A State Space Approach To Extracting The Signal From Uncertain Data

Alastair Cunningham[*]

Jana Eklund[**]

Christopher Jeffery[†]

George Kapetanios[§]

and

Vincent Labhard[¶]

Working Paper no. ?

[*] Bank of England, Email: Alastair.Cunningham@bankofengland.co.uk

[**] Bank of England, Email: Jana.Eklund@bankofengland.co.uk

[†] Bank of England, Email: Chris.Jeffery@bankofengland.co.uk

[§] Bank of England and Queen Mary University of London, Email: George.Kapetanios@bankofengland.co.uk

[¶] European Central Bank, Email: Vincent.Labhard@ecb.int

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England. We have benefited from helpful comments from Andrew Blake, Spencer Dale, Kevin Lee, Lavan Mahadeva, Shaun Vahey and Tony Yates. This paper was finalised on 23 July 2007.

The Bank of England's working paper series is externally refereed.

Information on the Bank's working paper series can be found at
www.bankofengland.co.uk/publications/workingpapers/index.htm.

Publications Group, Bank of England, Threadneedle Street, London, EC2R 8AH; telephone
+44 (0)20 7601 4030, fax +44 (0)20 7601 3298, email mapublications@bankofengland.co.uk.

ISSN 1749-9135 (on-line)

Contents

Abstract

Most macroeconomic data are uncertain - they are estimates rather than perfect measures. Use of these uncertain data to form an assessment of current activity can be viewed as a problem of signal extraction. 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 to extracting the signal from uncertain data that takes the experience of past revisions as representative of the uncertainties surrounding the latest published estimates. Specifically, it describes a two-step estimation procedure in which the history of past revisions (real-time data) are first used to estimate the parameters of a measurement equation describing the official published estimates; and these parameters are then imposed in a maximum likelihood estimation of a state space representation of the 'true' profile of the macroeconomic variable.

Summary

Most macroeconomic data are uncertain – they are estimates rather than perfect measures. Measurement errors arise because data are typically based on incomplete samples. And they arise because many variables – for example, in-house software investment – are not easily observable; necessitating the use of proxies. Such uncertainty poses challenges for both forecasting and economic analysis. Where it is material, economists must decide how much weight to place on apparent 'news' in the published data. But how can the extent of the problem be judged and what can be done about it?

One symptom of data uncertainty is the propensity of statistical agencies to revise their estimates in light of new information (bigger samples) or methodological advances (better proxies). In the United Kingdom, the National Accounts are subject to a rich revisions process and as a result, the scale of the ensuing revisions may give a clear indication of the extent of data uncertainty in the past. And to the extent that past revisions give a good guide to the likely scale of revisions in the future, they can also be used to gauge the uncertainty associated with the latest data.

Recognition of this uncertainty leads naturally to a probabilistic view of the past. Estimation of a confidence interval around the official published data is a first step; giving an indication of the potential scale of revisions. Going further, economists can gather additional evidence about the current economic conjuncture; using that evidence to assess the likely impact of future revisions on the profile of growth.

Treating uncertain data in this way is neither new nor unique to the Bank. A 2004 study by the Statistics Commission concluded that "the main users of the [official] 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." However, most attempts to allow for potential revisions are informal. Approaching the issue more formally can add rigour to the exercise of combining such diverse source of information – this sort of exercise is known as a 'signal extraction problem'.

This paper describes a formal ('state space') model of uncertain (revisable) data that can be used to extract the signal from uncertain data. The model draws on the experience of past revisions to proxy the uncertainty surrounding the latest vintage of the official data published by the Office for

National Statistics. It estimates how far to update a simple estimate of how the data would evolve – based on past values of the variable in question – in the light of those data and any alternative indicators (such as business surveys). The model's output is an estimate of the 'true' value of the variable of interest – a 'backcast' – that cab be used as a cross-check of the latest published data, or even to substitute for those data in any economic applications. Since we assume that official estimates get better with time the resulting backcasts amount to a prediction of the cumulative impact of revisions.

In using the model to predict the cumulative impact of revisions, economists should, however, be alert to a number of caveats. In particular, the model relies on past revisions being a good indicator of current uncertainty. It is, however, possible that revisions may become less predictable in the future. For example, successful delivery of the Office for National Statistic's Statistical Modernisation Programme will enable faster balancing of National Accounts data from differing sources and facilitate internal reviews of collation procedures. And some significant methodological revisions in the past – such as the introduction of the ESA-95 accounting framework – may not be representative of current uncertainty. It is also quite possible that alternative indicators that have provided a good mapping to mature ONS data in the past will offer a worse indication in future – for example if the sample of respondents to a particular business survey becomes unrepresentative.

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. And measurement errors may also arise because many variables – for example, in-house software investment – are not easily observable; necessitating use of proxies. Where such uncertainty is material, economists should recognise the potential for measurement error when gauging how much weight to place on apparent 'news' in the published data. But how can we judge the importance of data uncertainty?

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, the experience of revisions gives an indication of the scale of past uncertainties. And, to the extent that the experience of past revisions gives a good guide to the likely incidence of revisions in the future, it provides 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 0.08pp over the period since 1993; compared with a variance of 0.07pp 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 macroeconomic variables now and in the past adds to the challenge of forming a forward-looking assessment of economic prospects and hence complicates policy formulation. Our understanding of the behavioural relationship between variables will be impaired if estimates of model parameters change as the underlying data are revised. And even where model parameters are not materially affected, revisions to the recent profile of macroeconomic data may affect the forecasts generated by those 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. For example, the August 2003 *Inflation Report* noted that "The MPC takes account of the likelihood that GDP data will be revised when deciding how much weight to put on the latest data". More generally, in reviewing revisions to the United Kingdom's National Accounts, 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 (2004) suggest that where measurement uncertainties are greatest in estimates of the recent past, models that downweight recent 'experience' may have a superior forecasting performance to models in which all observations are weighted equally. In a similar vein, Jaaskela and Yates (2005) explore the implications of uncertain data for performance of competing simple policy rules. The intuition they develop is that the more measurement error there is in the output gap data, and the worse current data are relative to lagged data; the greater the weight on inflation compared to output gap terms; and the greater the weight on lagged output gap terms relative to current ones.

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 model output 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 to interpretation of differing economic developments; as described in Bank of England (1999). An alternative strategy is to unbundle 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. In other words, focusing directly on the signal extraction problem posed by uncertain data.

This paper explores that signal extraction problem more formally. So 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 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 vintages – 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.

Role for alternative indictors. 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.

Persistence of data uncertainty. Howrey (1978) restricts 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 experience 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 vintages 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 vintage 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.

Serial correlation in measurement errors. 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 vintage 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.

Correlation between measurements errors and the 'true' state. 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 under-weight uncertain data relative to prior information. Jacobs and Van Norden (2006) propose a model that captures both 'noise' and 'news' elements. Annex A expands on the distinction between 'noise' and 'news' in revisions to better locate the approach taken in this paper.

The model developed in this paper seeks to capture these various 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.

In contrast with the treatment in much of the antecedent literature, we exclude earlier vintages from the set of measures used to estimate 'true' activity (see, for example, Garratt *et al* (2005)). Ignoring earlier vintages amounts to little more than assuming that the statistical agency processes new information efficiently, so that the information set on which the latest published estimate is based encompasses that used for earlier releases. This intuition is developed more formally in Annex B.

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.

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, the model comprises a transition law and separate measurement equations describing the latest official estimates and each of the alternative indicators considered. The measurement equation describing the official published estimates is designed to be sufficiently general to capture the patterns in revisions observed in historical revisions to a variety of United Kingdom National Accounts aggregates.

For ease of future generalisation, the model is presented in vector notation, for m variables of interest. However, as discussed below, 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.(1)

(1) 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 Kalman system 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 true value of the economic concepts of interest, but is not observed.

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

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

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 $\mathrm{E}\left(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t'\right) = \Sigma_\epsilon$ where we denote the main diagonal of Σ_ϵ by $\boldsymbol{\sigma}_\epsilon^2 = \left(\sigma_{\epsilon_1}^2, \ldots, \sigma_{\epsilon_m}^2\right)'$. 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. For future reference we define $\mathrm{E}(\mathbf{y}_t \mathbf{y}_t') = \Sigma_{\mathbf{y}}$ and its main diagonal $\boldsymbol{\sigma}_{\mathbf{y}}^2 = (\sigma_{y_1}^2, \ldots, \sigma_{y_m}^2)'$, where $\sigma_{y_i}^2 = \mathrm{E}(y_{it} - \mu_i)^2$.

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 nonlinearity 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. This treatment is common across the antecedent literature.

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

$$\mathbf{y}_t^{t+n} = \mathbf{y}_t + \mathbf{c}^n + \mathbf{v}_t^{t+n}. \qquad \textbf{(2)}$$

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 $\left(\mathbf{y}_{T-i}^{T-i+1}, \ldots, \mathbf{y}_{T-1}^{T}\right)'$ 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.

In principle, the model in Equation **(2)** could be applied to previous vintages as well as the latest estimates. One natural question is whether data-users should consider these previous vintages as competing measures of the truth – that is, using $\mathbf{y}_t^{t+n-j}$ alongside $\mathbf{y}_t^{t+n}$ as measures of $\mathbf{y}_t$. This treatment does not sit easily with our assumption that revisions tend to improve estimates. So long as the statistical agency processes new information efficiently – in other words, does not discard useful information – the latest release should entirely subsume earlier estimates and the data-user should ignore all earlier vintages. Annex B establishes this intuition more formally. We assume that the statistical agency does process information efficiently and hence the remainder of this Section develops the model summarised in Equation **(2)** for the latest release alone.

