IHS Economics Series
Working Paper 58
September 1998

# Unit Roots, Change, and Decision Bounds

Robert M. Kunst

Institut für Höhere Studien
Institute for Advanced Studies
Vienna

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Reihe Ökonomie / Economics Series No. 58

# Unit Roots, Change, and Decision Bounds

Robert M. Kunst

# Unit Roots, Change, and Decision Bounds

**Robert M. Kunst**

Reihe Ökonomie / Economics Series No. 58

**September 1998**

Robert M. Kunst
Abteilung Ökonomie
Institut für Höhere Studien
Stumpergasse 56, A-1060 Wien
Phone: +43/1/599-91-255
Fax: +43/1/599 91-163
e-mail: kunst@ihs.ac.at

**Institut für Höhere Studien (IHS), Wien**
**Institute for Advanced Studies, Vienna**

# Abstract

The problem of optimal decision among unit roots, trend stationarity, and trend stationarity with structural breaks is considered. Each class is represented by a hierarchically random process whose parameters are distributed in a non-informative way. The prior frequency for all three processes is the same. Observed trajectories are classified by two information condenser statistics $\zeta_1$ and $\zeta_2$. $\zeta_1$ is the traditional Dickey-Fuller *t*-test statistic that allows for a linear trend. $\zeta_2$ is a heuristic statistic that condenses information on structural breaks. Two loss functions are considered for determining decision contours within the $(\zeta_1, \zeta_2)$ space. Whereas quadratic discrete loss expresses the interest of a researcher attempting to find out the true model, prediction error loss expresses the interest of a forecaster who sees models as intermediate aims. For both loss functions and the empirically relevant sample sizes of $T = 50, 100, 150, 200$, optimal decision contours are established by means of Monte Carlo simulation.

**Keywords**

Time series, integrated processes, structural breaks, loss function

**JEL-Classifications**

C22, C44

**Comments**

The author wishes to thank Manfred Deistler, Philip Hans Franses, David Hendry, Waike Moos, and others for helpful comments. The usual proviso applies. This paper was written while the author was visiting the Center for Economic Studies (CES) in Munich whom the author wants to thank for their kind hospitality.

# Contents

# 1 Introduction

The contributions on tests for unit roots by FULLER (1996) and DICKEY AND FULLER (1979) have led to a persistent interest in integrated processes for the modeling of economic time series. Most econometric researchers have found evidence on unit roots in the main national accounts variables (cf., e.g., NELSON AND PLOSSER, 1982). These findings were particularly convenient as unit-root models allow identifying long-run equilibrium structures in multivariate dynamic systems in the context of cointegration. This would not be possible if economic variables were generated by stationary disturbances around time trends with a fixed pattern.

However, acceptance of the unit-root hypothesis and its consequences — persistence of shocks to variables and no return to growth trends, even in the long run — was not unanimous. In particular, the possibility of broken trend lines with added stationary processes ('structural breaks') was considered by some authors, such as PERRON (1989). If structural breaks are permitted in the deterministic part of the statistical models, the reported evidence on unit roots tends to weaken and stationary models with added structural breaks and deterministic trends (SBDT) appear to be preferred. However, this evidence in turn may not be convincing for a number of reasons.

Firstly, whereas it is fairly obvious that time series generated by an SBDT model lead to apparently spurious evidence on unit roots if structural breaks are not allowed in the applied statistical frame, the reverse specification error may also occur. A trajectory of finite length taken from a random walk or, more generally, an integrated process can be approximated with arbitrary accuracy by an SBDT model if the admitted number of breaks is increased. CHRISTIANO (1992) and ZIVOT AND ANDREWS (1992) have taken up a similar point, although they have been concerned primarily with the effects of estimating the timing of the break from the data. Also, these authors have remained strictly in the classical paradigm of 'null hypotheses', 'size distortion' and 'low power'. The SBDT model was generally seen as a variant of the trend-stationary model.

