

# Toward a theory of evaluating predictive accuracy

Robert M. Kunst  
University of Vienna  
and  
Institute for Advanced Studies Vienna

Adusei Jumah  
University of Vienna

September 3, 2004

## **Abstract**

We offer a suggestion for a theoretical basis for the comparative evaluation of forecasts. Instead of generally assuming the data to be generated from a stochastic model, we classify three stages of prediction experiments: pure non-stochastic prediction of given data, stochastic prediction of given data, double stochastic simulation. The concept is demonstrated using an empirical example of UK investment data.

# 1 Introduction

Currently, there is a growing interest in establishing an econometric theory of forecasting that satisfies the needs of practitioners as well as the demand for a firm basis in statistics. The popular books by CLEMENTS AND HENDRY (1998), CHATFIELD (2000), and ŠTULAJTER (2002) are some examples for the remarkable achievements in this field. An important message that is repeatedly stated by these researchers is that prediction methods, including their auxiliary models if the methods are model-based, should be separated carefully from the concept of true data-generating processes. Consequently, in the field of econometric forecasting there may be more promising tasks and also less promising ones. As examples for the former set, we may mention the search for methods that are valuable in repeatedly encountered empirical applications and projects, such as macroeconomic forecasts or stock-market predictions for asset allocation. If a method is found to dominate a rival method in the majority of comparable case studies, that one may become the recommended method in further applications. Similarly, it makes sense to compare rival predictions conditional on certain classes of assumed data-generating mechanisms. If convincing evidence on the nature of a probability law exists, either based on large amounts of comparable data and statistical testing methods or on grounds of substance-matter theory, such studies and results may motivate the usage of particular forecasting methods that are tuned to specific processes. Unfortunately, there are also research projects with a poorer perspective, such as the testing of subject-matter theories with the aid of prediction methods. Testing for correct statistical specification in a forecasting model is another less well motivated aspect, particularly if the forecasting model evolves as a good tool for prediction in empirical evaluations. Similar remarks apply to testing for the significance of differences across forecasting procedures in one given data set.

In this paper, it is attempted to provide a formal basis for the evaluation of forecasts. We consider three stages of increasing sophistication in prediction evaluations. Firstly, a time series of data is compared with a time series of point predictions. In such applications, only a single vector of observations is available and there is no real need to reach a final verdict on its statistical properties. The measurement of the distance of two vectors in Euclidean space is a mathematical task rather than a statistical one. It becomes statistical in repeated experiments or panels, for example for comparable data sets, such as gross production series from various countries or

different industries. Secondly, a given data series is compared with a hypothesized statistical data-generating process. This is commonly known as stochastic prediction and requires measuring the proximity of a vector and a probabilistic entity. We support this view of the task rather than the popular reversed view, which assumes a probabilistic nature for the data and a merely numerical one for point forecasts. That view may rely too much on equating a tentatively hypothesized model—the basis for the prediction tool—and reality. Thirdly, we consider the comparison of a stochastic data-generating process and a stochastic forecasting model. This third specification matches the aforementioned need for tuning a forecasting method to an *a priori* assumed generation model.

An empirical application is used to highlight the features that are presented in the first, theoretical part of this paper. We use a multivariate data set on investment and output in the UK economy. There is some economic theory that can be used *a priori*, such as the tendency of an investment-output ratio to be approximately constant in the long run, and there is also some econometric evidence that has evolved from decades of empirical research on macroeconomic aggregates, such as the general approximate validity of first-order integrated time-series models for the logarithms of national accounts data, such as investment and output. There is also some information on institutions, such as the strong influence of exogenous government policy on construction investment and its relative share in output. In order to focus on the more mechanical aspects of the issue, the latter information set is set aside for our application, which is an acceptable strategy for medium-run projections, where the prospects of future government policies are largely unknown. Based on these pieces of information, though without using a fully developed model of economic behavior, a small set of forecasting procedures is applied and compared.

The remainder of this paper is structured as follows. Section 2 introduces the concepts of a predictor, which essentially constructs a sequence of predicted numbers from a sequence of given numbers, and of a loss function, which measures the distance of the predicted values and the targeted time shift of the original sequence. Section 3 considers the extension of the concept to stochastic prediction, which uses a stochastic process to forecast a given sequence. Section 4 addresses the task of double stochastic evaluation, which compares an assumed data-generating process and a stochastic predictor. Section 5 considers an empirical example, in which the application of all concepts is demonstrated. Section 6 concludes.

## 2 Simple prediction

### 2.1 Basic definitions

**Definition 1** A predictor is a function  $\varphi$

$$\begin{aligned} \mathbb{R}^{\mathbb{N}} &\rightarrow \mathbb{R}^{\mathbb{N}} \\ x &\mapsto \varphi(x) \end{aligned} \tag{1}$$

with the property that the pre-image of any cylinder set  $(x_1, \dots, x_k) \times \mathbb{R}^{\mathbb{N}_{(k)}}$  is a superset of a cylinder set  $E \times \mathbb{R}^{\mathbb{N}_{(k)}}$ , where  $\mathbb{N}_{(k)}$  denotes the set  $\{k + 1, k + 2, \dots\}$ .

The condition guarantees that the image depends on the first  $k$  sequence elements only. In principle, prediction can also be used in a non-temporal environment and for finite spaces, such as  $\mathbb{R}^n$ . In this way, the concept is used for cross validation, where an observation  $x_j$  is predicted using all remaining observations and this exercise is repeated for all  $j \in \{1, \dots, n\}$ . It is also simple to extend the definition to non-real data. Here, we restrict attention to time series and to real data.

Whereas a predictor is defined as *any* function of  $(x_1, \dots, x_k)$ , we now need a criterion for discriminating good from bad predictions. Such assessment naturally depends on the target of prediction. This target is to approximate an observation  $x_j$ , with  $j \notin \{1, \dots, k\}$ . If  $j = k + h$ , this is an *h-step prediction*. The ideal situation would be that  $\varphi(x)$  be the sequence  $(x_{1+h}, x_{2+h}, \dots)$ , i.e. a time shift of the original sequence  $x$ . Generally, the quality of predictions can be measured by loss functions:

**Definition 2** A loss function for *h-step prediction* is a function  $G : \mathbb{R}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}} \rightarrow \mathbb{R}^+$ , which is symmetric in its arguments and fulfills  $G(\varphi(x), F^h x) = 0$  for  $\varphi(x) = F^h x$ , where  $\varphi$  is a predictor and  $F$  is the forward shift operator. Furthermore, it is required that  $G$  is monotonous in the sense that  $G(y, F^h x) \geq G(z, F^h x)$  if  $|y_j - x_{h+j}| \geq |z_j - x_{h+j}|$  for all  $j \in \mathbb{N}$ .

Usually, it may make sense to require *strict monotonicity* by demanding that  $G(y, F^h x) > G(z, F^h x)$  if  $|y_j - x_{h+j}| > |z_j - x_{h+j}| + \varepsilon$  for some  $\varepsilon > 0$  for a non-zero share of indices  $j \in \mathbb{N}$ . For a set  $J \subset \mathbb{N}$ , a share of indices may be defined as the  $\lim_{n \rightarrow \infty} n^{-1} \text{card} \{j \in J : j \leq n\}$ . Thus, one can also avoid trivial loss  $G \equiv 0$ . We point out, however, that while many common loss

functions are indeed strictly monotonous, strict monotonicity excludes some interesting cases of weighted loss, which we will consider in the following. Often, it may be convenient to view  $G$  as a function of  $x, h$ , and  $\varphi$ , thus one may refer to ‘the loss of the predictor  $\varphi$  for  $h$ -step prediction of  $x$ ’.

**Example.** Typical loss functions depend on single-observation loss functions in a ‘stationary’ way, such as the mean-squared error loss function

$$G_2(x, y) = \lim_{n \rightarrow \infty} n^{-1} \sum_{j=1}^n (x_j - y_j)^2 \quad (2)$$

or the mean-absolute error loss function

$$G_1(x, y) = \lim_{n \rightarrow \infty} n^{-1} \sum_{j=1}^n |x_j - y_j|$$

or, in general,

$$G(x, y) = \lim_{n \rightarrow \infty} n^{-1} \sum_{j=1}^n g(x_j, y_j).$$

Such a single-observation loss function  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^+$  corresponds to the function  $g$  that is commonly used in econometrics (see, e.g., DIEBOLD AND MARIANO, 1995). The general function  $G$  allows for ‘robust’ loss functions in the spirit of ROUSSEUW (1984), where summation is modified, such as median squared loss

$$G(x, y) = \lim_{n \rightarrow \infty} \text{median} \{(x_j - y_j)^2, j \leq n\}.$$

The general definition of loss also admits weighing observation loss for certain ranges or as  $n \rightarrow \infty$ , for example in the sense of discounted squared loss

$$G(x, y) = \lim_{n \rightarrow \infty} N(n, \rho)^{-1} \sum_{j=1}^n \rho(j) (x_j - y_j)^2,$$

with  $N(n, \rho)$  chosen such that non-trivial limits are obtained, typically as a sub-linear function of  $n$ . A simple example is  $\rho(j) = \rho^j$  and  $N(n, \rho) = (1 - \rho)^{-1}$  for  $0 < \rho < 1$ . In contrast to mean-squared error loss, single observations have a non-zero impact on the loss. Particularly, the first part

of the sample affects the loss critically. In order to capture the usual aims of empirical forecasters,  $\rho(j)$  may reach a maximum for  $j$  in the range of typically available sample sizes and may discount large  $j$  in the sense that  $\lim_{j \rightarrow \infty} \rho(j) = 0$ .  $\square$

The distinction between squared loss and discounted squared loss can be motivated in a simple example.

**Example.** The data  $x$  is a trajectory from an autoregression with unknown  $\phi$  and  $\sigma^2$ . Prediction relies on ‘correctly specified’ autoregressions with estimated  $\hat{\phi}_j$  and  $\hat{\sigma}_j^2$ . For small  $j$ , these estimates fluctuate strongly and therefore forecasters will allot  $\rho(j) \approx 0$ . Larger weights will be allotted to  $20 \leq j \leq 100$ . In this range,  $\hat{\phi}_j$  is known to be affected by the Hurwicz bias, which does not only impair estimation or approximation of  $G(\varphi(x), Fx)$  but also the relative performance as compared to rival predictors  $\tilde{\varphi}$  that may be constructed from other (‘mis-specified’) models or technical algorithms such as exponential smoothing. For  $j \approx 1000$ , the Hurwicz bias has almost disappeared, and as  $j$  increases further, the autoregressive  $\varphi$  procedure will ‘win’ all horse races against rival models unless  $\rho(j)$  declines. Such discounting appears to be appropriate, if the forecaster is unlikely to meet such large samples in applications. The common practice of evaluating predictive accuracy on the basis of the final part of the available sample conforms to  $\rho(j) = 0$  for  $j = 1, \dots, [\lambda n]$  and for  $j > n$ , and  $\rho(j) = 1$  for  $j = [\lambda n] + 1, \dots, n$ , with  $\lambda \in (0, 1)$ .  $\square$

Note that loss functions are usually defined by limit operations and thus are reminiscent of statistical expectations. However, no stochastic framework has been introduced. This reflects the usual aim of forecasting, where an unknown *observation* is targeted and not a stochastic process. The investigated time series *may be* a trajectory from a stochastic process, yet even knowledge of the probability law of the underlying stochastic process does not lead to perfect or ideal prediction. Alternative concepts would be loss functions for finite-length data  $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^+$ , or loss functions adapted to the sample size, such as  $n^{-1} \sum_{j=1}^n (x_j - y_j)^2$ . The latter concept can be embedded into our framework, while the former one abandons asymptotic arguments.

