A State Space Approach To The Policymaker's Data Uncertainty Problem*

Alastair Cunningham, Chris Jeffery,
George Kapetanios† & Vincent Labhard‡
Bank of England
†Queen Mary and Westfield College and Bank of England
‡European Central Bank

PAPER PRESENTED AT THE 38th ANNUAL MONEY MACRO AND FINANCE CONFERENCE
13 September 2006

Abstract

The paper describes the challenges that uncertainty over the true value of key macroeconomic variables poses for the policymaker - or, indeed, other data-users - and the way in which she may form and update her view of the evolution of economic time-series in light of a range of indicators and models. Specifically, it casts the data uncertainty challenge in state space form and describes a two-step estimator for the resulting signal extraction problem. Real-time data are first used to estimate the statistical properties of any measurement errors embedded in published estimates of macroeconomic variables; and these properties are then imposed in maximum likelihood estimation of the full state space model. The paper also considers how the data-user's signal extraction solution might be related to any constraints that the statistical agency faces in treating uncertain data.

*This paper represents the views and analysis of the authors and should not be thought to represent those of the Bank of England, Monetary Policy Committee, or any other organisation to which the authors are affiliated. We have benefited from helpful comments from Andrew Blake, Spencer Dale, Jana Eklund, Lavan Mahadeva and Tony Yates.

1 Introduction

Most data used in macroeconomic analysis are estimates of the 'true' outcome. One symptom of the resulting data uncertainties is the propensity of statistical agencies to revise their estimates in light of new information (bigger samples) or methodological advances (better proxies). In practice, these revisions have often appeared large relative to the variation observed in the published data. So in the UK for example, between 1993 and 2003 the mean absolute revision to quarterly real GDP growth was 0.2pp over the first three years from the initial release; relative to average GDP growth of 0.6%. This issue is by no means unique to the UK: see Mitchell (2004) for a review of work establishing the scale of historical revisions and Öller and Hansson (2002) for a representative cross-country comparison.

Uncertainty about the true value of economic series now and in the past adds to the challenge of forming a forward-looking assessment of economic prospects and hence complicates policy formulation[1]. More specifically, naïve use of economic data, abstracting from data uncertainties, can worsen policymaking and forecast performance in two ways. First, the policymaker - or, indeed, any data-user - may misunderstand the nature of the relationship between economic variables. Parameter estimates may change as data are revised and so too might model selection. Second, data-users may be ill-informed about the recent and current values of those economic variables. Model outputs may change as input data are revised. One symptom of naïve use of uncertain data is that revisions sometimes lead to material swings in economic assessment. Another is that, as data are revised, policy actions differ substantially from the recommendations that might follow from the revised data. Kozicki (2004) describes this phenomenon as "policy regret". There is a sizeable literature seeking to estimate the extent of past policy regret. In a representative paper, Nelson and Nikolov (2003) estimate that in the UK during the late 1980s, an ex-post Taylor rule would have advocated a nominal interest rate 500bp higher than a real-time Taylor rule.

The data-user need not, however, treat uncertain data in such a naïve way. And, indeed, there is some evidence that data-users have allowed for data uncertainties in

[1]This uncertainty stems from difficulties in estimating the true value of economic series at all, rather than from the timeliness of those estimates. Lags in estimating the recent evolution of economic variables also pose challenges to data-users. The options available in dealing with these lags are explored in a growing literature on "nowcasting", of which Evans (2005) is a proponent. The model developed in this paper is better described as "backcasting".

interpreting macroeconomic data. For example, in reviewing revisions to the UK'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."

One strategy that the data-user might adopt in the face of uncertain estimates of the past evolution of macroeconomic variables is to consider robustness to such data uncertainty as an additional criterion in choosing between competing models or policy rules - for a more detailed discussion, see Kozicki (2004) *op cit.* A number of papers test the performance of competing rules and estimators in a real-time setting. For example, Orphanides (2003) finds that, given the revisions experience in the US, nominal income growth targets may outperform output-gap based policy rules. And Harrison, Kapetanios, and Yates (2004) suggest that where measurement uncertainties are greatest in estimates of the recent past, models in which recent experience is downweighted may have a superior forecasting performance to models in which all observations are weighted equally.

A second, complementary strategy is to process uncertain data more effectively by reviewing an array of competing indicators and/ or by assigning some weight to expectations of how data would evolve; rather than taking the latest estimate at face value. In other words, to treat interpretation of uncertain data as a signal extraction problem. Lomax (2004) describes the UK Monetary Policy Committee's approach to uncertainty in data and highlights both the use of an array of quantitative and qualitative indicators and the careful attention paid to the quality of competing indicators. And 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".

This paper explores the signal extraction problem, seeking to formalise the current practice of many macroeconomic commentators. Specifically, we set up a state space representation of the measurement errors surrounding published estimates and any alternative indictors, and use the model to estimate the 'true' value around which the measures are taken. Real-time data describing historical experience of revisions are used to estimate the properties of any measurement errors attaching to published estimates.

The paper has 5 further sections. The next section briefly describes the existing literature in the area. Section 3 represents the signal extraction problem in state space, with the objective of capturing many of the features of the antecedent literature. Section 4 describes the use of historical revisions experience to calibrate some parameters of

the state space model, while Section 5 explores the implications for signal extraction of differing assumptions - or prior views - about the source of uncertainty in the data and about the actions of the statistical agency in extracting its own signal. The final section provides a practical illustration.

2 An Overview of the Literature

This paper focuses on the signal extraction problem faced by the data-user in interpreting uncertain data - assigning some weight to face value data, some to alternative indicators, and some to her expectations of how the data would evolve (i.e. the output of some transition law). In doing so, it follows a long-standing literature, of which Howrey (1978) was an early proponent. The common strand of the literature is to estimate 'true' data using some form of state space model. A variety of estimators have been applied to the problem. The authors also differ in the features of measurement errors exploited in the signal extraction solution, as summarised in Table 1 for a selection of papers. The simplest possible setting would be to assume that: measurements were unbiased; measurement errors iid; data of differing maturities equally uncertain; and that both earlier releases and alternative indicators were subsumed in the latest published estimates. Then, the solution of the signal extraction problem is simply a matter of estimating the signal-noise ratio across the latest data release. All the papers cited enrich the model in some ways.

Table 1: Features of the Signal Extraction Solution Covered by Differing Estimators

	Bias in estimates	Serial correlation in measurement errors	Correlation between measurement errors and economic shocks	Persistence of measurement errors in mature data	Differing vintages of data as competing measures	Differing indicators as competing measures
Howrey (1978)	✓	✓	×	×	×	×
Sargent (1989)	✓	✓	✓	✓	×	×
Patterson (1994)	✓	✓	×	✓	×	×
Garratt et al (2005)	✓	×	×	✓	✓	×
Ashley et al (2005)	✓	×	×	×	×	✓
This paper	✓	✓	✓	✓	×	✓

- *Bias in estimates.* All of the authors cited correct for any systematic biases apparent in previous preliminary estimates. Such biases appear to have been endemic in National Accounts data in the UK and elsewhere, as documented, for example, in Akritidis (2003), and Garratt and Vahey (2004).

- *Serial correlation in measurement errors.* Many authors allow for serial correlation in the revisions process, which also appears to have been a common feature of macroeconomic data (see, for example, Howrey (1984)). There is less common ground in the treatment of the other identifying assumptions set out in the Table.

- *Correlation between measurement errors and economic shocks.* Most authors consider measurement errors to be uncorrelated with economic shocks (the disturbance term in the state equation). The exception is Sargent (1989) whose model permits the statistical agency to filter data prior to publication.

- *Persistence of measurement errors in mature data.* Howrey (1978) and Ashley, Driver, Hayes, and Jeffery (2005) restrict attention to revisions occurring in the first few quarters after the preliminary release. Ashley, Driver, Hayes, and Jeffery (2005) justify this on the grounds that in the UK, National Accounts data are fully balanced by the second Blue Book after the initial release (i.e. after around two years). Patterson (1994) and Garratt, Lee, Mise, and Shields (2005) consider revisions to more mature data, on the grounds of historical revisions experience.

- *Role of differing vintages of data as competing indicators.* Most authors assume that the latest estimate of economic activity at any particular point in time subsumes all previous estimates. In other words, that there is no need to consider differing vintages of data as competing measures. The exception is Garratt, Lee, Mise, and Shields (2005).

- *Role of measures other than those published by the statistical agency.* Most authors consider only the statistical agency's estimates as candidate measures. Ashley, Driver, Hayes, and Jeffery (2005) suggest augmenting those estimates with alternative indicators available to the data-user. That would be consistent with the wide array of indicators monitored by policymakers (see Lomax (2004)) and is the approach pursued in this paper.

3 A General State Space Model of Uncertain Data

In this section, we present a relatively general state space representation of the signal extraction problem; designed to capture many of the features explored in the antecedent literature. However, although the objective is to retain flexibility, even at this stage we make a number of identifying assumptions whose violation might cause the performance of the estimation algorithm to deteriorate. In Section 5, we explore the use of prior views of the source of data uncertainties and the actions of the statistical agency, to motivate further restrictions on the model.

