Department of Economics

Estimating Time-Variation in Measurement Error from Data Revisions: An Application to Forecasting in Dynamic Models

George Kapetanios and Tony Yates

Working Paper No. 520 October 2004 ISSN 1473-0278

Queen Mary
University of London

Estimating Time-Variation in Measurement Error from Data Revisions: An Application to Forecasting in Dynamic Models

George Kapetanios
Queen Mary, University of London and Bank of England
and
Tony Yates[*]
Bank of England

September 2004

Abstract

Over time, economic statistics are refined. This means that newer data is typically less well measured than old data. Time variation in measurement error like this influences how forecasts should be made. We show how modelling the behaviour of the statistics agency generates both an estimate of this time variation and an estimate of the absolute amount of uncertainty in the data. We apply the method to UK aggregate expenditure data, and illustrate the gains in forecasting from exploiting our model estimates of measurement error.

Keywords: Forecasting, Data Revisions.
JEL codes: C32, C53

1 Introduction

Over time, official statistics agencies revise and improve estimates of data as they collect more information. It is likely that they measure some data better than others, and that they hone their estimates of data more quickly for some data than for others. If we could quantify the measurement error in different variables and vintages, this information would be useful for estimation, forecasting and for policy analysis. The more noise in data, the less weight we should put on them in forming an estimate of the state of the economy, or the right setting for economic policy. These points are well known. They have been emphasised recently in papers by Aoki (2003), Swanson (2000), Svensson and Woodford (2003) - who look at optimal monetary policy and measurement error - and, Harrison, Kapetanios, and

[*]This work grew out of previous joint work with Richard Harrison. We are also grateful for comments to Ian Bond, Alex Brazier, Todd Clark, Spencer Dale, Melissa Davey, Simon Hayes, Graham Elliott, Athanasios Orphanides and Jo Paisley.

Yates (2004) - who studied how optimal forecasts are affected by measurement errors that vary across vintages. However, since we never observe the true value for any data, we can therefore never observe the measurement error in any data vintage. Some data are based at least in part on surveys. But information about the sampling error in these surveys is typically not available. More importantly, many data are constructed from many different sources other than surveys. There may not exist readily quantifiable concepts of the measurement error contained in these other information sources: judgement is sometimes an important element in reconciling information from conflicting sources, and there is no simple way of quantifying the measurement error associated with that kind of information.

What we do observe are successive vintages of data. One option is to approximate the measurement error in a variable by assuming that the final release of a variable is equal to the true value. Then, the variance of early releases about the final release is an estimate of the variance of those releases about the truth. This approach has been taken by Harrison, Kapetanios, and Yates (2004), Coenen, Levin, and Weiland (2001) and many others. Rather obviously, the closer the final release is to the true value, the better this is as an approximating method. Ominously, since we don't know how good an estimate the final release is of the true value for some data, we can't assess how good an approximation this method is to the ideal. Another option is to use a state space model. That method uses an assumption about the economic process driving movements in the true variable to get an estimate of the truth. That estimate can then be used to go back and compute how, on average, early releases of data vary about a Kalman Filter estimate of the truth, based on final or near final data.

Our paper provides an alternative, exploiting behavioural assumptions about the statistics agency generating the real time data. We show how a few assumptions can generate an estimate of the measurement error in different vintages of data from observations on the variance of revisions to data. These assumptions are the following. First, a conjecture about the scheme that the statistics agency uses to weight new information about an observation together with old information. We start out assuming that the statistics agency weights the surveys up optimally as they arrive. We also experiment with another scheme we describe as 'naive'. Our second assumption is that the arrival of incremental information about a data point can be modelled as a sequence of independent random draws on a population. Crucially, our hypothetical agency does not engage in filtering and forecasting itself in the way that Sargent (1989) described was possible, and Mankiw, Runkle, and Shapiro (1984) investigated for releases of US GNP. This assumption is consistent with our assumption about an agency that does optimal weighting. Rational data collection can reasonably be taken to include making use of all future information that is not independent from current information in the current release. Therefore, that incremental information is independent. A third assumption relates to the evolution of the quantity of incremental information that arrives about an observation over time. The basic model assumes that the rate at which the flow of new information in surveys changes is fixed. However, we relax this and derive a model that allows the data to reveal how the flow of new information evolves.

These assumptions together build a model of a hypothetical statistics agency. The model is not a literal description of the real-world data collection process. Many data are constructed and subsequently revised using information that has many sources. Some of it is literally surveys. Other information is based on censuses. Still other information is based on economic models that are used to corroborate one information source with another, or judgement, or even filtering and forecasting. Our model will be useful if the flow of information from these many sources can, to a good enough approximation, be described *as if* it were a sequence of independent random surveys.

This model can be used to compare the predicted variance of revisions to data with the observed variance of revisions. Imagine a statistics agency that conducted ever smaller independent random samples from a population, and weighted them optimally. The revisions to the data that this agency would make would get smaller over time. The reliability of the new information would shrink relative to the ever larger weighted combinations of older information, and so would get less and less weight each time. By being specific about the weighting scheme the statistics agency uses, and about how the sample sizes of these surveys evolve, we can use the variance of the revisions in this way to uncover estimates of measurement error in the published data. As we will show, it turns out that the variance of revisions can be written as a function of the variance of the measurement error. In our simplest model there are only two unknowns: the initial period measurement error, and the rate of decay or growth of the arrival rate of information. We simply choose the combination of these two unknowns that best fits the predicted to the observed variance of revisions, for all possible revisions. With that done we can construct estimates of the measurement error surrounding any variable, and for any vintage of that variable.

That information can then be used in some optimal forecasting or optimal policy procedure. Armed with our proposed technique, we apply it to some data to illustrate that it can work. This involves first estimating the time-variation in measurement error, and second applying this information to an optimal forecasting problem. We use the UK real-time data set of Castle and Ellis (2002). We compare results for the real growth of consumption and imports expenditure, and find that the first release for imports is about six times worse measured than that for consumption.

We illustrate how the outputs from our procedure can be used, showing how a univariate forecasting model for UK consumption growth can be improved by using the estimated time-variation in the measurement error of UK consumption growth vintages. We estimate a univariate forecasting model for consumption growth, and improve on it, given estimated information about time-varation in data uncertainty coming from our statistics agency model.

Of course, there are many series which are never revised. In the United Kingdom, the Retail Price Index is one example. The mere fact that series don't get revised is not an indication that those series are measured perfectly. Our method is powerless to uncover anything about measurement error in these cases.

The paper is structured as follows: Section 2 provides a literature review. Section 3 presents the simple model of the statistics agency. Section 4 applies the model to a real time dataset. Section 5 extends the model to provide a more realistic setup. Section 6 carries out a real time forecasting exercise using the theoretical model. Finally, section 7 concludes.

2 Related literature

There is a growing literature studying the use of real-time data and the effect of data uncertainty on monetary policy and forecasting. Here we attempt briefly to locate our work amongst it.

A research effort has sprung up aimed at compiling consistent sets of real-time data: Castle and Ellis (2002) and Eggington, Pick, and Vahey (2002) present data for the United Kingdom. Croushore and Stark (2001) compile a real time dataset for the United States. Bernhardsen, Eitrheim, Jore, and Roisland (2004) and Gayen and VanNorden (2004) do the same for Norway and Canada respectively. Aoki (2003), Swanson (2000), Svensson and Woodford (2003), establish the theory of the consequences of data uncertainty for optimal monetary policy. Orphanides and VanNorden (2002) study the relative contribution of revisions to output data and the changability of typical detrending methods' estimates of the output gap at the end of the sample. Other papers that study the consequences of data uncertainty in real time for output gap mismeasurement in this vein are: Bernhardsen, Eitrheim, Jore, and Roisland (2004), (Norway), Gayen and VanNorden (2004), (Canada) and Kamada (2004) (Japan). Still others have used real-time data to assess the conduct of monetary policy in the past. The most celebrated example here is Orphanides (2001), a study of Fed policy. More recent examples of this genre studying the United Kingdom and Germany are Nelson and Nikolov (2003) and Gerberding, Worms, and Seitz (2004) respectively.

Our paper sits within a set of papers that study how real-time data can be used in forecasting. This literature goes back at least as far as Howrey (1978) and perhaps further. The papers most closely related to this paper are: Harvey, Mckenzie, Blake, and Desai (1983); Coenen, Levin, and Weiland (2001); Busetti (2001); Harrison, Kapetanios, and Yates (2004) and Jaaskela and Yates (2004). All these papers consider, among other things, how optimal forecasts (or optimal policy) depend on the amount of uncertainty in the data, and how that uncertainty varies across vintages. The methods suggested in those papers should, in principle, take as input what the statistical agency model, described below, gives as an output.

3 A benchmark model of the statistics agency

In this section, we extract estimates of measurement error from real time data using two models: a model of a rational statistics agency, and a model of a 'naive' statistics agency. A rational agency will be one that weights together samples optimally given information about the sample size. A naive agency will be one that uses equal weights regardless.