**Example.** If  $x$  is a trajectory from an autoregressive process  $x_t = \phi x_{t-1} + \varepsilon_t$  with  $\varepsilon_t \sim NID(0, \sigma^2)$  and both parameters  $\phi$  and  $\sigma^2$  are known, a natural predictor for  $h = 1$  is

$$\varphi(x) = \varphi(x_1, x_2, \dots) = (\phi x_1, \phi x_2, \dots) = \phi x,$$

with  $G(\varphi(x), Fx) = \sigma^2$  due to the law of large numbers. If parameters are

not known, a common alternative is the least-squares predictor

$$\varphi(x) = \left( \hat{\phi}_1 x_1, \hat{\phi}_2 x_2, \dots \right)$$

for

$$\hat{\phi}_t = \frac{\sum_{j=2}^t x_j x_{j-1}}{\sum_{j=1}^{t-1} x_j^2}, t \geq 2,$$

and  $\hat{\phi}_1 = 0$ . Replacing the upper summation bound in the denominator by  $t$  yields the Yule-Walker predictor. A loss close to  $\sigma^2$  may also be obtained by radically different methods, hence  $\varphi$  should not be interpreted as being model-determined necessarily.  $\square$

Next, note that the loss function  $G$  is not equivalent to a pseudo-metric on  $\mathbb{R}^N$ , i.e. a metric without the requirement that  $d(x, y) = 0$  only if  $x = y$  (see, e.g., DAVIDSON, 1994, p.75). While any pseudo-metric on  $\mathbb{R}^N$  defines a loss function, typical loss functions do not satisfy the triangle inequality. It would be possible to re-define mean-squared loss by taking square roots, though such transformations may be much harder to achieve for other usual loss functions. Restricting loss functions to pseudo-metrics or to monotonous functions of pseudo-metrics appears to be unnecessarily restrictive.

In the literature, there is some disagreement on the importance of parameter uncertainty in forecast evaluation. For example, CLEMENTS AND HENDRY (1995) opine that parameter uncertainty has only moderate effects in the presence of correct specification. Let us re-consider the typical calculations in support of this view, for the case of the least-squares autoregressive predictor and a data trajectory from an autoregressive process. Mean squared loss is

$$\begin{aligned} & \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \left( \hat{\phi}_t x_t - \phi x_t - \varepsilon_t \right)^2 \\ = & \lim_{n \rightarrow \infty} n^{-1} \left\{ \sum_{t=1}^n \left( \hat{\phi}_t - \phi \right)^2 x_t^2 + \sum_{t=1}^n \varepsilon_t^2 + 2 \sum_{t=1}^n \left( \hat{\phi}_t - \phi \right) x_t \varepsilon_t \right\} \end{aligned}$$

The third term is small and converges to 0, while the second term converges to  $\sigma^2$ . For large  $t$ , the term  $\hat{\phi}_t - \phi$  can be approximated by  $t^{-1/2} w_t$  with  $w_t \sim N(0, 1 - \phi^2)$  because of root consistency of the least-squares estimator.

Therefore, one may write

$$n^{-1} \sum_{t=1}^n (\hat{\phi}_t - \phi)^2 x_t^2 = n^{-1} \left\{ \sum_{t=1}^N (\hat{\phi}_t - \phi)^2 x_t^2 + \sum_{t=N+1}^n t^{-1} (w_t^2 x_t^2 + o_p(1)) \right\}$$

As  $n \rightarrow \infty$ , the first term and the remainder term disappear, while the second term behaves like  $n^{-1}\sigma^2(C + \ln n)$ , where  $C$  is the sum of Euler's constant and a correction for the starting index  $N$ . These calculations imply that mean squared loss will be  $\sigma^2$ , as if  $\phi$  were known. However, note that the argument does not apply any more if discounted loss functions are used. The influence of the second term will not disappear asymptotically, and the first term will keep loss substantially above the benchmark loss for known parameters.

## 2.2 Loss functions suggested in practice

In recent work, CHEN AND YANG (2004) attempt to optimize loss functions for prediction accuracy assessments. To this aim, they survey specifications that are in current usage. Apart from the classical MAE and MSE criteria, they consider the MAPE (mean absolute percentage error)

$$n^{-1} \sum_{t=1}^n \frac{|\hat{x}_t - x_t|}{|x_t|}.$$

This MAPE is described completely by single-observation loss  $g(x, y) = |x/y - 1|$ , which is asymmetric. Asymmetry is a first disadvantage and implies that  $x$  viewed as a forecast of  $y$  is treated differently from  $y$  as a forecast of  $x$ . Another one is the fact that  $g(x, y) \rightarrow \infty$  as  $y \rightarrow 0$  and thus requires infinite predictive accuracy for small true values. The popularity of MAPE stems from a concern for scale invariance in the sense that predicting  $y$  by  $x$  should be assessed equivalently to predicting  $\lambda y$  by  $\lambda x$  for any  $\lambda \neq 0$ . There seems to be little reason for such a requirement within a single series. Scale invariance may make more sense for a comparison across a set of prediction experiments for various similar variables. In that case, however, one could simply standardize the MSE or MAE values after conducting the experiment, although it is less obvious whether scaling by dispersion is adequate. One could also argue that the assessment of forecasting algorithms should account for the degree of predictability.

Another criterion is the symmetric MAPE (sMAPE) that was suggested by MAKRIDAKIS AND HIBON (2000)

$$n^{-1} \sum_{t=1}^n \frac{|\hat{x}_t - x_t|}{(|\hat{x}_t| + |x_t|) / 2}.$$

The sMAPE is based on a symmetric single-observation loss  $g(x, y) = |x - y| / (|x| + |y|)$ . From MAPE, it inherits the divergence as  $|x| + |y| \rightarrow 0$ , while it is still defined if only one of the arguments approaches 0. In order to cope with this problem, CHEN AND YANG suggest adding a positive number to the denominator, for example a sequentially defined measure of dispersion

$$S(t) = (t-1)^{-1} \sum_{s=1}^{t-1} |x_s - \bar{x}_{s-1}|, \quad \bar{x}_s = s^{-1} \sum_{j=1}^s x_s,$$

which is defined for  $t > 2$ , while  $S(1)$  is undefined and  $S(2) = 0$ . This suggestion leads to an asymmetric loss function, which is at odds with the idea of sMAPE. The modified sMAPE is more complex than MAPE and sMAPE, as it cannot be described by a single-observation loss function alone. sMAPE and MAPE are monotonous in the sense of our basic definition.

In Table 1, we compare the four variants of absolute-error criteria by plotting the areas of equal distance to given data points  $(x_1, x_2) = (1, 3)$  and  $(x_1, x_2) = (-0.5, -1.5)$ . MAE yields the familiar diamond shapes of the  $L_1$  norm. MAPE yields shrinking diamonds as data approach 0. It also distorts the axes due to its inherent asymmetry. Thus, area size changes with distance to the origin, while asymmetry increases with the angle to the  $x = y$  median line. sMAPE and the modified sMAPE yield non-convex shapes, which only vaguely are reminiscent of the original diamonds. For the modified sMAPE,  $S(t)$  had to be redefined in order to allow a graphical representation. While both variants look similar for the selected data points, differences increase as the variation of the data increases. The modified sMAPE tolerates large forecasting errors for volatile data sets.

Analogous modifications can also be conducted for  $L_2$  criteria. CHEN AND YANG consider an NMSE (normalized MSE)

$$\frac{\sum (\hat{x}_t - x_t)^2}{\sum (x_t - \bar{x})^2}.$$

Like MAPE, NMSE is invariant to the scaling of the true data and it is asymmetric. Unlike the modified MAPE criteria, NMSE is not ‘causal’ in

the sense that the denominator uses the whole data up to  $n$ . While we made no assumption concerning such a causal structure of loss functions, NMSE has the drawback of being non-monotonous. A simple example demonstrates that NMSE can increase if an observation  $x_t$  moves closer to its prediction  $\hat{x}_t$ . Table 2 shows areas of equal distance for true data points  $(1, 3)$ ,  $(-1, 3)$ , and  $(-0.5, -1.5)$ . The radii of the circles increase with the distance from the origin and also with dispersion. The data point  $(-1, 3)$  is more ‘volatile’ than  $(1, 3)$  and causes more tolerance with regard to its prediction. A slight modification of the denominator would yield a symmetric loss function that would correspond to our basic definitions.

**Example.** Suppose  $n = 2$  and  $(\hat{x}_1, \hat{x}_2) = (1, 3)$  while  $(x_1, x_2) = (\xi, 5)$ . The NMSE takes its minimum for  $\xi = 0$ , not for  $\xi = 1$ , and increases for  $\xi \in (0, 1)$ .  $\square$

**Discussion.** Because many loss functions in current usage do not obey our definition, one could consider relaxing assumptions.

1. *Symmetry* follows the idea that prediction is viewed as approximation and that therefore  $x$  as a forecast of  $y$  should be treated equivalently to  $y$  as a forecast of  $x$ . If loss is determined by cost arguments, it makes sense to distinguish the costs of under- and over-prediction. We note that none of the above loss functions follows from cost arguments. Their asymmetry is a side effect rather than a deliberate specification.
2. *Monotonicity* is a logical requirement. If a forecast ‘improves’ as the true observation moves away from it, the loss concept should be revised.
3. Criteria that are not defined in important areas of the sample space and therefore are not functions  $\mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$  cannot be recommended. This applies to most variants of the MAPE concept.

It appears that arguments of logical consistency support restricting focus to loss functions within the limits of our definition. This supports the usage of traditional loss functions  $G_1$  and  $G_2$ , while robust and weighted variants may also deserve attention.

### 3 Stochastic prediction

Forecasters can choose among a variety of different predictors. While in theory loss functions may be determined by economic cost arguments, such

externally determined loss functions are rare in empirical applications. An exception is the field of empirical finance and portfolio selection, where loss may be determined by asset returns, transaction costs, and risk premia. This field is of considerable importance but it is not typical. In most applications of forecasting, including macroeconomic forecasts, meteorological and hydrological forecasts as well as predictions of election results, the forecaster is free to choose a loss function as well as a predictor. Often, different loss functions are considered in the same exercise, such as mean absolute *and* mean squared error functions.