Secondly, in order to create a valid descriptive model for extended samples, an assumption on the break-generating mechanism in the trend would have to be imposed. Ironically, such assumptions typically lead to some sort of integrated model with infinite persistence of shocks to trend lines, although this persistence only concerns the supposedly rare trend shocks and excludes the non-persistent supposedly near-Gaussian innovations. Even those authors who use asymptotically reasonable models, such as CHEN AND TIAO (1990), see the (correct) classification of an SBDT process as a unit-root process as a 'misspecification'. This point has also been made by HANSEN (1992).

Thirdly, the effects of misclassifying variables on prediction should be evaluated, accounting for the fact that forecasts based on misclassified processes and finite samples may dominate those based on correctly classified processes, due to the sampling variation in estimated parameters. In short, the aims of correct specification and optimal prediction may be conflicting. This is mirrored in the theoretical debate between *realism* and *instrumentalism* (see POIRIER, 1995, p. 1).

This paper attempts to incorporate the classification problem in a multiple-decision framework. Decision bounds are calculated in order to optimize decisions on two discrete parameters of interest, the order of integration and the presence of structural breaks. The sensitivity of decision bounds with respect to the choice of the loss function is also evaluated. In particular, quadratic loss and minimum quadratic prediction error are considered as criterion functions.

A related approach was adopted by STOCK (1994) who, however, focuses on the binary decision between I(0) (which may include the trend-stationary case) and I(1) models, i.e., models integrated of order zero and one. STOCK (1994) concentrates on the asymptotic properties of his procedure and implicitly treats the SBDT case as I(0). In contrast, I focus on optimization in finite samples and treat SBDT as a separate class, although I am aware that it is really part of a slightly generalized I(2) class. Interestingly, the statistical properties of an obvious extension, I(1) models with structural breaks, are very similar to the SBDT models and hence this extension will not be treated separately here.

The remainder of this paper is organized as follows. Section 2 considers the three different models (classes of data-generation processes) among which a decision is searched. Section 3 highlights the distinction of several types of parameters that are used in the modeling approach. Section 4 considers the information-condenser statistics that are used for discriminating among the model classes. Section 5 introduces two different loss functions for assessing classification decisions. Section 6 presents and analyzes decision contour maps for the decision setup introduced in the previous sections. An illustrative example is also discussed. Section 7 concludes.

## 2 The data-generation processes

Three potential data-generating mechanisms are considered. The processes A, B, C are stochastic processes (models) that are seen as candidates for the true mechanism by the researcher who envisages the observed data as parts of trajectories from one of the three processes. The lag operator is denoted by $B$ and the first-difference operator by $\Delta = 1 - B$.

Process A is a real-valued unit-root process started in $t = 0$. Its difference-equation representation

$$\Delta X_t = \tilde{\mu} + \sum_{i=1}^{2} \varphi_i \Delta X_{t-i} + \varepsilon_t \tag{1}$$

is completed by a standard assumption on the innovation process

$$\varepsilon_t \quad \text{iid} \quad N(0, \sigma^2) \tag{2}$$

and by a weighting-prior assumption on the parameter space

$$\begin{aligned} \mu &= \tilde{\mu}/(1 - \varphi_1 - \varphi_2) \sim N(0,1) & (3) \\ (\varphi_1, \varphi_2)' &\sim U(S_2) & (4) \end{aligned}$$

with $\mu$ and $(\varphi_1, \varphi_2)$ mutually independent. Here, $U(A)$ denotes the uniform distribution on the set $A$ and $S_2$ is the stability region of the second-order difference equation coefficients within $\mathsf{R}^2$. It is known (see, e.g., Box et al. 1994) that $S_2$ is the triangular area $\{(\varphi_1, \varphi_2) | \varphi_1 + \varphi_2 < 1, -\varphi_1 + \varphi_2 < 1, \varphi_2 > -1\}$. One could also impose a weighting prior on the scaling parameter $\sigma$ but the decision problem is likely nearly invariant to scaling. In all reported simulations, $\sigma = 1$ is used for simplicity. Similarly, starting values $X_0, X_1, X_2$ are assumed as known to the researcher and are set at zero for the bounds simulations.