The constant term $\mathbf{c}^n$ is included in Equation **(2)** to permit consideration of biases in the statistical agency's dataset. As noted above, the n superscript allows for observations of different maturities to be differently biased. Specifically, we model $\mathbf{c}^n$ as

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

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 $(-1 \leq \lambda \leq 0)$. This representation imposes structure on the bias in official published estimates – we assume that any bias tends monotonically towards zero as those estimates become more mature. The particular functional form chosen is arbitrary and it is possible that other richer forms might fit the revisions experience of specific variables better; albeit at some cost in terms of generality of application.

We assume that the measurement errors, $\mathbf{v}_t^{t+n}$, are distributed normally with finite variance. We allow that measurement errors be serially correlated, heteroscedastic with respect to maturity, and correlated with economic activity.

Serial correlation.

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

$$\mathbf{v}_t^{t+n} = \sum_{i=1}^{p} \mathbf{B}_i \mathbf{v}_{t-i}^{t+n} + \varepsilon_t^{t+n}, \qquad \textbf{(4)}$$

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 $\varepsilon_t^{t+n} = \left(\varepsilon_{1t}^{t+n}, \ldots, \varepsilon_{mt}^{t+n}\right)'$, with $\mathrm{E}\left(\varepsilon_t^{t+n}\left(\varepsilon_t^{t+n}\right)'\right) = \Sigma_{\boldsymbol{\varepsilon}}^n$ as we are allowing for heteroscedasticity in measurement errors with respect to n. The representation picks up serial correlation between errors attaching to the various observations within each data release. In other words, errors in today's estimates of yesterday may be correlated with error's in today's estimate of last week. 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. 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.

Heteroscedasticity.

We allow that ε_t^{t+n} and therefore $\mathbf{v}_t^{t+n}$ has heteroscedasticity with respect to n. Specifically, we model the main diagonal of $\Sigma_{\boldsymbol{\varepsilon}}^n$ as $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^n}^2 = \left(\sigma_{\varepsilon_1^n}^2, \ldots, \sigma_{\varepsilon_m^n}^2\right)'$, where $\sigma_{\varepsilon_i^n}^2 = \mathrm{E}\left(\varepsilon_{it}^{t+n}\right)^2$. For future reference we also define $\mathrm{E}(\mathbf{v}_t^{t+n}(\mathbf{v}_t^{t+n})') = \Sigma_{\mathbf{v}}^n$ and its main diagonal $\boldsymbol{\sigma}_{\mathbf{v}^n}^2 = (\sigma_{v_1^n}^2, \ldots, \sigma_{v_m^n}^2)'$, where $\sigma_{v_i^n}^2 = \mathrm{E}\left(v_{it}^{t+n}\right)^2$. The model for $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^n}^2$ is given by

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

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 $(-1 \leq \delta \leq 0)$. 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).

Correlation with economic activity.

Over-and-above any serial correlation in revisions, we allow that measurement errors be correlated with the underlying true state of the economy, $\mathbf{y}_t$. In doing so, we approximate the degree of 'news' as opposed to 'noise' inherent in the published estimates – addressing the challenge posed by Mankiw and Shapiro (1986). Annex A expands on the distinction between 'noise' and 'news' in revisions to better locate the approach taken.

In order to separately identify serial correlation and correlation, we allow that ε_t^{t+n} be correlated with shock ϵ_t to the transition law in Equation **(1)**, so that, for any variable of interest

$$\operatorname{cov}\left(\epsilon_{it}, \varepsilon_{it}^{n}\right) = \rho_{\epsilon\varepsilon}\sigma_{\epsilon_i}\sigma_{\varepsilon_i^n}. \tag{6}$$

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 $\Sigma_{\mathbf{v}^s}$. This, of course, is more restrictive than the model for the official data.[2] 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

To summarise the model, we give its complete state space form for the latest available release. The model treats the most recent vintage of data published by the statistical agency and any alternative

[2] 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.

indicators as measures of the variable of interest. The state space representation of the model is

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

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

3 Estimation of the State Space Model

In this section, we discuss the strategy adopted in estimating the model. The estimation is performed in two steps: first using the experience of revisions to past published data to estimate Equations **(2)** through **(6)**; and then, as a second step, estimating the remaining parameters *via* maximum likelihood using the Kalman filter. Section 3.1 gives a brief discussion of the motivation for this approach and Section 3.2 describes the use of real-time data describing the experience of past revisions to estimate bias and measurement error parameters. The form of the Kalman filter is standard and is given in Annex C for ease of reference. 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.

3.1 Rationale for two-step estimation of the model

The state space problem represented by Equations **(8)** and **(9)** is a simple linear model. Extensive previous work (see, for example, Harvey (1989) and Durbin and Koopman (2001)) has shown that the Kalman filter and smoother algorithms prove a robust estimator for this class of models so long as identification conditions are satisfied. In principle, all the parameters of the model could be estimated *via* maximum likelihood using the Kalman filter.

At this point, it is worth noting that the model treats the latest official estimate as a substitute for all earlier official estimates – making no reference to vintage data. Estimation *via* the Kalman filter would, therefore, only exploit the patterns apparent in the latest data in estimating Equations **(2)** to **(6)**. As discussed above, our assumption that the statistical agency processes its information efficiently motivates disregarding earlier vintages as competing measures of economic activity. It does not, however warrant ignoring any evidence of the statistical properties of past measurement errors. All that is required for this past experience to be informative about the parameters in Equations **(2)** to **(6)** is that (i) the parameters of the process driving bias and measurement errors be constant between vintages; and (ii) revisions evaluated over a finite window be a reasonable proxy for measurement errors. Assuming that these conditions hold enables us to exploit the past experience of revisions to estimate the paramters of the measurement equations describing the latest published data.[3]

3.2 Use of the past experience of revisions to estimate bias and measurement error parameters

In recent years, a number of 'real-time' datasets have been developed – describing the evolution of estimates through successive data releases (vintages). Using this real-time data to estimate the parameters in **(2)** to **(6)** requires us to first manipulate the real-time dataset to derive a matrix of revisions to published data of differing maturities. The parameters describing the bias and measurement errors associated with the latest official published estimates can then be estimated over that matrix.

[3] Approaching estimation in two steps 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).

3.2.1 Manipulation of the real-time dataset

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

Table A: Quarterly Growth of Whole Economy Investment - Extract From the Real-time Database

		Vintage date						
		2003 Q1	2003 Q2	2003 Q3	….	2006 Q2	2006 Q3	2006 Q4
Reference date	2002 Q4	-0.15	0.16	0.67	⋮	3.51	3.51	3.51
	2003 Q1		-1.13	-0.73	⋮	-3.18	-3.18	-3.18
	2003 Q2			1.48	⋮	-1.49	-1.49	-1.49
	⋮				⋮	⋮	⋮	⋮
	2006 Q2						1.31	1.21
	2006 Q3							1.32

Table B: Stylised Real-time Database - Maturity of Observations

		Vintage date						
		2003 Q1	2003 Q2	2003 Q3	….	2006 Q2	2006 Q3	2006 Q4
Reference date	2002 Q4	1	2	3	⋮	11	12	13
	2003 Q1		1	2	⋮	10	11	12
	2003 Q2			1	⋮	9	10	11
	⋮				⋮	⋮	⋮	⋮
	2006 Q2						1	2
	2006 Q3							1

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

$$w_t^{j,n} = y_t^{t+n+j} - y_t^{t+n} \qquad \textbf{(10)}$$

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.

If the real-time dataset contains W vintages 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)**. 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 $N = J = 20$.

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. In describing the properties of bias and measurement errors, our interest is in tracing out any relationship between data uncertainties attaching to observations, as described below.

3.2.2 Estimating bias

We can use the sample of historical revisions in matrix $\mathbf{W}^J$ to estimate c^1 and λ quite trivially.[4] 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 $\left(w^{J,n}\right)$, the parameters c^1 and λ are then estimated from

$$\text{mean}\left(w^{J,n}\right) = c^1 \left(1 + \lambda\right)^{n-1} + \psi_n \qquad \textbf{(11)}$$

where $-1 \leq \lambda \leq 0$ and ψ_n denotes a remainder term.

3.2.3 Estimating the correlation between measurement and transition errors

We cannot use historical revisions to estimate $\rho_{\epsilon\varepsilon}$ directly, because we do not observe either ϵ_t, or ε_t^{t+n} in real time. But we can use the sample of historical revisions to form an approximation of ρ_{yv} – denoted ρ_{yv}^*. Assuming that there is no intertemporal correlation between ϵ and ε, we can express cov $\left(\epsilon_t, \varepsilon_t^{t+n}\right)$ as a function of ρ_{yv}^* and the variances of ϵ_t and ε_t^{t+n}. We can, then, substitute this expression into the relevant state space model covariance matrices. The manipulation involved is summarised in Annex C.

[4] 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.

A first step is to estimate ρ^*_{yv}. We can readily calculate the correlation between revisions to data of any maturity (n) and published estimates of maturity $J + n$, denoted by $\rho^n_{yv} = \text{corr}\left(y_t^{t+J+n}, w_t^{J,n}\right)$, at each maturity. 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.[5]

3.2.4 *Estimating heteroscedasticity and serial correlation*

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 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 of exposition, 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

$$\mathbf{V} = \frac{\sigma^2_{\varepsilon^1}}{1-(1+\delta)\beta_1^2}\begin{pmatrix} 1 & (1+\delta)\beta_1 & \cdots & (1+\delta)^{N-1}\beta_1^{N-1} \\ (1+\delta)\beta_1 & (1+\delta) & \cdots & (1+\delta)^{N-1}\beta_1^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ (1+\delta)^{N-1}\beta_1^{N-1} & (1+\delta)^{N-1}\beta_1^{N-2} & \cdots & (1+\delta)^{N-1} \end{pmatrix}. \quad \textbf{(12)}$$

A sample estimate of the variance-covariance matrix $\widehat{\mathbf{V}}$ can be calculated trivially from the matrix of historical revisions $\mathbf{W}^J$. Taking the variance-covariance matrix to the data, we can estimate β_1, $\sigma^2_{\varepsilon^1}$ and δ via GMM by minimising

$$\left(\text{vec}\left(\mathbf{V}\right) - \text{vec}\left(\widehat{\mathbf{V}}\right)\right)'\left(\text{vec}\left(\mathbf{V}\right) - \text{vec}\left(\widehat{\mathbf{V}}\right)\right). \quad \textbf{(13)}$$

Higher lag-orders of p require some further manipulation to derive the variance-covariance matrix, as outlined in Annex C.