3.1 A model of a rational statistics agency

It is assumed that the statistics agency faces a sequence of problems over time. In the first period, a survey is conducted and an estimate published. In the second period, new information, equivalent to a second survey of the same population, comes in, and the statistics agency has to weight the first and second surveys together to form a combined estimate. The statistics agency never observes the true value of anything (and neither do we as econometricians). But it starts out by trying to minimise the expected variance of its estimates about the true value.

At time t, the agency simply publishes the results of the first survey, on a variable dated t which we will call y. At time $t+1$ the problem is to minimise the expected variance of a weighted average of the first two surveys about the true value, $y_{t|T}$. At time t+1 the problem is:

$$\text{Min}E\left[\lambda_{1|1}y_{t|t+1}+(1-\lambda_{1|1})y_{t|t}-y_{t|T}\right]^2 \tag{1}$$

where $\lambda_{1|1}$ should be read as follows: the second subscript tells you which survey release is being weighted; the first subscript tells you when it is being weighted. At time $t+2$ the agency has to choose weights to minimise the weighted sum of the first three releases:

$$\text{Min}E\left[\lambda_{2|0}y_{t|t}+\lambda_{2|1}y_{t|t+1}+\lambda_{2|2}y_{t|t+2}-y_{t|T}\right]^2 \tag{2}$$

such that $\sum_{k=0}^{2}\lambda_{2|k}=1$. Generally, at time $t+n$ the problem is:

$$\text{Min}E\left[\sum_{k=0}^{n}\lambda_{n|k}y_{t|t+k}-y_{t|T}\right]^2 \tag{3}$$

such that $\sum_{k=0}^{n}\lambda_{n|k}=1$.

This is a stylised assumption to make about the process that generates data revisions. It is illustrative only. We need only to make an assumption about how the statistics agency weights new information. This could be based on solving optimal signal extraction problems like the one set out here. Alternatively, it could be based on assuming the agency follows a rule of thumb. We use the 'optimising' agency as a benchmark, but the general method does not depend on it. Note that although taken literally our model assumes that all the agency does is conduct surveys, the model's usefulness will depend on whether the more complicated data collection and revision process (involving judgements, corroboration, forecasting, filtering) can be taken as if it behaved like a sequence of independent surveys. The assumption of independence is not overly restrictive as it relates to incremental information which under any concept of rationality is bound to be uncorrelated (independent under normality) from past information.

We proceed as follows: first, we solve for the weights that the statistics agency will place on new information over time. Second, we use this to solve for expressions that link the variance of data revisions with the variance of the underlying measurement error, and the rate at which this measurement error decays over time.

Writing a survey on a data observation as a function of the true value and some measurement error, note that:

$$y_{t|t+k}=y_{t|T}+v_{t|t+k} \tag{4}$$

where $v_{t|t+n}$ is the measurement error contained in a survey based estimate of y_t carried out at some time $t + n$. We can then write:

$$\text{Min}E\left[\sum_{k=0}^{n}\lambda_{n|k}(y_{t|T} + v_{t|t+k}) - y_{t|T}\right]^2 \tag{5}$$

such that $\sum_{k=0}^{n}\lambda_{n|k} = 1$. Since the weights sum to 1, the true y's cancel out and we get

$$\text{Min}E\left[\sum_{k=0}^{n}\lambda_{n|k}v_{t|t+k}\right]^2 \tag{6}$$

Expressing this in terms of the variance of the measurement error gives:

$$\text{Min}\sum_{k=0}^{n}(\lambda_{n|k})^2\sigma^2_{v|k} \tag{7}$$

For the sake of notational simplicity, $\sigma^2_{v|k} = var(v_{t|t+k})$. Assuming that each successive survey can be thought of as an independent draw does not imply that the published data are independent. However, it implies that incremental surveys, from which published estimates are formed, are independent.

The next important assumption we make is that the variance of the measurement error around successive surveys changes at a fixed rate. In terms of the 'survey' metaphor, this means assuming that each period, the incremental surveys get smaller and smaller. To be more concrete, we assume the following:

$$\sigma^2_{v|k} = (1 + i)^k\sigma^2_{v|0} \tag{8}$$

If i were zero, each incremental survey estimate would have the same variance about the truth, and the statistics agency would weight them equally. Nothing is assumed about i at this stage. Data on the variance of revisions will be used to estimate i. Intuitively, i is some positive, constant number. This leads to revisions getting smaller and smaller over time, which is a broad feature of the data. If i is positive, each incremental survey will be a more noisy estimate of the truth than the preceding one. But each published release, which will include more and more surveys, will be a *less* noisy estimate than the preceding one.

Conceptually, it would be straightforward to allow for a degree of decay i that varies over k, or to test whether it does or does not. Doing this would have the advantage that it must be more realistic. But doing so would involve a trade-off, as we shall see. The more i varies over k, the fewer observations there are to estimate it. So although the model would be more realistic, estimates of it would be more imprecise. It is also possible to experiment with more complex functional forms that describe the decay of the measurement error than the one we have here. But for the moment we illustrate the basic approach with a constant i.

In our model, the only discrepancy that arises between surveys, published data and the truth are those that come from the surveys being conducted using samples smaller than the population. There are no additional 'data quality' issues that we model. The real-life data generating process includes the task of detecting and correcting errors in completing or processing surveys, errors of interpretation, and much more. The model abstracts from that. The minimand for the statistics agency can be written as:

$$\text{Min} \sum_{k=0}^{n} (\lambda_{n|k})^2 (1+i) \sigma_{v|0}^2 \tag{9}$$

We write the problem out as:

$$\lambda_{n|0}^2 \sigma_{v|0}^2 + \lambda_{n|1}^2 (1+i) \sigma_{v|0}^2 + \lambda_{n|2}^2 (1+i)^2 \sigma_{v|0}^2 + \lambda_{n|3}^2 (1+i)^3 \sigma_{v|0}^2 + ...+ \tag{10}$$

$$\lambda_{n|n-1}^2 (1+i)^{n-1} \sigma_{v|0}^2 + (1 - \lambda_{n|0} - \lambda_{n|1} - \lambda_{n|2} - ... - \lambda_{n|n-1})(1+i)^n \sigma_{v|0}^2 \tag{11}$$

We drop $\sigma_{v|0}^2$ and so we minimise the following object:

$$\lambda_{n|0}^2 + \lambda_{n|1}^2 (1+i) + \lambda_{n|2}^2 (1+i)^2 + \lambda_{n|3}^2 (1+i)^3 + ...+ \tag{12}$$

$$\lambda_{n|n-1}^2 (1+i)^{n-1} + (1 - \lambda_{n|0} - \lambda_{n|1} - \lambda_{n|2} - ... - \lambda_{n|n-1})^2 (1+i)^n \tag{13}$$

The first order conditions for this problem are as follows:

$$\lambda_{n|j} - (1 - \lambda_{n|0} - \lambda_{n|1} - \lambda_{n|2} - ... - \lambda_{n|n-1})(1+i)^n = 0, \quad j = 0, \ldots, n-1 \tag{14}$$

It is clearly the case from the FOC's that:

$$\lambda_{n|0} = \lambda_{n|1}(i+1) = ... = \lambda_{n|n-1}(i+1)^{n-1} \tag{15}$$

So, we go to (14) for $j = 0$ and substitute to get

$$\lambda_{n|0} - (1 - \lambda_{n|0} - \lambda_{n|0}/(1+i) - \lambda_{n|0}/(1+i)^2 - ... - \lambda_{n|0}/(1+i)^{n-1})(1+i)^n = 0 \tag{16}$$

This implies the following:

$$\lambda_{n|0} = \frac{i(1+i)^n}{(1+i)^{n+1} - 1} \tag{17}$$

Then, the expression for the optimal weight is as follows:

$$\lambda_{n|j} = \frac{i(1+i)^{n-j}}{(1+i)^{n+1} - 1}, j = 0, ...n \tag{18}$$

This expression implies (assuming a positive value of i), that later surveys get smaller weights. This is intuitive: later surveys, under a positive i, are subject to larger measurement error.

