time performance is evaluated on the basis of forecast errors - often using the RMSE as a summary statistic. Evaluation of backcasts is more complex because we do not have observations of the 'truth' as a basis for evaluation. Instead, we evaluate performance of backcasting the profile of investment revealed 14 releases after the official data were published. That is, we compare the value of the smoothed backcast at time t of maturity n with the data release at time t of maturity $n+14$ to derive an RMSE-type metric

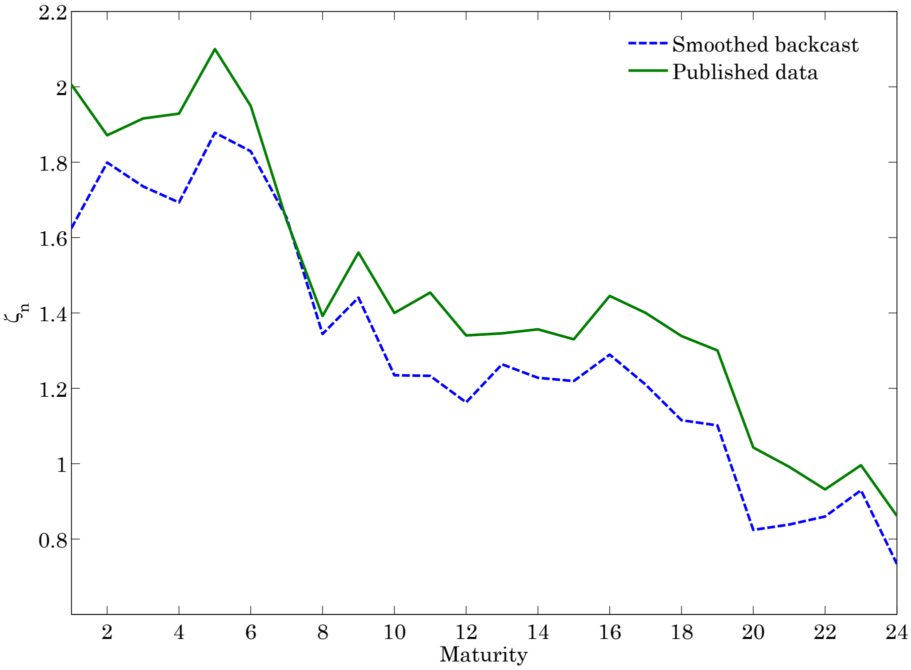
$$\varsigma^n = \sqrt{\frac{1}{s_1 - s_0 + 1}\sum_{t=s_0}^{s_1}(\hat{y}_t - y_t^{t+n+14})^2}.$$

where $\hat{y}_t$ is the smoothed backcast of y_t made at maturity n in the case of the smoothed data and is the published data otherwise.

Figure 3 plots ς^n for published data and smoothed backcasts for maturities 1 to 24. The backcasting errors appear smaller than the errors attaching to the official published estimates.

Table 8 reports the results of Diebold-Mariano tests, S_{DM}, (Diebold and Mariano (1995)) of the significance of the difference in performance between backcasts and official published estimates for maturities 1 to 12. Harvey, Leybourne, and Newbold (1997) have proposed a small-sample correction for the above test statistic, S^*_{DM}. The table reports the test statistics for the null hypothesis that the two alternative backcasts are equally good. We also report probability values for these statistics. Probability values below 0.05 indicate rejection of the null hypothesis in favour of the hypothesis that the state space model backcast is better than the early release in estimating the truth. Note that in a number of cases the Diebold-Mariano statistics are reported as missing. This is because in these cases the estimated variance of the numerator of the statistic is negative as is possible in small samples. The results show that the Diebold-Mariano test rejects the null hypothesis of equal forecasting ability in all available cases. On the other hand the modified test never rejects. We choose to place more weight on the results

Figure 3 RMSE for maturities 1 to 24 for smoothed backcast (dashed line) and published data (solid line)



of the original, and more widely used, Diebold-Mariano test and conclude that there is some evidence to suggest that the state space model backcast is superior to the early release in estimating the truth.