Notice the reparameterization from $\tilde{\mu}$ to $\mu$ for the drift constant. This is motivated by the solution of the difference equation (1)

$$X_t = (1 - \varphi_1 B - \varphi_2 B^2)^{-1} \sum_{s=1}^{t} \varepsilon_s + \mu t \tag{5}$$

The parameter $\mu$ is the slope of the added linear trend term and does not change its meaning if the short-run coefficients $\varphi_1, \varphi_2$ are varied.

Process B is a stationary autoregression with an added linear deterministic trend. Because it competes with the unit-root process A, it has a higher autoregressive lag order. The difference-equation representation of process B reads

$$X_t = \tilde{a} + \tilde{b}t + \sum_{i=1}^{3} \varphi_i X_{t-i} + \varepsilon_t \tag{6}$$

which again is completed by the standard assumption on the innovation process (2) and by a weighting-prior assumption on the parameter space

$$\begin{aligned} (a,b)' &\sim N(0, I_2) & (7) \\ (\varphi_1, \varphi_2, \varphi_3)' &\sim U(S_3) & (8) \end{aligned}$$

where $(a, b)$ are defined from $(\tilde{a}, \tilde{b})$ by $a = \tilde{a}/(1 - \sum_{i=1}^{3} \varphi_i)$, $b = \tilde{b}/(1 - \sum_{i=1}^{3} \varphi_i)$ as before. The symbol $I_n$ denotes an $n \times n$–identity matrix and $S_3$ is the stability region for third-order difference equations, which is displayed graphically as Figure 1. Unlike the simple shapes of $S_1 = (-1, 1)$ and $S_2$, the shape of $S_3$ is surprisingly complicated and it is not even convex. It is again conjectured that the decision problem is nearly invariant to changes in $\sigma$, and it is assumed that the starting values $X_0, X_1, X_2$ are known to the researcher. For simulating decision bounds, I set $\sigma = 1$ and $X_0 = X_1 = X_2 = 0$.

Again notice that solving (5) yields

$$X_t = (1 - \sum_{i=1}^{3} \varphi_i B)^{-1} \varepsilon_t + a + bt \tag{9}$$

and the meaning of $b$ is retained under variations of the coefficients $\varphi_i$. Equations (5) and (9) are the counterparts to the difference equations (1) and (6) and permit decompositions of the observed data into a stochastic component and a deterministic trend component. Thus, the models resemble the 'structural' models by HARVEY (1989) or the Bayesian time series models by WEST AND HARRISON (1989). These analogies even become more pronounced for the SBDT process C.

Process C is the most complicated stochastic mechanism, as it tries to incorporate change or 'structural breaks'. It is a modification of the trend-stationary process B

$$X_t = \tilde{a}_t + \tilde{b}_t t + \sum_{i=1}^{3} \varphi_i X_{t-i} + \varepsilon_t \tag{10}$$

again with the standard innovation assumption (2) and the same weighting prior on the autoregressive coefficients

$$(\varphi_1, \varphi_2, \varphi_3)' \sim U(S_3) \quad . \tag{11}$$

However, the normalized trend coefficients $(a,b)' = (1 - \sum \varphi_i)^{-1}(\tilde{a}, \tilde{b})'$ are not fixed in one trajectory but are themselves processes that obey

$$\begin{aligned} b_t &= b_{t-1} + \eta_t & (12) \\ a_t &= a_{t-1} - t\eta_t & (13) \end{aligned}$$

with $(\eta_t)$ n.i.d. $(0,1)$ such that the local trends form a linked chain sequence. In detail, the solution of (10)–(13) yields

$$\begin{aligned} X_t &= (1 - \sum_{i=1}^{3} \varphi_i B^i)^{-1} \varepsilon_t + t \sum_{s=1}^{t} \eta_t - \sum_{s=1}^{t} s\eta_t + a_0 + b_0 \\ &= \Phi^{-1}(B)\varepsilon_t + \sum_{s=1}^{t} (t-s)\eta_s + a_0 + b_0 & (14) \end{aligned}$$