An alternative to pre-specified predictors, such as  $\varphi(x) = \phi x$  for given  $\phi$ , or estimation-based predictors, such as  $\varphi(x) = \left(\hat{\phi}_j x_j\right)_{j \in \mathbb{N}}$ , are stochastic predictors. The simplest form of a stochastic predictor is a randomized version of an estimation-based predictor, such as  $\varphi(x) = \left(\hat{\phi}_j x_j + \varepsilon_j\right)_{j \in \mathbb{N}}$ , where  $\varepsilon_j$  is independently drawn from a distribution. Parametric randomization relies on convenient error laws, such as the normal distribution  $N(0, \hat{\sigma}_j^2)$  with  $\hat{\sigma}_j^2$  determined from the available data  $\{x_1, \dots, x_j\}$  by

$$\hat{\sigma}_j^2 = j^{-1} \sum_{k=1}^j \left(x_{k+1} - \hat{\phi}_k x_k\right)^2.$$

Non-parametric randomization relies on draws from empirical distributions, such as a uniform distribution on

$$\left\{x_{k+1} - \hat{\phi}_j x_k, k = 1, \dots, j\right\}.$$

Even when  $x$  is indeed a trajectory from the stochastic process that underlies the estimation method—in this backdrop case, an autoregressive process with coefficient  $\phi$ —such randomizations are unable to achieve any improvement in accuracy as measured by the loss function. Their advantage may be the more realistic visual approximation of some properties of the sequence  $x$ . For example, a sequence of predictions for varying step size  $h$  starting from any time point  $t$  looks like an asymptote for deterministic prediction, while it looks like the remainder of  $x$  for stochastic prediction.

Stochastic prediction becomes more attractive if a large number of trajectories are drawn from the parametric or non-parametric distribution. This defines  $\varphi(x)$  as a random sequence and allows to evaluate its moments, which may be used, in turn, as a basis for accuracy assessments.

**Definition 3** A stochastic predictor is defined as a random function  $\varphi$

$$\begin{aligned} & \mathbb{R}^{\mathbb{N}} \times \Omega \rightarrow \mathbb{R}^{\mathbb{N}} \\ (x, \omega) & \mapsto \varphi(x, \omega) = \zeta(\kappa(x), \delta(\omega)) \end{aligned}$$

with the property that  $\kappa : \mathbb{R}^{\mathbb{N}} \rightarrow \mathbb{R}^{\mathbb{N}}$  is a predictor and  $\delta : \Omega \rightarrow \mathbb{R}^{\mathbb{N}}$  defines a stochastic process. Moreover, the distribution of the finite sub-sequence  $(y_1, \dots, y_m)$  of  $y = \varphi(x, \cdot)$  is completely determined from the finite sub-sequence  $(x_1, \dots, x_m)$ . The function  $\zeta : \mathbb{R}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}} \rightarrow \mathbb{R}^{\mathbb{N}}$  is a time-constant continuous linking function, which can be equivalently seen as  $\zeta : \mathbb{R}^2 \rightarrow \mathbb{R}$ . Often,  $\zeta$  will just be a binary operator. The function  $\kappa(\cdot)$  determines the skeleton of the stochastic model and corresponds to the function  $\varphi(\cdot)$  for non-stochastic predictors. Often, the stochastic process  $\delta : \Omega \rightarrow \mathbb{R}^{\mathbb{N}}$  will be specified as random noise.

The expression ‘skeleton’ for the non-stochastic part of a dynamic model is due to TONG (1990). The conditions can be motivated as follows. Violation of the predictor property of  $\kappa(\cdot)$  means using unknown (future) parts of the sequence  $x$  for forecasts, maybe even in a deterministic way (the ‘perfect forecast’ of economic theory). Violation of the stated property of  $\delta(\cdot)$  would allow, e.g., to estimate the error variance for predicting  $x_k$  from future (unknown) observations. Its statement is deliberately vague and does not exclude knowing that variance *a priori*. Note that the distribution of  $(y_1, \dots, y_m)$  must be determined from the *observations* and not from an unknown stochastic process that may have generated  $x$ . All stochastic constructions are under the forecaster’s control and there is no unknown stochastic component here. The only unknown object is the remainder of the non-stochastic real sequence  $x$ .

A typical example is an autoregressive forecasting model with unknown  $\phi$  and  $\sigma^2$ . Then,  $\zeta(x, y) = x + y$ ,  $\kappa(x) = (\hat{\phi}_1 x_1, \hat{\phi}_2 x_2, \dots)$ , and  $\delta(\omega)$  is an infinite trajectory of normal random variables with mean zero and time-changing variance  $\hat{\sigma}_j^2$  that is ‘estimated’ from  $(x_1, \dots, x_j)$  according to any of the rules described before. In this case, the meaning of the skeleton is clear, as the  $m$ -th element of  $\kappa(x)$  is a center of the conditional distribution of  $(y_m | x_1, \dots, x_m)$ . This is not so for non-linear prediction models.

Like deterministic predictions, also stochastic predictions may be subjected to an evaluation criterion. For single trajectories, the framework of the last section can be used. For example,  $x$  may be a trajectory from a

random walk with known variance and independent increments  $\nu_t$ . A deterministic predictor aiming at single-step forecasting could be just  $\varphi(x) = x$ . The quadratic loss is  $G_2(\varphi(x), Fx) = G_2(x, Fx) = \sigma^2$ . A stochastic predictor may be  $\varphi(x, \omega) = x + \varepsilon$ , where  $\varepsilon_t \sim NID(0, \sigma^2)$ . Quadratic loss is

$$\begin{aligned} G_2(\varphi(x, \omega), Fx) &= \lim n^{-1} \sum_{j=1}^n (x_j + \varepsilon_j - x_j - \nu_j)^2 \\ &= \lim n^{-1} \sum_{j=1}^n (\varepsilon_j - \nu_j)^2 = 2\sigma^2, \end{aligned}$$

which is twice the loss of deterministic prediction. If more trajectories are generated, it seems natural to focus on the limit of

$$\lim_{K \rightarrow \infty} K^{-1} \sum_{k=1}^K G(\varphi(x, \omega_k), F^h x) = \mathbb{E} G(\varphi(x, \cdot), F^h x),$$

assuming that expectations are finite. If loss functions depend on single-observation loss in a linear way, one obtains the simple expression

$$\mathbb{E} G(\varphi(x, \cdot), F^h x) = \lim_K \lim_n K^{-1} n^{-1} \sum_{k=1}^K \sum_{j=1}^n g(y_j + \varepsilon_j(\omega_k), x_{j+h}),$$

where  $y_j$  is the prediction according to the skeleton. Note that this criterion is usually different from

$$G(\mathbb{E} \varphi(x, \cdot), F^h x),$$

i.e. the loss of the mean predictor, and also from

$$G(\kappa(x), F^h x),$$

the loss of the skeleton predictor. For example, if a random walk is predicted at single steps using a random-walk model, both the mean predictor and the skeleton predictor incur a loss of  $\sigma^2$ , while the expected loss of the stochastic predictor is  $2\sigma^2$ .

**Definition 4** *The expected loss of a stochastic predictor for  $h$ -step prediction is defined as  $\mathbb{E} G(\varphi(x, \cdot), F^h x)$ , where  $G(\cdot, \cdot)$  is a loss function and*

$\varphi(\cdot, \cdot)$  is a stochastic predictor. Expectation is taken with respect to the distribution of  $\varphi(x, \cdot)$ . The mean predictor loss is defined as  $G(\mathbb{E} \varphi(x, \cdot), F^h x)$ . The skeleton predictor loss is defined as  $G(\kappa(x), F^h x)$ , where  $\kappa(x)$  is the skeleton.

These three loss concepts should be distinguished carefully. The *expected loss* describes the average distance of generated stochastic-predictor trajectories from the true and given data  $x$ . The *mean-predictor loss* evaluates the quality of the mean of the stochastic predictor distribution as a point forecast for  $x$ . The *skeleton-predictor loss* reduces the stochastic predictor to its core and evaluates the quality of that prediction. In empirical applications, the skeleton-predictor loss is the easiest to construct, while it usually may be the least reliable version. Theoretical guidelines may recommend to rule out skeleton prediction as a useful concept. This recommendation, however, may be too severe for prediction models that are dominated by their linear component. For linear prediction models, the mean predictor and the skeleton predictor coincide. It appears more important to focus on the distinction of expected loss and the mean-predictor loss.

**Example.** If the data  $x$  is a trajectory from a random walk with incremental variance  $\sigma^2$  and the prediction model specifies the incremental variance at  $\tau^2$ , the expected squared loss at single steps is  $\sigma^2 + \tau^2$ . A small  $\tau^2$  reduces expected loss but gives unrealistic scenarios. A large  $\tau^2$  increases expected loss, while mean-predictor loss remains at  $\sigma^2$ . Similarly, if  $x$  is really a trajectory from the autoregressive process  $x_t = \phi x_{t-1} + \nu_t$ , expected loss becomes

$$\begin{aligned} & \lim_{K \rightarrow \infty} \lim_{n \rightarrow \infty} K^{-1} n^{-1} \sum_{k=1}^K \sum_{j=1}^n (x_j + \varepsilon_j(\omega_k) - \phi x_j - \nu_{j+1})^2 \\ &= \tau^2 + (1 - \phi)^2 \mathbb{E} x_t^2 + \sigma^2 = \tau^2 + \frac{2}{1 + \phi} \sigma^2 \end{aligned}$$

and a ‘correct’ specification of the incremental variance at  $\tau^2 = 2\sigma^2(1 + \phi)^{-1}$  yields again an expected loss of  $4\sigma^2(1 + \phi)^{-1}$ , twice the loss of the mean predictor. For  $h = 2$ , expected loss becomes

$$\begin{aligned} & \lim_{K \rightarrow \infty} \lim_{n \rightarrow \infty} K^{-1} n^{-1} \sum_{k=1}^K \sum_{j=1}^n (x_j + \varepsilon_j(\omega_k) + \varepsilon_{j+1}(\omega_k) - \phi^2 x_j - \phi \nu_{j+1} - \nu_{j+2})^2 \\ &= 2\tau^2 + (1 - \phi^2)^2 \mathbb{E} x_t^2 + (\phi^2 + 1) \sigma^2 = 2\tau^2 + 2\sigma^2. \end{aligned}$$

Clearly, the mean predictor loss is the second component  $2\sigma^2$ . This doubling property is typical for a wide range of models with linear errors  $\zeta(x, y) = x + y$ , including non-linear autoregressions.  $\square$

We do not consider the possibility of correlation of  $\varepsilon$  (noise in stochastic prediction) and  $\nu$  (noise in data generation), for its lack of intuitive basis. We recall that it was not generally assumed that  $x$  is a trajectory from a stochastic process or even from a time-homogeneous or stationary process. Therefore, objects like  $\mathbb{E} x_t^2$  do not exist in general. In some cases, it may be convenient to consider pseudo-moments, such as  $\lim n^{-1} \sum_{t=1}^n x_t = \lim \bar{x}(n)$  for the pseudo-mean and  $\lim n^{-1} \sum_{t=1}^n (x_t - \bar{x}(n))^2$  for the pseudo-variance. For example, suppose  $x = (1, -1, 1, -1, \dots)$ , then the pseudo-mean is 0 and the pseudo-variance is 1, even though the sequence has a clearly recognizable deterministic pattern. The predictor  $(0, 0, 0, \dots)$  with added  $N(0, 1)$  noise yields a mean-prediction loss of 1 and an expected loss of 2. The random-walk predictor  $\varphi(x) = x$  yields a mean-prediction loss of 4 and an expected loss of 8.

We summarize the doubling property as a theorem. The vector information of  $(x_1, \dots, x_t)$  will be denoted as  $X_1^t$ .