5 Conclusions

We have articulated a state space representation of the signal extraction problem faced when using uncertain data to form a conjunctural assessment of economic activity. The model draws on the revisions history to proxy the uncertainty surrounding the latest published estimates. Therefore it establishes the extent to which prior views on economic activity should evolve in light of new data and any other available measures, such as business surveys. The model produces estimates of the 'true' value of the variable of interest,

Table 8 Diebold-Mariano test results for maturities 1 to 12

n	S_{DM}	p-value	S^{*}_{DM}	p-value
1	-1.694	0.045	-1.600	0.084
2	–	–	–	–
3	-1.775	0.038	-1.315	0.296
4	–	–	–	–
5	–	–	–	–
6	–	–	–	–
7	–	–	–	–
8	–	–	–	–
9	–	–	–	–
10	–	–	–	–
11	-3.452	0.000	-0.597	0.279
12	-1.908	0.028	-0.247	0.404

a backcast, that can be used as a cross-check of the latest published official data, or even to substitute for those data in any economic applications. Since we assume that official estimates asymptote to the truth as they become more mature, our backcasts amount to a prediction of the cumulative impact of revisions to official estimates.

In using backcasts to predict the cumulative impact of revisions, one should, however, be alert to a number of caveats. First, we assume that the revisions history provides a good indication of past uncertainties. This assumption is likely to be violated where statistical agencies do not revise back data in light of new information or changes in methodology - in other words, the model is only applicable where statistical agencies choose to apply a rich revisions process. Second, we assume that the structures of both the data generating process (the transition law) and the data production process (measurement equations) are stable. Finally, the model is founded on a number of simplifying assumptions. In particular, the model is linear and stationary; measurement errors are assumed to be normally distributed; and the driving matrices are diagonal so that we can neither exploit any behavioural relationship between the variables of interest nor any correlation in measurement errors across variables.

A The Role of Early Releases Once More Mature Estimates Are Available

The model of Section 2 uses the latest published estimates as a measure but makes no reference to earlier data releases. We do this largely for pragmatic reasons, but it begs the question: why should the data-user ignore all earlier published estimates? The focus on the latest release is justified if the statistical agency processes new information effectively so that the information set driving the latest release encompasses that driving all earlier releases. This appendix develops this notion. In doing so, we need to model the evolution of measurement errors across releases.

Consistent with the notation in the main paper, denote the true value of the variable of interest by y_t. The model for the published data is then $y_t^{t+n} = c^n + y_t + v_t^{t+n}$, where y_t^{t+n} is the n-th release of published data for the truth at time t and c^n is a bias term which depends on n. We model $v_t^T, t = 1, \ldots, T-1$ by assuming that it is an AR process over t. We have $B(L)v_t^T = \varepsilon_t^T$.

We can also consider the process describing the evolution of ε_t^{t+i} over i - that is the evolution of errors through successive releases. Recognising that the statistical agency's information set grows through time, we can write ε_t^{t+i} as follows

$$\varepsilon_t^{t+i} = \eta_t^{t+i} + \eta_t^{t+i+1} + \ldots = \sum_{j=0}^{\infty} \eta_t^{t+i+j}.$$

As maturity increases, the statistical agency receives incremental information. That information is used to successively remove bits, η_t^{t+i}, of error from ε_t^{t+i}. As long as the statistical agency does not throw away information and new information helps, the variance of the measurement errors will decline with maturity. We formalise this below.

Assume that η_t^{t+i} can be treated as independently, but not identically, distributed (i.ni.d). By the i.ni.d assumption on η_t^{t+i} we then know that $\mathrm{var}(\varepsilon_t^{t+i}) = \sum_{j=0}^{\infty} \sigma^2_{\eta^{i+j}}$ where $\mathrm{var}(\eta_t^{t+i}) = \sigma^2_{\eta^i}$.