Having solved for the weights the statistics agency places on successive surveys, the next step is to write down an expression for the revision between any two periods. The data releases between t and $t+n$ are given by:

$$y_{t|P,t+n} = \lambda_{n|0}y_{t|t} + \lambda_{n|1}y_{t|t+1} + \lambda_{n|2}y_{t|t+2} + ... + \lambda_{n|n}y_{t|t+n} \tag{19}$$

$$= \sum_{k=0}^{n} \lambda_{n|k}y_{t|t+k} \tag{20}$$

The revisions between successive periods will be given by:

$$Ry_{t|n} = y_{t|P,t+n} - y_{t|P,t+n-1} \tag{21}$$

$$= (\lambda_{n|0} - \lambda_{n-1|0})y_{t|t} + (\lambda_{n|1} - \lambda_{n-1|1})y_{t|t+1} + (\lambda_{n|2} - \lambda_{n-1|2})y_{t|t+2} + \tag{22}$$

$$... + (\lambda_{n|n-1} - \lambda_{n-1|n-1})y_{t|t+n-1} + \lambda_{n|n}y_{t|t+n} \tag{23}$$

$$= \sum_{k=0}^{n-1}(\lambda_{n|k} - \lambda_{n-1|k})y_{t|t+k} + \lambda_{n|n}y_{t|t+n} \tag{24}$$

The revision between any two periods, say an $n-l$ period revision, is given by:

$$Ry_{t|n,l} = y_{t|P,t+n} - y_{t|P,t+l} \tag{25}$$

$$= (\lambda_{n|0} - \lambda_{l|0})y_{t|t} + (\lambda_{n|1} - \lambda_{l|1})y_{t|t+1} + (\lambda_{n|2} - \lambda_{l|2})y_{t|t+2} + \tag{26}$$

$$... + (\lambda_{n|l} - \lambda_{l|l})y_{t|t+l} + \lambda_{n|l+1}y_{t|t+l+1} + ... + \lambda_{n|n}y_{t|t+n} \tag{27}$$

$$= \sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})y_{t|t+k} + \sum_{k=l+1}^{n} \lambda_{n|k}y_{t|t+k} \tag{28}$$

We can get an expression for the variance of revisions by substituting in the relation between the observation and the true data:

$$y_{t|t+n} = y_{t|T} + v_{t|t+n} \tag{29}$$

to give

$$\sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})(y_{t|T} + v_{t|t+k}) + \sum_{k=l+1}^{n} \lambda_{n|k}(y_{t|T} + v_{t|t+k}) \tag{30}$$

Since the weights on individual surveys in any vintage sum to one, or, formally noting that:

$$\sum_{k=0}^{l} \lambda_{l|k} = \sum_{k=0}^{l} \lambda_{n|k} + \sum_{k=l+1}^{n} \lambda_{n|k} = 1 \tag{31}$$

we can see that the terms in the true value of y cancel, implying that the revision between any two dates can be written as:

$$Ry_{t|n,l} = \sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})v_{t|t+k} + \sum_{k=l+1}^{n} \lambda_{n|k}v_{t|t+k} \tag{32}$$

Setting all covariance terms to zero, we write the variance of a revision between any two dates as:

$$var(Ry_{t|n,l}) = \sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})^2\sigma^2_{v|k} + \sum_{k=l+1}^{n} (\lambda_{n|k})^2\sigma^2_{v|k} \tag{33}$$

Recalling (8), we get that this revision variance can be written as:

$$var(Ry_{t|n,l}) = \sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})^2(1+i)^k\sigma^2_{v|0} + \sum_{k=l+1}^{n} (\lambda_{n|k})^2(1+i)^k\sigma^2_{v|0} \tag{34}$$

Noting (15), we can write:

$$var(Ry_{t|n,l}) = \sigma^2_v \left[\sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})^2(1+i)^k + \sum_{k=l+1}^{n} (\lambda_{n|0})^2(1+i)^{-k}\right] \tag{35}$$

We can expand these summations through some long-winded but basic algebra to get the following expression:

$$var(Ry_{t|n,l}) \tag{36}$$

$$= \sigma^2_{v|0}\left[\frac{\left((1+i)^{n-l} - 1\right)\left(i(1+i)^l\right)}{\left((1+i)^{n+1} - 1\right)^2}\right]\left[\frac{(1+i)^{n-l} - 1}{(1+i)^{l+1} - 1} + (1+i)^{n-l}\right] \tag{37}$$

We can now form systems of equations in the variance of revisions (which we observe), the decay parameter i and the variance of the measurement error $\sigma^2_{v|0}$. For example, given values for revisions over two different periods say (e.g the variance of six and ten period revisions) we could form two equations in our two unknowns, and solve. In fact, the data may allow us to collect many observations on the variance of revisions at different periods, and this will allow us to construct an estimator for $\sigma^2_{v|0}$ and i. This is simply a standard GMM estimator. Formally, we solve:

$$Min\sum_{n}\sum_{l}(var(Ry_{t|n,l})) - \widetilde{var(Ry_{t|n,l})})^2 \tag{38}$$

where the tilda denotes the variance estimated from the data. Note that the estimator we propose is unweighted GMM and is therefore not optimal. However, as Altonji and Segal (1996) and especially Clark (1996) discuss, the use of the optimal GMM estimator is likely to be problematic for estimating covariance structures. In particular, Clark (1996) shows via Monte Carlo experiments that the optimal estimator for a variety of nonlinear estimators of covariance structures, is biased whereas the unweighted one is not.

Once we estimate $\sigma^2_{v|0}, i$, we can get an expression for the variance of the measurement error in any published release. This variance is given by the following expression:

$$var(y_{t|P,t+n} - y_{t|T}) = \sigma^2_{v|0} \sum_{k=0}^{n} (\lambda_{n|0})^2 (1+i)^{-k} \tag{39}$$

We can expand the summation term on the right hand side, to yield

$$var(y_{t|P,t+n} - y_{t|T}) = \sigma^2_{v|0} \frac{i(1+i)^n}{(1+i)^{n+1} - 1} \tag{40}$$

3.2 A model of a naive or rule of thumb statistics agency

So far it is assumed that the statistics agency solves an optimisation problem when it chooses how to weight the incremental surveys together. We turn next to a model where we instead assume that incremental surveys are weighted equally. We do this for three reasons. First, we want to demonstrate that our method does not depend on the 'rational' agency assumption: it just relies on making some behavioural assumption. Second, it's plausible that the 'rational' agency model is not the best one to capture the real world. And this might not be because actual agencies are not rational, but that they may solve more complicated problems than the one described here. Third, we want to take our method to the data, and we want some way of exploring how robust the estimates coming from this method are to choosing alternatives to our basic behavioural model.

Assuming that the statistics agency weights surveys incrementally implies that a revision to a data release between any two periods l and n will be given by:

$$R_t y_{t|t+n,t+l} = y_{t|P,t+n} - y_{t|P,t+l} \tag{41}$$

$$= (\lambda_{t+n} - \lambda_{t+l})(\sum_{k=0}^{n-1} y_{t|t+k}) + \lambda_{t+n}(\sum_{k=l+1}^{n} y_{t|t+k}) \tag{42}$$

$$= \left(\frac{1}{n+1} - \frac{1}{l+1}\right)\left(\sum_{k=0}^{n-1} y_{t|t+k}\right) + \frac{1}{n+1}(\sum_{k=l+1}^{n} y_{t|t+k}) \tag{43}$$

We can then show that the variance of revisions is given by:

$$var(R_t y_{t|t+n,t+l}) = (\frac{1}{n+1} - \frac{1}{l+1})^2(\sum_{k=0}^{n-1} \sigma^2_{v|k}) + (\frac{1}{n+1})^2(\sum_{k=l+1}^{n} \sigma^2_{v|k}) \tag{44}$$

Substituting in our familiar expression for the relationship between the variance of measurement error around successive surveys, and expanding these summation terms, we can show that the variance of revisions is given by:

$$var(R_t y_{t|t+n,t+l}) = \sigma^2_{v|0}[(\frac{1}{n+1} - \frac{1}{l+1})^2(\frac{(1+i)^{l+1}-1}{i}) \tag{45}$$

$$+ (\frac{1}{n+1})^2(\frac{(1+i)^{n+1}-(1+i)^{l+1}}{i})] \tag{46}$$

Note that this expression differs from its counterpart in the model of the optimising statistics agency in the previous section, contained in equation (36). The new expression for the variance of the measurement error in different releases will be different. It is given by:

$$var(y_{t|P,t+n} - y_{t|T}) = \sigma^2_{v|0}\frac{(1+i)^{n+1}-1}{i(n+1)^2} \tag{47}$$

This, by inspection, is different from our earlier expression under an optimising statistics agency, given in equation (40).

4 Estimating the term structure of measurement error in a real time data set

By way of an illustration, we next take the model to some data. We use the real time data set compiled by Castle and Ellis (2002) based on data published by the UK's Office for National Statistics. The data set is described in more detail in that article. We will estimate the measurement error contained in the initial release,$\sigma^2_{v|0}$and the rate of decay of that measurement error i for real (ie constant price) growth in private consumption and imports expenditure. The real time data we use cover releases between 1985 and 2001. For each observation on a series, we have typically around 24 releases. (There are roughly two releases per quarter over this period, although the exact frequency of releases has changed from time to time.) So we can compile a set of variances that records the average variance of revisions between the first and the second release, and between the first and the third, and between the first and fourth, and so on. We compute these averages over observations.