3.1 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

$$\mathbf{A}(L)\left(\mathbf{y}_t - \boldsymbol{\mu}\right) = \boldsymbol{\epsilon}_t \tag{1}$$

where $\boldsymbol{A}(L) = 1 - \boldsymbol{A}_1 L - \ldots - \boldsymbol{A}_q L^q$ is a lag polynomial whose roots are outside the unit circle; $\boldsymbol{\mu}$ is a vector of constants; and $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}_\epsilon$. We further assume that $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_q$ are diagonal, so that the true value of each variable of interest is related only to its own historical values. This representation has a number of limiting features in practical application:

- 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.
- 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.
- 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.

3.2 The statistical agency's proprietary information

Let $\mathbf{y}_t^{s|t+n}$ denote a noisy estimate of $\mathbf{y}_t$ obtained by the statistical agency at time $t+n$, $n = 1, \ldots T-t$, but not observable by other economic agents. As discussed in Section 5.1, the statistical agency's measure might be obtained through statistical returns covering a sample of activity, or through indirect measurement (use of proxies). The model for $\mathbf{y}_t^{s|t+n}$ is given by

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

The constant term $\mathbf{c}^{s|n}$ is included to permit consideration of biases in the statistical agency's data-set. The n superscript allows for observations of different maturities to be differently biased. We assume that measurement errors $\mathbf{v}_t^{s|t+n}$ are distributed normally with finite variance. Consistent with this assumption, the properties of the error term $\mathbf{v}_t^{s|t+n}$ depend on the measurement technology applied by the statistical agency in two ways:

- *Whether the statistical agency receives more information as data become more mature.* Were the statistical agency to receive additional information subsequent to its initial estimate $\mathbf{y}_t^{s|t+n+1}$ would be based on a larger sample than $\mathbf{y}_t^{s|t+n}$ and the variance of measurement errors might decline as maturity increases; in line with the intuition described in Kapetanios and Yates (2004).

  Similarly, in the latest data release - published at time T>t - the statistical agency's observation of the value of the variable that prevailed at time t ($\mathbf{y}_t^{s|T}$) will be based on a smaller sample than its observation of period t-1 ($\mathbf{y}_{t-1}^{s|T}$). Importantly, this recognises that any data release will include observations of differing maturities; ranging from preliminary estimates of the most recent past through more mature observations of data points that were first observed some years previously. In the interests of generality, we therefore assume that $\mathbf{v}_t^{s|t+n}$ has heteroscedasticity with respect to n, so that $E\left(\mathbf{v}_t^{s|t+n}\left(\mathbf{v}_t^{s|t+n}\right)'\right) = \mathbf{\Sigma}_v^n$. Homoscedastic errors, such as might arise were the statistical agency to receive no further information after its initial estimate, nest within this representation with $\mathbf{\Sigma}_v^i = \mathbf{\Sigma}_v^j$ for all maturities i, j.

- *Whether the statistical agency observes* $\mathbf{y}_t$ *directly or not.* Were the statistical agency's measure to cover a randomly drawn and representative sample of economic activity, $\mathbf{v}_t^{s|t+n}$ would be distributed independently of previous measurement errors. But other measurement technologies might generate serially correlated errors. So,

in the interests of generality, we allow that $\mathbf{v}_t^{s|t+n}$ is serially correlated, so that $E\left(\mathbf{v}_t^{s|t+i}\left(\mathbf{v}_t^{s|t+j}\right)'\right) = \mathbf{\Sigma}_v^{ij}$ for any measurement errors of maturities i and j. Non-serially correlated errors nest within this general representation, with $\mathbf{\Sigma}_v^{ij} = 0$.

3.3 The statistical agency's published estimate

The statistical agency publishes an estimate of $\mathbf{y}_t$, at time $t+n$, denoted by $\tilde{\mathbf{y}}_t^{s|t+n}$. This estimate is, of course, observed by the other economic agents. The distinction between the measures observed by the statistical agency and the estimates that it publishes is introduced to permit consideration of the way in which the statistical agency's actions in the face of data uncertainty affect the data-user's signal extraction solution (see Section 5).

The model for these published data is

$$\tilde{\mathbf{y}}_t^{s|t+n} = \mathbf{y}_t + \tilde{\mathbf{c}}^{s|n} + \tilde{\mathbf{v}}_t^{s|t+n} \tag{3}$$

where the properties of the error term $\tilde{\mathbf{v}}_t^{s|t+n}$ depend on the error term $\mathbf{v}_t^{s|t+n}$ and on the modelling choices of the statistical agency; and where the properties of $\tilde{\mathbf{c}}^{s|n}$ depend on the extent of bias in the statistical agency's observation $\left(\mathbf{c}^{s|n}\right)$ and on the statistical agency's modelling choices [i.e. whether any biases are adjusted for].

We make a number of modelling assumptions regarding the form of serial correlation, heteroscedasticity and bias in the published estimates:

- *Serial correlation.* Consistent with the treatment of $\mathbf{v}_t^{s|t+n}$, we allow that $\tilde{\mathbf{v}}_t^{s|t+n}$ is serially correlated. Specifically, we model serial correlation in the errors attaching to the data in any data release published at t+n, as

$$\mathbf{B}\left(L\right)\tilde{\mathbf{v}}_t^{s|t+n} = \tilde{\boldsymbol{\varepsilon}}_t^{s|t+n} \tag{4}$$

where $\mathbf{B}(L) = 1 - \mathbf{B}_1 L - \ldots - \mathbf{B}_p L^p$ is a lag polynomial whose roots are outside the unit circle and $E\left(\tilde{\boldsymbol{\varepsilon}}_t^{s|t+n}(\tilde{\boldsymbol{\varepsilon}}_t^{s|t+n})'\right) = \mathbf{\Sigma}_{\tilde{\varepsilon}}^{t+n}$ as we are allowing for heteroscedasticity in measurement errors. This representation picks up serial correlation between errors attaching to the various observations within each data release. Equation (4) imposes some structure on $\tilde{\mathbf{v}}_t^{s|t+n}$ 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 measurement errors in the statistical agency's published estimates of each variable are related only to historical measurement errors in published estimates of that variable rather than to measurement errors in estimates of other variables.

- *Heteroscedasticity.* Consistent with the treatment of $\mathbf{v}_t^{s|t+n}$, we allow that $\tilde{\mathbf{v}}_t^{s|t+n}$ and therefore $\tilde{\boldsymbol{\varepsilon}}_t^{s|t+n}$ has heteroscedasticity with respect to n, so that $E\left(\tilde{\boldsymbol{\varepsilon}}_t^{s|t+n}(\tilde{\boldsymbol{\varepsilon}}_t^{s|t+n})'\right) = \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varepsilon}}}^n$ and $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varepsilon}}}^n$ depends on n. Specifically, we model $(\boldsymbol{\sigma}^2_{\tilde{\boldsymbol{\varepsilon}}^{s|n}} = \boldsymbol{\sigma}^2_{1\tilde{\varepsilon}^{s|n}}, \ldots, \boldsymbol{\sigma}^2_{m\tilde{\varepsilon}^{s|n}})$ where $\boldsymbol{\sigma}^2_{i\tilde{\varepsilon}^{s|n}} = E\left(\tilde{\boldsymbol{\varepsilon}}_{it}^{s|t+n}\right)^2$ and $\tilde{\boldsymbol{\varepsilon}}_t^{s|t+n} = (\tilde{\boldsymbol{\varepsilon}}_{i1}^{s|t+n}, \ldots, \tilde{\boldsymbol{\varepsilon}}_{im}^{s|t+n})'$. Then the model for $\boldsymbol{\sigma}^2_{\tilde{\boldsymbol{\varepsilon}}^{s|n}}$ is given by

$$\boldsymbol{\sigma}^2_{\tilde{\boldsymbol{\varepsilon}}^{s|n}} = \boldsymbol{\sigma}^2_{\tilde{\varepsilon}^{s|1}} \tilde{\boldsymbol{\omega}}_h^s(n) \tag{5}$$

where $\tilde{\boldsymbol{\omega}}_h^s(n)$ denotes some vector of monotonically declining functions of n such that $\tilde{\boldsymbol{\omega}}_h^s(1) = (1, \ldots, 1)'$; and $\boldsymbol{\sigma}^2_{\tilde{\varepsilon}^{s|1}}$ is the variance of measurement errors at maturity $n = 1$. This representation imposes some structure on the variance of measurement errors because we assume that that variance declines monotonically as the statistical agency's observations become more mature. Monotonicity 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) *op cit.* We further assume that the matrix $\tilde{\boldsymbol{\omega}}_h^s$ is diagonal so that the variance of measurement errors for each variable of interest at maturity n is related only to the variance of measurement errors attaching to earlier maturities of that variable.

- *Bias.* Consistent with the treatment of $\mathbf{c}^{s|n}$, we allow that bias in published estimates may vary with maturity. Specifically, we model $\tilde{\mathbf{c}}^{s|n}$ as

$$\tilde{\mathbf{c}}^{s|n} = \tilde{\mathbf{c}}^{s|1} \tilde{\boldsymbol{\omega}}_b^s(n) \tag{6}$$

where $\tilde{\boldsymbol{\omega}}_b^s(n)$ denotes some vector of monotonically declining functions of n such that $\tilde{\boldsymbol{\omega}}_b^s(1) = (1, \ldots, 1)'$; and $\tilde{\mathbf{c}}^{s|1}$ is the bias in published data of maturity $n = 1$. This representation imposes some structure on the bias in measurement, because we assume that the bias tends monotonically towards zero as the statistical agency's observations become more mature. In contrast with monotonicity in the variance of measurement errors, this treatment is not motivated by any view of the statistical agency's practices. We are not aware of any convincing explanation of the potential sources of bias in initial estimates. In common with the rest of the model, we assume that the matrix $\tilde{\boldsymbol{\omega}}_b^s$ is diagonal so that bias in published estimates of each variable of interest is considered independent of bias to other variables of interest.