[5] 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 vintages (such as those described in Annex B) which we do not wish to impose on the model.

3.3 Monte Carlo simulation properties

As a check on the small-sample performance of our estimator, we run a small Monte Carlo simulation exercise. The focus of the exercise is on the model's performance in fitting the true state, $\mathbf{y}_t$, rather than on the estimation of specific parameters. In particular, we want to establish whether filtering of the data is an improvement on taking the latest published estimate, $\mathbf{y}_t^{t+n}$, at face value.

3.3.1 Simulation assumptions

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, ie $q = 1$. 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, ie $p = 1$. We assume no additional indicators are available. The output of the model is an estimate of the true state prevailing in each period, denoted $\hat{y}_t$. 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 parameterisation and the results presented are averages over the replicates.

Parameterisation. We evaluate simulation properties 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 differing 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 experience of revisions 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 $\sigma^2_{\mathbf{v}_T^{T+1}} = 1$ implying that the signal-to-noise ratio is also one at maturity one.

3.3.2 Simulation results

We use the simulation results to gauge the degree to which using the model improves relative to taking the latest published estimate at face value. The metric used is the standard deviation of model errors across replications, relative to the standard deviations of errors in the latest published estimate. 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 outperforms the published data for all maturities up to 58 quarters. Thereafter, the measurement errors attaching to the published estimates have declined sufficiently that any gains from filtering are more than offset by parameter uncertainties.

Chart 1: Simulation results

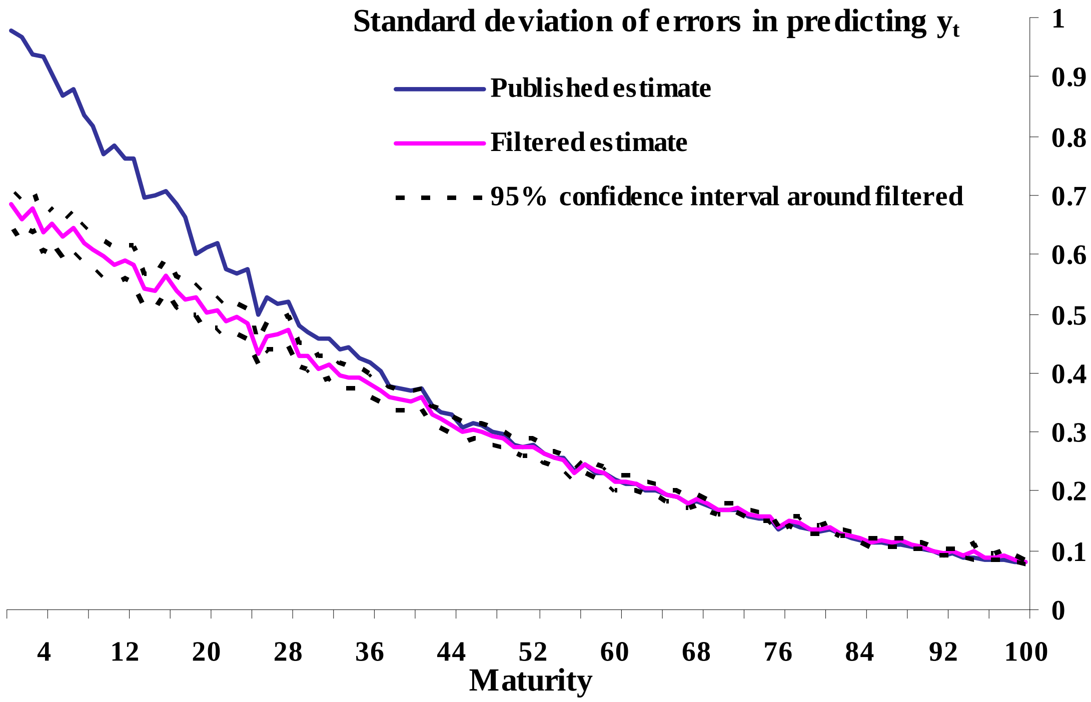


Table C summarises the results for the other parameterisations. In all cases, the estimated model outperforms published estimates with maturities below 18 quarters.

Table C: Gains from filtering for simulation results

			Gain from filtering at maturity:		Earliest maturity at which published estimates outperform filter
$\rho_{\epsilon\varepsilon}$	α	β	1	9	
0.5	0.1	0.1	47.7	43.6	–†
0.5	0.1	0.6	47.4	41.2	80
0.5	0.6	0.1	51.2	46.4	–†
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

† For these parameter settings, the filter outputs outperform the published data at all maturities.

4 An Illustrative Example

As an illustrative example, we apply the model to quarterly growth of whole economy investment. The Bank of England's real-time dataset was described in Castle and Ellis (2002) and includes published estimates of investment from 1961.[6] We consider the British Chambers of Commerce's Quarterly Survey as an indicator – specifically, the balance of service sector respondents reporting an upward change to investment plans over the past 3 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*.

[6] The Bank of England's real-time dataset is available at www.bankofengland.co.uk/statistics/gdpdatabase.

4.1 Characterising the revisions history

Table D sets out some summary statistics describing the experience of revisions to published data of differing maturities – evaluating revisions over a 20 quarter window as discussed in Section 3.2.

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

	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 upward revision	1.70	1.49	1.25	1.13	0.85	0.96
Mean downward revision	−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 3.09pp for estimates with a maturity of 1 quarter. That is similar to the variance of whole economy investment growth (3.12pp). 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.

4.2 Estimating the bias and measurement error parameters

As outlined in Section 3, the model is estimated in two stages: first estimating the parameters of Equations **(2)** though **(6)** – across real-time data and second applying those parameters in estimation of the state space model *via* the Kalman filter. Table E reports these estimated parameters.

Table E: Quarterly Growth of Investment - Estimated Parameters

		Parameter	Standard Error
Initial variance	$\sigma^2_{v^1}$	3.584	0.296
Variance decay	δ	-0.058	0.013
Serial correlation - 1st order	β_1	-0.220	0.055
Correlation with mature data	ρ^*_{yv}	-0.315	0.162

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

4.3 Estimating the state space model

Once Equations **(2)** though **(6)** have been estimated, the remaining model parameters are estimated *via* maximum likelihood using the Kalman filter. Table F reports the estimated parameters, while Table G 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 dataset. The model sets $q = 0$ so that the transition equation does not include an autoregressive component. Higher orders of q were not found to be statistically significant.

Table F: Quarterly Growth of Investment - Estimated Transition Law and Indicator Parameters

		Parameter	Standard Error
Transition law			
Constant	μ	1.177	0.238
Error variance	σ^2_ϵ	3.217	0.673
Indicator measurement			
Constant	c^s	1.177	0.219
Slope	Z^s	0.369	0.138
Error variance	$\sigma^2_{v^s}$	2.629	0.567

Both the prediction errors(7) 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. The errors surrounding predictions for the indicator variable are less well-behaved. In particular, there is evidence of significant serial correlation in these residuals.(8)

Table G: Quarterly Growth of Investment - Model Residual Diagnostics

	Prediction: y_t	Prediction: y_t^s	Transition
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

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 on 5% significance level.

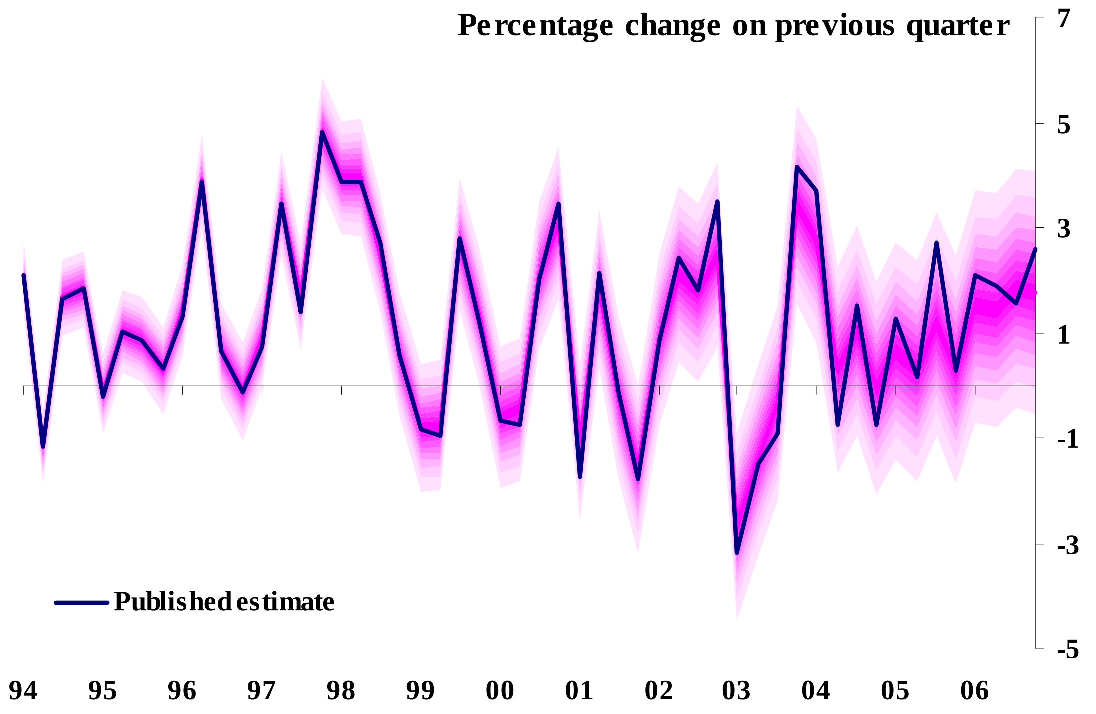
We next turn to the estimate for quarterly growth of whole economy investment – that is, the smoothed backcast. Given the focus on data uncertainty, it makes sense to view the backcast in probabilistic terms rather than focussing on point estimates of past growth. The standard error surrounding the smoothed backcast of the most recent quarter is less than $\sigma_{\upsilon 1}^{2}$, suggesting some in-sample gain from applying the model to the signal extraction problem posed by uncertainty in measurement. Figure 2 reports the estimates of quarterly growth of whole economy investment. Following the convention of the GDP and inflation fan charts plotted in the Bank of England's *Inflation Report* each band contains 10% of the distribution of possible outcomes. In this application, because we assume normality, the outer (90%) band is equivalent to a +/- 1.6 standard error bound.