**Theorem 1** (*Doubling property*) *If the following conditions hold*

1. *The data  $x$  are a trajectory from a time-homogeneous process with additive errors.*
2. *The stochastic predictor is defined according to the data-generating time-series process, as the conditional expectation with added noise  $\varepsilon$ , where  $\mathbb{E} \varepsilon_t = 0$  and  $\mathbb{E} \varepsilon_t^2$  is the true variance of  $x_{t+h} - \mathbb{E}(x_{t+h} | X_1^t)$ .*
3. *The loss function  $G$  is defined by single-observation squared loss as  $G(x, y) = \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n (x_t - y_t)^2$*

*then the expected loss  $\mathbb{E} G(\varphi(x, \cdot), F^h x)$  is twice the mean-predictor loss  $G(\mathbb{E} \varphi(x, \cdot), F^h x)$ .*

Proof: If  $x$  is generated from a time-series process with additive errors, then it holds that

$$x_{t+1} = \mathbb{E}(x_{t+1} | X_1^t) + \nu_{t+1},$$

if  $X_1^t = \{x_1, \dots, x_t\}$  and, for general  $h$ , that

$$x_{t+h} = \mathbb{E}(x_{t+h}|X_1^t) + \sum_{j=1}^h \theta_j \nu_{t+j},$$

with time-constant weights  $\theta_j$  and white-noise ‘prediction’ errors  $\nu_t$ . These are the Wold-decomposition errors for stationary processes and are defined analogously for non-stationary time-homogeneous processes. Because of condition 2, prediction relies on white-noise random numbers  $\varepsilon_t$  with the same variance as  $\sum \theta_j \nu_{t+j}$ . In obvious notation, we use  $\varphi_t(X_1^t) + \varepsilon_t$  for the prediction of  $x_{t+h}$ , such that  $\varphi(x) = (\varphi_1(X_1^1) + \varepsilon_1, \varphi_2(X_1^2) + \varepsilon_2, \dots)$ . Because of condition 3,

$$\begin{aligned} \mathbb{E} G(\varphi(x, \cdot), F^h x) &= \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \mathbb{E}(\varphi_t(X_1^t) + \varepsilon_t - x_{t+h})^2 \\ &= \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \mathbb{E}\{\mathbb{E}(x_{t+h}|X_1^t) + \varepsilon_t - \mathbb{E}(x_{t+h}|X_1^t) - \sum_{j=1}^h \theta_j \nu_{t+j}\}^2 \\ &= \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \mathbb{E}\left(\varepsilon_t - \sum_{j=1}^h \theta_j \nu_{t+j}\right)^2 = 2 \mathbb{E} \varepsilon_t^2, \end{aligned}$$

while the mean-predictor loss is determined as

$$\begin{aligned} G(\mathbb{E} \varphi(x, \cdot), F^h x) &= \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \{\mathbb{E}(\mathbb{E}(x_{t+h}|X_1^t) + \varepsilon_t) - x_{t+h}\}^2 \\ &= \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \{\mathbb{E}(x_{t+h}|X_1^t) - x_{t+h}\}^2 \\ &= \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \left(\sum_{j=1}^h \theta_j \nu_{t+j}\right)^2 \\ &= \sum_{j=1}^h \theta_j^2 \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \nu_{t+j}^2 + \sum_{j=1}^h \sum_{k \neq j, k=1}^h \theta_j \theta_k \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \nu_{t+j} \nu_{t+k} \\ &= \sum_{j=1}^h \theta_j^2 \mathbb{E} \nu_t^2 = \mathbb{E} \varepsilon_t^2. \square \end{aligned}$$

While the result *per se* appears rather trivial, we note that the conditions cannot be relaxed substantially. If the predictor does not correspond to the generating model, either variance expressions will differ or a bias will invalidate the equivalence. As our examples show, the doubling property also holds for many ‘mis-specified’ predictors, as long as the variance of the simulated errors  $\varepsilon$  corresponds to the true one and the prediction error has zero unconditional expectation. If non-quadratic criteria are used, the adding-up property fails. Loss functions of the squared type with discounting sometimes retain the doubling property—such as different weights for odd and even observations, which may represent ‘winter’ and ‘summer’ in an application. More often they do not do so, as the ergodicity argument in the last part of the proof will not work, particularly if a finite ‘window of interest’ is allotted a non-zero weight. Even when the true model is substituted by a correctly specified model with a consistent estimator, the equivalence depends on the rate of convergence of the estimator and is not generally true. Also note that independence of the Wold-type errors  $\nu_t$  is not required. Independence holds if the process is linear and its errors are Gaussian.

Clearly, unconstrained minimization of expected loss does not make sense for stochastic predictors. A useful requirement may be to search for the minimum expected loss in a class of stochastic predictors that conform to the doubling-property rule. That condition is tuned to the case of squared loss, due to the equivalence of squared-error minimization and taking expectations, and has to be replaced for different loss functions. For squared loss, stochastic prediction with  $\mathbb{E}G(\varphi(x, \cdot), F^h x) > 2G(\mathbb{E}\varphi(x, \cdot), F^h x)$  displays *over-dispersion* in the sense that the prediction model is more volatile than the data. Conversely,  $\mathbb{E}G(\varphi(x, \cdot), F^h x) < 2G(\mathbb{E}\varphi(x, \cdot), F^h x)$  implies *under-dispersion* in the sense that the prediction model does not capture the variation of the data fully.

In practice, even for ‘correctly specified’ prediction models in the sense that the stochastic predictor uses a model class that contains the mechanism that generates  $x$ , model parameters are unknown and have to be estimated. Therefore, the doubling property will not hold exactly. In the more realistic situation that the prediction model does not contain the data-generating mechanism and such a mechanism may even fail to exist or at least to be time-varying, the rule may serve as a rough guideline.

It is worth while discussing whether stochastic prediction is a useful tool, as it appears that optimization of the predictor algorithm can be conducted on the basis of a single trajectory ( $n \rightarrow \infty$ ) and mean or deterministic pre-

diction. The point is that typical macroeconomic (and other) data samples are rather small and ‘ $n \rightarrow \infty$ ’ is a daring idealization. For fixed  $n$ ,  $K \rightarrow \infty$  reveals some of the properties of predictor loss, as long as  $x$  features properties of a time-homogeneous process trajectory. Note that working with fixed  $n$  violates strict monotonicity of the loss function  $G$ .

## 4 Double stochastic evaluation

Many existing forecasting evaluation studies explicitly or tacitly assume that the observed data  $x$  has been generated by a time-homogeneous stochastic process, i.e. either a stationary process or a non-stationary process that becomes stationary after differencing. Such a stochastic process, of which only one trajectory is observed for a finite time range, is then called a data-generating process (DGP). If the predictor is model-based and its model class contains the DGP, the prediction model is called ‘correctly specified’, otherwise it is called ‘mis-specified’. If the predictor is not model-based, this distinction does not make sense.

Although we suggest to adopt such an approach only with the utmost caution, it may be convenient to assume, for the sake of an experiment, that  $x$  indeed is the trajectory of such a DGP. Such exercises may be useful for answering the question, how well a model-based predictor would perform, if its model really were the DGP. One may compare the loss from such an exercise to the loss (or expected loss) from other experiments, such as entertaining rival predictors based on mis-specified models. If the rival predictor yields a comparable loss but wins the reverse horse race, which assumes that the model for the rival predictor were the DGP and again calculates the loss for both methods, usage of the rival predictor may be recommended for the sake of ‘robustness’.

In such double stochastic evaluations, a model is estimated from  $x$  using the whole available time range. This model is then assumed as the DGP. An expression such as

$$\begin{aligned} & \lim_{K_2 \rightarrow \infty} \lim_{K_1 \rightarrow \infty} K_2^{-1} K_1^{-1} \sum_{j=1}^{K_2} \sum_{k=1}^{K_1} G(\varphi(x(\omega_{2j}), \omega_{1k}), F^h x(\omega_{2j})) \\ &= \mathbb{E}_x \mathbb{E} G(\varphi(x, \cdot), F^h x) \end{aligned}$$

may serve as a criterion of accuracy. Here, two kinds of expectations are

taken: firstly, the expectation w.r.t. the generation law of the stochastic predictor; secondly, the expectation w.r.t. the generation law of the DGP. For ‘correct specification’, both laws are members of the same model class, while they differ for ‘mis-specification’. Even when an  $x$  trajectory of infinite length is available, however, the laws do not coincide, as  $x(\omega_2)$  is drawn from the true distribution in this case, while  $\varphi(x, \omega_1)$  stands for gradual approximations of that true distribution by estimates from finite segments of the trajectory. Note that computation time increases by a factor of  $K_2$  relative to regular stochastic prediction evaluation. For non-stochastic prediction, the experiment simplifies considerably and its computer time requirements are again comparable to stochastic prediction.

The examples introduced in the last section also apply to this case. Approximating expectations by averaging across random draws instead of across time is convenient if the sample is short. However, note that double stochastic evaluation is not constrained to the available sample size. Samples of large size can be generated artificially. Large-sample double stochastic prediction evaluations serve as guidelines for ‘asymptotic behavior’ for hypothetical cases when large samples could become available. Usually, consistent estimators for the predictors imply that the ‘true’ model, i.e. a match between the DGP and the predictor model, achieves the lowest (double) expected loss, though slow convergence toward true parameter values may counteract this standard result. In particular, misspecified models may beat correctly specified prediction for some non-linear models with poorly identified parameters and for discounted loss functions.

Double stochastic evaluation is not really a procedure that measures predictive accuracy for a given data sample. It rather is a simulation exercise that evaluates the relative accuracy of predictors against the background of an assumed true model class.

## **5 An empirical project: investment components in gross output**

### **5.1 The data**

We use a data set from the national accounts of the United Kingdom. Quarterly total fixed investment (or gross fixed capital formation, GFCF) is defined as the sum of investment in equipment and machinery, investment in

residential construction, investment in non-residential construction, and some minor positions. Additionally to these investment subaggregates, we use a time series of quarterly gross domestic product (GDP). All variables are at constant prices.

Data series were taken from the UK quarterly national accounts for the time range 1965:1 to 2002:3. Figure 1 shows the evolution over time of ratios of total GFCF and of some raw investment components to GDP. The investment components were aggregated to the two main components later. It is seen that the ratio of total GFCF over GDP has remained fairly stable over the whole time range, at around 17–18%. By contrast, the share of equipment investment has increased from less than 5% to around 8%, while the share of residential construction has fallen from 5–6% to less than 3% over the same time range. The three shown subaggregates do not sum to total investment. Some smaller components and discrepancies and a fourth major position of ‘transport equipment’ add to the overall increase in equipment investment.

For the prediction experiment, we simplify the breakdown of GFCF as follows. In order to keep the historical distinction of the two major components of investment (see, e.g., BERNDT, 1996), we form the two subaggregates ‘construction investment’ from the residential and non-residential series and summarize the remainder, i.e. the difference of total GFCF and construction investment, in a variable ‘non-construction investment’, which we will identify with ‘equipment investment’ in the following. The share of these two components in GDP output is shown in Figure 2. We denote the logarithms of construction and of non-construction investment by  $z_1$  and  $z_2$ , while  $Y$  denotes the logarithm of GDP. Trivariate models will be developed for  $X = (z_1, z_2, Y)'$ , while the variables in focus are the logarithmic component ratios  $z_1 - Y$ ,  $z_2 - Y$ , and  $z - Y$ , with  $\exp(z) = \exp(z_1) + \exp(z_2)$ . In the notation of the theoretical part of this paper, these variables define  $x$ .