In the model described in Section 2, we assume that v_t^{t+n} has heteroscedasticity with respect to n, with $\sigma^2_{\varepsilon^n} = \sigma^2_{\varepsilon^1}(1+\delta)^{n-1}$. This exponential decay in measurement error variance would be consistent with an exponential decay in $\sigma^2_{\eta^i}$ with maturity - the intuition being that the increments to the statistical agency's information set decrease in size as estimates become more mature. Thus $\sigma^2_{\eta^i} = \sigma^2_{\eta^1}(1+\zeta)^{i-1}$, where $-1 < \zeta < 0$. To establish the expectation of y_t, we need to determine the covariance between measurement errors of differing releases within this model set-up; that is $\mathrm{E}(v_t^{t+i}v_t^{t+j})$. Assuming for simplicity, that v_t^{t+i} is given by an AR(1) process of the form $v_t^{t+i} = \beta v_{t-1}^{t+i} + \varepsilon_t^{t+i}$. Using standard algebra gives

$$\sigma^2_{v^k} = \mathrm{E}(v_t^{t+i}v_t^{t+j}) = \frac{\sigma^2_{\varepsilon^1}(1+\delta)^{k-1}}{1-\beta^2(1+\delta)} \tag{A.1}$$

where $k = \max(i,j)$. The covariance between measurement errors attaching to differing releases is equal to the variance of the most recent, that is the least mature release. Given a model for the covariance of revisions across releases, we can derive an expectation of y_t conditional on the entire set of available releases. Assume we have N available releases of data. Then, in forming our expectation of y_t, we want to find the coefficients that minimise the mean-square error in the following expectations function

$$\mathrm{E}(y_t|y_t^{t+1},\ldots,y_t^{t+N}) = \mathrm{E}(y_t|\mathbf{y}_t^N) = \mu + \gamma_1 y_t^{t+1} + \ldots + \gamma_N y_t^{t+N}.$$

Using standard results on conditional expectations the $\boldsymbol{\gamma}$ parameters in this expression will be given by $(\mathrm{var}(\mathbf{y}_t^N))^{-1}\,\mathrm{cov}(y_t,\mathbf{y}_t^N)$.

In the framework of the model developed in the previous section of this appendix, it can be shown that the optimal coefficients are zero for all releases but the most recent. We assume that the underlying shocks (the η_t^{t+j}s) are uncorrelated with the true data so $\mathrm{var}(\mathbf{y}_t) = \boldsymbol{\iota}_N\sigma_y^2\boldsymbol{\iota}_N' + \boldsymbol{\Sigma}_{\mathbf{v}}^N$ and $\mathrm{cov}(y_t,\mathbf{y}_t) = \boldsymbol{\iota}_N\sigma_y^2$ where $\boldsymbol{\iota}_N$ is a $N\times 1$ vector of ones, σ_y^2 is $\mathrm{E}(y_t - \mathrm{E}(y_t))^2$ and $\boldsymbol{\Sigma}_{\mathbf{v}}^N = \mathrm{E}(\mathbf{v}_t^N\mathbf{v}_t^{N\prime})$ is the variance-covariance matrix of measurement errors, where $\mathbf{v}_t^N = (v_t^{t+1}, v_t^{t+2},\ldots,v_t^{t+N})$.

We can then use equation (A.1) to build this variance-covariance matrix as

$$\mathrm{E}(\mathbf{v}_t^N\mathbf{v}_t^{N\prime}) = \boldsymbol{\Sigma}_{\mathbf{v}}^N = \left(\begin{array}{cccc} \sigma^2_{v^1} & \sigma^2_{v^2} & \cdots & \sigma^2_{v^N} \\ \sigma^2_{v^2} & \sigma^2_{v^2} & \cdots & \sigma^2_{v^N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{v^N} & \sigma^2_{v^N} & \cdots & \sigma^2_{v^N} \end{array}\right).$$

Putting these elements together

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_N \end{pmatrix} = (\boldsymbol{\iota}_N \sigma_y^2 \boldsymbol{\iota}_N' + \boldsymbol{\Sigma}_{\mathbf{v}}^N)^{-1} \boldsymbol{\iota}_N \sigma_y^2 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{\sigma_y^2}{\sigma^2{}_{v^N} + \sigma_y^2} \end{pmatrix}. \tag{A.2}$$

Hence, under the assumptions made above, we can legitimately focus on just the most recent release of data.