with the invertible operator $\Phi(B) = 1 - \sum \varphi_i B^i$. It is convenient to assume a standard $N(0, I_2)$ weighting prior for the first draw $(a_0, b_0)$. Without special assumptions on the break-generating process $(\eta_t)$, process C is a unit-root process with two unit roots at 1, an 'I(2)' process similar to the integral of process A. Similar definitions of unit-root processes with added short-term variation and trends are common in the literature on so-called 'structural' models in the econometric sense of the word (cf. HARVEY, 1989). A characteristic property of the break model C, however, is the mixed distribution of the break process. In particular, I consider

$$\begin{aligned} \eta_t \ \text{iid} &\sim \lambda \mathbf{0} + (1-\lambda) N(0,1) & (15) \\ \lambda &\sim U([c,1]) & (16) \end{aligned}$$

The symbol **0** denotes a degenerate distribution with unit point mass at zero. A mixed distribution with weights of $\lambda$ on $p_1$ and $1-\lambda$ on $p_2$ is denoted by $\lambda p_1 + (1-\lambda)p_2$. It turns out that the obvious exhaustive choice $c = 0$ results in a too high frequency of non-zero realizations of $\eta_t$, which would not correspond to the idea of *rare* breaks. Hence, I set $c = 0.9$, which reflects the *a priori* idea that a probability of breaking higher than 0.1 is definitely outside the scope of those who think that the economy follows a linear trend with rare breaks.

Although the abbreviation SBDT is used for processes of type C in the following, one should be aware of the fact that the underlying broken trend $a_t + b_t t$ is not 'deterministic' in the sense that $f(\theta(\omega), t, \omega)$ would depend on $\omega$ only through $\theta(\omega)$. I hold the view that truly deterministic broken trends are difficult to imagine. If the breaking point is deterministic and fixed, additional sampling from the process at the end results in the pre-break parameters being transient. Keeping the break at $[T\nu]$ with fixed $\nu \in (0,1)$ results in the asymptotics of a sequence of processes rather than of processes, as the timing of the break changes with additional sampling.

Thus, the potential data-generating processes A, B, and C are uniquely defined, among which a decision is searched that is optimal in some sense. The decision maker is assumed to be informed about the generating mechanisms of candidates A and B but not about the parameter values within the candidate models that have to be approximated from partially observing a single trajectory. The decision maker is not fully informed about the generating law of process C and, in line with the literature, views this candidate as a sort of modification of B with rare breaks in the linear trend that occur approximately once in every 100 observations.

# 3 A classification of parameters

The parameters appearing in the definition of processes A, B, and C have different statistical properties. Most of them are denoted by Greek letters, such as $\mu$, $\lambda$, or $\varphi_1$. These *primary parameters* are typically defined on subsets of $\mathsf{R}$ and information on them accumulates even by observing a single process trajectory of increasing length, assuming they are 'identified'. Primary parameters are usually random and are drawn from distributions that exhaust the *data-admissible* region. The criterion of data admissibility (cf.

HENDRY, 1995) reflects *a priori* restrictions due to substance-matter theory.

A second type of parameters may be used to describe the weighting-prior distributions within parametric families of distributions. An example is the left corner point $c$ of the interval in (15). Information on such *hyperparameters* can only accumulate from observing an increasing number of trajectories. In econometrics, where frequently only one trajectory of the whole process is revealed by the data, the specification of these hyperparameters is determined by further *a priori* considerations. In contrast to the primary parameters, hyperparameters are typically fixed real numbers.

A third type of parameters is only implicit in the previous definitions. All three processes A, B, and C contain special assumptions on the length of their respective memory. The lag length — here 2 or 3 — is an example of a *discrete auxiliary parameter.* The nature of these auxiliary parameters is ambiguous. Because information about them may accumulate from observing a single trajectory, they may be viewed as a discrete version of primary parameters. They may be randomized on the basis of discrete distributions. However, they may also be viewed as fixed components of the *a priori* specified design of the model processes. Alternatively, they may be targeted by the decision problem.