For  $z_1 - Y$ ,  $z_2 - Y$ , and  $z - Y$ , some descriptive unit-root test statistics are summarized in Table 3. While for the Dickey-Fuller tests the lag order was determined by the AIC information criterion, a window length of 4 was generally used for the Phillips-Perron version of the test. In summary, unit roots are never formally rejected for any variable, although the share of total GFCF comes closer to a rejection than the shares of the subaggregates. This result is slightly at odds with visual impression and is likely due to the long swings and high volatility of the total share series. The result is confirmed by a multivariate cointegration test on the three variables. A search for the coin-

tegrating rank according to the JOHANSEN method (see JOHANSEN, 1995, for a detailed description) yields a rank of zero and hence no cointegrating vector. This excludes the possibility of self-cointegration and stationarity of any of the individual variables.

In summary, stationarity of the investment quota is not supported statistically, although one may wish to impose it for longer-run prediction (large  $h$ ), for reasons of plausibility. By contrast, stationarity of subaggregate quotas is unsupported by statistics as well as by plausibility.

## 5.2 Evaluations conditional on observed data

As candidates for stochastic forecasts, we use model-based predictions from five models:

1. an unrestricted VAR in differences  $\Phi(B)(\Delta X_t - \mu) = u_t$  with  $u_t \sim NID(0, \sigma^2)$ . Here,  $B$  denotes the backshift operator and  $\mu$  is the mean of the assumedly stationary  $\Delta X$ . This model serves as a convenient simple benchmark model. The degree of the lag polynomial  $\Phi(\cdot)$  is determined by the AIC information criterion. Parameters  $\Phi(\cdot)$ ,  $\mu$ ,  $\sigma^2$  are determined by least squares.
2. a non-linear error-correction model that is specified as

$$\begin{aligned} \Delta X_t = & \mu + \alpha [\ln \{ \exp(z_{1,t-1} - Y_{t-1}) + \exp(z_{2,t-1} - Y_{t-1}) \} - \delta] \\ & + \Gamma \Delta X_{t-1} + \varepsilon_t. \end{aligned} \tag{3}$$

This model implies difference stationarity of  $X$  and stationarity of the investment-output ratio, though it imposes no restriction on the component ratios  $\exp(z_1 - Y)$  and  $\exp(z_2 - Y)$ . Thus, it reflects prior information from theory and from empirical evidence. Lag orders beyond one were not considered. Some cursory residual analysis supported this decision. Note that, while additional sophistication may be required if modeling aims at retrieving precise parameter estimates, correct specification is not required for the definition of a forecasting tool. All free parameters are determined by least squares. (see ESCRIBANO AND MIRA, 2001, for general properties of non-linear error-correction models)

3. a non-linear error-correction model as (3) with the additional constraint that  $E \Delta X$  is a scalar vector. This variant avoids (or at least mitigates) the tendency of the freely estimated model that the share of one of the components approaches 0. While the long-run behavior of the component ratios is still left unrestricted, deterministic divergence is ruled out. This condition is called *growth homogeneity* in the following. Parameters are determined by a first-stage least-squares identification step and a secondary adjustment of the constant  $\mu$ .
4. a linear error-correction model with stationary subcomponent ratios. This model imposes stationarity on  $\Delta X$  as well as on  $z_1 - Y$  and on  $z_2 - Y$ .
5. a VAR in differences with growth homogeneity. This variant of the benchmark model avoids a particular feature of that model that may look most implausible in medium-run (large  $h$ ) predictions, though it still assumes no equilibrium relation for any combination (linear or non-linear) of the ‘level variable’  $X$ .

Models #1 to #4 can be viewed as assuming an increasing amount of restrictions on long-run behavior. Model #1, which is formally supported by the in-sample statistics reported in Table 3, assumes non-stationarity of  $z - Y$ ,  $z_1 - Y$ ,  $z_2 - Y$ . Model #2 has stationary  $z - Y$ , but non-stationary  $z_j - Y$ ,  $j = 1, 2$ . This is also true for model #3, which imposes a zero-drift restriction on the difference-stationary subcomponents ratios and is supported by plausibility considerations. Finally, model #4 assumes stationarity for all quotas. Model #5 restricts the deterministic part of model #1 but not the stochastic part. Thus, it is more restrictive than model #1 but less restrictive than model #3, while such ranking is not possible between models #2 and #3.

Each model was used to generate  $h$ -step predictions for the last 50 observations of the available sample and for  $1 \leq h \leq 40$ . In all error-correction models, instability was excluded by changing unstable influences of the error-correction terms to zero. This criterion was used separately at each time point, such that the experiment is out-of-sample in all regards. In terms of Section 2, all predictors conform to the cylinder-set condition. While various other models could be used for a comparison, note that it is not necessary to impose growth homogeneity on the linear cointegration model, as it is fulfilled automatically.

For the function  $g_2(x, y) = (x - y)^2$ , i.e. mean squared errors, skeleton prediction yields Figures 3–5. The graphs show loss as a function of  $h$ , where  $G$  corresponds to (2), with  $n$  stopped at  $n = 50$ . The three ratios are targeted separately, which we considered the most informative variant. Similarly, one may consider joint measures on the vector series  $X$  or on a range of horizons  $h \in \{1, \dots, h_{\max}\}$ . See CLEMENTS AND HENDRY (1998, Ch.3) for such suggestions. Stochastic mean prediction—all stochastic prediction experiments were conducted with 200 replications—yields similar results to the skeleton, even for the non-linear models #2 and #3. Results for expected loss in stochastic prediction are displayed in Figures 6–8. For  $g_1(x, y) = |x - y|$ , i.e. mean absolute errors, the ranking of forecasts is similar. The benchmark model #1 in differences without any further restriction clearly yields inferior forecasts. Contrary to the simulations of ENGLE AND YOO (1987), cointegrating models dominate at almost all horizons for all series, not only at larger horizons. Note, however, that we do not use the VAR in levels as a benchmark that was used by ENGLE AND YOO but in differences, and that we evaluate predictive accuracy for the (stationary or at least bounded) ratios and not for the assumedly integrated variables, such as  $Y$ .

Between the ‘deterministic’ and the ‘stochastic’ evaluations, the ranking of forecasting procedures is generally similar, with two exceptions.

Firstly, the benchmark model in differences *with* growth homogeneity shows a much stronger relative performance in the deterministic than in the stochastic evaluation. This puzzle may have a simple explanation. Growth homogeneity removes a major obstacle for acceptable performance of a model without long-run equilibrium conditions *on average*, while trajectories from the corresponding processes diverge. To quote a common application of cointegration, predicting income and consumption to grow at the same speed implies parallel behavior of the two variables for the conditional expectation but not for arbitrarily drawn trajectories. Stochastic prediction reveals the missing equilibrium condition.

Secondly, the linear cointegration model performs better in the stochastic experiment. For small  $h$ , linear cointegration achieves the best stochastic forecasts of the equipment quota. This observation corresponds to a widespread advantage of linear structures, which tend to generate reasonable trajectories at shorter horizons, while their point predictions are not optimal. At longer horizons, their performance deteriorates because of the possibly spurious equilibrium conditions that are imposed in model #4.

### 5.3 Evaluations conditional on simulated data

These evaluations assume that a specified model class is the correct one and determines the free parameters by estimation from the full available sample. From this estimated ‘pseudo-true’ model, artificial samples are generated (‘parametric bootstrap’), which are then ‘predicted’ using all of the previously specified methods. One of the methods corresponds to the class used for generating the data. These evaluations provide information on the relative merits with regard to the accuracy of forecasts from correctly specifying the model class. Because of sampling variation in parameter estimation, the true model class is not necessarily the best one at all forecast horizons.

Figures 9–11 rely on 200 replications both for the stochastic predictors and for the pseudo-true model. The assumed true model class is the nonlinear cointegration model with the growth homogeneity restriction (#3). Figures are drawn for mean absolute errors rather than  $g_2$ , to allow a more instructive visual separation of curves. For both criteria, the ranking is identical. Predictions based on the true model dominate at all horizons, while the ranking of the other predictors varies. For the equipment investment quota, the nonlinear model without the homogeneity restriction falls behind the linear error-correction model, while for the construction and total quotas, linear cointegration performs worse than the unrestricted nonlinear model. The primitive models without any error-correction restriction are worst for the construction and total quotas, whereas the differences VAR with growth homogeneity achieves a similar performance as the unrestricted nonlinear model for the equipment investment quota.

These simulations offer an informal test of whether the assumed model is a likely data-generating mechanism for the British data, even though such tests are not in the focus of our investigation. If a nonlinear error-correction model actually had generated the British investment data, Figures 9–11 should roughly match the features seen in the observational counterparts, Figures 6–8. While that correspondence is acceptable in general, there are some noteworthy differences. The empirical plots support the linear cointegration model as a forecasting tool at some prediction horizons, while this model is not among the preferred ones for the simulation graphs. This mismatch may indicate that true data behavior is ‘in between’ the linear and the nonlinear model, in the sense that the persistence of subcomponent quotas is stronger than would be implied by the nonlinear error-correction model, though not as strong as would be implied by the linear error-correction model. For both

data sets, prediction errors increase monotonously for the bootstrap version, while they deviate from monotonicity for the empirical version. This may indicate that longer-run cycles play a larger role in empirical data than in all suggested model classes. These longer-run cycles may reflect cycles in political attitudes, as particularly construction investment is severely influenced by policy decisions. Finally, the numerical values of mean absolute and mean squared errors show noteworthy differences, which however is to be expected due to sampling variation, if the data is viewed as a single observation of a trajectory from a time-series process.

The different criteria are summarized graphically in Figures 12–14 for prediction model #3, with model #3 assumed as the DGP for double stochastic prediction. Although the model is non-linear, there is hardly any visible difference of the skeleton and the mean forecast. Stochastic prediction increases the reported  $g$  values, though the correspondence to the doubling property is not precise. The relative effect of double stochastic simulation varies across series. For the total ratio and the equipment ratio, the expected loss from double stochastic simulation is *below* that from stochastic prediction. This may indicate some amount of incorrect specification in model #3 for the observed data. In other words, assuming the data as being generated from model #3 *simplifies* the prediction task artificially. Conversely, for the construction ratio, double stochastic evaluation yields larger loss than stochastic prediction. Generally, a mismatch of the true data-generating process and the forecasting model should not discourage the usage of the forecasting procedure, as long as no better prediction method has been found.

## 6 Summary and conclusion

The present paper is an attempt at forming a theoretical basis for measuring relative prediction accuracy in empirical applications. A general framework for such measurement has been defined for the cases of pure point prediction, stochastic prediction, and checking forecasting properties by parametric bootstrapping. The definitions have been applied in a practical macroeconomic forecasting experiment.

The suggested definitions are designed to make up for the lack of a rigorous foundation that is, e.g., mentioned by CHATFIELD (2000), who points to the uncertainty surrounding the statistical sampling model in the application of popular tests for predictive accuracy, such as the test by DIEBOLD AND

MARIANO (1995). We see this contribution as a beginning of more research in this direction, while we also think that our approach removes some of the problems that are mentioned in the literature. We also feel that the presented approach comes closer to the needs of empirical forecasters than some of the statistical approaches that are in current usage.