Note that (A.2) is obtained as follows. For $n = 2$ the result follows from elementary calculations. Then, the $n = 2$ result may be used to show that $\mathrm{E}(y_t|y_t^{t+1}, y_t^{t+n}) = \mathrm{E}(y_t|y_t^{t+n})$. Given this it follows that $\mathrm{E}(y_t|y_t^{t+1}, y_t^{t+2}, y_t^{t+n}) = \mathrm{E}(y_t|y_t^{t+n})$ if $\mathrm{E}(y_t|y_t^{t+2}, y_t^{t+n}) = \mathrm{E}(y_t|y_t^{t+n})$. But this can be shown by appealing to the $n = 2$ result. Proceeding inductively and by repeated use of the $n = 2$ result, the general n case is obtained.

B The Mapping Between $\rho_{\epsilon\varepsilon}$ and ρ_{yv}

Section 3.2 describes the use of the real-time data set to estimate $\sigma^2_{v^1}$, δ and $\beta_1, \ldots, \beta_p$ and the manipulation of these estimates to derive an estimate of $\sigma^2_{\varepsilon^1}$. This manipulation is trivial for low orders of p. For $p = 1$ we have, from equation (12) $\sigma^2_{\varepsilon^1} = \sigma^2_{v^1}(1 - (1+\delta)\beta_1^2)$. For higher orders of p, following the model of serial correlation in measurement errors described in Section 2, the model for measurement errors in period t is

$$v_t^{t+n} = \beta_1 v_{t-1}^{t+n} + \beta_2 v_{t-2}^{t+n} + \ldots + \beta_p v_{t-p}^{t+n} + \varepsilon_t^{t+n}. \tag{B.1}$$

To derive $\mathbf{V}$ we need to build up the matrix in $p \times p$ blocks. We can do this by writing equation (B.1) in companion form as $\mathbf{v}_t = \mathbf{B}\mathbf{v}_{t-1} + \boldsymbol{\varepsilon}_t$, where $\mathbf{v}_t = (v_t^{t+n}, v_{t-1}^{t+n}, \ldots, v_{t-p}^{t+n})'$,$\boldsymbol{\varepsilon}_t = (\varepsilon_t^{t+n}, 0, \ldots, 0)'$ and

$$\mathbf{B} = \begin{pmatrix} \beta_1 & \beta_2 & \ldots & \beta_{p-1} & \beta_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix}.$$

Taking the variance of both sides gives $\operatorname{var}(\mathbf{v}_t) = \mathbf{B}\operatorname{var}(\mathbf{v}_{t-1})\mathbf{B}' + \operatorname{var}(\boldsymbol{\varepsilon}_t)$. Recognising from equation (12) that $\operatorname{var}(\mathbf{v}_t) = (1+\delta)\operatorname{var}(\mathbf{v}_{t-1})$ and using the identity $\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\operatorname{vec}(\mathbf{B})$, we have $\operatorname{vec}(\operatorname{var}(\mathbf{v}_t)) = (1+\delta)(\mathbf{B} \otimes \mathbf{B})\operatorname{vec}(\operatorname{var}(\mathbf{v}_t)) + \operatorname{vec}(\operatorname{var}(\boldsymbol{\varepsilon}_t))$. Rearranging gives $\operatorname{vec}(\operatorname{var}(\mathbf{v}_t)) = (\mathbf{I}_{p^2} - (1+\delta)\mathbf{B} \otimes \mathbf{B})^{-1}\operatorname{vec}(\operatorname{var}(\boldsymbol{\varepsilon}_t))$. We can then build up the full $\mathbf{V}$ matrix in a similar fashion to equation (12)