A fourth type of parameters is also discrete but is right in the focus of the decision problem. These are the *secondary parameters* that are representative of the — usually finite — number of candidate processes (or, generally, models) among which a decision is desired. One may define the set of secondary parameters as $\{\mathrm{A},\mathrm{B},\mathrm{C}\}$ but it is often convenient to code the secondary parameters as $n$ –tuples of integer numbers. Like the primary parameters, the secondary parameters are really random. In the absence of a possible *a priori* inclination toward one of the considered processes, the distribution on the set of secondary parameters $\Xi$ will always be assumed as uniform discrete, as long as $\Xi$ is finite. The typical element of $\Xi$ will be denoted by $\kappa$. Using the equivalence of $\kappa$ and the process $X_t(\omega,\kappa)$, note that this view defines a super-structure by

$$\begin{aligned} X_t^F(\omega) &= X_t(\omega,\kappa) && (17) \\ \kappa &\sim U(\Xi) && (18) \end{aligned}$$

which is again a stochastic process. This process is the most general model (MGM) considered, it acts like a window on the world of data (cf. POIRIER,1995) and it will also be called the *frame* of the decision problem.

Notice that the words 'process' and 'model' are used interchangeably and many researchers may view each of the processes A, B, C as a 'class of processes with a prior distribution'. However, abandoning this distinction permits a comprehensive evaluation of the decision problem.

The point of view expressed by the MGM or the frame differs from traditional econometric inference, as it is seen, e.g., by WHITE (1996). Traditional inference views the data as being actually generated from an arbitrarily general probability distribution with possibly time-changing properties, which is also called the *Haavelmo distribution* in the British econometric literature (cf. SPANOS, 1986). However, nothing proves the existence of such a distribution and, even if it exists, nothing can be learned about it from the data without further restrictions, hence the concept appears pretty useless. SPANOS (1986), WHITE (1996), and others view inference as a problem of approximating the Haavelmo distribution by parameterized families of model distributions and of minimizing the Kullback-Leibler distance between the Haavelmo and the best model distribution. In contrast, the 'reality' that is considered here consists of the observed data only, whereas all models are products of the researcher's imagination whose angle of vision is summarized in the frame. All that can be learned from data is the relative validity of certain components of the frame with respect to whether hypothesizing them to be data-generating mechanisms is useful and plausible.

# 4 Condensing information

## 4.1 Principles of condensing information

It is a statistical principle to base decisions on the evidence in the data. Bayesian methodology may use the complete information in the data by way of 'on-line' integration and calculation of posterior densities. This procedure can be cumbersome and time-consuming but it is guaranteed not to miss any relevant information. If time series of real numbers are observed, the data is an element of $\mathsf{R}^T$. It is usually convenient to summarize the information in a vector of smaller dimension, with its elements specifically tuned to the aim of the decision problem. Such numbers that are functions of the data are *statistics*.

If statistics are *sufficient*, there is no loss in relevant information if the statistics are used in lieu of the original data. A Gaussian frame does not admit any information but first- and second-order moment estimates to be relevant for the decision problem, hence the maximum-likelihood estimates of all primary parameters are bijective functions of these moment estimates and form a vector of sufficient statistics (cf. GOURIEROUX AND MONFORT, 1995, Ch.3). If there are only a few parameters, these parameter estimates can be used immediately. For any given combination of parameter estimates, the expected loss can be evaluated and the decision corresponding to a minimum loss should be taken. A *map of optimal decisions* can be drawn, and a potential user who is given parameter estimates can arrive at the optimal decision at once using the map. These maps depend on the sample size $T$ and have to be re-drawn if $T$ increases from 100 to 120, say. In some decision problems, asymptotic approximations to true maps are helpful in larger samples.