As mentioned above, the distinction between the measures observed by the statistical agency and the estimates that it publishes is introduced to permit consideration of the statistical agency's actions in the face of data uncertainty - in other words for the possibility that the statistical agency recognises its own signal extraction problem. A spectrum of possible behaviours can be envisaged for the statistical agency (see Sargent (1989)). At one extreme, the agency might be thought of as simply a reporting agency which compiles the information it collects via statistical returns. At the other, the agency might apply its own economic models to enhance signal extraction and might draw on a variety of alternative indicators to complement the statistical returns. In between those poles, Mankiw and Shapiro (1986) argue that when statistical staff "meet to evaluate and adjust the estimates before they are released" they are implicitly applying some sort of filtering model.

Any filtering activities by the statistical agency have the potential to affect the data-user's signal extraction solution. For illustrative purposes, Annex B works through the impact of filtering by the statistical agency in a highly simplified version of the general model described here. In that example, the statistical agency's filtering model is the same as the data-user's and hence fully resolves the signal extraction problem - the data-user's estimate of the truth is identical to the published data. More generally, where the statistical agency applies any sort of filtering model we might expect some correlation between measurement errors in published estimates and economic shocks ($\boldsymbol{\epsilon}_t$). The intuition is that, in filtering their proprietary information, statistical staff attach some weight to their own transition law so that shocks to that transition law will be reflected in the published data. If the statistical agency's filters are set up well, this correlation will be negative and the published estimate will appear closer to the state space solution than would be the case with no correlation.

Absent any view of the way in which the statistical agency applies filtering models, we do not impose any structure on the correlation. So, for any variable of interest, we write

$$E(\tilde{\varepsilon}_t^{s|t+n}\epsilon_t') = \rho_{\tilde{\varepsilon}\epsilon}^{s}\sigma_{\epsilon}\sigma_{\tilde{\boldsymbol{\varepsilon}}^{s|t+n}} \tag{7}$$

In line with the treatment in the rest of the model, we assume that any covariance matrix across variables is diagonal so that measurement errors attaching to published estimates of one variable are independent of shocks to the transition law driving other variables. Given this assumption, there is no need to generalise equation (7) to a multivariate setting.

3.4 Alternative indicators

In addition to the statistical agency's estimate, the data-user can observe a range of alternative indicators of the variable of interest that were not exploited by the statistical agency; such as private sector business surveys. We denote the set of these indicators by $\mathbf{y}_t^i$, $t = 1, \ldots, T$. Unlike published estimates, the alternative indicators are not direct measures of the underlying variables. And, indeed, in practical application many of the alternative indicators available to us are not measured in the same units as the variable of interest - for example unlike macroeconomic variables, private sector business surveys typically report the proportion of respondents answering in a particular category. The alternative indicators are therefore assumed to be only linearly related to the true data rather than being direct measures of them

$$\mathbf{y}_t^i = \mathbf{c}^i + \mathbf{Z}^i \mathbf{y}_t + \mathbf{v}_t^i \tag{8}$$

The error term $\mathbf{v}_t^i$ is assumed to be i.i.d. with variance $\mathbf{\Sigma}_i$. This, of course, is more restrictive than the model for $\tilde{\mathbf{v}}_t^{s|t+n}$. In particular, the model does not exploit:

- Any heteroscedasticity or serial correlation in measurement errors associated with the indicators;
- Any correlation between transition shocks and the measurement errors surrounding the alternative indicators;
- Any correlation between the measurement errors attaching to the alternative indicators and those attaching to the published estimates. This is a restrictive assumption, as it requires that the statistical agency not consider the alternative indicators when solving its own signal extraction problem. Section 5.1 gives a qualitative discussion of the reasons why a statistical agency might not make use of available alternative indicators.

To summarise the model, we give its complete state space form for the latest available release; where the two equations describe measurement and transition. The model treats the latest vintage of data published by the statistical agency and the latest vintage of any alternative indicators as competing measures of the variable of interest.

$$\begin{pmatrix} \tilde{\mathbf{y}}_t^{s|T} \\ \mathbf{y}_t^i \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{c}}^{s|n} \\ \mathbf{c}^i \end{pmatrix} + \begin{pmatrix} \mathbf{I} & \dots & \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} \\ \mathbf{Z}^i & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y}_t \\ \dots \\ \mathbf{y}_{t-q+1} \\ \tilde{\mathbf{v}}_t^{s|T} \\ \dots \\ \tilde{\mathbf{v}}_{t-p+1}^{s|T} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{v}_t^i \end{pmatrix} \tag{9}$$

$$\begin{pmatrix} \mathbf{y}_t \\ \dots \\ \mathbf{y}_{t-q+1} \\ \tilde{\mathbf{v}}_t^{s|T} \\ \dots \\ \tilde{\mathbf{v}}_{t-p+1}^{s|T} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1 & \dots & \dots & \mathbf{A}_q & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{B}_1 & \dots & \dots & \mathbf{B}_p \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-1} \\ \dots \\ \mathbf{y}_{t-q} \\ \tilde{\mathbf{v}}_{t-1}^{s|T} \\ \dots \\ \tilde{\mathbf{v}}_{t-p}^{s|T} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_t \\ \dots \\ \mathbf{0} \\ \tilde{\boldsymbol{\varepsilon}}_t^{s|T} \\ \dots \\ \mathbf{0} \end{pmatrix} \tag{10}$$

The state space problem represented by equations (9) and (10) 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. Details about the Kalman filter may be found in the above references. The form of the filter is also given in Appendix A.

In principle, all the parameters of the model could be estimated via maximum likelihood using the Kalman filter - so long as the functions $\tilde{\boldsymbol{\omega}}_b^s$ and $\tilde{\boldsymbol{\omega}}_h^s$, the lag orders q and p, and the dimensions of the vector of alternative indicators $\mathbf{y}_t^i$ are set sufficiently parsimoniously to be soluble over the sample of data available. But in doing so, we exploit only the properties of the latest release of published data and hence ignore any information about the properties of measurement errors embedded in previous releases. In other words, the estimator may not make the most efficient use of the evidence available to us. One symptom of the inefficiency of using only the latest release is the authors' previous experience of severe numerical problems in the maximisation of the log likelihood in models of this type.

In the face of this estimation challenge, one possibility would be to model previous data releases within the state space setting; with some cross-variable restrictions on the properties of the measurement errors. However, setting up the model in this way would require estimation of the variance-covariance matrix across the measurement errors under differing releases. In the absence of any model describing the data release process, the number of parameters to estimate would increase materially and the estimation burden could well prove intractable.

One model would be to assume that the latest vintage subsumes all earlier estimates - in other words that the statistical agency processes data efficiently. But in that case, there is no efficiency gain from modelling earlier vintages alongside the latest vintage. Instead, the approach taken in this paper is to estimate the model in two steps - trading off the inefficiency of two-step estimation against the increase in the sample of measurement errors. In the first step, the properties of measurement errors in the statistical agency's published estimates - that is the parameters driving equations (4) through (6) - are estimated across a real-time data set. In the second step, the remaining parameters are estimated via maximum likelihood using the Kalman filter as outlined in Appendix A[2]. In other words, patterns in the historical revisions dataset are assumed to be representative of the statistical properties of the measurement errors surrounding the latest data release, but the data-user does not assign any weight to previous data releases in forming a view of $\mathbf{y}_t$. The next Section describes the use of real-time data to estimate properties of measurement errors in published data.

4 Using real-time data to estimate the properties of measurement errors

As described above, in the application of the Kalman filter, we impose parameters for equations (4) through (6) drawn from analysis of historical revisions experience - in

[2]This estimation strategy leaves the potential correlation between measurement and economic shocks ($\rho^s_{\tilde{\varepsilon}\epsilon}$) estimated via maximum likelihood within the Kalman Filter. Because the covariance between measurement and economic shocks will be time-varying, there may be computational issues in maximum likelihood estimation. Where computational issues arise, an alternative is to exploit the properties of the real-time data - in other words exploiting the correlation observed between revisions and mature published estimates. Another alternative, to be pursued in further work, is to consider whether properties of real-time data can be used to inform Kalman estimation through selection of starting values or bounding ranges for numerical solution.

other words, exploiting the vintage (or real-time) data set. In drawing on historical revisions experience in this way, we assume that the properties of measurement errors are fully reflected in statistical properties of historical revisions to the statistical agency's published estimates. This mapping need not hold in practice, as the statistical agency may consider factors other than measurement uncertainty in reaching any policy decision over whether to make revisions to published back data in light of new evidence received or methodological advances made. For some macroeconomic aggregates - such as the UK's CPI - the statistical agency's policy is not to revise.

Using real-time data to estimate the properties of measurement errors requires us to first manipulate the real time dataset to derive a matrix of revisions to published data of differing maturities and then to estimate the parameters describing the measurement errors in the statistical agency's latest published release over those vectors. As a preliminary, recall that we have assumed $\tilde{\boldsymbol{\omega}}_h^s(n)$, $\mathbf{B}_1, \ldots, \mathbf{B}_p$, and $\tilde{\boldsymbol{\omega}}_b^s(n)$ to be diagonal. As a result, the functions can be calibrated for individual variables rather than for the system of all variables of interest. In the remainder of this section, we therefore consider calibration for a single variable and discard vector notation.