Choosing a prediction model from a set of candidates over a test range, in the hope that the ranking for larger  $n$  will correspond to the in-sample ranking, is a natural approach. Comparable procedures are commonly applied to various statistical problems in time series. For an example, we name the choice of bandwidth in nonparametric modeling, as it is suggested by FAN AND YAO (2003, p. 323). An alternative approach, which suggests to establish more sophisticated prediction models by gauging them against a given ‘primitive’ benchmark model on the basis of ‘significance’, appears to be less attractive, although it is often emphasized in the econometric literature.

Building on the framework of this paper, future research should address the empirical significance of recurrent features in the comparison among the prediction methods. The differences between the pure stochastic forecast and the double stochastic simulation are of particular interest. We conjecture that double stochastic simulation yields a variant of measuring the distance between processes. Double stochastic simulation ignores specific features of the sample in favor of the assumed parametric structure. In contrast, stochastic prediction fully reflects the sample-specific features of the given data set. The match or mismatch of the two approaches could add to the existing toolkit for developing forecast models.

## References

- [1] BERNDT, E. (1996) *The Practice of Econometrics: Classic and Contemporary*. Addison-Wesley.
- [2] CHATFIELD, C. (2000) *Time-Series Forecasting*. Chapman & Hall.
- [3] CHEN, Z., AND YANG, Y. (2004) ‘Assessing Forecast Accuracy Measures’, Working Paper, Iowa State University.
- [4] CLEMENTS, M.P., AND HENDRY, D.F. (1998) *Forecasting Economic Time Series*. Cambridge University Press.
- [5] DAVIDSON, J. (1994) *Stochastic Limit Theory*. Oxford University Press.

- [6] DIEBOLD, F.X., AND MARIANO, R.S. (1995) ‘Comparing Predictive Accuracy’ *Journal of Business and Economic Statistics* **13**, 253–263.
- [7] ENGLE, R.F., AND YOO, B.S. (1987) ‘Forecasting and Testing in Co-integrated Systems’, *Journal of Econometrics* **35**, 143–159.
- [8] ESCRIBANO, A., AND MIRA, S. (2001) ‘Nonlinear Error Correction Models’, Documento de Trabajo 2001-03, Universidad Carlos III de Madrid.
- [9] FAN, J, AND YAO, Q. (2003) *Nonlinear Time Series*. Springer-Verlag.
- [10] JOHANSEN, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.
- [11] MAKRIDAKIS, S., AND HIBON, M. (2000) ‘The M3 Competition: results, conclusions, and implications’ *International Journal of Forecasting* **16**, 451–476.
- [12] ROUSSEEUW, P. (1984) ‘Least median of squares regression’ *Journal of the American Statistical Association* **79**, 871–880.
- [13] ŠTULAJTER, F. (2002) *Predictions in Time Series Using Regression Models*. Springer-Verlag.
- [14] TONG, H. (1990) *Non-linear time series: a dynamical system approach*. Oxford University Press.

## Tables and figures

Table 1: Areas of same distance for MAE-type criteria. Curves give points at a distance of 0.2 and 0.4 for true data  $(1, 3)$  and  $(-0.5, -1.5)$ . From top left, criteria are MAE, MAPE, sMAPE, and modified sMAPE.

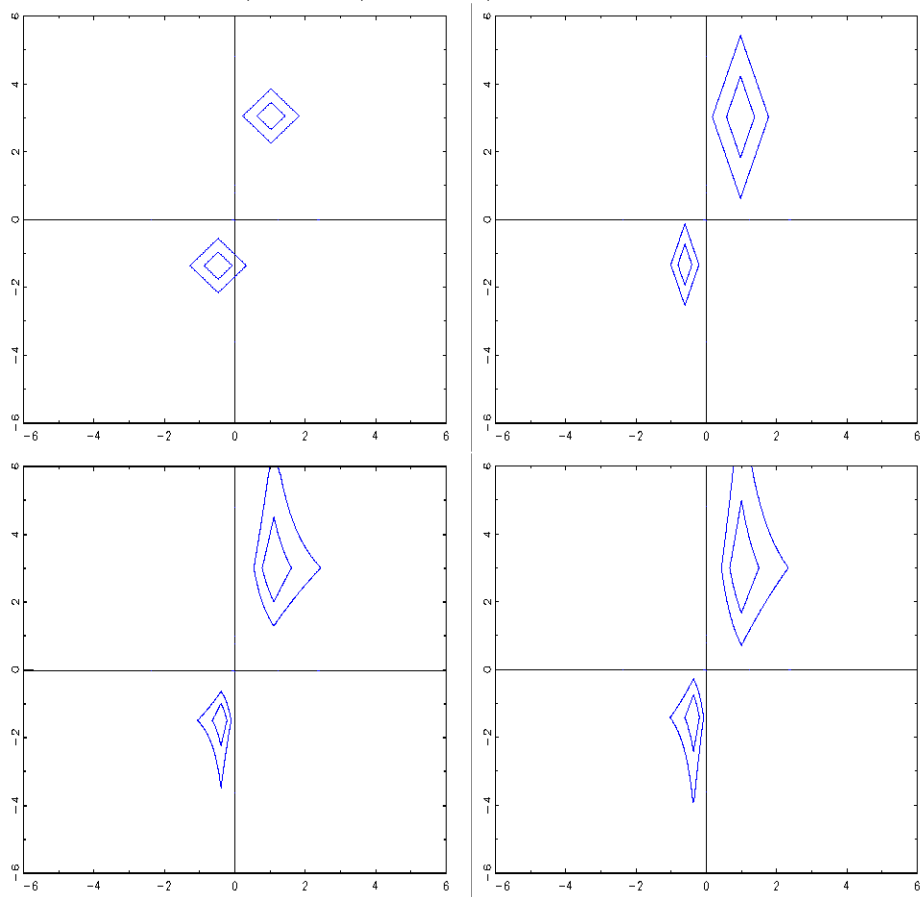


Table 2: Areas of same distance for MSE-type criteria. Curves give points at a distance of 2.5 for true data  $(1, 3)$ ,  $(-0.5, -1.5)$ , and  $(-1, 3)$ . Criteria are MSE and NMSE.

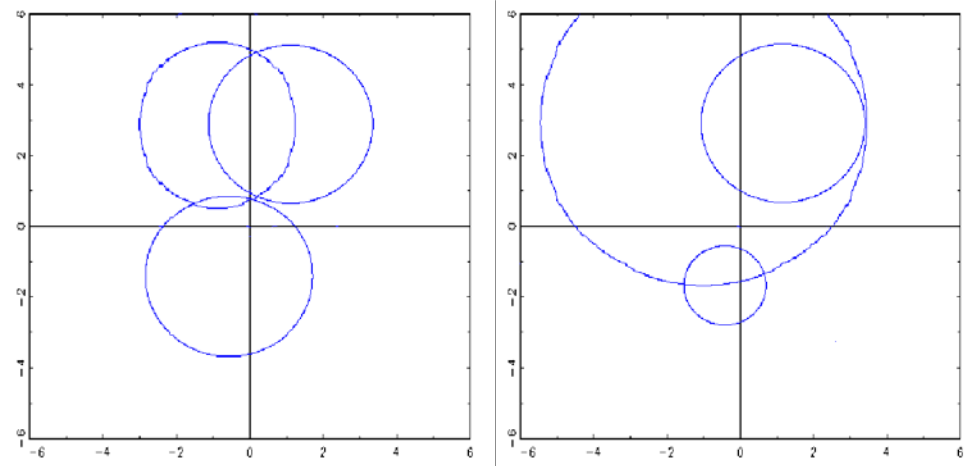


Table 3: Unit root tests on British series

	$\log(\text{IFC}/\text{GDP})$	$\log(\text{IFE}/\text{GDP})$	$\log(\text{IF}/\text{GDP})$
Dickey-Fuller tests			
augmenting lags	2	1	1
$\mu$ -statistics	-0.752	-1.234	-2.003
Phillips-Perron tests			
window length	4	4	4
statistics	-1.121	-1.406	-2.310

IFC is construction investment, IFE is equipment investment, IF is total fixed investment, GDP is gross domestic product.

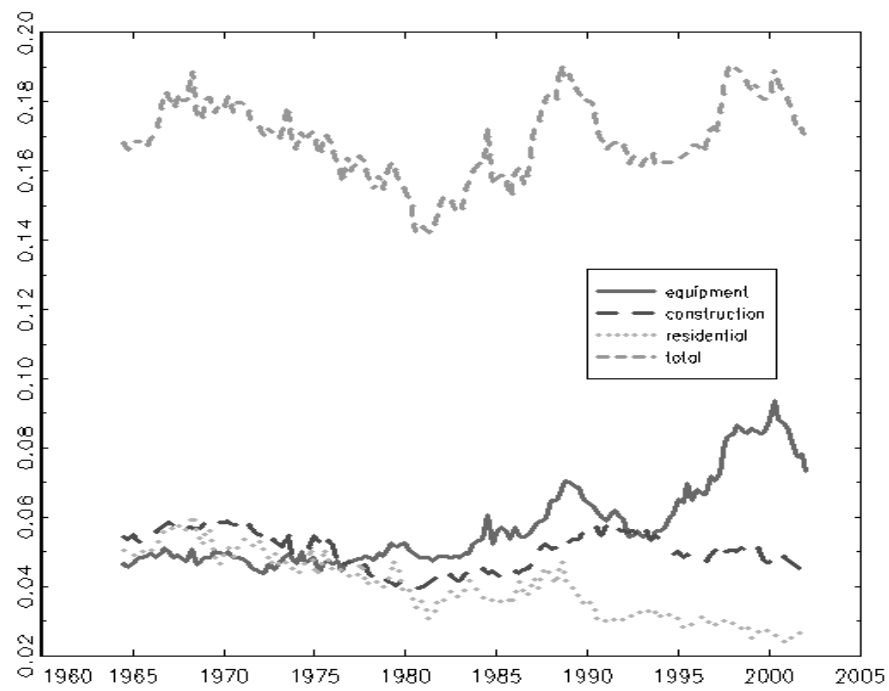


Figure 1: Shares of investment components in British GDP. Quarterly data 1965:1–2002:3.