Table 1 shows our estimates for the measurement error and the rates of decay in the two series. The estimates tell us that the variance of the measurement error in the first release of the growth of imports is a little under six times that of consumption growth. And the information flow falls off faster (i is higher) for consumption than for imports. To interpret the magnitude of these measurement errors, suppose that the steady-state consumption growth rate is about 2.5 per cent a year, or about 0.6 per cent a quarter (0.006 in the units in the table). That means that the variance of the estimate of the growth rate is is a little over 1/100th the average growth rate itself; and the standard deviation of the growth rate (about 0.008, or 0.8 per cent) is therefore roughly of the same order of magnitude as the growth rate itself.

growth in:	$\sigma^2_{v\|0}$	**i**
imports	4.0E-04	1.1E-06
consumption	7.1E-05	1.1E-02

Table 1: Initial release measurement errors, and rates of decay

In the Figures that follow, we use the estimated $\sigma^2_{v|0}$ and i to compute what the model says about the measurement error in different releases of the two series. Figure 1 is for consumption and Figure 2 is for imports.

The measurement error (on the y-axis) shrinks, of course, as we move to later releases (along the x-axis). The Figures compare estimates that come from assuming a rational agency with those when we assume a 'naive' one in the sense set out earlier in the paper. The shape of the curves for consumption and imports are the same - that is because we are using the functional forms coming from the same model in each case. The differences in the two series are apparent from the different y-axis scales. The naive estimates differ in ways that have some intuition. Note that for the first few releases, the estimates of the measurement error in the two series are pretty close. For that period at least our method is in some measure robust to polar assumptions about the behaviour of the statistics agency. But the naive estimates are lower. This seems counter-intuitive but is not. The model looks at the variance of revisions between two early releases. For the naive agency, it assumes that some of the variance of revisions is due to poor weighting, implying that the underlying sampling error in the early surveys is lower than in the case when it tries to fit an optimal agency through that same observed variance of revisions. Further out, the measurement error for the naive estimates increases exponentially. This is because later surveys, whose measurement error is growing exponentially, are weighted equally. At infinity, the variance of the published estimate for a naive agency will tend to infinity.

We can get some indication of how well the model fits the data by displaying Figures of some model predicted variance of revisions against actual data. Figures 3 and 4 do this.

These are not the only revisions to which the model is fitted. The model is fitted solving the minimisation problem set out above. That involves looking at revisions between all possible pairs of releases. The Figures plot the variance of revisions between some release and the final release on the x-axis. But these fits tell us something. Crudely, the optimal agency model seems to do better, just, at capturing the slope of revisions for consumption growth; but the naive agency does better for imports.

5 A model with a variable rate of decay

5.1 Theory

Up to this point, we have assumed a fixed rate of decay i, the quantity that determines the size of the sample of next period's incremental survey relative to this period's sample. We now want to relax this assumption and allow the data to determine how i evolves over time. The flow of information may not decline at a fixed rate over time in reality. This is a straightforward extension of the model, though the algebra becomes a little involved. An appendix presents the derivation in full, but here we state the key points of departure only.

The relationship between the measurement error of one survey relative to another, the counterpart to equation (8) in the fixed-decay model, is now given by:

$$\sigma^2_{v|k} = \prod_{j=0}^{k}(1+i_j)\sigma^2_{v|0}, k = 0 \tag{48}$$

To simplify notation we introduce $\prod_{j=l}^{k}(1+i_j) \equiv \eta_{k,l}$. We can now write out the statistics agency's minimisation problem, given by (7) for the fixed decay model, in terms of the first period's survey measurement error:

$$\text{Min}\sum_{k=0}^{n}(\lambda_{n|k})^2\eta_{k,1}\sigma^2_{v|0} \tag{49}$$

After some straightforward algebra, we get the following expression for the relationship between our observables (the variance of revisions) and our unknowns (the i_js and $\sigma^2_{v|0}$):

$$var(Ry_{t|n,l}) = \sigma^2_{v|0}\frac{\begin{array}{c}\eta^2_{n,i}\left(\sum_{k=l+1}^{n}\eta^{-1}_{k,1}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)^2 + \\ \left(1+\sum_{i=0}^{n-l-2}\eta_{n,n-i}\right)^2\eta_{l,1}\sum_{k=0}^{l}\eta_{l,k+1}\end{array}}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)^2\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)^2} \tag{50}$$

In the same way as before, we note that this gives us a system of equations in our observables and unknowns, and find the optimal choice of the i_js and $\sigma^2_{v|0}$ that minimises an objective function involving the observed and predicted variance of revisions. We turn next to apply this model to the data.

5.2 Time-varying decay: an application

Allowing for time-varying i_js makes the model more realistic, but there is a cost: we will have more parameters to estimate, and will inevitably uncover estimates that are correspondingly less precise. The modelling framework allows us to fix some of the i_js to be equal if we want to. To speed up the numerical maximisation, we decided to fix the i_js beyond the 15th release.

Figures 5 and 6 show, for consumption growth, the implied model based estimates of the variance of measurement error about different releases, and a plot of the fit of the model based revisions to data on actual revisions.

The time-varying model based estimate of the variance of measurement error in the first release of consumption growth (σ_v^2) is $7.5 * 10^{-5}$, not much different from the time-invariant case ($7.1 * 10^{-5}$). But the i's, the implied rates of decay of the incremental sample sizes are very different, both from the time-invariant case, and from release to release. Starting from the decay between release 1 and 2, and moving on, the first few of these i's are given by $\{3.3, 4.8 * 10^{-4}, 4.8 * 10^{-3}, 7.5 * 10^{-5}, 1.4 * 10^{-3}....\}$, which compares with the fixed i estimate of $1.1 * 10^{-2}$. These figures generate substantially different estimates of the variance of measurement error for consumption growth, differences that get larger as we move through to later releases.

The Figures plotting how the models look against (some of) the data on the variance of revisions indicates that the time-varying i model does better at describing the data.

6 Data uncertainty and optimal forecasting

We move on now to provide an illustration of how the estimates based on models like ours could be used for practical purposes. We choose a forecasting example that derives from some earlier work. In Harrison, Kapetanios, and Yates (2004) (section 4 of that paper), we presented an example of how to compute optimal forecasts in a dynamic model subject to the constraint that the forecast only uses data as old as the longest lag in the forecasting equation. We will describe that procedure very briefly here. Readers interested in a fuller description should go back to the original paper. Note that the outputs of the procedure we described above would have many more general uses: as an input into state space model based forecasts (see Kapetanios and Yates (2004)), or models of optimal monetary policy under uncertainty. We present a particular application merely to illustrate that our estimates of the variation in data uncertainty across vintages can make a material difference in a real setting.

The basic set up is in the context of a univariate model for some true data, where we denote true data y_t^*using asterisks. The univariate model is given by:

$$y_t^* = \sum_{i=1}^{p} a_i y_{t-i}^* + e_t \tag{51}$$

where e_t is a white noise shock, and the a_i's are coefficients. We write the measurement model as:

$$\mathbf{y}_t = \mathbf{y}_t^* + \mathbf{v}_t, (\mathbf{y}_t = \{y_t, y_{t-1}, ...y_{t-p+1}\}) \tag{52}$$

As before, v_t is the white noise measurement error. The problem is to minimise the one step ahead forecast error:

$$y_{t+1}^* - \hat{y}_{t+1} \tag{53}$$

where the hat superscript denotes the forecast.

$$= \mathbf{A}_1 \mathbf{y}_t^* + \epsilon_t - \tilde{\mathbf{A}}_1 \mathbf{y}_t^* + \tilde{\mathbf{A}}_1 \mathbf{v}_t = (\mathbf{A}_1 - \tilde{\mathbf{A}}_1)\mathbf{y}_t^* + \tilde{\mathbf{A}}_1 \mathbf{v}_t + \epsilon_{t+1}$$

Here, the $\mathbf{A}_1$ is a vector of the a_i's used to compute the 1 (hence the subscript) step ahead forecast. The choice variable in this minimisation problem is the matrix of coefficients $\tilde{\mathbf{A}}_1$. As we showed in Harrison, Kapetanios, and Yates (2004), and state briefly here, the optimal choice for $\mathbf{A}_1$ involves weighting a variable according to its signal, and according to the measurement error variance.

The mean squared error is written as:

$$(\mathbf{A}_1 - \tilde{\mathbf{A}}_1)\mathbf{\Gamma}(\mathbf{A}_1 - \tilde{\mathbf{A}}_1)' + \tilde{\mathbf{A}}_1 \mathbf{\Sigma}_v^T \tilde{\mathbf{A}}_1' + \sigma_\epsilon^2$$

where $\mathbf{\Gamma} = E(\mathbf{y}_t^* \mathbf{y}_t^{*'})$, and the elements of this we can draw from the matrix $\sigma_\epsilon^2[\mathbf{I}_{p^2} - \mathbf{A} \otimes \mathbf{A}]^{-1}$.$\mathbf{\Sigma}_v^T$ denotes the variance of the measurement error $\boldsymbol{v}_T$.