4.1 Manipulation of the real time data-set

The real time database for each variable of interest is an upper-triangular data matrix with two axes: publication (or vintage) dates along the horizontal axis and reference dates down the vertical axis. 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 2 shows an extract of the real-time database for distribution output used in the illustrative example developed in Section 6; and Table 3 describes the maturity of the various observations.

Table 2: Quarterly Growth of Distibution Output - extract from the real-time database

		Vintage date						
		2003 Q1	2003 Q2	2003 Q3	….	2005 Q3	2005 Q4	2006 Q1
Reference date	2002 Q4	0.6	0.3	0.9	⋮	1.3	1.3	1.3
	2003 Q1		0.1	-0.5	⋮	0.0	0.0	0.0
	2003 Q2			0.8	⋮	1.3	1.3	1.3
	⋮				⋮	⋮	⋮	⋮
	2005 Q3						0.2	0.1
	2005 Q4							1.1

Table 3: Stylised Real-Time Database - Maturity of Observations

		Vintage date						
		2003 Q1	2003 Q2	2003 Q3	….	2005 Q3	2005 Q4	2006 Q1
Reference date	2002 Q4	1	2	3	⋮	11	12	13
	2003 Q1		1	2	⋮	10	11	12
	2003 Q2			1	⋮	9	10	11
	⋮				⋮	⋮	⋮	⋮
	2005 Q3						1	2
	2005 Q4							1

Define the revisions to published estimates of an individual variable of interest $\tilde{y}_t^s$ between maturities n and j as

$$w_t^{s|j,n} = \tilde{y}_t^{s|t+j} - \tilde{y}_t^{s|t+n} \tag{11}$$

For calibration purposes, we take revisions over the J quarters subsequent to each data-point to be representative of the uncertainty surrounding that estimate. So for example, with $J = 24$, we evaluate uncertainties surrounding data of maturity 1 by considering revisions between the 1st and 25th release; and we evaluate uncertainties surrounding data of maturity 12 by considering revisions between the 12th and 37th release. 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 by $(W - J)$ matrix of revisions ($\mathbf{W}^{s|J}$) over which to estimate the parameters of equations (4) through (6). N and J are both choice variables and should be selected to maximise the efficiency of estimation of the parameters driving equations (4) through (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 to be representative. In the illustrative example in Section 6, we arbitrarily set $N = J = 16$.

The nt^{th} element of $\mathbf{W}^{s|J}$ is the revision to the published estimate of the value taken by $\tilde{y}^s$at time t between the vintage published at time $t + n$ and the "mature" estimate published at time $t + n + J$. Each column of the matrix therefore contains observations of revisions to data within a single data release. And each row describes revisions to data of a specific maturity n. In describing the properties of measurement errors, our interest is in tracing out any relationship between data uncertainties attaching to observations within a data release, as described below.

4.2 Calibrating heteroscedasticity and serial correlation in measurement errors

The variance-covariance matrix of historical revisions may be used to jointly estimate both the heteroscedasticity in measurement errors and their serial correlation. As a first step, we impose an arbitrary functional form for $\tilde{\omega}_h^s(n)$ - the function describing the decline in measurement error variance as maturity increases:

$$\tilde{\omega}_h^s(n) = (1+\delta)^{n-1} \tag{12}$$

Then the variance covariance matrix can be modelled as a function of $B_1, \ldots, B_p$, $\sigma_{\tilde{\varepsilon}}^{2s|1}$ and δ, as outlined below. Once specified, the parameters are estimated through application of (restricted) GMM.

Modelling the variance covariance matrix is trivial for errors that are first-order serially correlated:

$$\mathbf{V}_{\tilde{v}} = \frac{\sigma^2_{\tilde{\varepsilon}^{s|1}}}{1-\beta^2} \begin{bmatrix} 1 & (1+\delta)\beta & (1+\delta)^2\beta^2 & \cdots & (1+\delta)^{J-1}\beta^{J-1} \\ (1+\delta)\beta & (1+\delta) & (1+\delta)^2\beta & \cdots & (1+\delta)^{J-1}\beta^{J-2} \\ (1+\delta)^2\beta^2 & (1+\delta)^2\beta & (1+\delta)^2 & \cdots & (1+\delta)^{J-1}\beta^{J-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (1+\delta)^{J-1}\beta^{J-1} & (1+\delta)^{J-1}\beta^{J-2} & (1+\delta)^{J-1}\beta^{J-3} & \cdots & (1+\delta)^{J-1} \end{bmatrix}$$

Higher orders of p require some further manipulation. Following the model of serial correlation in measurement errors described in Section 3, the model is

$$\tilde{v}_t^{s|T} = \beta_1 \tilde{v}_{t-1}^{s|T} + \beta_2 \tilde{v}_{t-2}^{s|T} + \ldots + \beta_p \tilde{v}_{t-p}^{s|T} + \tilde{\varepsilon}_t, \; t = 1, \ldots, T$$

where we allow for heteroscedasticity in $\tilde{\varepsilon}_t$, i.e. $E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}') = \boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_T^2)$ where $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_T)'$. We want to derive the variance covariance matrix of $\tilde{\mathbf{v}} = (\tilde{v}_1^{s|T}, \ldots, \tilde{v}_T^{s|T})'$.

We proceed as follows:

- Let the model be written in companion form as

$$\tilde{\mathbf{v}}_t = \mathbf{B}\tilde{\mathbf{v}}_{t-1} + \tilde{\boldsymbol{\varepsilon}}_t$$

where $\tilde{\mathbf{v}}_t = (\tilde{v}_t^{s|T}, \tilde{v}_{t-1}^{s|T}, \ldots, \tilde{v}_{t-k}^{s|T})'$, $\tilde{\boldsymbol{\varepsilon}}_t = (\tilde{\varepsilon}_t, 0, \ldots, 0)'$ and $\mathbf{B} = \begin{pmatrix} \beta_1 & \beta_2 & \ldots & \beta_p \\ 1 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 \end{pmatrix}$.

- Stacking observations gives

$$\hat{\mathbf{v}} = \mathbf{x}\mathbf{B}' + \tilde{\boldsymbol{\varepsilon}}$$

where $\hat{\mathbf{v}} = (\tilde{\mathbf{v}}_{\mathbf{T}}, \ldots, \tilde{\mathbf{v}}_{\mathbf{p+1}})'$, $\mathbf{x} = (\tilde{\mathbf{v}}_{\mathbf{T-1}}, \ldots, \tilde{\mathbf{v}}_{\mathbf{p}})'$ and $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\boldsymbol{\varepsilon}}_{\mathbf{T}}, \ldots, \tilde{\boldsymbol{\varepsilon}}_{\mathbf{p+1}})'$.

- Then, using the identity $\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\,\text{vec}(\mathbf{B})$, we have

$$\text{vec}(\hat{\mathbf{v}}) = (\mathbf{B} \otimes \mathbf{I}_{T-p})\,\text{vec}\,(\mathbf{x}) + \text{vec}(\tilde{\boldsymbol{\varepsilon}})$$

and

$$\text{var}(\text{vec}(\hat{\mathbf{v}})) = (\mathbf{B} \otimes \mathbf{I}_{T-p})\,\text{var}(\text{vec}(\mathbf{x}))(\mathbf{B}' \otimes \mathbf{I}_{T-p}) + \text{var}(\text{vec}(\tilde{\boldsymbol{\varepsilon}})) \tag{13}$$

where var is the variance operator. Let $\tilde{\boldsymbol{\Sigma}} = \text{vec}(\text{vec}(\tilde{\boldsymbol{\varepsilon}}))$.

- Assume that $\mathbf{V} \equiv \text{var}\,(\text{vec}(\hat{\mathbf{v}})) = \text{var}(\text{vec}(\mathbf{x}))$. Note that the variance we are looking for is the top LH corner $T \times T$ submatrix of $\mathbf{V}$. Then, it follows from Equation (13) that

$$\text{vec}(\mathbf{V}) = ((\mathbf{B} \otimes \mathbf{I}_{T-p}) \otimes (\mathbf{B} \otimes \mathbf{I}_{T-p}))\,\text{vec}(\mathbf{V}) + \text{vec}(\tilde{\boldsymbol{\Sigma}})$$

or

$$\text{vec}(\mathbf{V}) = \left(\mathbf{I}_{((T-p)p)^2} - ((\mathbf{B} \otimes \mathbf{I}_{T-p}) \otimes (\mathbf{B} \otimes \mathbf{I}_{T-p}))\right)^{-1} \text{vec}(\tilde{\boldsymbol{\Sigma}}) \tag{14}$$

The above gives an expression of the variance covariance matrix of $\tilde{\mathbf{v}}$, denoted $\mathbf{V}_{\tilde{v}}$, in terms of of the parameters $B_1, \ldots, B_p$, $\sigma^2_{\tilde{\varepsilon}^{s|1}}$ and δ. Let a suitably truncated version of $\mathbf{V}_{\tilde{v}}$ be denoted by $\mathbf{V}^{\tau}_{\tilde{v}}$ where the truncation allows only the first k autocovariances of $\tilde{v}_1^{s|T}$ to enter $\mathbf{V}^{\tau}_{\tilde{v}}$.