$$\mathbf{V} = \begin{pmatrix} \mathbf{I}_p & (1+\delta)^p\mathbf{B} & \cdots & (1+\delta)^{kp}\mathbf{B}^k \\ (1+\delta)^p\mathbf{B} & (1+\delta)^p\mathbf{I}_p & \cdots & (1+\delta)^{kp}\mathbf{B}^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ (1+\delta)^{kp}\mathbf{B}^k & (1+\delta)^{kp}\mathbf{B}^{k-1} & \cdots & (1+\delta)^{kp}\mathbf{I}_p \end{pmatrix} \times (\mathbf{I}_{k+1} \otimes \operatorname{var}(\mathbf{v}_t)). \tag{B.2}$$

Taking the variance-covariance matrix to the data, we can estimate $\beta_1, \ldots, \beta_p, \sigma^2_{\varepsilon^1}$ and δ via GMM by minimising $(\operatorname{vec}(\mathbf{V}) - \operatorname{vec}(\hat{\mathbf{V}}))'(\operatorname{vec}(\mathbf{V}) - \operatorname{vec}(\hat{\mathbf{V}}))$.

We can apply a similar set of manipulations to express $\rho_{\epsilon\varepsilon}$ as a function of ρ_{yv}, the variance of measurement errors σ^2_ε and the parameters of the transition law - assuming there is no intertemporal correlation between ϵ_t and ε_t^{t+n}. We can write the transition equation (1) in companion form $\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t$, where $\mathbf{y}_t = (y_t, \ldots, y_{t-p})', \boldsymbol{\epsilon}_t = (\epsilon_t, 0, \ldots, 0)'$ and $\mathbf{A} = \begin{pmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_{q-1} & \alpha_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix}$.

By similar manipulations, we obtain that

$$\operatorname{vec}\left(\operatorname{var}\left(\mathbf{y}_t\right)\right) = \left(\mathbf{I}_{q^2} - \mathbf{A} \otimes \mathbf{A}\right)^{-1} \operatorname{vec}\left(\operatorname{var}\left(\boldsymbol{\epsilon}_t\right)\right). \tag{B.3}$$

The covariance between $\mathbf{y}_t$ and $\mathbf{v}_t$ can be written as $\operatorname{cov}(\mathbf{y}_t, \mathbf{v}_t) = \operatorname{cov}(\boldsymbol{\epsilon}_t, \boldsymbol{\varepsilon}_t) +$ $\mathbf{A}\operatorname{cov}(\mathbf{y}_{t-1}, \mathbf{v}_{t-1})\mathbf{B}'$. Recognising that $\operatorname{cov}(\mathbf{y}_{t-1}, \mathbf{v}_{t-1}) = \sqrt{(1+\delta)}\operatorname{cov}(\mathbf{y}_t, \mathbf{v}_t)$ we can rearrange to obtain

$$\operatorname{vec}(\operatorname{cov}(\mathbf{y}_t, \mathbf{v}_t)) = (\mathbf{I}_{pq} - \sqrt{(1+\delta)}\mathbf{B} \otimes \mathbf{A})^{-1} \operatorname{vec}(\operatorname{cov}(\boldsymbol{\epsilon}_t, \boldsymbol{\varepsilon}_t)). \tag{B.4}$$

The first element in the vector on the right-hand side rescales the covariance between $\mathbf{y}_t$ and $\mathbf{v}_t$ to the covariance between $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\varepsilon}_t$. To uncover the rescaled correlation we also need to take account of the differences in variance between the dynamic series and the respective shocks.

Putting (B.3) - (B.4) together reveals the mapping between ρ_{yv} and $\rho_{\epsilon\varepsilon}$. In the case when $p = q = 1$, it can be shown quite easily that $|\rho_{\epsilon\varepsilon}| \geq |\rho_{yv}|$.

References

Akritidis, L. (2003): “Revisions to Quarterly GDP Growth and Expenditure Components,” *Economic Trends*, 601, 69–85.