Importantly, we assume that the covariances of the signal ϵ and the noise v are assumed to be zero. Differentiating the expression for the mean squared forecast error with respect to $\tilde{\mathbf{A}}_1$, and setting equal to zero, we get:

$$\tilde{\mathbf{A}}_1^{opt\ \prime} = (\mathbf{\Gamma} + \mathbf{\Sigma}_v^T)^{-1} \mathbf{\Gamma} \mathbf{A}_1'$$

Note that the greater the measurement error surrounding a particular vintage, the lower the implied corresponding element in $\tilde{\mathbf{A}}_1^{opt\ \prime}$. Or, in short, the more noise in a variable, the less weight it has in an optimal forecast.

We apply this procedure to a univariate model for the quarterly growth rate of UK private consumption. The two tables below set out the results.

Table 2: Whole Period Coefficients in AR forecasting models for UK consumption growth: standard and uncertainty corrected

	AR(1)	AR(2)	AR(3)	AR(4)
standard	$-0.063_{(0.123)}$			
	$-0.057_{(0.122)}$	$0.156_{(0.122)}$)		
	$-0.118_{(0.118)}$	$0.169_{(0.117)}$	$0.291_{(0.120)}$	
	$-0.073_{(0.124)}$	$0.196_{(0.119)}$	$0.272_{(0.120)}$	$-0.158_{(0.125)}$
optimal	$-0.059_{(0.114)}$			
	$-0.053_{(0.110)}$	$0.146_{(0.113)}$		
	$-0.100_{(0.101)}$	$0.153_{(0.106)}$	$0.267_{(0.111)}$	
	$-0.054_{(0.089)}$	$0.166_{(0.101)}$	$0.241_{(0.108)}$	$-0.150_{(0.113)}$

This first table shows the effect on the forecasting equations of carrying out the procedure we have just described. These are estimates on data from 1980-1998. Most coefficients fall. The model is a model of demeaned consumption growth, so this implies putting more weight on the mean. There is some slight tendency to put more weight on older data relative to newer data. For example, take the AR(3). The ratio of the AR(1) to the AR(3) coefficients in the standard model is about 1:2.5. In the uncertainty adjusted case that ratio is 1:2.7.

Table 3: MSE ratios and Diebold-Mariano tests

	Whole period		First subperiod		Second subperiod	
Model	MSE Ratio	D-M Test	MSE Ratio	D-M Test	MSE Ratio	D-M Test
AR(1)	0.987	2.46^*	0.987	1.73^*	0.987	2.93^*
AR(2)	0.974	2.48^*	0.968	2.33^*	0.989	1.57^*
AR(3)	0.977	1.55	0.975	1.26	0.983	1.38^*
AR(4)	0.965	1.73^*	0.959	1.52	0.980	1.08^*

* denotes significance at the 5% level

This second table shows recursive out-of-sample Diebold-Mariano[1] forecast evaluation tests on our two forecasting models. The whole period refers to out-of-sample tests for 1988-1998: the two sub-periods divide that sample into two. The Diebold-Mariano test compares the adjusted and the unadjusted root mean squared errors, and looks to see whether the forecasts can be said to be statistically significantly different from one another. The test results show that many of them are. The ratio of the mean squared forecast errors are all less than 1, for all the models we considered, implying that the measurement error corrected

[1] See, for details, Diebold and Mariano (1995)

forecasts are better. The Diebold-Mariano test statistics have a critical value of 1.96 at the 5 per cent level. A majority of these in table 3 are greater than that value.

7 Conclusions

Knowing how well one series is measured relative to another, or how much more reliable older, revised data is than more recent data is useful in many situations: estimation, forecasting and policy-making. We have presented a method for extracting estimates of measurement error from observations on the variance of revisions in a data series. This method involves a conjecture about how the reliability of the incremental information a statistics agency obtains declines over time, and about how the agency weights the information together to form a new estimate of a data point. We chose to illustrate our method using an assumption that the measurement error in incremental surveys grows exponentially, to capture the idea that each period less and less new information arrives; and that this information is weighted optimally to form new estimates of the data.

But our method doesn't depend on this precise assumption. We showed that by deriving our results for a variable rate of decay and also by assuming a 'naive' statistics agency that gives as much weight to later, less well-measured surveys as to earlier, better ones. Applying our method to real time data on quarterly growth rates of UK private consumption and imports, we get estimates that suggest that the measurement error in the growth of imports is almost six times larger than that for consumption. Finally, we used our estimates of measurement error in a simple forecasting exercise described in Harrison, Kapetanios, and Yates (2004). Using AR models for the quarterly growth in private consumption, we showed how the out-of-sample forecasting performance of model-based, measurement-error-corrected forecasts significantly outperform forecasts coming from unadjusted OLS equations.

References

ALTONJI, J. G., AND L. M. SEGAL (1996): "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, 14(3), 353–366.

AOKI, K. (2003): "On the optimal monetary policy response to noisy indicators," *Journal of Monetary Economics*, 50(3), 501–23.

BERNHARDSEN, T., O. EITRHEIM, A. JORE, AND O. ROISLAND (2004): "Real-time Data for Norway: Challenges for Monetary Policy," Paper presented at the Bundesbank conference on real-time data, Frankfurt.

Busetti, F. (2001): “The use of preliminary data in econometric forecasting: an application with the Bank of Italy Quarterly Model,” Bank of Italy Discussion Paper.

Castle, J., and C. Ellis (2002): “Building a real-time database for GDP(E),” *Bank of England Quarterly Bulletin, Spring*, 42(1), 42–49.

Clark, T. E. (1996): “Small Sample Properties of Estimators of Nonlinear Models of Covariance Structures,” *Journal of Business and Economic Statistics*, 14(3), 367–373.

Coenen, G., A. Levin, and W. Weiland (2001): “Data uncertainty and the role of money as an information variable in monetary policy,” *ECB Working Paper no 84*.

Croushore, D., and T. Stark (2001): “A Real Time Dataset for Macroeconomists,” *Journal of Econometrics*, 105, 111–130.

Diebold, F. X., and R. S. Mariano (1995): “Comparing predictive accuracy,” *Journal of Business and Economic Statistics*, 13(3), 253–62.

Eggington, D., A. Pick, and S. Vahey (2002): “Keep it real!: A real-time data set for macroeconomists: does the vintage matter?,” *Economics Letters*, forthcoming.

Gayen, J., and S. VanNorden (2004): “The Reliability of Canadian Output Gap Estimates,” Paper presented at the Bundesbank conference on real-time data, Frankfurt.

Gerberding, C., A. Worms, and F. Seitz (2004): “How the Bundesbank really conducted monetary policy: An analysis based on real-time data,” Bundesbank, Paper presented at the Bundesbank conference on real-time data, Frankfurt.

Harrison, R., G. Kapetanios, and T. Yates (2004): “Forecasting with measurement errors in dynamic models,” *Bank of England Working Paper.*

Harvey, A., C. R. Mckenzie, D. P. C. Blake, and M. J. Desai (1983): “Irregular Data Revisions,” *A Zellner (ed), Proceedings of ASA-CENSUS-NBER Conference on Applied Time Series Analysis of Economic Data. US Department of Commerce, Washington, DC*, pp. 329–347.

Howrey, E. P. (1978): “The use of preliminary data in econometric forecasting,” *Review of Economic Statistics*, 60, 193–200.

Jaaskela, J., and T. Yates (2004): “Monetary policy and data uncertainty,” *Mimeo, Bank of England.*

Kamada, K. (2004): “Real-Time Estimation of the Output Gap in Japan and its Usefulness for Inflation Forecasting and Policymaking,” Bank of Japan, Paper presented at the Bundesbank conference on real-time data, Frankfurt.

Kapetanios, G., and T. Yates (2004): “Using a Real-Time Dataset for Refining Preliminary Data and Forecasting,” *Mimeo, Bank of England.*

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

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

Orphanides, A. (2001): “Monetary policy rules based on real-time data,” *American Economic Review*, pp. 964–985.

Orphanides, A., and S. VanNorden (2002): “The unreliability of output gap estimates in real time,” *Review of Economics and Statistics*, 84, 569–583.

Sargent, T. (1989): “Two models of measurement and the investment accelerator,” *Journal of Political Economy*, 97(2), 251–87.

Svensson, L. E. O., and M. Woodford (2003): “Indicator variables for monetary policy,” *Journal of Monetary Economics*, 50, 691–20.