If the number of parameters is high — and any number beyond 3 must be viewed as being 'high' — the usefulness of decision maps based on parameter estimates is limited. Therefore, in many statistical decision problems, nearly sufficient or asymptotically sufficient statistics are formed from the primary parameter estimates. Ratios of local likelihoods between pairs of (secondary) decision parameters define *likelihood-ratio* (LR) *statistics*, which are commonly used in classical statistics even for multiple decision problems. These LR statistics are typically not sufficient for the decision problem, hence the possible deteriorating influence of the remaining information must be evaluated carefully. The remaining information can be summarized conveniently in the estimates of *nuisance parameters*, which are usually defined as that subvector of the primary parameter vector that is irrelevant for the theoretical definition of the 'hypotheses'. Care must be taken in applying such definitions. Not all decision problems can be formulated in such a way that the primary parameter set for *all hypotheses jointly* $\Theta$ is some (subset of the) $\mathsf{R}^n$ and that the hypotheses can be defined by restrictions on some $\theta_i$ components. It is perhaps more natural to see $\Theta$ as the union of primary parameter sets $\Theta_i$, each of which may be isomorphic to some (subset of) $\mathbf{R}^{n_i}$ with the dimension varying over $i$. Seen this way, some problems of hypothesis testing, such as the 'presence of a nuisance parameter only under the alternative' (see ANDREWS AND PLOBERGER, 1994, and DAVIES, 1977, 1987), may be partly caused by the problem formulation that is often inspired by the instinctive urge of researchers to make $\Theta$ isomorphic to (an interval-type subset of) $\mathbf{R}^n$.

In many circumstances, the calculation of LR statistics is numerically demanding and heuristic approximations are used. These approximations may also be more 'robust' to modifications of the frame than exact LR statistics. For example, if decisions on the existence of change are aimed at, researchers use a wide variety of information condensers, such as the CUSUM, CUSUM–squared, MOSUM, Chow statistics with some variation in corresponding decision rules. For a good survey of several aspects of testing for change in econometrics, see HACKL AND WESTLUND (1991).

## 4.2 Condensing information on unit roots

A problem that was much in the focus of time-series statistics recently is reaching a decision between processes of type A and of type B. Difficulties arise as process B has more parameters than process A and hence A is not simply nested in B by imposing a unit root on the lag polynomial defined by $(\varphi_1, \varphi_2, \varphi_3)$. Some authors have suggested defining A 'in B' by the additional restriction $b = 0$, which defines $a = \mu$, whereas others have been concerned with the corresponding modification in the meaning of the parameters. The asymmetric treatment of 'null' and 'alternative' hypothesis in classical statistics has incited some authors to re-define the frame in order to reverse the order of hypotheses. Finally, some authors have developed attempts to see the problem in a decision-theoretic context. For this last point, see STOCK (1994) and HATANAKA (1996). See also the latter citation and the book by BANERJEE ET AL. (1993) for good surveys of the relevant literature.

The most commonly used statistic for reaching a decision between process A and B is the DF–$\tau$ statistic suggested by DICKEY AND FULLER (1979). It is derived from a regression

$$\Delta X_t = \tilde{a} + \tilde{b}t + \rho X_{t-1} + \sum_{i=1}^{p} \Delta X_{t-i} + u_t \qquad (19)$$

and calculating the $t$–value of the coefficient estimate $\hat{\rho}$. The lag length $p$ is designed to achieve approximate white-noise errors under both processes, hence in view of the defined MGM it appears natural to set $p = 3$. Although the term $\tilde{b}t$ is not present in process A, its inclusion is supported due to the better invariance properties of the distribution of DF–$\tau$ with respect to the remaining nuisance parameters, in particular $\tilde{a}$. The finite-sample behavior of DF–$\tau$ under process A has been analyzed thoroughly in the literature, and it has turned out that the (analytically tractable) asymptotic distribution approximates it satisfactorily. For process B, only local simulation results with fixed parameters are available, whereas for process C one may use the results of PERRON (1989) and those who built on his work as a guideline (see also ZIVOT AND ANDREWS, 1992)

Frequency histograms of the finite-sample distribution of $\zeta_1$ =DF–$\tau$ are given in Figure 2 for each of the processes A–C. The figures are based on 10,000 replications and $T = 100$. Evidently, the statistic $\zeta_1$ possesses satisfactory discriminatory power between the unit-root process A and the trend-stationary process B, as the probability mass overlapping between the two finite-sample distributions is rather small. For process C the distribution is similar to that of process A, although the dispersion is much higher. The higher dispersion results in a substantial overlap with process B, which represents those cases where the breaks are few and not very pronounced. In summary, however, process C is recognized as a unit-root process. However, this is not a weakness of the test, as process C *is* a second-order unit-root process.