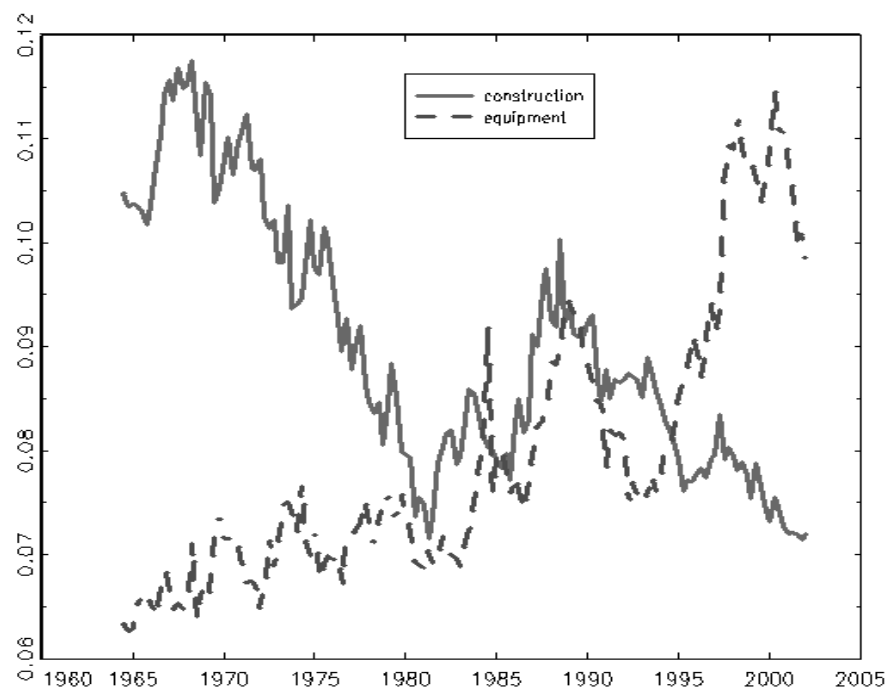


Figure 2: Shares of constructed investment components in British GDP. Here, “construction” comprises residential and non-residential construction, while “equipment” comprises all categories of total fixed investment excluding construction.

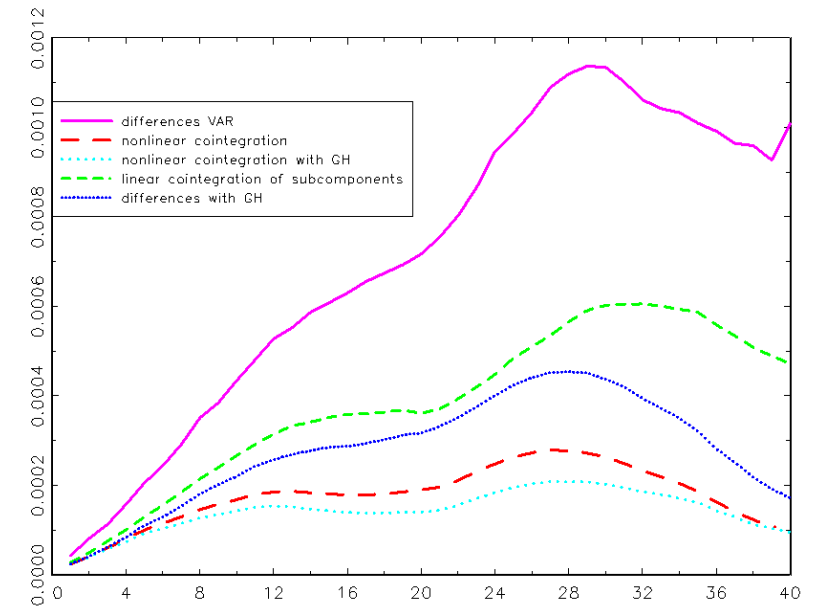


Figure 3: Mean squared error for the skeleton prediction for the total investment quota.

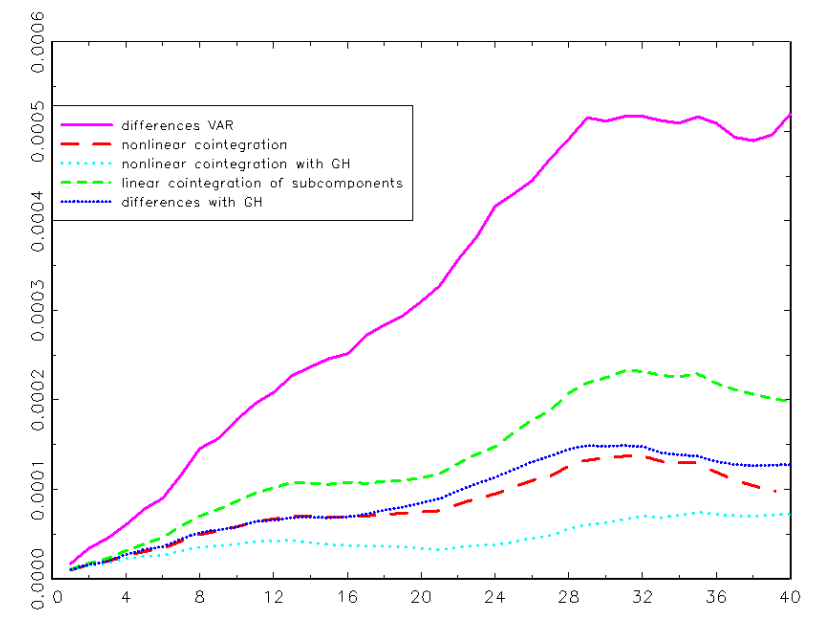


Figure 4: Mean squared error for the skeleton prediction for the construction investment quota.

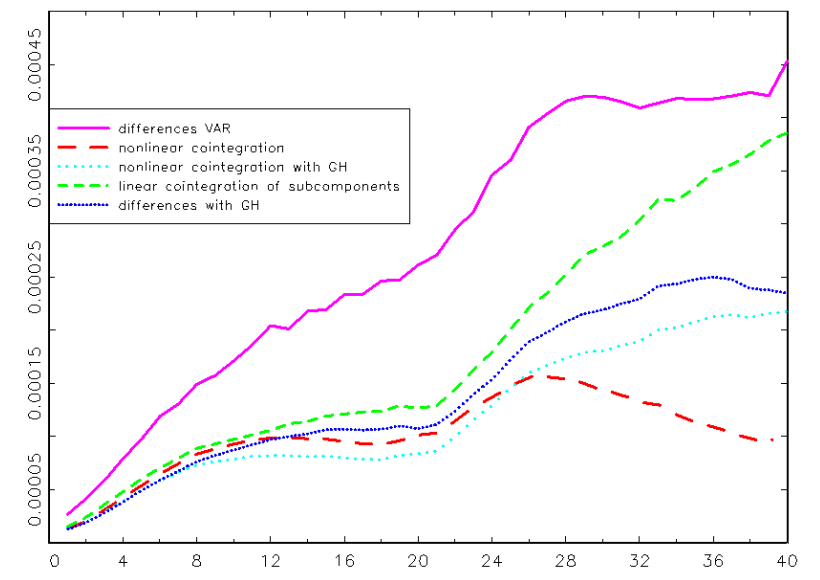


Figure 5: Mean squared error for the skeleton prediction for the equipment investment quota.

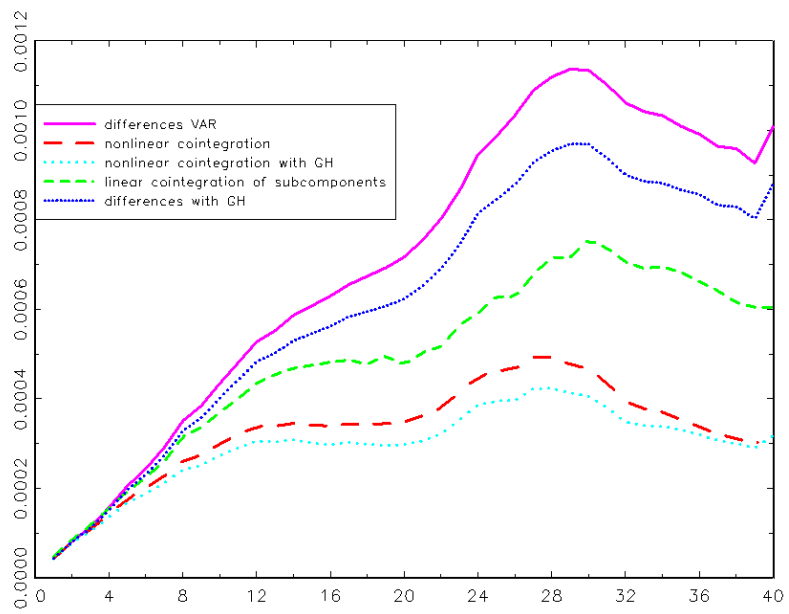


Figure 6: Expected mean squared errors for stochastic prediction, calculated from averaging across 200 replications. Predicted variable is the UK total investment quota.

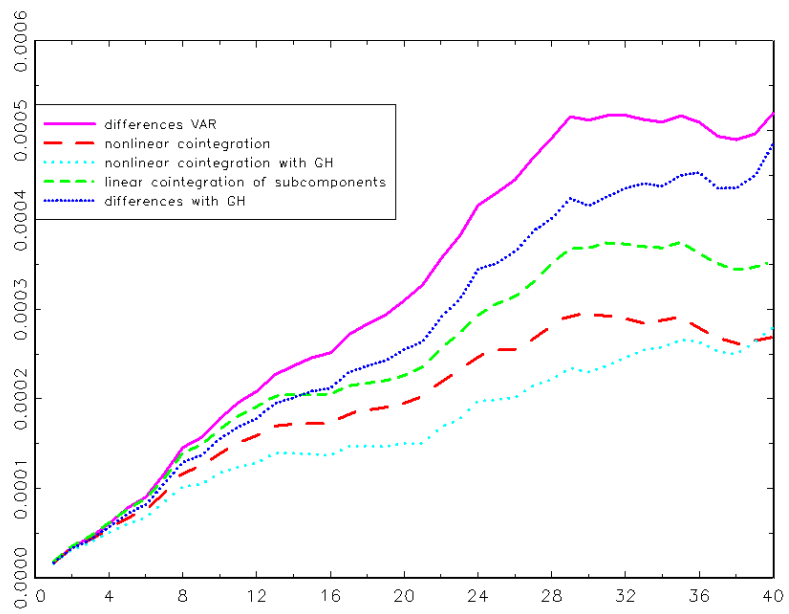


Figure 7: Expected mean squared errors for stochastic prediction, calculated from averaging across 200 replications. Predicted variable is the UK construction investment quota.

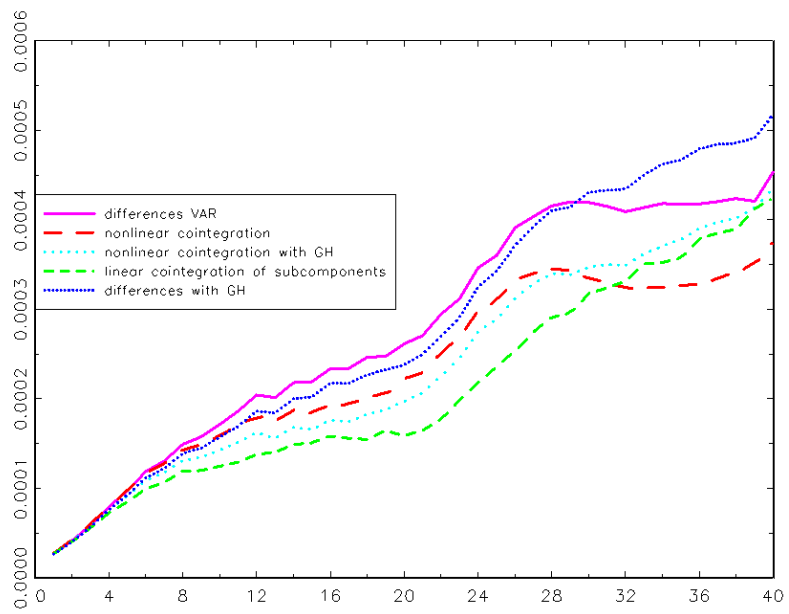


Figure 8: Expected mean squared errors for stochastic prediction, calculated from averaging across 200 replications. Predicted variable is the UK equipment investment quota.