Swanson, E. (2000): “On signal extraction and non-certainty equivalence in optimal monetary policy rules,” *Econometric Society World Congress.*

8 Appendix: a naive or rule of thumb statistics agency, measurement error, and data revisions

In this appendix, we derive the relationship between observed revisions and the unobserved parameters of the measurement error function, under the assumption that data observations are weighted equally, rather than optimally. Suppose that measured and true variables are related by (4). The variance of the measurement error around successive surveys is given by (8). If the statistics agency weights observations equally, then successive releases will be given by:

$$t : y_{t|P,t} = y_{t|t} \tag{54}$$

$$t+1 : y_{t|P,t+1} = \lambda_{t+1} y_{t|t} + \lambda_{t+1} y_{t|t+1}, \lambda_{t+1} = 1/2 \tag{55}$$

$$t+2 : y_{t|P,t+2} = \lambda_{t+2} y_{t|t} + \lambda_{t+2} y_{t|t+1} \lambda_{t+2} y_{t|t+2}, \lambda_{t+2} = 1/3 \tag{56}$$

$$t+n : y_{t|P,t+n} = \sum_{k=0}^{n} \lambda_{t+n} y_{t|t+k}, \lambda_{t+n} = \frac{1}{n+1} \tag{57}$$

One period revisions to published data will then be given by:

$$R_t y_{t|t+1} = y_{t|P,t+1} - y_{t|P,t} = (\lambda_{t+1} - 1) y_{t|t} + \lambda_{t+1} y_{t|t+1} = \tag{58}$$

$$1/2 y_{t|t} + 1/2 y_{t|t+1} \tag{59}$$

More generally, an l period revision will be given by:

$$R_t y_{t|t+n,t+l} = y_{t|P,t+n} - y_{t|P,t+l} \tag{60}$$

$$= (\lambda_{t+n} - \lambda_{t+l})(\sum_{k=0}^{n-1} y_{t|t+k}) + \lambda_{t+n}(\sum_{k=l+1}^{n} y_{t|t+k}) = \tag{61}$$

$$(\frac{1}{n+1} - \frac{1}{l+1})(\sum_{k=0}^{n-1} y_{t|t+k}) + \frac{1}{n+1}(\sum_{k=l+1}^{n} y_{t|t+k}) \tag{62}$$

Substituting in the fact that the published data equals the true data plus the measured data, it turns out that the terms in the true value of y cancel out to give an expression solely in terms of the measurement error. This is shown below:

$$R_t y_{t|t+n,t+l} = (\frac{1}{n+1} - \frac{1}{l+1})(l+1) y_{t|T} + \sum_{k=0}^{n-1} v_{t|t+k}) \tag{63}$$

$$+ \frac{1}{n+1}(n-l) y_{t|T} + (\sum_{k=l+1}^{n} v_{t|t+k}) \tag{64}$$

$$=(\frac{1}{n+1} - \frac{1}{l+1})(\sum_{k=0}^{n-1} v_{t|t+k}) + \frac{1}{n+1}(\sum_{k=l+1}^{n} v_{t|t+k}) \tag{65}$$

The variance of an l period revision can be written as:

$$var(R_t y_{t|t+n,t+l}) = (\frac{1}{n+1} - \frac{1}{l+1})^2(\sum_{k=0}^{n-1} \sigma^2_{v^k}) + (\frac{1}{n+1})^2(\sum_{k=l+1}^{n} \sigma^2_{v^k}) \tag{66}$$

Substituting in (8), we get that:

$$var(R_t y_{t|t+n,t+l}) = (\frac{1}{n+1} - \frac{1}{l+1})^2(\sum_{k=0}^{n-1}(1+i)^k \sigma^2_{v|0}) \tag{67}$$

$$+ (\frac{1}{n+1})^2(\sum_{k=l+1}^{n} (1+i)^k \sigma^2_{v|0}) \tag{68}$$

This implies

$$var(R_t y_{t|t+n,t+l}) = \sigma^2_{v|0}[(\frac{1}{n+1} - \frac{1}{l+1})^2(\sum_{k=0}^{n-1}(1+i)^k) + (\frac{1}{n+1})^2(\sum_{k=l+1}^{n}(1+i)^k)] \tag{69}$$

We can expand the summation terms as follows:

$$\sum_{k=0}^{n-1}(1+i)^k = \frac{(1+i)^{l+1}-1}{i}, \sum_{k=l+1}^{n}(1+i)^k = \frac{(1+i)^{n+1}-(1+i)^{l+1}}{i} \tag{70}$$

to give

$$var(R_t y_{t|t+n,t+l}) = \sigma^2_{v|0}[(\frac{1}{n+1} - \frac{1}{l+1})^2(\frac{(1+i)^{l+1}-1}{i}) \tag{71}$$

$$+ (\frac{1}{n+1})^2(\frac{(1+i)^{n+1}-(1+i)^{l+1}}{i})] \tag{72}$$

We can also get the variance of the published data as

$$var(y_{t|P,t+n} - y_{t|T}) = \sigma^2_{v|0}\frac{(1+i)^{n+1}-1}{i(n+1)^2} \tag{73}$$

9 Appendix: a variable rate of decay model

Here we derive our results for the main paper under the assumption that the rate of decay of the size of the sample of incremental surveys changes over time. Recall that in the main paper this quantity, i was assumed to be fixed. In particular, we will assume that the measurement error surrounding incremental surveys is given by the following expression:

$$\sigma^2_{v|k} = \prod_{j=0}^{k}(1+i_j)\sigma^2_{v|0}, k = 0 \tag{74}$$

To simplify notation we introduce $\prod_{j=l}^{k}(1+i_j) \equiv \eta_{k,l}$. We can now write out the statistics agency's minimisation problem in terms of the first period's survey measurement error:

$$\text{Min}\sum_{k=0}^{n}(\lambda_{n|k})^2\eta_{k,1}\sigma^2_{v|0} \tag{75}$$

We write the problem out as

$$\lambda^2_{n|0}\sigma^2_{v|0} + \lambda^2_{n|1}\eta_{1,1}\sigma^2_{v|0} + \lambda^2_{n|2}\eta_{2,1}\sigma^2_{v|0} + \lambda^2_{n|3}\eta_{3,1}\sigma^2_{v|0} + ...+ \tag{76}$$

$$\lambda^2_{n|n-1}\eta_{n-1,1}\sigma^2_{v|0} + (1 - \lambda_{n|0} - \lambda_{n|1} - \lambda_{n|2} - ... - \lambda_{n|n-1})\eta_{n,1}\sigma^2_{v|0} \tag{77}$$

We drop $\sigma^2_{v|0}$ and so minimise

$$\lambda^2_{n|0} + \lambda^2_{n|1}\eta_{1,1} + \lambda^2_{n|2}\eta_{2,1} + \lambda^2_{n|3}\eta_{3,1} + ... \tag{78}$$

$$+ \lambda^2_{n|n-1}\eta_{n-1,1} + (1 - \lambda_{n|0} - \lambda_{n|1} - \lambda_{n|2} - ... - \lambda_{n|n-1})^2\eta_{n,1} \tag{79}$$

The FOCs for this problem are

$$\lambda_{n|j} - (1 - \lambda_{n|0} - \lambda_{n|1} - \lambda_{n|2} - ... - \lambda_{n|n-1})\eta_{n,1} = 0, j = 0, \ldots, n-1 \tag{80}$$

Clearly

$$\lambda_{n|0} = \lambda_{n|1}\eta_{1,1} = ... = \lambda_{n|n-1}\eta_{n-1,1} \tag{81}$$

So using (80) for $j = 0$ we get

$$\lambda_{n|0} - \left(1 - \lambda_{n|0} - \lambda_{n|0}/\eta_{1,1} - \lambda_{n|0}/\eta_{2,1} - ... - \lambda_{n|0}/\eta_{n-1,1}\right)\eta_{n,1} = 0 \tag{82}$$

or

$$\lambda_{n|0} - \left(\eta_{n,1} - \lambda_{n|0}\eta_{n,1} - \lambda_{n|0}\eta_{n,2} - \lambda_{n|0}\eta_{n,3} - ... - \lambda_{n|0}\eta_{n,n}\right) = 0 \tag{83}$$

Grouping the $\lambda_{n|0}$ gives

$$\lambda_{n|0} = \frac{\eta_{n,1}}{1 + \eta_{n,n} + \eta_{n,n-1} + ... + \eta_{n,1}} \tag{84}$$

Then

$$\lambda_{n|j} = \frac{\eta_{n,j}}{1 + \eta_{n,n} + \eta_{n,n-1} + ... + \eta_{n,1}}, j = 1, \ldots n \tag{85}$$

Recalling from the main body of the paper that the variance of revisions is given by:

$$var(Ry_{t|n,l}) = \sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})^2\sigma^2_{v|k} + \sum_{k=l+1}^{n}(\lambda_{n|k})^2\sigma^2_{v|k} \tag{86}$$

which, given the relationship between the $\sigma^2_{v|k}$'s, is given by:

$$var(Ry_{t|n,l}) = \sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})^2\left(\eta_{k,1}\sigma^2_{v|0}\right) + \sum_{k=l+1}^{n}(\lambda_{n|k})^2\left(\eta_{k,1}\sigma^2_{v|0}\right) \tag{87}$$