(7) Prediction errors are defined above Equation **(C-8)** in Annex C as the 'surprise' in the observable variables (ie 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 parameterisation of the Kalman filter (and the resulting smoothed backcasts) will be incorrect.

(8) We have assumed that residuals associated with the indicator variables are i.i.d. This assumption could be relaxed in future work.

The centre-point of the fan chart tracks the statistical agency's published estimates quite closely once those estimates are mature. This is a corollary of the heteroscedasticity in measurement errors. Over the most recent past, the centre-point differs more materially: reflecting both the higher measurement error variance attaching to earlier releases and the difference between the large apparent changes in the published estimates and the stability of the transition law.

Chart 2: Quarterly Growth of Investment: Full Model



4.4 Real-Time Evaluation of the State Space Model

As noted above, the variance surrounding the estimates of quarterly growth in investment is less than the variance of past revisions to the statistical agency's published estimates, suggesting some in-sample gain from the modelling exercise. However, to the extent that estimated variances ignore parameter uncertainty they are likely to overstate the gain from filtering. To assess the importance of this, we evaluate the real-time performance of the model.

For this experiment, the evaluation period starts at $s_0 = 1998Q1$ and ends at $s_1 = 2004Q4$. 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 at the following time period, ie $1998Q2$. This is repeated until $2002Q4$ to give 20 evaluation observations. For each run, we compare the performance of the backcast with that of the official published estimates available at the time the backcast was formed. Because each official data release includes data points of differing maturities, we evaluate performance in backcasting 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 in forecasting 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

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

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

Figure 3 plots ς^n for published data and backcasts for maturities 1 to 24. The backcasting errors are smaller than the errors attaching to the official published estimates. Table H 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 'forecasts' 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 forecasting 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.

Chart 3: RMSE for whole economy investment for maturities 1 to 24

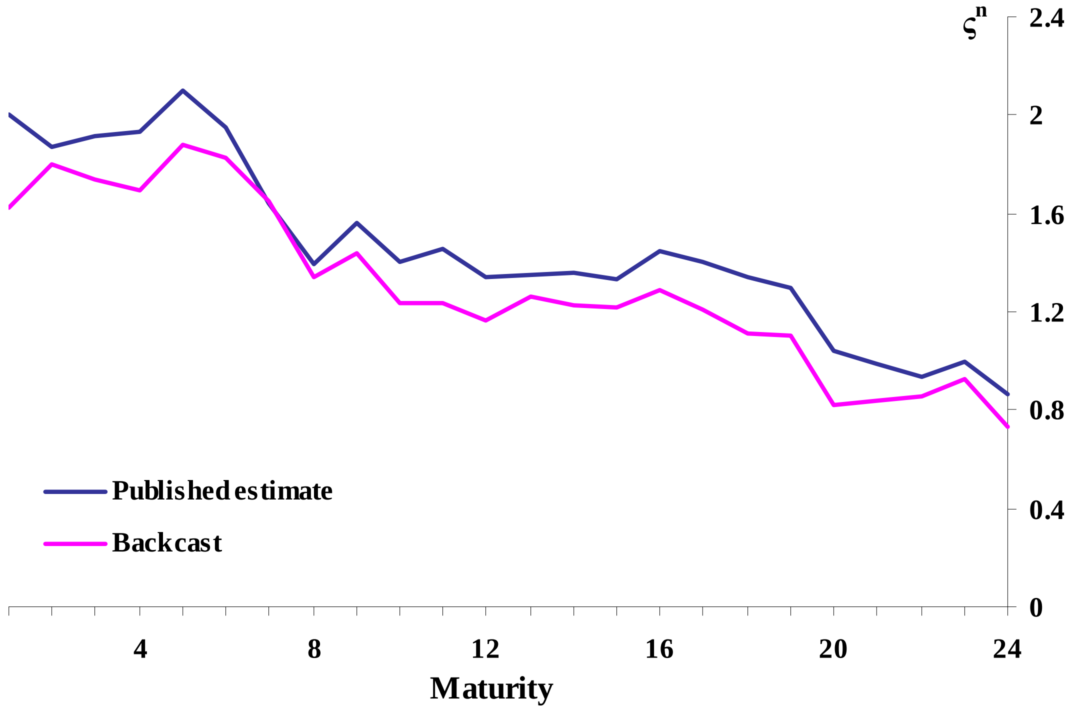


Table H: Diebold-Mariano test results for model of whole economy investment

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

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 experience of past revisions to proxy the uncertainty surrounding the latest published estimates and hence establish how far to update a prior view of how economic activity would evolve in light of those data and any other measures (such as business surveys). The model's output is an estimate of the 'true' value of the variable of interest – 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 past experience of revisions 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.

References

Akritidis, L (2003), ‘Revisions to quarterly GDP growth and expenditure components’, *Economic Trends*, Vol. 601, pages 69–85.

Ashley, J, Driver, R, Hayes, S and Jeffery, C (2005), ‘Dealing with data uncertainty’, *Bank of England Quarterly Bulletin*, Vol. 45, No. 1, pages 23–30.

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

Castle, J and Ellis, C (2002), ‘Building a real-time database for GDP(E)’, *Bank of England Quarterly Bulletin*, Vol. 42, No. 1, pages 42–48.

Diebold, F X and Mariano, R S (1995), ‘Comparing predictive accuracy’, *Journal of Business and Economic Statistics*, Vol. 13, pages 253–263.

Durbin, J and Koopman, S J (2001), *Time Series Analysis by State Space Methods*, Oxford University Press.

Garratt, A, Lee, K, Mise, E and Shields, K (2005), ‘Real-time representations of the output gap’, *University of Leicester Discussion Paper No 130.*

Garratt, A and Vahey, S P (2006), ‘UK real-time macro data characteristics’, *The Economic Journal*, Vol. 116, No. 509, pages F119–F135.

Harrison, R, Kapetanios, G and Yates, T (2004), ‘Forecasting with measurement errors in dynamic models’, *Bank of England Working Paper*, Vol. 237.

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

Harvey, D I, Leybourne, S J and Newbold, P (1997), ‘Testing the equality of prediction mean square errors’, *International Journal of Forecasting*, Vol. 13, pages 273–281.

Howrey, E P (1978), ‘The use of preliminary data in econometric forecasting’, *Review of Economics and Statistics*, Vol. 60, No. 2, pages 193–200.

Howrey, E P (1984), ‘Data revision, reconstruction, and prediction: An application to inventory investment’, *The Review of Economics and Statistics*, Vol. 66, No. 3, pages 386–393.

Jaaskela, J and Yates, T (2005), ‘Monetary policy and data uncertainty’, *Bank of England Working Paper*, Vol. 281.

Jacobs, J and Sturm, J-E (2006), ‘A real-time analysis of the swiss trade account’, Unpublished.

Jacobs, J P A M and Van Norden, S (2006), ‘Modelling data revisions: Measurement error and dynamics of "true" values’, *CCSO Working Papers*, Vol. 2006/07.

Kapetanios, G and Yates, T (2004), ‘Estimating time-variation in measurement error from data revisions; an application to forecasting in dynamic models’, *Bank of England Working Paper*, Vol. 238.

Lomax, R (2004), ‘Stability and statistics’, Speech at the North Wales Business Club, 23 November 2004.

Mankiw, N G and Shapiro, M D (1986), ‘News or noise: An analysis of GDP revisions’, *Survey of Current Business*, Vol. 66, pages 20–25.

Mitchell, J (2004), ‘Review of revisions to economic statitistics: A report to the statistics commission’, *Statistics Commission Report No 17*, Vol. 2.

Öller, L-E and Hansson, K-G (2002), ‘Revisions of swedish national accounts 1980-1998 and an international comparison’, 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*, Vol. 46, pages 215–22.

Statistics Commission (2004), ‘Revisions to economic statistics’, *Statistics Commission Report 17*.

Weale, M (1985), ‘Testing linear hypothesis on national accout data’, *Review of Economics and Statistics*, Vol. 67, No. 4, pages 685–89.