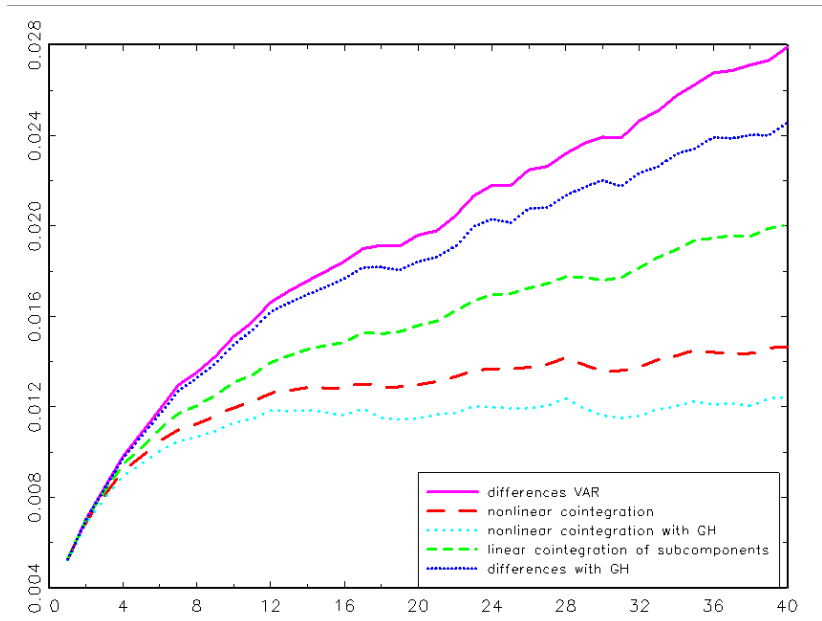


Figure 9: Expected mean absolute errors for double stochastic simulation based on 200 replications. Predicted variable is a parametric bootstrap version of the British total investment quota, assuming a non-linear error-correction model with growth homogeneity as the DGP.

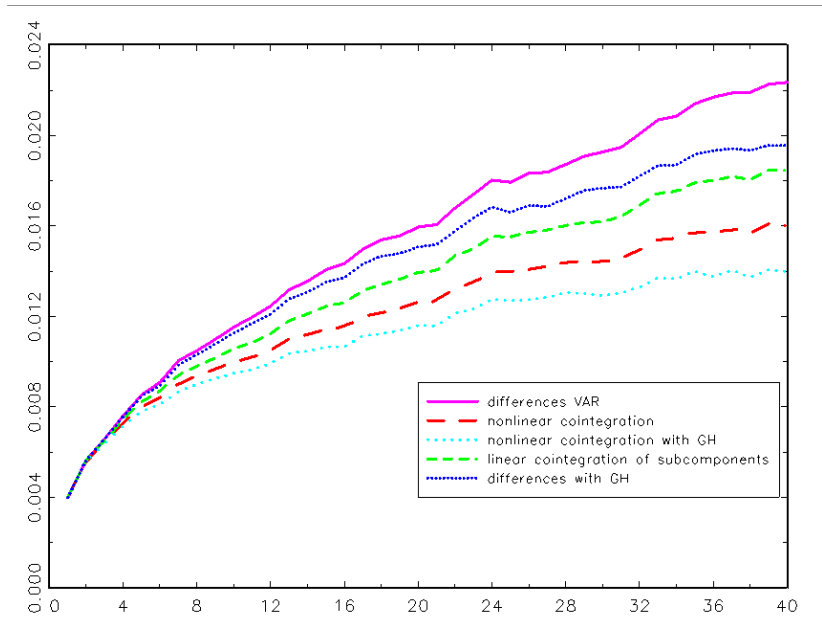


Figure 10: Expected mean absolute errors for double stochastic simulation based on 200 replications. Predicted variable is a parametric bootstrap version of the British construction investment quota, assuming a non-linear error-correction model with growth homogeneity as the DGP.

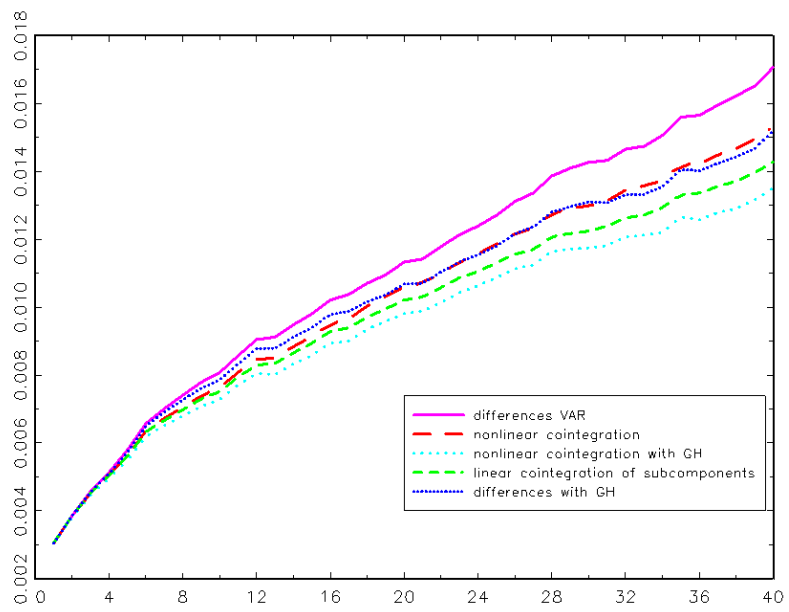


Figure 11: Expected mean absolute errors for double stochastic simulation based on 200 replications. Predicted variable is a parametric bootstrap version of the British equipment investment quota, assuming a non-linear error-correction model with growth homogeneity as the DGP.

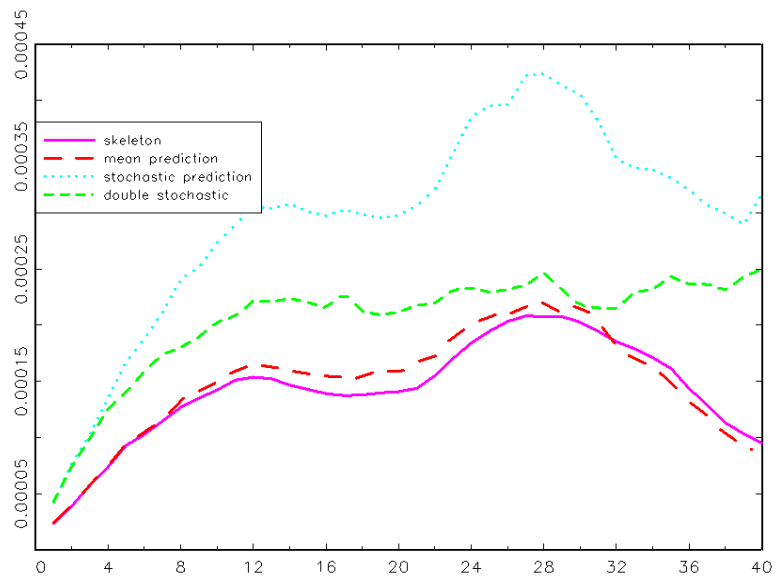


Figure 12: Effect of the stochastic nature of mean-squared error loss on forecasting performance for the total investment to output ratio. Predictor (and DGP for double stochastic evaluation) is based on the non-linear error-correction model with growth homogeneity.

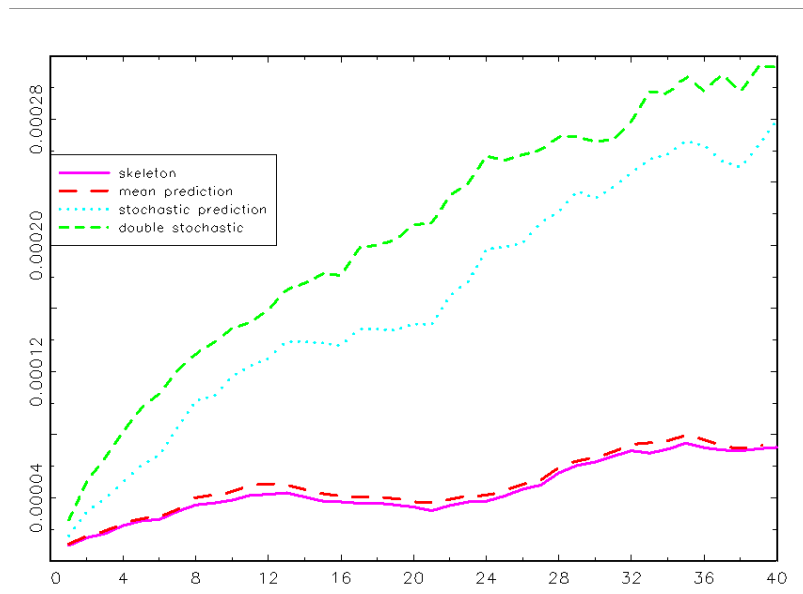


Figure 13: Effect of the stochastic nature of mean-squared error loss on forecasting performance for the construction investment to output ratio. Predictor (and DGP for double stochastic evaluation) is based on the non-linear error-correction model with growth homogeneity.

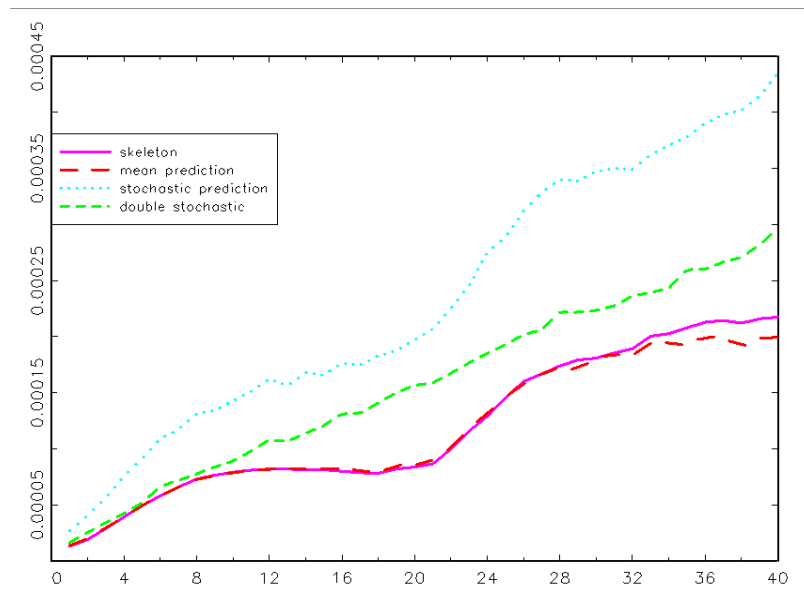


Figure 14: Effect of the stochastic nature of mean-squared error loss on forecasting performance for the equipment investment to output ratio. Predictor (and DGP for double stochastic evaluation) is based on the non-linear error-correction model with growth homogeneity.