Noting (81), we can write:

$$Var(Ry_{t|n,l}) = \sigma^2_{v|0}\left[\sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})^2\eta_{k,1} + \sum_{k=l+1}^{n}(\lambda_{n|0})^2\eta^{-1}_{k,1}\right] \tag{88}$$

We next set about expanding the summations in these expressions. First we get

$$\sum_{k=l+1}^{n}(\lambda_{n|0})^2\eta^{-1}_{k,1} = \left(\frac{\eta_{n,1}}{1 + \eta_{n,n} + \eta_{n,n-1} + ... + \eta_{n,1}}\right)^2\sum_{k=l+1}^{n}\eta^{-1}_{k,1} \tag{89}$$

Next the term $\sum_{k=0}^{l}(\lambda_{n|k} - \lambda_{l|k})^2\eta_{k,1}$. First examine

$$(\lambda_{n|k} - \lambda_{l|k})^2 = \left(\frac{\eta_{n,k+1}}{1 + \sum_{i=0}^{n-1}\eta_{n,n-i}} - \frac{\eta_{l,k+1}}{1 + \sum_{i=0}^{l-1}\eta_{l,l-i}}\right)^2 = \tag{90}$$

$$\left(\frac{\eta_{n,k+1}\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)-\eta_{k+1,l}\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)}\right)^2 = \tag{91}$$

$$\left(\frac{\eta_{l,k+1}\left(\eta_{n,l+1}\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)-\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\right)}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)}\right)^2 = \tag{92}$$

$$\left(\frac{\eta_{l,k+1}\left(\left(\sum_{i=-1}^{l-1}\eta_{n,l-i}\right)-\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\right)}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)}\right)^2 \tag{93}$$

$$= \left(\frac{-\eta_{l,k+1}\left(1+\sum_{i=0}^{n-l-2}\eta_{n,n-i}\right)}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)}\right)^2 \tag{94}$$

So

$$\sum_{k=0}^{l}(\lambda_{n|k}-\lambda_{l|k})^2\eta_{k,1} = \sum_{k=0}^{l}\left(\frac{-\eta_{l,k+1}\left(1+\sum_{i=0}^{n-l-2}\eta_{n,n-i}\right)}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)}\right)^2\eta_{k,1} \tag{95}$$

Finally

$$var(Ry_{t|n,l}) = \sigma_v^2\left[\begin{array}{c}\sum_{k=0}^{l}\left(\frac{-\eta_{l,k+1}\left(1+\sum_{i=0}^{n-l-2}\eta_{n,n-i}\right)}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)}\right)^2\eta_{k,1}\\ +\left(\frac{\eta_{n,1}}{1+\sum_{i=0}^{n-1}\eta_{n,n-i}}\right)^2\sum_{k=l+1}^{n}\eta_{k,1}^{-1}\end{array}\right] \tag{96}$$

or

$$var(Ry_{t|n,l}) = \sigma_{v|0}^2\frac{\begin{array}{c}\eta_{n,i}^2\left(\sum_{k=l+1}^{n}\eta_{k,1}^{-1}\right)\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)^2\\ +\left(1+\sum_{i=0}^{n-l-2}\eta_{n,n-i}\right)^2\eta_{l,1}\sum_{k=0}^{l}\eta_{l,k+1}\end{array}}{\left(1+\sum_{i=0}^{n-1}\eta_{n,n-i}\right)^2\left(1+\sum_{i=0}^{l-1}\eta_{l,l-i}\right)^2} \tag{97}$$

This gives the system of equations linking the observed variance of revisions to the unknown variable rates of decay i_j and the initial period's measurement error $\sigma_{v|0}^2$.