The matrix $\mathbf{V}^{\tau}_{\tilde{v}}$ describes the variance-covariance matrix of revisions as a function of the parameters of interest. A sample estimate of the variance-covariance matrix $\hat{\mathbf{V}}^{\tau}_{\tilde{v}}$ can also be calculated trivially from the matrix of historical revisions $\mathbf{W}^{s|J}$. Then GMM estimation of $B_1, \ldots, B_p$, $\sigma^{2s|1}_{\tilde{\varepsilon}}$ and δ amounts to minimising

$$\left(\text{vec}\,(\mathbf{V}^{\tau}_{\tilde{\mathbf{v}}}) - \text{vec}\left(\hat{\mathbf{V}}^{\tau}_{\tilde{\mathbf{v}}}\right)\right)' \left(\text{vec}\,(\mathbf{V}^{\tau}_{\tilde{\mathbf{v}}}) - \text{vec}\left(\hat{\mathbf{V}}^{\tau}_{\tilde{\mathbf{v}}}\right)\right) \tag{15}$$

with respect to $B_1, \ldots, B_p$, $\sigma^2_{\tilde{\varepsilon}^{s|1}}$ and δ.

4.3 Calibrating bias in measurement errors

We can use the sample of historical revisions in matrix $\mathbf{W}^{s|J}$ to calibrate $c_{\tilde{v}}^{s|1}$and $\tilde{\omega}_b^s(n)$. The sample means of revisions to estimates of each maturity $n = 1$ to N are simply the average of observations in each row of $\mathbf{W}^{s|J}$. Denote the average revision to data of maturity n by mean$(w^{s|J,n})$.

The data-user could use these mean revisions directly in modelling the bias attaching to estimates of differing maturities. That is, setting $\tilde{c}^{s|n} = \text{mean}(w^{s|J,n})$. The pitfalls of this approach are that: first, the small-sample properties of the real-time dataset may not match the data-user's prior view of the functional form of $\tilde{\omega}_b^s(n)$ if she has one; and second, that the estimator is not particularly efficient; requiring N parameters. The alternative, pursued in production of the illustrative example in Section 6, is to specify a functional form for $\tilde{\omega}_b^s(n)$. As outlined above, the desirable features we seek to enforce are first, that bias tends to zero as maturity tends to infinity - so that the statistical agency eventually arrives at an unbiased estimate of y_t - and second, that the decline is monotonic.

In, practical application, we impose the following arbitrary functional form, consistent with these features:

$$\text{mean}(w^{s|J,n}) = \tilde{c}^{s|1}(1+\lambda)^{n-1} + \psi_n \tag{16}$$

where $-1 \leq \lambda \leq 0$ and ψ_n denotes a disturbance term. The parameters $\tilde{c}^{s|1}$ and λ are then estimated over the vector of average revisions using (restricted) non-linear least squares.

5 Prior views about the nature of data uncertainty

The state space representation articulated in Section 3 is quite general in its treatment of the measurement errors associated with published estimates of economic variables. This generality enables us to exploit many of the patterns in historical revisions that may be apparent in real-time datasets. However, the data-user may not view this historical experience as representative. She may, for example, have a prior view that any serial correlation apparent in the real-time data is an accidental property of the small sample of vintage data available. The contention developed in this Section is that such priors might be informed by consideration of the sources of uncertainty in the data and of the actions taken by the statistical agency in the face of those uncertainties.

5.1 Sources of data uncertainty

There are two main sources of uncertainty in economic data; with differing implications for the set-up of any signal extraction problem. First, data uncertainties may arise where estimates are based on samples rather than complete information. Assuming sampling methodologies to be robust, it would be reasonable to expect measurement errors in published data not to be serially correlated. Alternatively, uncertainties may arise because the underlying economic concept is not directly observable - as is the case, for example, when seeking to measure value added in financial services. In this case, the statistical agency may make use of indirect measures, or proxies; modelling the relationship between observable concepts and the variable of interest, as described in Cook (2004). Depending on the methodologies chosen to construct indirect measures, it is possible that measurement errors will prove serially correlated - in other words the proxy cannot necessarily be considered as a noisy indicator of the variable of interest.

Were sample size to increase as data becomes more mature, we would expect the variance of measurement errors to fall as the statistical agency's proprietary information set grows. For example, as outlined above, Kapetanios and Yates (2004) develop one model in which the variance of measurement errors declines as data become more mature and the statistical agency receives more information. Both direct and indirect measures are typically constructed from samples of statistical returns.

The discussion suggests two routes through which consideration of the sources of data uncertainty might be used to motivate simplifying assumptions on the general model. The two priors can be treated independently.

- *If use of indirect measures is seen to be negligible* then the data-user may choose to ignore serial correlation and set $\mathbf{B}_1, \ldots, \mathbf{B}_p = 0$ from Equation (4).
- *If the statistical agency's proprietary information set is not seen to grow as data become more mature* then measurement errors will be homoscedastic with respect to maturity - for example, by setting $\delta = 0$ in equation (12).

5.2 The statistical agency's actions in the face of data uncertainty

One natural challenge is to ask why there might be a signal extraction problem for the data-user to address once the statistical agency has processed its proprietary information. In other words, to ask why the statistical agency does not solve its own signal extraction problem and publish results on that basis - identifying the constraints under which the statistical agency operates. Cook (2004) outlines a number of practical constraints on statistical measurement. Our contention is that as a provider of data, the statistical agency has to balance the potential for statistical inference to improve on face value treatment of data against the impact of any modelling approximations made on the transparency, coherence and credibility of the National Statistics as a whole. In contrast, as users of data, economists are free to make approximations and apply statistical inference. The threshold for economic analysis is simply whether the results have a better-than-evens chance of improving on what went before.

In practice in the UK, the ONS follow a detailed rulebook, conforming to international standards, when collecting and compiling data. That rule book may constrain the statistical staff in two main ways:

- **Available information set**. The foundations of any National Statistics are a range of statistical returns that record the experience of individuals and firms. The statistical agency may feel constrained in looking beyond these statistical returns and hence ignore some available indicators. The ONS prefer not to make use of private sector surveys, such as the CBI's survey of manufacturing trends, where it has direct measures available in the form of its own statistical returns (see Mai and Richardson (2004)). And, more generally, statistical agencies might not wish to consider behavioural economic relationships as measures.

- **Use of models**. If the statistical agency wishes to preserve a transparent mapping from individual statistical returns to aggregate estimates then it may be constrained in its use of economic models to manipulate those estimates. In other words, the statistical agency may feel constrained in attaching any weight to prior views of how data will evolve. In practice, once statistical returns are available, the ONS do not make extensive use of 'top level' adjustments based on econometric models (see Clements and Hendry (2003)).

The discussion suggests two routes through which consideration of the statistical agency's actions in the face of data uncertainty might be used to motivate simplifying assumptions. In contrast with the discussion of sources of data uncertainty, the implications of the two priors are not independent. There are four scenarios:

a). *The statistical agency uses all available information (so that there are no additional indicators and* $\mathbf{y}_t^i$ *is empty) and uses sophisticated models to assign some weight to its expectation of how* $\mathbf{y}_t$ *would evolve (so that* $\tilde{\mathbf{y}}_t^{s|t+n} \neq \mathbf{y}_t^{s|t+n}$*).* If the statistical agency is able to use the same modelling technology and indicators as the data-user then the data-user cannot "add value" through her own filtering of the published data or through separate consideration of any alternative indicators - as illustrated in Annex B. If the statistical agency's model is seen to differ from the full state space model - for example, if the agency's modelling is better approximated by application of some qualitative guidelines or rules of thumb - then the data-user can add value. Ideally, she would take account of any modelling already applied by the statistical agency. In practice, however, the statistical agency's model is not known and we adopt the full model in Section 3 leaving $\mathbf{y}_t^i$ empty.

b). *The statistical agency is constrained in its use of alternative indicators (*$\mathbf{y}_t^i$ *is not empty), but uses sophisticated models to assign some weight to its expectation of how* $\mathbf{y}_t$ *would evolve (so that* $\tilde{\mathbf{y}}_t^{s|t+n} \neq \mathbf{y}_t^{s|t+n}$*).* The data-user cannot take the published estimates at face value without discarding a part of her information set. And, because the alternative indicators will affect her expectation of the dynamics of $\mathbf{y}_t$ (i.e. her estimate of the parameters $\mathbf{A}_1, \ldots, \mathbf{A}_q$), she must solve the full state space problem. One corollary of this prior is that we should expect measurement errors to the statistical agency's published estimate to be correlated with economic shocks so that it is *not* appropriate to set $\boldsymbol{\rho}_{\tilde{\varepsilon}\epsilon}^s = 0$.

c). *The statistical agency uses all available information (so that there are no additional indicators and* $\mathbf{y}_t^i$ *is empty) but is constrained in its use of models (so that* $\tilde{\mathbf{y}}_t^{s|t+n} = \mathbf{y}_t^{s|t+n}$*).* The data-user cannot take published data at face value without discarding her modelling technology and hence should solve the signal extraction problem with $\mathbf{y}_t^i$ left empty[3]. Because the statistical agency is not seen to be pre-filtering, measurement shocks will not correlate with economic shocks and $\boldsymbol{\rho}_{\tilde{\varepsilon}\epsilon}^s = 0$.