Appendix A: 'News' versus 'Noise' in Revisions

Early papers discussing the revisions process had little to say about the actions of the statistical agency in producing its estimates. In an influential paper, Mankiw and Shapiro (1986) highlighted the potential impact of this omission. They argued that at one extreme, provisional estimates could be thought of as small-sample observations containing a random measurement error. In this world, the revisions process should then be viewed as the elimination of 'noise' from preliminary estimates. At the other extreme, the statistical agency could apply its own filtering models prior to release. Provisional estimates would then be best viewed as rational forecasts, with revisions representing the arrival of unpredictable 'news'. In between these poles, Mankiw and Shapiro (1986) argued that when statistical staff "meet to evaluate and adjust the estimates before they are released" they are implicitly applying some sort of filtering model and we should expect a mixture of 'news' and 'noise' revisions. These differing views of the data production process have strikingly different implications for the statistical properties of the revisions.

Following the notation of Section 2, let y_t^{t+n} denote a noisy estimate of the 'true' data, y_t published by the statistical agency at time $t+n$, where $n = 1, \ldots, T-t$. Assuming for simplicity that measurement is unbiased, this published estimate is assumed to be equal to the true data plus some measurement error, υ_t^{t+n},

$$y_t^{t+n} = y_t + \upsilon_t^{t+n}. \qquad \textbf{(A-1)}$$

Under the 'noise' characterisation, measurement errors are independent of the true data, but correlated with the published data. Any correlation with published data will be positive because high initial estimates would tend to be revised down/ low initial estimates revised up. So

$$\begin{aligned} \mathrm{E}\left(y_t \upsilon_t^{t+n}\right) &= 0, \qquad \textbf{(A-2)} \\ \mathrm{E}\left(y_t^{t+n} \upsilon_t^{t+n}\right) &= \sigma_{\upsilon^n}^2. \end{aligned}$$

Under the 'news' characterisation, measurement errors are correlated with the true data, but independent of the published data. Any correlation with the true data will be negative: for a given value of y_t^{t+n}, positive shocks to the true data will be matched by negative shocks to the measurement error. So

$$\mathrm{E}\left(y_t v_t^{t+n}\right) = -\sigma_{v^n}^2, \qquad \textbf{(A-3)}$$
$$\mathrm{E}\left(y_t^{t+n} v_t^{t+n}\right) = 0.$$

The impact of these differences is readily apparent in the variance of the published data. Under the 'noise' case

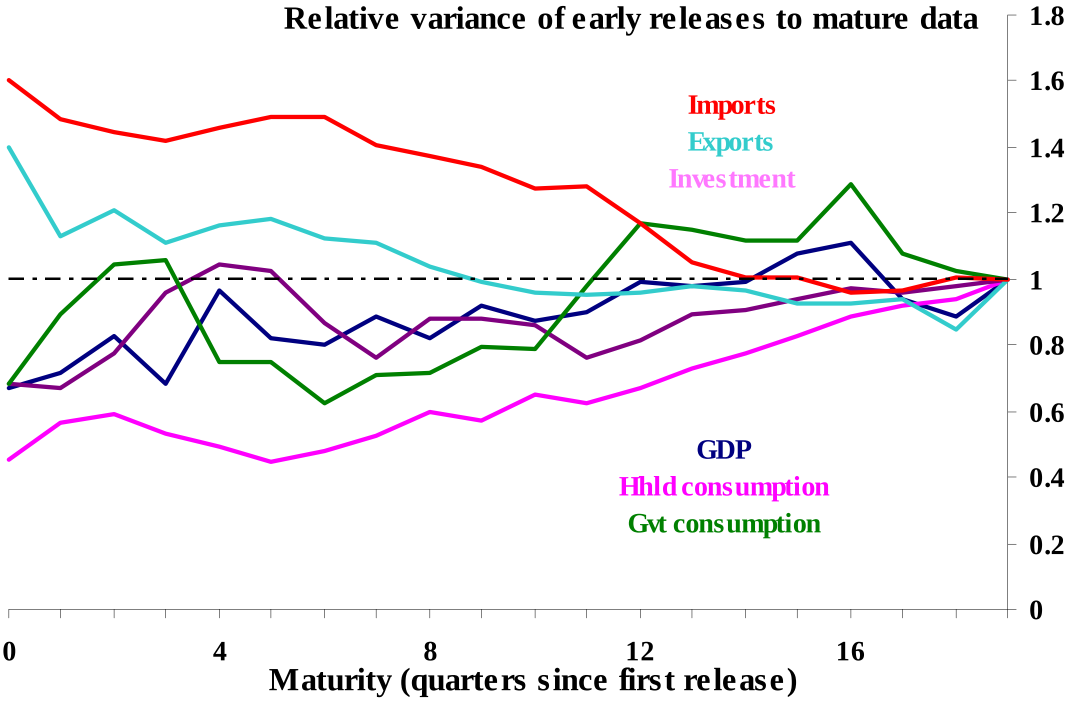
$$\sigma_{y^n}^2 = \sigma_y^2 + \sigma_{v^n}^2. \qquad \textbf{(A-4)}$$

whereas under the 'news' case

$$\sigma_{y^n}^2 = \sigma_y^2 - \sigma_{v^n}^2. \qquad \textbf{(A-5)}$$

We can exploit the difference between Equations **(A-4)** and **(A-5)** to gauge the degree of 'noise' and 'news' in past published estimates. In the model of Section 2, we assume that the variance of measurement errors, $\sigma_{v^n}^2$, declines with maturity. The variance in the published data should therefore be declining in maturity if revisions represent 'noise' **(A-4)** and increasing if they are best characterised as 'news' **(A-5)**. Figure 4 plots the variance of low maturity estimates relative to data with a maturity of 20 quarters for a range of United Kingdom National Accounts variables; evaluated over vintages released since 1993.

Chart 4: Relative variance of early estimates and mature data



The impression given is that some variables contain more 'news' whereas others contain more 'noise'. In any general representation of data uncertainty, we therefore need to be able to impose both noise and news processes. To do this, we can introduce an extra parameter, κ, that reflects the proportion of 'news' revisions for any variable, where $\kappa = 1$ in the case of pure 'news' and $\kappa = 0$ in the case of pure 'noise'. Then, revisions are correlated with both preliminary and mature estimates

$$\mathrm{E}\left(y_t v_t^{t+n}\right) = -\kappa\sigma_{v^n}^2, \tag{A-6}$$

$$\mathrm{E}\left(y_t^{t+n} v_t^{t+n}\right) = (1-\kappa)\,\sigma_{v^n}^2. \tag{A-7}$$

Moreover, the variance of published data will decrease (increase) with maturity when κ is less than (greater than)

$$\sigma_{y^n}^2 = \sigma_y^2 + (1-2\kappa)\,\sigma_{v^n}^2 \tag{A-8}$$

The model developed in Section 2 is based on a similar principle. By manipulating Equation **(A-6)**, we can uncover the correlation between the measurement error and the 'truth'

$$\operatorname{corr}\left(y_t, v_t^{t+n}\right) = -\kappa\frac{\sigma_{v^n}}{\sigma_y}. \tag{A-9}$$

When revisions contain *any* 'news' component, we would expect them to be negatively correlated with mature data. Moreover, if $\sigma_{v^n}^2$ is declining with maturity, we would expect any negative correlation to attenuate towards zero with maturity.

In Section 2, such a correlation is introduced through Equation **(6)**. If we were confident that this correlation was purely a function of the 'news' vs 'noise' distinction we would expect it to be non-positive, and attenuating with maturity. In practical application other influences may be important (for example, difficulties in measurement at different stages of the economic cycle), so we have instead adopted an agnostic approach with a freely-estimated constant correlation across maturities. Reassuringly, experimentation with United Kingdom national accounts data suggests that positive correlations are very rarely observed.

Appendix B: 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. This begs the question: why should the data-user ignore all earlier published estimates? The intuition supporting our focus on the latest release is that so long as the statistical agency processes new information effectively the information set driving the latest release will encompass that driving all earlier vintages. This annex develops that intuition more formally, drawing out any additional assumptions necessary to support the model of Section 2. In doing so, we need to model the evolution of measurement errors across releases and then consider whether previous releases enter the expectation of y_t.

B.1 A model of measurement errors across successive vintages

Consistent with the notation in the main paper, denote the true value of the variable of interest by y_t. Ignoring any bias in estimates, the model for the published data is then

$$y_t^{t+n} = y_t + \upsilon_t^{t+n}, \qquad \textbf{(B-1)}$$

where y_t^{t+n} is the n-th vintage of published data for the truth at time t. We model $\upsilon_t^T, t = 1, \ldots, T-1$ by assuming that it is an AR process over t. We have

$$B\left(L\right)\upsilon_t^T = \varepsilon_t^T. \qquad \textbf{(B-2)}$$

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

$$\varepsilon_t^{t+i} = \eta_t^{t+i} + \eta_t^{t+i+1} + \ldots = \sum_{j=0}^{\infty} \eta_t^{t+i+j}. \qquad \textbf{(B-3)}$$

As maturity increases, the statistical agency receives incremental information. That information is used to remove bits of error from ε_t^{t+i}, the η_t^{t+i} represent these bits of error that are successively removed from ε_t^{t+i}. So 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

$$\operatorname{var}\left(\varepsilon_t^{t+i}\right) = \sum_{j=0}^{\infty} \sigma_{\eta^{i+j}}^2 \qquad \textbf{(B-4)}$$

where $\operatorname{var}\left(\eta_t^{t+i}\right) = \sigma_{\eta^i}^2$.