Figure 1: Measurement error variance for consumption growth

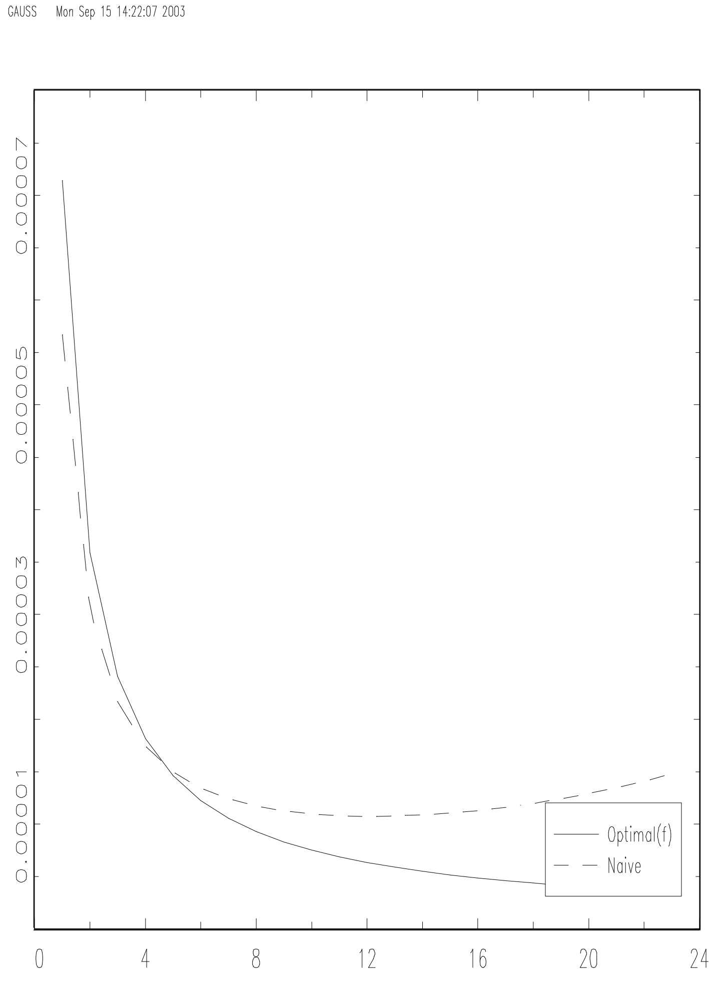


Figure 2: Measurement error variance for import growth

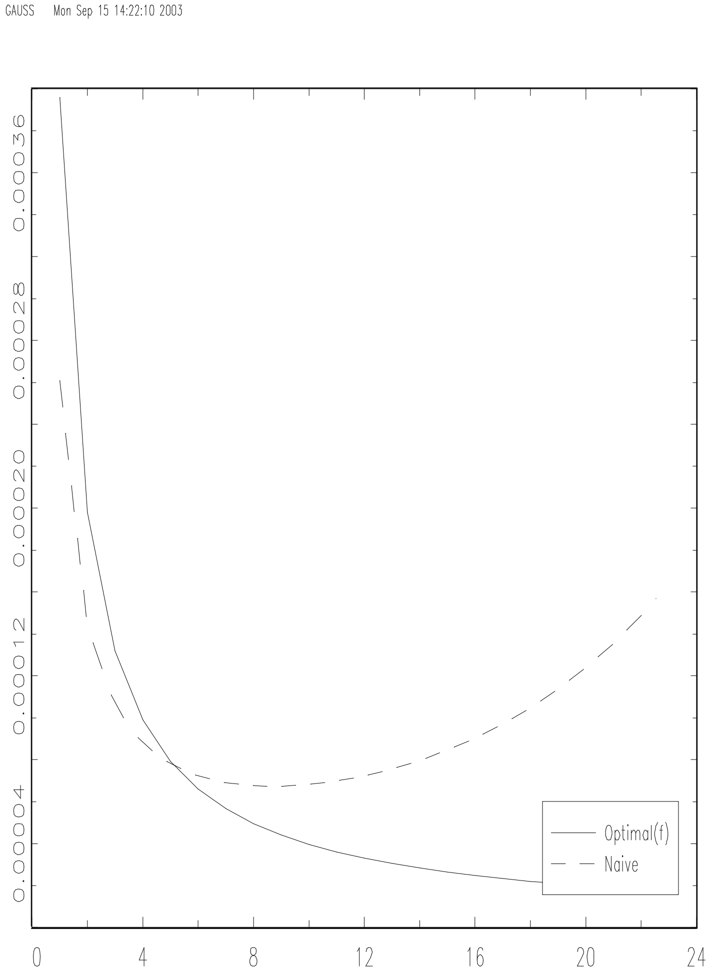


Figure 3: Variance of $Ry_{t|24,l}$, $l = 1, \ldots, 23$ for consumption growth

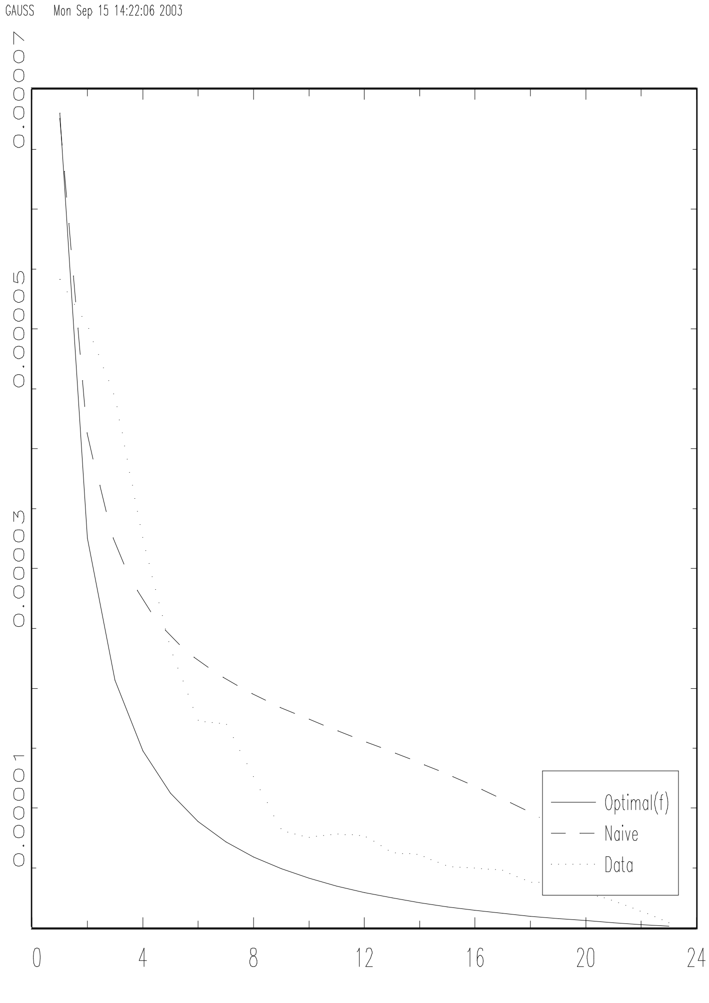


Figure 4: Variance of $Ry_{t|24,l}$, $l = 1, \ldots, 23$ for import growth

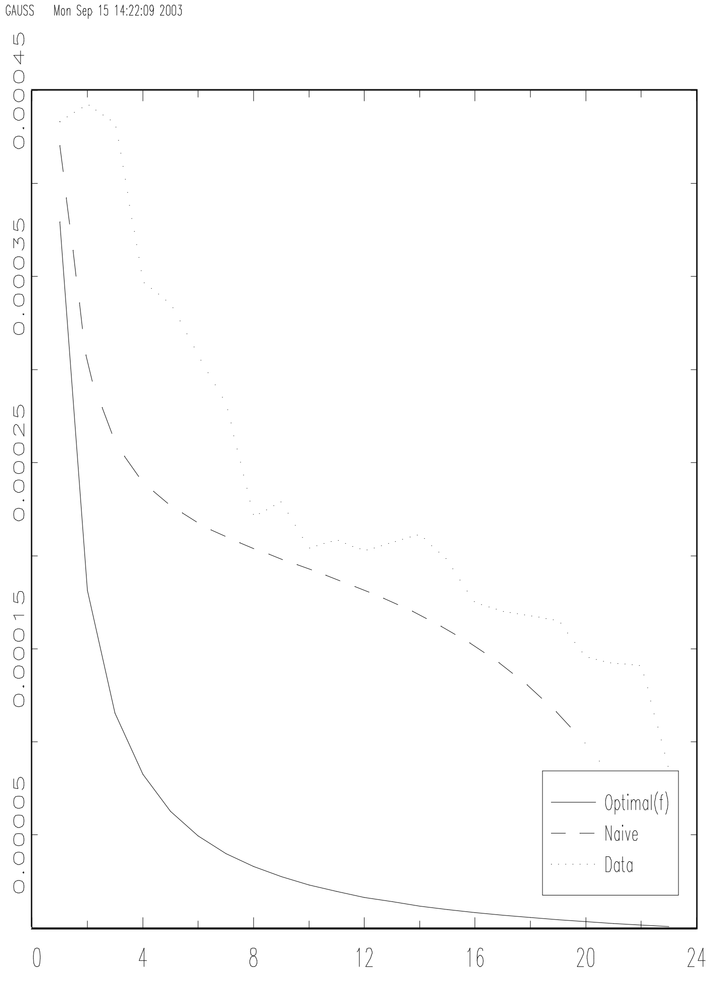


Figure 5: Measurement error variance for consumption growth

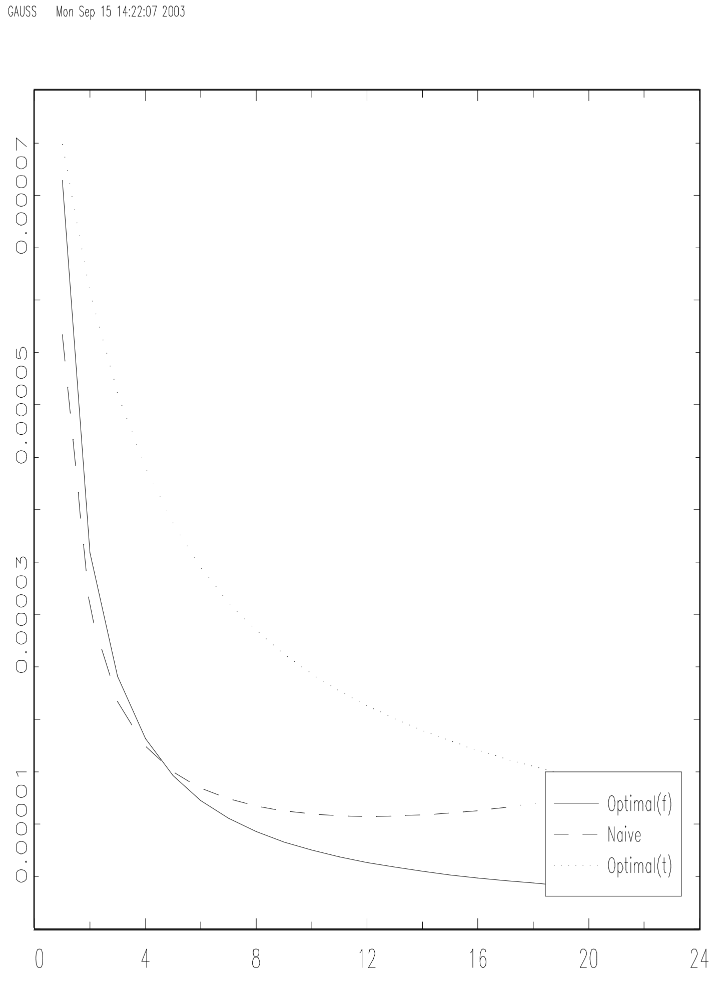
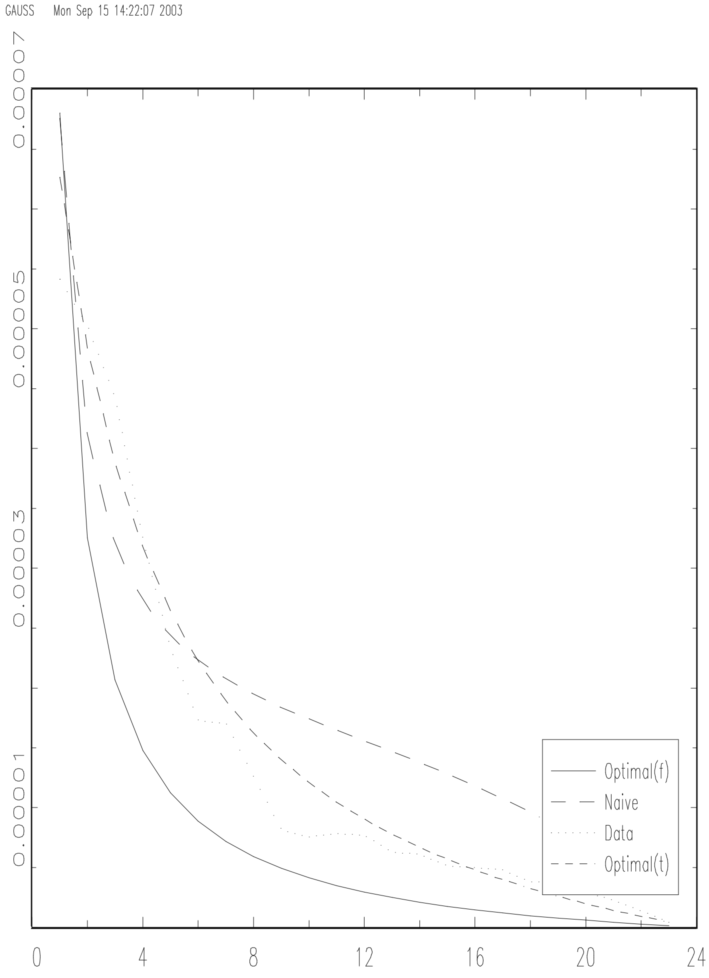


Figure 6: Variance of $Ry_{t|24,l}$, $l = 1, \ldots, 23$ for consumption growth



Queen Mary
University of London

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

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