[3]It is here that our simplifying assumption that measurement errors attaching to alternative indicators are uncorrelated with those attaching to the statistical agency's estimates becomes restrictive. Were this assumption to be relaxed, the policymaker would benefit from consideration of alternative indicators (which she can observe) as part of the signal extraction solution.

d). *The statistical agency is constrained in its use of alternative indicators ($\mathbf{y}_t^i$ is not empty) and is constrained in its use of models (so that $\tilde{\mathbf{y}}_t^{s|t+n} = \mathbf{y}_t^{s|t+n}$).* This is the general representation as articulated in Section 3. Because the statistical agency is not seen to be pre-filtering, measurement shocks will not correlate with economic shocks and $\boldsymbol{\rho}_{\tilde{\varepsilon}\epsilon}^s = 0$.

6 An illustrative example

As an illustrative example, we apply the model to quarterly growth of distribution output - estimating the state space model for a single variable of interest. The real-time data-set used is an extension of the Bank of England's real-time database for GDP(E) described in Castle and Ellis (2002). It includes 52 vintages of distribution output, with reference dates running from 1989 Q1 to 2006 Q1. We consider the CBI's distributive trade survey as an indicator - specifically, the proportion of all respondents reporting good sales for the time of year. 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.

6.1 Characterising the revisions history

Table 4 sets out some summary statistics describing the experience of revisions to published data of differing maturities - evaluating revisions over a 16 quarter window as discussed in Section 4.1.

The summary statistics suggest that revisions have been upwards more often than downwards and that, on average, upward revisions have had a larger magnitude than have downward revisions. As a result the mean revision is upward. There is no firm evidence that the mean revision declines as maturity increases. The null that mean revisions are zero can be rejected at the 90% level for 13 out of 16 maturities.

Table 4: Quarterly Growth of Distibution Output - Revisions Model Parameters

	Maturity				
	1	**4**	**8**	**12**	**16**
Mean	0.06	0.18	0.22	0.24	0.15
P-value[1]	(0.30)	(0.09)	(0.02)	(0.02)	(0.02)
Proportion of revisions >0	58%	53%	61%	72%	64%
Mean upward revision	0.51	0.80	0.63	0.51	0.39
Mean downward revision	-0.57	-0.51	-0.43	-0.47	-0.27
Mean absolute revision	0.54	0.66	0.55	0.50	0.35
Standard deviation of revisions	0.70	0.81	0.62	0.67	0.42
Variance	0.49	0.65	0.39	0.45	0.18
P-value[2]	N/A	(0.80)	(0.24)	(0.40)	(0.00)
Median	0.10	0.18	0.28	0.13	0.13
Skew	-0.24	0.06	-0.19	0.62	-0.14
Kurtosis	4.04	2.69	2.05	4.53	2.67
Memo – characteristics of the latest estimate of distribution output					
Mean growth	0.85				
Standard deviation of growth	0.70				

[1] Probability that mean revision is zero at each maturity
[2] Probability that revisions variance at each maturity is smaller than revisions variance at first release

The mean absolute revision is 0.54pp for estimates with a maturity of 1 quarter. That compares with average growth of 0.85pp in distribution output. 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 14 quarters - the null that the variance of revisions is equal to that at maturity 1 is rejected at the 90% level for all maturities beyond 14 quarters.

6.2 Calibrating heteroscedasticity, serial correlation and bias

As outlined in Section 3, the model is estimated in two stages: first estimating the properties of the measurement errors - that is equations (4) though (6) - across real-time data and second applying those properties in estimation of the state space model via the Kalman Filter. And, because we do encounter numerical difficulties in estimation of $\rho^s_{\tilde{\varepsilon}\epsilon}$ within the Kalman Filter, we also calibrate that correlation using the real-time data. Table 5 reports the parameters driving bias, heteroscedasticity and serial correlation. Because the selection of J and N - the maturities over which to calibrate and the window over which to calculate revisions - is arbitrary, we report results for J=N=12 and J=N=20 alongside the estimates used in the remainder of this Section.

The model of bias is very simple and maps easily from the summary statistics quoted in Table 4. Because the mean revision is similar across most maturities, the bias decay parameter (λ) is very close to zero. The variance decay parameter (δ) is more negative giving the variance of measurement errors a half life of 14 quarters. Note that calibration of this parameter is sensitive to the choice of N - the range of maturities over which to calibrate the model parameters. This is not surprising given that Table 4 shows the variance of revisions not to decline much before maturity 14. The summary statistics do not give an indication of the serial correlation in measurement errors. As discussed in Section 4.2, the models for serial correlation and heteroscedasticity are estimated jointly. There is some negative serial correlation across revisions, with parameter values not particularly sensitive to the choice of J and N.

Table 5: Quarterly Growth of Distibution Output - Revisions Model Parameters

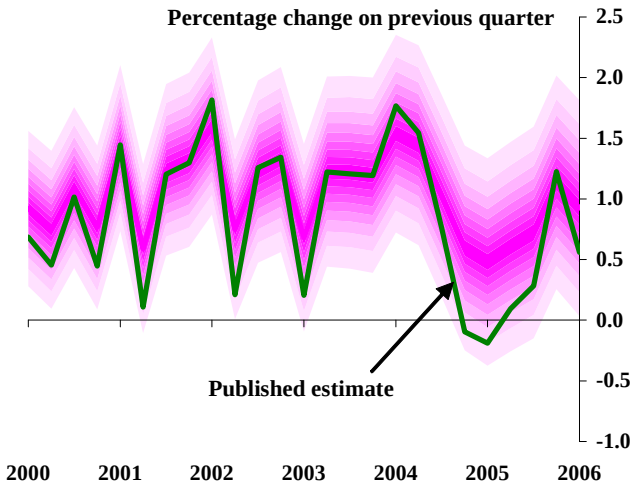
	J=N=12	**J=N=16**	J=N=20
$c_{\tilde{v}}^{s\|1}$	-0.1428	**-0.1782**	-0.1940
λ	0.0000	**0.0000**	0.0000
$\sigma^2_{v\varepsilon^{s\|1}}$	0.4339	**0.6015**	0.6663
δ	-0.0158	**-0.0474**	-0.0841
β_1	-0.2240	**-0.2182**	-0.2410
β_2	-0.0680	**-0.1034**	-0.0776
β_3	0.0478	**-0.0674**	-0.0713
β_4	0.0686	**0.1070**	0.2532

6.3 Estimating the state space model

Once equations (4) though (7) have been calibrated, the remaining model parameters are estimated via maximum likelihood using the Kalman Filter. Examination of the various residuals of the Kalman Filter gives some indication of the degree to which modelling assumptions are violated in the data-set. Both the prediction errors for the published ONS data and the smoothed estimates of the errors on the transition equations pass standard tests for stationarity, homoscedasticity and absence of serial correlation at the 5% level. There is some evidence of non-normality in the smoothed residuals on our transition equation - largely driven by outliers at the beginning of the esimation window. The errors surrounding predictions for the indicator variable are less well-behaved. In particular, there is evidence of significant serial correlation in these residuals.

Figure 1 reports the estimates of quarterly growth of distribution output. Following the convention of the GDP and inflation fancharts 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.

Figure 1: Quarterly Growth of Distibution Output: Full Model



The statistical agency's published estimate is below the centre-point of the fanchart across much of the sample - unsurprisingly given the estimate of bias in published estimates. The centre-point of the fanchart tracks the published estimates quite closely once those estimates are mature. This is a corollary of the heteroscedasticity in measurement error variance. 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.

6.4 Prior views about the nature of data uncertainty

In Section 5, we introduce the possibility that the data-user will have some prior view of the nature of data uncertainties. Suppose that the data-user is confident that any bias and serial correlation apparent in the real-time dataset is not representative. This prior view of serial correlation might follow from her view that the statistical agency does not make material use of indirect measures in forming its estimate of the variable of interest. For

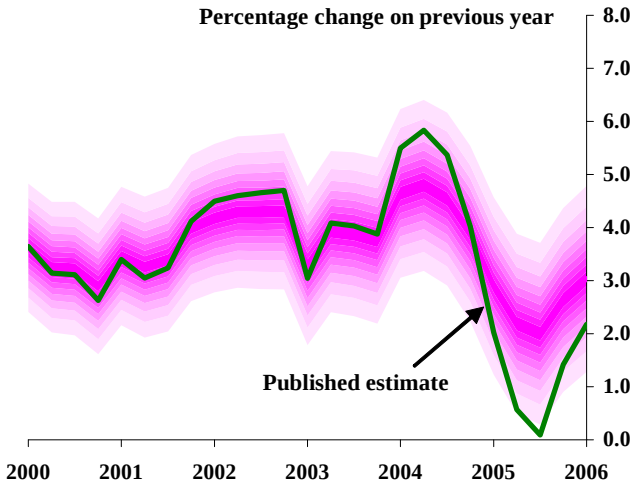
illustrative purposes, we examine the impact of this prior on the backcast for distribution output. We do not present any evidence for or against this prior.

- Because the equations describing serial correlation and heteroscedasticity are estimated jointly, setting $\mathbf{B}_1, \ldots, \mathbf{B}_p = 0$ affects the estimates of $\sigma^2_{\tilde{\varepsilon}^{s|1}}$ and δ. The half-life of measurement uncertainties surrounding the published estimates appears slightly shorter at 13 quarters compared with 14 when serial correlation is included.