In the model described in Section 2, we assume that v_t^{t+n} has heteroscedasticity with respect to n, with $\sigma_{\varepsilon^n}^2 = \sigma_{\varepsilon^1}^2 (1+\delta)^{n-1}$. This exponential decay in measurement error variance would be consistent with an exponential decay in $\sigma_{\eta^i}^2$ with maturity – the intuition being that the increments to the statistical agency's information set decrease in size as estimates become more mature. Thus

$$\sigma_{\eta^i}^2 = \sigma_{\eta^1}^2 (1+\zeta)^{i-1}. \qquad \textbf{(B-5)}$$

B.1.1 The covariance of errors across vintages

To establish our expectation of y_t, we need to determine the covariance between measurement errors of differing vintages within this model set up; that is $\mathrm{E}\left(v_t^{t+i} v_t^{t+j}\right)$. As a first step, we express $\sigma_{\varepsilon^1}^2$ as a function of the increments to the statistical agency's information set. We have[9]

$$\sigma_{\varepsilon^1}^2 (1+\delta)^{i-1} = \sum_{j=0}^{\infty} \sigma_{\eta^1}^2 (1+\zeta)^{i+j-1} = \frac{\sigma_{\eta^1}^2 (1+\zeta)^{i-1}}{-\zeta}. \qquad \textbf{(B-6)}$$

Thus, we get $\delta = \zeta$ and

$$\sigma_{\varepsilon^1}^2 = \frac{\sigma_{\eta^1}^2}{-\zeta}. \qquad \textbf{(B-7)}$$

Then $\operatorname{cov}\left(\varepsilon_t^{t+i}, \varepsilon_t^{t+j}\right)$ is given by

$$\mathrm{E}\left(\varepsilon_t^{t+i} \varepsilon_t^{t+j}\right) = \sum_{r=k}^{\infty} \sigma_{\eta^r}^2 = \frac{\sigma_{\eta^1}^2}{-\zeta} (1+\delta)^{k-1} = \sigma_{\varepsilon^1}^2 (1+\delta)^{k-1} = \sigma_{\varepsilon^k}^2 \qquad \textbf{(B-8)}$$

where $k = \max(i, j)$.

Finally we would like to determine $\operatorname{cov}\left(v_t^{t+i}, v_t^{t+j}\right)$; that is, to recognise the serial correlation in measurement errors across reference periods. For simplicity, we take this serial correlation to be described by an AR(1) process so that

$$v_t^{t+i} = \beta v_{t-1}^{t+i} + \varepsilon_t^{t+i}. \qquad \textbf{(B-9)}$$

[9] Using the standard rule for infinite geometric series that $\sum_{i=1}^{\infty} Ar^i = A/(1-r)$ where $|r| < 1$.

Then, we get

$$\sigma_{v^i}^2 = \sum_{p=0}^{\infty} \beta^{2p} \operatorname{E}\left(\varepsilon_{t-p}^{t+i}\right)^2 = \frac{\sigma_{\varepsilon^1}^2 (1+\delta)^{i-1}}{1-\beta^2(1+\delta)}. \qquad \textbf{(B-10)}$$

Then,

$$v_t^{t+i} v_t^{t+j} = \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \beta^{p+q} \varepsilon_{t-p}^{i+p} \varepsilon_{t-q}^{j+q}. \qquad \textbf{(B-11)}$$

So

$$\operatorname{E}\left(v_t^{t+i} v_t^{t+j}\right) = \sum_{p=0}^{\infty} \beta^{2p} \operatorname{E}\left(\varepsilon_{t-p}^{i+p} \varepsilon_{t-p}^{j+p}\right) = \sum_{p=0}^{\infty} \beta^{2p} \sigma_{\varepsilon^1}^2 (1+\delta)^{p+k-1} = \frac{\sigma_{\varepsilon^1}^2 (1+\delta)^{k-1}}{1-\beta^2(1+\delta)} = \sigma_{v^k}^2 \qquad \textbf{(B-12)}$$

where $k = \max(i, j)$. The covariance between measurement errors attaching to differing vintages is equal to the variance of the most recent, that is the least mature release.

B.2 Expectations of y_t

Given a model for the covariance of revisions across vintages, we can derive an expectation of y_t conditional on the entire set of available vintages. Assume we have N available vintages 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

$$\operatorname{E}\left(y_t | y_t^{t+1}, \ldots, y_t^{t+N}\right) = \operatorname{E}(y_t | \mathbf{y}_t) = \mu + \gamma_1 y_t^{t+1} + \ldots + \gamma_N y_t^{t+N}. \qquad \textbf{(B-13)}$$

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

It can be shown that the optimal coefficients are zero for all releases but the most recent. This conclusion holds under both 'noise' and 'news' characterisations of the data production process.

According to the 'noise' hypothesis, the underlying shocks (the η_t^{t+j}'s) are uncorrelated with the true data so

$$\operatorname{var}(\mathbf{y}_t) = \boldsymbol{\iota}_N \sigma_y^2 \boldsymbol{\iota}_N' + \Sigma_v, \qquad \textbf{(B-14)}$$

$$\operatorname{cov}(y_t, \mathbf{y}_t) = \boldsymbol{\iota}_N \sigma_y^2, \qquad \textbf{(B-15)}$$

where $\boldsymbol{\iota}_N$ is a $N \times 1$ vector of ones, σ_y^2 is $\operatorname{E}(y_t - \operatorname{E}(y_t))^2$ and $\Sigma_v = \operatorname{E}\left(\mathbf{v}_t \mathbf{v}_t'\right)$ is the variance-covariance matrix of measurement errors, where $\mathbf{v}_t = \left(v_t^{t+1}, v_t^{t+2}, \ldots, v_t^{t+N}\right)$.

We can then use Equation **(B-12)** to build this variance-covariance matrix as

$$\mathrm{E}(\mathbf{v}_t\mathbf{v}_t') = \Sigma_v = \begin{pmatrix} \sigma^2_{v^1} & \sigma^2_{v^2} & \sigma^2_{v^3} & \cdots & \sigma^2_{v^N} \\ \sigma^2_{v^2} & \sigma^2_{v^2} & \sigma^2_{v^3} & \cdots & \sigma^2_{v^N} \\ \sigma^2_{v^3} & \sigma^2_{v^3} & \sigma^2_{v^3} & \cdots & \sigma^2_{v^N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{v^N} & \sigma^2_{v^N} & \sigma^2_{v^N} & \cdots & \sigma^2_{v^N} \end{pmatrix}. \qquad \textbf{(B-16)}$$

Putting these elements together[10]

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_N \end{pmatrix} = \left(\boldsymbol{\iota}_N\sigma^2_y\boldsymbol{\iota}_N' + \Sigma_v\right)^{-1}\boldsymbol{\iota}_N\sigma^2_y = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{\sigma^2_y}{\sigma^2_{v^N}+\sigma^2_y} \end{pmatrix}. \qquad \textbf{(B-17)}$$

Given the data structure of Equations **(B-3)** and **(B-4)** - ie assuming that the statistical agency never discards useful information and that increments to the agency's information set are independent – we can legitimately focus on just the most recent vintage of data. Intuitively, the information contained in earlier releases is entirely subsumed in the latest available release. The optimal expectation of y_t involves smoothing through the noise in the latest available release (as $\frac{\sigma^2_y}{\sigma^2_{v^N}+\sigma^2_y} \leq 1$).

A similar result holds under the 'news' hypothesis, where the underlying shocks (the η_t^{t+j}'s) are negatively correlated with the true data. Equations **(B-14)** and **(B-15)** become marginally more complicated

$$\mathrm{var}\,(\mathbf{y}_t) = \boldsymbol{\iota}_N\sigma^2_y\boldsymbol{\iota}_N' + \Sigma_v - \boldsymbol{\iota}_N\,\mathrm{diag}\,(\Sigma_v)' - \mathrm{diag}\,(\Sigma_v)\,\boldsymbol{\iota}_N', \qquad \textbf{(B-18)}$$

$$\mathrm{cov}\,(y_t, \mathbf{y}_t) = \boldsymbol{\iota}_N\sigma^2_y - \mathrm{diag}\,(\Sigma_v)\,. \qquad \textbf{(B-19)}$$

[10] To see the second equality of **(B-17)** we note that for $n = 2$ the result follows from elementary calculations. To show the result for general n we proceed by induction. The $n = 2$ result may be used to show that $\mathrm{E}\left(y_t|y_t^{t+1}, y_t^{t+n}\right) = \mathrm{E}\left(y_t|y_t^{t+n}\right)$. Given this it follows that $\mathrm{E}\left(y_t|y_t^{t+1}, y_t^{t+2}, y_t^{t+n}\right) = \mathrm{E}\left(y_t|y_t^{t+n}\right)$ if $\mathrm{E}\left(y_t|y_t^{t+2}, y_t^{t+n}\right) = \mathrm{E}\left(y_t|y_t^{t+n}\right)$. 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.

Putting the elements together as before[11]

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_N \end{pmatrix} = \left(\boldsymbol{\iota}_N \sigma_y^2 \boldsymbol{\iota}_N' + \Sigma_\upsilon - \boldsymbol{\iota}_N \operatorname{diag}(\Sigma_\upsilon)' - \operatorname{diag}(\Sigma_\upsilon)\, \boldsymbol{\iota}_N'\right)^{-1} \left(\boldsymbol{\iota}_N \sigma_y^2 - \operatorname{diag}(\Sigma_\upsilon)\right) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \qquad \textbf{(B-20)}$$

Again, given the data structure of Equation **(B-3)** – we can legitimately focus on just the most recent vintage of data. The difference between the 'news' and 'noise' cases lies in the amount of weight we should place on the latest available release, not in the different treatment of alternative vintages.

[11] The second equality of **(B-20)** follows similarly to **(B-17)** by using the argument of footnote 10.