Ashley, J., R. Driver, S. Hayes, and C. Jeffery (2005): “Dealing with Data Uncertainty,” *Bank of England Quarterly Bulletin*, pp. 23–30.

Bank of England (1999): *Economic Models at the Bank of England*. Bank of England.

Cameron, A. C., and P. K. Trivedi (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Castle, J., and C. Ellis (2002): “Building A Real-Time Database for GDP(E),” *Bank of England Quarterly Bulletin*, pp. 42–48.

Diebold, F. X., and R. S. Mariano (1995): “Comparing Predictive Accuracy,” *Journal of Business and Economic Statistics*, 13, 253–263.

Donald, S. G., and W. K. Newey (2001): “Choosing the Number of Instruments,” *Econometrica*, 69, 1161–1191.

Garratt, A., K. Lee, E. Mise, and K. Shields (2005): “Real-Time Representations of the Output Gap,” *University of Leicester Discussion Paper No. 130.*

Garratt, A., and S. P. Vahey (2006): “UK Real-Time Macro Data Characteristics,” *The Economic Journal*, 116(509), F119–F135.

Harrison, R., G. Kapetanios, and T. Yates (2005): “Forecasting with Measurement Errors in Dynamic Models,” *International Journal of Forecasting*, 21(3), 595–607.

Harvey, A. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

HARVEY, D. I., S. J. LEYBOURNE, AND P. NEWBOLD (1997): “Testing the Equality of Prediction Mean Square Errors,” *International Journal of Forecasting*, 13, 273–281.

HOWREY, E. P. (1978): “The Use of Preliminary Data in Econometric Forecasting,” *Review of Economics and Statistics*, 60(2), 193–200.

——— (1984): “Data Revision, Reconstruction, and Prediction: An Application to Inventory Investment,” *The Review of Economics and Statistics*, 66(3), 386–393.

JÄÄSKELÄ, AND T. YATES (2005): “Monetary Policy and Data Uncertainty,” *Bank of England Working Paper*, 281.

JACOBS, J. P. A. M., AND J.-E. STURM (2006): “A Real-Time Analysis of the Swiss Trade Account,” Unpublished.

JACOBS, J. P. A. M., AND S. VAN NORDEN (2006): “Modelling Data Revisions: Measurement Error and Dynamics of ”True” Values,” *CCSO Working Papers*, 2006/07.

KAPETANIOS, G., AND T. YATES (2004): “Estimating Time-Variation in Measurement Error from Data Revisions; an Application to Forecasting in Dynamic Models,” *Bank of England Working Paper*, 238.

——— (2008): “Estimating Time-Variation in Measurement Error from Data Revisions; an Application to Forecasting in Dynamic Models,” Bank of England mimeo.

LOMAX, R. (2004): “Stability and Statistics,” Speech at the North Wales Business Club, 23 November 2004.

MANKIW, N. G., AND M. D. SHAPIRO (1986): “News or Noise: An Analysis of GDP Revisions,” *Survey of Current Business*, 66, 20–25.

MITCHELL, J. (2004): “Review of Revisions to Economic Statistics: A Report to the Statistics Commission,” *Statistics Commission Report No 17*, 2.

Öller, L.-E., and K.-G. Hansson (2002): “Revisions of Swedish National Accounts 1980-1998 and an International Comparison,” Discussion paper, Statistics Sweden.

Patterson, K. D. (1994): “A State Space Model for Reducing the Uncertainty Associated with Preliminary Vintages of Data with an Application to Aggregate Consumption,” *Economics Letters*, 46, 215–222.

Statistics Commission (2004): “Revisions to Economic Statistics,” *Statistics Commission Report 17.*

Weale, M. (1985): “Testing Linear Hypothesis on National Account Data,” *Review of Economics and Statistics*, 67(4), 685–689.

Queen Mary
University of London

This working paper has been produced by
the Department of Economics at
Queen Mary, University of London

Department of Economics
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/papers/wp.htm