- The Kalman Filter prediction errors for the published official data still pass standard diagnostic tests for serial correlation at the 5% level. In other words, serial correlation does not appear to be a particularly significant feature of the dataset.

- Imposing this prior may affect both the point-estimate and the estimated standard errors surrounding it. Figure 2 shows the estimates of year-on-year growth of distribution output consistent with the quarterly model setting $\mathbf{B}_1, \ldots, \mathbf{B}_p = 0$. The centre-point of the distribution is affected by the assumption that the published estimates are not biased. And the standard errors surrounding the year-on-year growth rates are some 14% wider than was the case in the model with negative serial correlation.

Figure 2: Year on Year Growth of Distibution Output: No bias, no serial correlation



7 Conclusions

We have represented the policymaker's - and, indeed any data-user's - data uncertainty problem as a signal extraction problem in which she seeks to establish the appropriate weight to attach to the latest published estimates, alternative indicators and her prior expectation of the how the data would evolve. The model developed is relatively general and permits us to consider both a relatively rich representation of the potential measurement errors in the statistical agency's published estimates and to consider alternative indicators alongside those published estimates. Expressing the model in the general form used in Section 3 provides a base on which to consider the implications of differing prior views about the nature of the uncertainties facing the statistical agency and its actions in dealing with them.

The model and its solution are founded on a number of 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. One obvious extension would be to recast the state space problem to ensure that accounting identities are satisfied - either following Doran (1992) in adding the accounting identities to the vector of measurements taken on each variable or following Weale (1985) in allocating any accounting identity 'residual' arising from estimation of the Kalman system across elements, to minimise some loss function.

A Kalman Filter Algorithm

The model developed in Section 3 is summarised in state space form as equations (9) and (10). Linear state space models of this form can be cast in the general representation given below, following the notation in Harvey (1989).

$$\mathbf{y}_t = \mathbf{d}_t + \mathbf{Z}_t\mathbf{b}_t + \mathbf{u}_t, \qquad \mathbf{u}_t \sim i.i.d.N(0, \mathbf{\Sigma}_{t,u}),\ t = 1, \ldots, T \tag{A.1}$$

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

and $E(\boldsymbol{\eta}_t\mathbf{u}_t') = \mathbf{G}_t$. Below, we abstract from issues arising from the estimation of the parameters of the model which enter the matrices $\mathbf{c}_t, \mathbf{Z}_t, \mathbf{\Sigma}_{t,u}, \mathbf{\Sigma}_{t,\eta}, \mathbf{d}_t, \mathbf{T}_t, \mathbf{G}_t$ and $\mathbf{R}_t$ and concentrate on the estimation of the state vector $\mathbf{b}_t$ conditional on the parameters being known. Let us denote the estimator 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 as 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:

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

known as the prediction equations, and

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

known as the updating equations, where

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

The set of smoother estimates and their respective covariance matrices are denoted by $\hat{\mathbf{b}}_{t|T}$ and $\mathbf{P}_{t|T}$ and are given by

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

and

$$\mathbf{P}_{t|T} = \hat{\mathbf{P}}_t + \mathbf{P}_t^*(\mathbf{P}_{t+1|T} - \hat{\mathbf{P}}_{t+1|t})\mathbf{P}_t^{*\prime} \tag{A.7}$$

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 observation equation (A.1), is denoted by $\mathcal{L}(\boldsymbol{\vartheta})$ where $\boldsymbol{\vartheta}$ denotes the vector of parameters with respect to which the log likelihood is maximised, can be written in terms of the prediction errors $\boldsymbol{\varpi}_t = \mathbf{y}_t - \mathbf{Z}_t'\hat{\mathbf{b}}_{t|t-1}$ as

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

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 $\mathbf{c}_t, \mathbf{Z}_t, \boldsymbol{\Sigma}_{t,u}, \boldsymbol{\Sigma}_{t,\eta}, \mathbf{d}_t, \mathbf{T}_t$, $\mathbf{G}_t, \mathbf{R}_t, \mathbf{b}_0$ and $\mathbf{P}_0$.

This representation and solution method 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 3. There, the parameter vector $\boldsymbol{\vartheta}$ comprises $= (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', .., \boldsymbol{\alpha}_q', \boldsymbol{\sigma}_{\tilde{\boldsymbol{\varepsilon}}}^{\mathbf{2s|1'}}, \boldsymbol{\delta}'$, $\boldsymbol{\beta}_{\mathbf{1}}', \ldots, \boldsymbol{\beta}_p', \mathbf{c}^{\mathbf{s|1'}}, \boldsymbol{\lambda}', \boldsymbol{\rho}_{\tilde{\varepsilon}\epsilon}^{s\prime}, \boldsymbol{\sigma}_i^{2\prime}, \boldsymbol{\mu}', \boldsymbol{\sigma}_\epsilon^{2\prime}, \mathbf{c}^i, \mathbf{Z}^i)$.

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

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

Then we have the following setup.

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

$$\mathbf{Z}_t = \begin{pmatrix} \mathbf{I} & \ldots & \mathbf{0} & \mathbf{I} & \ldots & \mathbf{0} \\ \mathbf{Z}^i & \ldots & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \end{pmatrix}$$

$$\boldsymbol{\Sigma}_{t,u} = \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{\Sigma}_i \end{pmatrix}$$

$$\boldsymbol{\Sigma}_{t,\eta} = \begin{pmatrix} \boldsymbol{\Sigma}_\epsilon & \mathbf{0} & \dots & \boldsymbol{\zeta}_{\tilde{\varepsilon}\epsilon}^{T-t} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \dots & \dots & \dots & \dots \\ \boldsymbol{\zeta}_{\tilde{\varepsilon}\epsilon}^{s} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\varepsilon}}}^{T-t} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \dots. & \dots & \dots & \mathbf{0} \end{pmatrix}$$

$$\mathbf{d}_t = \begin{pmatrix} \tilde{\mathbf{c}}^{s|n}(\mathbf{1} + \boldsymbol{\lambda})^{T-t-1}] \\ \mathbf{c}^i \end{pmatrix}$$

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

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

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

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

and

$$\mathbf{P}_0 = \left(\mathbf{I} - \mathbf{T}_0(\boldsymbol{\vartheta})\right)^{-1} \boldsymbol{\Sigma}_{0,\eta}(\boldsymbol{\vartheta})$$

This is the most general setup possible for the estimation of the state space model of Section 3. However, as described in the main text, in estimation we set some parameters to constants having obtained suitable values for them via prior estimation (as we discuss in Section 4). 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 $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1', \boldsymbol{\vartheta}_2')'$ is some suitable partition of $\boldsymbol{\vartheta}$. With the heteroscedasticity, serial correlation and bias parameters estimated by GMM, the partition is $\boldsymbol{\vartheta}_1 = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, .., \boldsymbol{\alpha}_q, \boldsymbol{\sigma}_i, \boldsymbol{\mu}, \boldsymbol{\sigma}_\epsilon, \boldsymbol{\rho}_{\tilde{\varepsilon}\epsilon}^s, \mathbf{c}^i, \mathbf{Z}^i)$ and $\boldsymbol{\vartheta}_2 = (\boldsymbol{\sigma}^2_{\tilde{\varepsilon}^{s|1}}, \boldsymbol{\delta}, \boldsymbol{\beta}_\mathbf{1}, \dots, \boldsymbol{\beta}_\mathbf{p}, \mathbf{c}_{\tilde{v}}^{s|1}, \boldsymbol{\lambda})$.

Finally, note that when $\mathbf{B}_1 = \mathbf{B}_2 = \dots = 0$ the measurement error has no serial correlation. Then, the above state space is equivalent to one where the measurement error enters the measurement rather than the transition equation.

B A stylised model comparing the implications of statistical agency as compiler and modeller

In Section 5, we assert that if the statistical agency applies the same modelling strategy as the data-user to the same dataset then there is no added value to be had. This annex expands on the intuition, using a much simplified representation of the model articulated in Section 3. Let there be a single variable of interest y_t and assume that the model for the true data is given by a simple 1st order autoregressive process

$$y_t = ay_{t-1} + \epsilon_t \qquad \text{where } \epsilon_t \sim N(0, \sigma_\epsilon^2). \tag{B.1}$$

Assume that past data values are observed with certainty so that at time t, both the statistical agency and the data-user know the value of y_{t-1}. The signal extraction problem is to form a view of y_t on the basis of y_{t-1} and any other information available.

Let y_t^s denote a proprietary and noisy estimate of y_t obtained by the statistical agency. Because we assume that past data values are known with certainty, there is no need to model heteroscedasticity in measurement errors and the $t+n$ superscript used in the general model becomes redundant. Similarly, there is no need to model any serial correlation in measurement errors associated with this proprietary information. We further assume that the measure is unbiased. Then, the model for the statistical agency's noisy measure is given by:

$$y_t^s = y_t + v_t \qquad \text{where } v_t \sim N(0, \sigma_v^2) \text{ and } E(\epsilon_t v_t) = 0. \tag{B.2}$$

The statistical agency publishes an estimate of y_t based on y_t^s, denoted by $\tilde{y}_t^s$. In doing so, the agency may either act as a data compiler (taking y_t^s at face value) or as a data modeller (assigning some weight to its prior view of y_t in line with the properties of the transition law). The data-user is tasked with forming her own view of y_t based on $\tilde{y}_t^s$, denoted by $\hat{y}_t^{pol}$ No alternative indicators are available.