Appendix C: Details of the Application of the Kalman Filter in Backcasting

C.1 General representation of the Kalman filter

The model developed in Section 2 is summarised in state space form as Equations **(8)** and **(9)**. Linear state space models of this form can be cast in the general representation given below, following the notation in Harvey (1989).

$$\mathbf{y}_t = \mathbf{d}_t + \mathbf{Z}_t \mathbf{b}_t + \mathbf{u}_t, \qquad \mathbf{u}_t \sim i.i.d.N\left(\mathbf{0},\, \Sigma_{t,u}\right), t = 1, \ldots, T \tag{C-1}$$

$$\mathbf{b}_t = \mathbf{c}_t + \mathbf{T}_t \mathbf{b}_{t-1} + \mathbf{R}_t \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim i.i.d.N\left(\mathbf{0},\, \Sigma_{t,\eta}\right) \tag{C-2}$$

and $\mathrm{E}\left(\boldsymbol{\eta}_t \mathbf{u}_t'\right) = \mathbf{G}_t$. Below, we abstract from issues arising from the estimation of the parameters of the model which enter the matrices $\mathbf{c}_t$, $\mathbf{Z}_t$, $\Sigma_{t,u}$, $\Sigma_{t,\eta}$, $\mathbf{d}_t$, $\mathbf{T}_t$,$\mathbf{G}_t$ and $\mathbf{R}_t$ and concentrate on the estimation of the state vector $\mathbf{b}_t$ conditioned on those known parameters. Denote the estimate of $\mathbf{b}_t$ conditional on the information set $\mathcal{I}_{t-1}$ as $\hat{\mathbf{b}}_{t|t-1}$ and that conditional on the information set up to and including time t by $\hat{\mathbf{b}}_t$. Denote the covariance matrices of the estimators $\hat{\mathbf{b}}_{t|t-1}$ and $\hat{\mathbf{b}}_t$ as $\hat{\mathbf{P}}_{t|t-1}$ and $\hat{\mathbf{P}}_t$, respectively. The Kalman filter is initialised by specifying $\mathbf{b}_0$ and $\mathbf{P}_0$. Then, estimation of $\hat{\mathbf{b}}_t$ by the Kalman filter comprises sequential application of the following two sets of Equations

$$\begin{aligned} \hat{\mathbf{b}}_{t|t-1} &= \mathbf{c}_t + \mathbf{T}_t \hat{\mathbf{b}}_{t-1}, \\ \hat{\mathbf{P}}_{t|t-1} &= \mathbf{T}_t \hat{\mathbf{P}}_{t-1} \mathbf{T}_t' + \mathbf{R}_t \Sigma_{t,\eta} \mathbf{R}_t', \end{aligned} \tag{C-3}$$

known as the prediction Equations, and

$$\begin{aligned} \hat{\mathbf{b}}_t &= \hat{\mathbf{b}}_{t|t-1} + \left(\hat{\mathbf{P}}_{t|t-1} \mathbf{Z}_t' + \mathbf{R}_t \mathbf{G}_t\right) \mathbf{F}_t^{-1} \left(\mathbf{y}_t - \mathbf{Z}_t \hat{\mathbf{b}}_{t|t-1} - \mathbf{d}_t\right), \\ \hat{\mathbf{P}}_t &= \hat{\mathbf{P}}_{t|t-1} - \left(\hat{\mathbf{P}}_{t|t-1} \mathbf{Z}_t' + \mathbf{R}_t \mathbf{G}_t\right) \mathbf{F}_t^{-1} \left(\mathbf{Z}_t \hat{\mathbf{P}}_{t|t-1} + \mathbf{G}_t' \mathbf{R}_t'\right), \end{aligned} \tag{C-4}$$

known as the updating Equations, where

$$\mathbf{F}_t = \mathbf{Z}_t \hat{\mathbf{P}}_{t|t-1} \mathbf{Z}_t' + \mathbf{Z}_t \mathbf{R}_t \mathbf{G}_t + \mathbf{G}_t' \mathbf{R}_t' \mathbf{Z}_t' + \Sigma_{t,u}. \tag{C-5}$$

The set of smoothed estimates (i.e. the estimate of $\mathbf{b}_t$ conditional on the information set $\mathcal{I}_T$) and their respective covariance matrices, denoted by $\hat{\mathbf{b}}_{t|T}$ and $\hat{\mathbf{P}}_{t|T}$, are given by

$$\hat{\mathbf{b}}_{t|T} = \hat{\mathbf{b}}_t + \mathbf{P}_t^* \left(\hat{\mathbf{b}}_{t+1|T} - \mathbf{T}_{t+1} \hat{\mathbf{b}}_t\right) \tag{C-6}$$

and

$$\hat{\mathbf{P}}_{t|T} = \hat{\mathbf{P}}_t + \mathbf{P}_t^* \left(\hat{\mathbf{P}}_{t+1|T} - \hat{\mathbf{P}}_{t+1|t}\right) \mathbf{P}_t^{*\prime}, \tag{C-7}$$

where $\mathbf{P}_t^* = \hat{\mathbf{P}}_t \mathbf{T}_{t+1}' \hat{\mathbf{P}}_{t+1|t}^{-1}$.

The log-likelihood function for the system, denoted by $\mathcal{L}(\boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta}$ denotes the vector of parameters with respect to which the log likelihood is maximised. It can be written in terms of the prediction errors $\boldsymbol{\varpi}_t = \mathbf{y}_t - \mathbf{Z}_t\hat{\mathbf{b}}_{t|t-1} - \mathbf{d}_t$ as

$$\mathcal{L}(\boldsymbol{\vartheta}) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\varpi}_t'\mathbf{F}_t^{-1}\boldsymbol{\varpi}_t. \qquad \textbf{(C-8)}$$

This log likelihood function $\mathcal{L}(\boldsymbol{\vartheta})$ can be used to estimate the unknown parameters of the model, $\boldsymbol{\vartheta}$. The matrices $\mathbf{F}_t$ and $\boldsymbol{\varpi}_t$ are dependent on the matrices $\mathbf{c}_t$, $\mathbf{Z}_t$, $\Sigma_{t,u}$, $\Sigma_{t,\eta}$, $\mathbf{d}_t$, $\mathbf{T}_t$, $\mathbf{G}_t$, $\mathbf{R}_t$, $\mathbf{b}_0$ and $\mathbf{P}_0$.

C.2 Representation of the data uncertainty model

The solution method described above is general to all linear state space models. In the remainder of this Annex, we give further details of its application to the model developed in Section 2. In that model, the parameter vector $\boldsymbol{\vartheta}$ comprises
$= \left(\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \ldots, \boldsymbol{\alpha}_q', \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^1}^2, \delta, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p', \mathbf{c}^{1'}, \lambda, \boldsymbol{\rho}_{\epsilon\varepsilon}, \boldsymbol{\sigma}_{\mathbf{v}^s}^{2'}, \boldsymbol{\mu}', \boldsymbol{\sigma}_{\epsilon}^{\mathbf{2}'}, \mathbf{c}^s, \mathbf{Z}^s\right).$

The model is multivariate with all the parameter matrices assumed diagonal, so:

- The parameters of the transition Equation, given by $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_q$, are defined by $\boldsymbol{\alpha}_i = \text{diag}(\mathbf{A}_i)$;
- The variance of the shocks to that Equation by $\boldsymbol{\sigma}_{\epsilon}^2 = \text{diag}(\Sigma_{\epsilon})$;
- The heteroscedastic variance of measurement errors in the published data by $\Sigma_{\boldsymbol{\varepsilon}}^{T-t}$ - a diagonal matrix whose diagonal elements are a function of $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}^1}^2$ and δ.
- Serial correlation in those measurement errors by $\boldsymbol{\beta}_i = \text{diag}(\mathbf{B}_i)$;
- The covariance between measurement errors of differing maturities and shocks to the transition Equation by $\boldsymbol{\zeta}_{\epsilon\varepsilon}^{T-t}$ - a diagonal matrix whose diagonal elements are a function of $\boldsymbol{\rho}_{\epsilon\varepsilon}$, $\Sigma_{\boldsymbol{\varepsilon}}^{T-t}$ and $\boldsymbol{\sigma}_{\epsilon}^2$.
- The variance of measurement errors attaching to indicators by $\boldsymbol{\sigma}_s^2 = \text{diag}(\Sigma_s)$.

Then we have the following setup

$$\mathbf{d}_t = \begin{pmatrix} \mathbf{c}^1 (1+\lambda)^{T-t-1} \\ \mathbf{c}^s \end{pmatrix},$$

$$\mathbf{c}_t = \boldsymbol{\mu},$$

$$\mathbf{Z}_t = \begin{pmatrix} \mathbf{I} & \dots & \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} \\ \mathbf{Z}^s & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix},$$

$$\mathbf{T}_t = \begin{pmatrix} \mathbf{A}_1 & \dots & \dots & \mathbf{A}_q & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{B}_1 & \dots & \dots & \mathbf{B}_p \\ \mathbf{0} & \ddots & \ddots & \mathbf{0} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} \end{pmatrix},$$

$$\Sigma_{t,u} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_s \end{pmatrix},$$

$$\Sigma_{t,\eta} = \begin{pmatrix} \Sigma_\epsilon & \mathbf{0} & \dots & \mathbf{0} & \zeta_{\epsilon\varepsilon}^{T-t} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \zeta_{\epsilon\varepsilon}^{T-t} & \mathbf{0} & \dots & \mathbf{0} & \Sigma_\varepsilon^{T-t} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix}$$

$$\mathbf{R}_t = \mathbf{I},$$

$$\mathbf{G}_t = \mathbf{0},$$

$$\mathbf{b}_0 = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix},$$

and

$$\mathbf{P}_0 = \Big(\mathbf{I} - \mathbf{T}_0(\boldsymbol{\vartheta})\Big)^{-1} \Sigma_{0,\eta}(\boldsymbol{\vartheta}).$$

C.3 2-step estimation: Imposition of parameters estimated over the real-time dataset

As described in the main text, in estimation we set some parameters to constants having obtained suitable values for them *via* prior estimation (Section 3.2). Then the maximum likelihood estimation problem becomes one where the log likelihood is maximised with respect to $\boldsymbol{\vartheta}_1$ keeping $\boldsymbol{\vartheta}_2$ constant; where $\boldsymbol{\vartheta} = \left(\boldsymbol{\vartheta}_1', \boldsymbol{\vartheta}_2'\right)'$ is some suitable partition of $\boldsymbol{\vartheta}$. With the heteroscedasticity, serial correlation, bias and correlation parameters estimated over the real-time dataset and assuming for simplicity that $m = 1$, the partition is $\boldsymbol{\vartheta}_1 = \left(\alpha_1, \alpha_2, \ldots, \alpha_q, \sigma^2_{\mathbf{v}^s}, \mu, \sigma^2_{\boldsymbol{\epsilon}}, c^s, \mathbf{Z}^s\right)'$ and $\boldsymbol{\vartheta}_2 = \left(\sigma^2_{\varepsilon^1}, \delta, \beta_1, \ldots, \beta_p, \rho_{\epsilon\varepsilon}, c^1, \lambda\right)'$.

The vector of imposed parameters, $\boldsymbol{\vartheta}_2$, includes two that are not directly observed but map in a one-for-one fashion from directly observed features of the real-time dataset. Because we do not observe revisions net of serial correlation, we observe $\sigma^2_{\mathbf{v}^1}$ rather than $\sigma^2_{\boldsymbol{\varepsilon}^1}$. Similarly, we cannot estimate $\rho_{\epsilon\varepsilon}$, but can estimate ρ^*_{yv} directly.

Section 3.2.4 describes the use of the real-time dataset to estimate $\sigma^2_{v^1}$, δ and $B_1, \ldots, B_p$ 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)**

$$\sigma^2_{\varepsilon^1} = \sigma^2_{v^1}\left(1 - (1+\delta)\, B_1^2\right). \qquad \textbf{(C-9)}$$

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

$$v_t^{t+n} = \beta_1 v_{t-1}^{t+n} + \beta_2 v_{t-2}^{t+n} + \ldots + \beta_p v_{t-p}^{t+n} + \varepsilon_t^{t+n}. \qquad \textbf{(C-10)}$$

To derive $\mathbf{V}$ we need to build-up the matrix in p by p blocks. We can do this by writing Equation **(C-10)** in companion form as

$$\mathbf{v}_t = \mathbf{B}\mathbf{v}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad \textbf{(C-11)}$$

where $\mathbf{v}_t = \left(v_t^{t+n}, v_{t-1}^{t+n}, \ldots, v_{t-p}^{t+n}\right)'$, $\boldsymbol{\varepsilon}_t = \left(\varepsilon_t^{t+n}, 0, \ldots, 0\right)'$ 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

$$\operatorname{var}(\mathbf{v}_t) = \mathbf{B}\operatorname{var}(\mathbf{v}_{t-1})\,\mathbf{B}' + \operatorname{var}(\boldsymbol{\varepsilon}_t), \qquad \textbf{(C-12)}$$

where var is the variance operator.

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

$$\operatorname{vec}(\operatorname{var}(\mathbf{v}_t)) = (1+\delta)(\mathbf{B}\otimes\mathbf{B})\operatorname{vec}(\operatorname{var}(\mathbf{v}_t)) + \operatorname{vec}(\operatorname{var}(\varepsilon_t)). \qquad \textbf{(C-13)}$$

Rearranging gives

$$\operatorname{vec}(\operatorname{var}(\mathbf{v}_t)) = \left(\mathbf{I}_{p^2} - (1+\delta)\,\mathbf{B}\otimes\mathbf{B}\right)^{-1}\operatorname{vec}(\operatorname{var}(\varepsilon_t)). \qquad \textbf{(C-14)}$$

We can then build-up the full $\mathbf{V}_t$ matrix in a similar fashion to Equation **(12)**

$$\mathbf{V}_t = \begin{pmatrix} \mathbf{I}_p & (1+\delta)^p\,\mathbf{B} & (1+\delta)^{2p}\,\mathbf{B}^2 & \cdots & (1+\delta)^{kp}\,\mathbf{B}^k \\ (1+\delta)^p\,\mathbf{B} & (1+\delta)^p\,\mathbf{I}_p & (1+\delta)^{2p}\,\mathbf{B} & \cdots & (1+\delta)^{kp}\,\mathbf{B}^{k-1} \\ (1+\delta)^{2p}\,\mathbf{B}^2 & (1+\delta)^{2p}\,\mathbf{B} & (1+\delta)^{2p}\,\mathbf{I}_p & \cdots & (1+\delta)^{kp}\,\mathbf{B}^{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (1+\delta)^{kp}\,\mathbf{B}^k & (1+\delta)^{kp}\,\mathbf{B}^{k-1} & (1+\delta)^{kp}\,\mathbf{B}^{k-2} & \cdots & (1+\delta)^{kp}\,\mathbf{I}_p \end{pmatrix} \qquad \textbf{(C-15)}$$

$$\times\left(\mathbf{I}_{k+1}\otimes\operatorname{var}(\mathbf{v}_t)\right).$$

Again, it is invariant with respect to time. Taking the variance-covariance matrix to the data, we can estimate $B_1,\ldots,B_p,\sigma^2_{\varepsilon^1}$ and δ via GMM by minimising $\left(\operatorname{vec}(\mathbf{V}) - \operatorname{vec}\left(\widehat{\mathbf{V}}\right)\right)'\left(\operatorname{vec}(\mathbf{V}) - \operatorname{vec}\left(\widehat{\mathbf{V}}\right)\right)$ with respect to $B_1,\ldots,B_p,\sigma^2_{\varepsilon^1}$ and δ.

For higher orders of p, Equation **(C-14)** gives

$$\operatorname{vec}(\operatorname{var}(\varepsilon_t)) = \left(\mathbf{I}_{p^2} - (1+\delta)\,\mathbf{B}\otimes\mathbf{B}\right)\operatorname{vec}(\operatorname{var}(\mathbf{v}_t)), \qquad \textbf{(C-16)}$$

where $\sigma^2_{\varepsilon^1}$ is the first element of $\operatorname{vec}(\operatorname{var}(\varepsilon_t))$.

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

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t, \qquad \textbf{(C-17)}$$

where $\mathbf{y}_t = \left(y_t, \ldots, y_{t-p}\right)'$, $\boldsymbol{\epsilon}_t = (\epsilon_t, 0, \ldots, 0)'$ and $\mathbf{A} = \begin{pmatrix} A_1 & A_2 & \ldots & A_{q-1} & A_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix}$.

The covariance between $\mathbf{y}_t$ and $\mathbf{v}_t$ can be written as

$$\operatorname{cov}\left(\mathbf{y}_t, \mathbf{v}_t\right) = \mathbf{A} \operatorname{cov}\left(\mathbf{y}_{t-1}, \mathbf{v}_{t-1}\right) \mathbf{B}' + \operatorname{cov}\left(\boldsymbol{\epsilon}_t, \boldsymbol{\varepsilon}_t\right), \qquad \textbf{(C-18)}$$

where cov is the covariance operator. Using the identity $\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A}) \operatorname{vec}(\mathbf{B})$, we have

$$\operatorname{vec}\left(\operatorname{cov}\left(\mathbf{y}_t, \mathbf{v}_t\right)\right) = (\mathbf{B} \otimes \mathbf{A}) \operatorname{vec}\left(\operatorname{cov}\left(\mathbf{y}_{t-1}, \mathbf{v}_{t-1}\right)\right) + \operatorname{vec}\left(\operatorname{cov}\left(\boldsymbol{\epsilon}_t, \boldsymbol{\varepsilon}_t\right)\right). \qquad \textbf{(C-19)}$$

Recognising that $\operatorname{cov}\left(\mathbf{y}_{t-1}, \mathbf{v}_{t-1}\right) = \sqrt{(1+\delta)} \operatorname{cov}\left(\mathbf{y}_t, \mathbf{v}_t\right)$ we can rearrange to give

$$\operatorname{vec}\left(\operatorname{cov}\left(\mathbf{y}_t, \mathbf{v}_t\right)\right) = \left(\mathbf{I}_{pq} - \sqrt{(1+\delta)}\mathbf{B} \otimes \mathbf{A}\right)^{-1} \operatorname{vec}\left(\operatorname{cov}\left(\boldsymbol{\epsilon}_t, \boldsymbol{\varepsilon}_t\right)\right). \qquad \textbf{(C-20)}$$

The first-element in the vector on the right-hand side re-scales the covariance between $\mathbf{y}_t$ and $\mathbf{v}_t$ to the covariance between $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\varepsilon}_t$. To uncover the re-scaled correlation we also need to take account of the differences in variance between the dynamic series and the respective shocks. From Equation **(C-14)** we know that

$$\operatorname{vec}\left(\operatorname{var}\left(\mathbf{v}_t\right)\right) = \left(\mathbf{I}_{p^2} - (1+\delta)\, \mathbf{B} \otimes \mathbf{B}\right)^{-1} \operatorname{vec}\left(\operatorname{var}\left(\boldsymbol{\varepsilon}_t\right)\right). \qquad \textbf{(C-21)}$$

By similar reasoning, we also know that

$$\operatorname{vec}\left(\operatorname{var}\left(\mathbf{y}_t\right)\right) = \left(\mathbf{I}_{q^2} - \mathbf{A} \otimes \mathbf{A}\right)^{-1} \operatorname{vec}\left(\operatorname{var}\left(\boldsymbol{\epsilon}_t\right)\right). \qquad \textbf{(C-22)}$$

Putting **(C-20)** to **(C-22)** 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 $\left|\rho_{\epsilon\varepsilon}\right| \geq \left|\rho_{yv}\right|$. Intuitively, the correlation between the two autoregressive processes is a diluted version of the correlation between the two underlying shocks.