This annex shows the implications of the two approaches open to the statistical agency for the data-user's estimate of true output.

Statistical agency as a 'data-compiler'

If the statistical agency acts as a 'data-compiler' it is assumed to simply publish its proprietary information without any adjustment

$$\tilde{y}_t^s = y_t^s. \tag{B.3}$$

Knowing that the statistical agency is acting in this way, the data-user looks to form an estimate of y_t on the basis of her *complete* information set. Notably, that information set includes the time-series forecast derived from the structural model in equation (B.1). So, with a linear expectations function, the data-user's estimate of y_t will be given by

$$\hat{y}_t^{pol} = \gamma a y_{t-1} + \beta \tilde{y}_t^s. \tag{B.4}$$

From first principles, the data-user should choose γ and β to minimise some kind of loss function - here assumed to be quadratic in the expected error. Given the assumptions about measurement errors and economic shocks, this loss function will uncover the maximum likelihood estimate of y_t given the available information.

$$\begin{aligned} L &= E(y_t - \gamma a y_{t-1} - \beta \tilde{y}_t^s)^2 \qquad \text{(B.5)} \\ &= E(a y_{t-1} + \epsilon_t - \gamma a y_{t-1} - \beta(a y_{t-1} + \epsilon_t + v_t))^2 \\ &= E((1-\gamma-\beta) a y_{t-1} + (1-\beta)\epsilon_t - \beta v_t)^2 \\ &= (1-\gamma-\beta)^2 a^2 y_{t-1}^2 + (1-\beta)^2 \sigma_\epsilon^2 + \beta^2 \sigma_v^2. \end{aligned}$$

Minimising that loss function with respect to γ and β allows us to uncover the following first-order conditions

$$\frac{\partial L}{\partial \gamma} = -2(1-\gamma-\beta) a^2 y_{t-1}^2 = 0, \tag{B.6}$$

$$\frac{\partial L}{\partial \beta} = -2(1-\gamma-\beta) a^2 y_{t-1}^2 - 2(1-\beta)\sigma_\epsilon^2 + 2\beta\sigma_v^2 = 0. \tag{B.7}$$

γ^* and β^* (the optimal weights to attach to the prior and the statistical agency's estimate) are then a function of the relative error variances, as given below

$$\gamma^* = \frac{\sigma_v^2}{\sigma_\epsilon^2 + \sigma_v^2} = \frac{1/\sigma_\epsilon^2}{1/\sigma_\epsilon^2 + 1/\sigma_v^2}, \tag{B.8}$$

$$\beta^* = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_v^2} = \frac{1/\sigma_v^2}{1/\sigma_\epsilon^2 + 1/\sigma_v^2}. \tag{B.9}$$

This is the relatively familiar result that indicators should be weighted according to their inverse standard errors: the smaller the standard error associated with a particular piece of information, the larger the weight that should be attached to it.

Statistical agency as a 'data-modeller'

If the statistical agency is not constrained in its use of models, it could make use of the prior view embodied in the transition law itself. In that case, the agency would already be taking account of the information in the structural model and the published data will use the optimal weights from the previous section and be given by

$$\tilde{y}_t^s = \frac{\sigma_v^2 a y_{t-1}}{\sigma_\epsilon^2 + \sigma_v^2} + \frac{\sigma_\epsilon^2 y_t^s}{\sigma_\epsilon^2 + \sigma_v^2}. \tag{B.10}$$

Exactly as in the previous section, the data-user is then tasked with choosing γ and β in her expectations function

$$\hat{y}_t^{pol} = \gamma a y_{t-1} + \beta \tilde{y}_t^s. \tag{B.11}$$

Using the same quadratic loss function as before gives

$$\begin{aligned} L &= E(y_t - \gamma a y_{t-1} - \beta \tilde{y}_t^s)^2 \qquad \text{(B.12)} \\ &= E\left[a y_{t-1} + \epsilon_t - \gamma a y_{t-1} - \beta\left(\frac{\sigma_v^2 a y_{t-1}}{\sigma_\epsilon^2 + \sigma_v^2} + \frac{\sigma_\epsilon^2 (a y_{t-1} + \epsilon_t + v_t)}{\sigma_\epsilon^2 + \sigma_v^2}\right)\right]^2 \\ &= E\left[(1-\gamma-\beta) a y_{t-1} + \left(1 - \frac{\sigma_\epsilon^2 \beta}{\sigma_\epsilon^2 + \sigma_v^2}\right)\epsilon_t - \left(\frac{\beta \sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_v^2}\right) v_t\right]^2 \\ &= (1-\gamma-\beta)^2 a^2 y_{t-1}^2 + \left(1 - \frac{\sigma_\epsilon^2 \beta}{\sigma_\epsilon^2 + \sigma_v^2}\right)^2 \sigma_\epsilon^2 + \left(\frac{\beta \sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_v^2}\right)^2 \sigma_v^2. \end{aligned}$$

Again, minimising that loss function with respect to γ and β allows us to uncover the first-order conditions

$$\frac{\partial L}{\partial \gamma} = -2(1-\gamma-\beta) a^2 y_{t-1}^2, \tag{B.13}$$

$$\frac{\partial L}{\partial \beta} = -2(1-\gamma-\beta) a^2 y_{t-1}^2 - \frac{2\sigma_\epsilon^4 \left(1 - \frac{\sigma_\epsilon^2 \beta}{\sigma_\epsilon^2 + \sigma_v^2}\right)}{\sigma_\epsilon^2 + \sigma_v^2} + \frac{2\beta \sigma_\epsilon^4 \sigma_v^2}{(\sigma_\epsilon^2 + \sigma_v^2)^2}. \tag{B.14}$$

γ^* and β^* (the optimal weights) are now very different

$$\gamma^* = 0, \tag{B.15}$$

$$\beta^* = 1. \tag{B.16}$$

Given that the statistical agency are already taking account of the structural forecast, it would be a mistake for the data-user to 'double-count' that information. Putting any additional weight on the time-series forecast would, in this case, lead to over-filtering.

References

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

Ashley, J., R. Driver, S. Hayes, and C. Jeffery (2005): "Dealing with Data Uncertainty," *Bank of England Quarterly Bulletin*, 45(1), 23–30.

Castle, J., and C. Ellis (2002): "Building A Real-Time Database for GDP(E)," *Bank of England Quarterly Bulletin*, 42(1), 42–48.

Clements, M., and D. Hendry (2003): "Report of a Scoping Study of Forecasting in the National Accounts at the Office for National Statistics," *Statistics Commission Report 12: Annex A*.

Cook, L. (2004): "Revisions to Statistics: Their Role in Measuring Economic Progress," *Economic Trends*, 603, 36–43.

Doran, H. E. (1992): "Constraining Kalman Filter and Smoothing Estimates to Satisfy Time-Varying Constraints," *The Review of Economics and Statistics*, 74, 568–572.

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

Evans, M. (2005): "Where are We Now? Real-Time Estimates of the Macro Economy," *International Journal of Central Banking*, 1, 127–75.

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

Garratt, A., and S. P. Vahey (2004): "UK Real-Time Macro Data Characteristics," *Birkbeck Working Paper*, No. 513.

Harrison, R., G. Kapetanios, and T. Yates (2004): "Forecasting with Measurement Errors in Dynamic Models," *Bank of England Working Paper*, 237.

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

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

Howrey, E. P. (1984): "Data Revision, Reconstruction, and Prediction: An Application to Inventory Investment," *The Review of Economics and Statistics*, 66(3), 386–393.

Kapetanios, G., and T. Yates (2004): "Estimating Time-Variation in Measurement Error from Data Revisions; an Application to Forecasting in Dynamic Models," *Bank of England Working Paper*, 238.

Kozicki, S. (2004): "How Do Data Revisions Affect the Evaluation and Conduct of Monetary Policy?," *Federal Reserve Bank of Kansas City Economic Review*, First Quarter, 5–38.

Lomax, R. (2004): "Stability and Statistics," Speech at the North Wales Business Club, 23 November 2004.

Mai, N., and C. Richardson (2004): "Using Revisions Information to Improve the National Accounts: A Discussion Paper," *Paper presented at OECD-ONS workshop: "Assessing and Improving Statistical Quality - Revisions Analysis for the National Accounts".*

Mankiw, N. G., and M. Shapiro (1986): "News or Noise: An Analysis of GDP Revisions," *Survey of Current Business*, pp. 20–25, No. 66.

Mitchell, J. (2004): "Review of Revisions to Economic Statitistics: A Report to the Statistics Commission," *Statistics Commission Report No 17*, 2.

Nelson, E., and K. Nikolov (2003): "UK Inflation in the 1970s and 1980s: The Role of Output Gap Mismeasurement," *Journal of Economics and Business*, 55, 353–370.

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

Orphanides, A. (2003): "The Quest for Prosperity Without Inflation," *Journal of Monetary Economics*, 50, 633–663.

Patterson, K. D. (1994): "A State Space Model for Reducing the Uncertainty Associated with Preliminary Vintages of Data with an Application to Aggregate Consumption," *Economics Letters*, 46, 215–22.

Sargent, T. J. (1989): "Two Models of Measurements and the Investment Accelerator," *Journal of Political Economy*, 97, 251–87.

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*, 67, 685–89.