If $T$ increases, the frequency plots remain virtually the same for processes A and C. The small-sample distributions of DF–$\tau$ under these two processes reach their asymptotic shape rather quickly. For process A, this is the Dickey-Fuller distribution, a mixture of Gaussian distributions that can also be expressed as the ratio of two Brownian motion integrals. For process C, it is another non-standard distribution that can also be expressed as a Brownian functional. For process B, the mode and the main mass of the distribution shifts leftward and discrimination against the other two processes becomes easier. To discriminate process C against processes A and B, another information condenser is needed.

## 4.3 Condensing information on change

Condensing information on change in order to discriminate process C from processes A and B is a more complex exercise than discriminating A and B. Following the idea of LR statistics, the ratio of the local likelihoods of B and C can be calculated, as C touches upon B as a borderline case. However, likelihood estimation of the primary parameters in C is costly. A good alternative information condenser can be calculated on the basis of the statistics proposed by PERRON (1989) who, however, assumes just one structural break. Here, a third alternative statistic will be used that builds on the Chow test and assumes that the potential number of breaks increases with increasing $T$.

A researcher who suspects that the data can be modeled more efficiently by breaking trends may split the sample in two or more parts and compare the resulting error variances of the unrestricted (split sample) and the restricted (one sample) estimation. For a fixed point of the sample split and least-squares regression, this yields the traditional Chow test (CHOW 1960). If the change point is unknown, the minimum of such comparison statistics over a grid of regressions may be chosen. A similar idea is known as the Quandt test in the econometric literature (QUANDT 1960). For classical statistics, the complexity of the resulting null distribution creates problems. This restriction is of no importance here, as decision bounds will be created on the basis of simulations anyway.

In detail, at first a 'restricted' regression

$$X_t = \tilde{a} + \tilde{b}t + \sum_{i=1}^{3} \varphi_i X_{t-i} + u_t \tag{20}$$

is run for $t = 4, \ldots, T$ and the resulting error variance estimate is denoted by $\hat{\sigma}_R^2$. Then, 'unrestricted' pairs of regressions are run for $t = 4, \ldots, T_1$ and $t = T_1 + 1, \ldots, T$ and calculate the resulting error variance estimates $\hat{\sigma}_U^2(T_1)$. The split point $T_1$ is varied between $\left[2\sqrt{T}\right] + 3$ and $T - \left[2\sqrt{T}\right]$. The symbol $[.]$ denotes the largest integer or entier function. Usage of the $2\sqrt{T}$ rule is due to HUBER and is common in time-series statistics. For example, if $T = 50$, $T_1$ is varied between 17 and 36. Notice that, for $T \to \infty$, the minimum sample size increases sublinearly. The procedure defines a convenient condenser by

$$\zeta_2 = \frac{\min \hat{\sigma}_U^2(T_1)}{\hat{\sigma}_R^2}$$

which will accompany $\zeta_1$ =DF$-\tau$ in the quest.

The process C admits a maximum of 10 break points and an average of 5 break points in the sample. Due to the Gaussian prior on the change process $\eta_t$, the majority of these breaks are small and one cannot expect more than

1 of changes that cross traditional statistical significance bounds. Therefore, only one break is admitted as long as $T \leq 150$. For $T \in (150, 250]$, two breaks at $T_1$, $T_2$ are allowed and one may proceed along similar lines for larger $T$. Because minimization has to be conducted over a growing number of regressions, the test may become time-consuming for large $T$ and one may prefer to calculate the minimum sequentially. For example, if $T = 200$, I first minimize over one break and two regressions to find $T_1$ and then run a sequence of triples of regressions over $t = 4, \ldots, \left[2\sqrt{T}\right] + 3$, $t = \left[2\sqrt{T}\right] + 4, \ldots, T_1$, $t = T_1 + 1, \ldots, T$ and then widen the first range on the cost of the second one until that one contain less than $2\sqrt{T}$ observations. Then, I proceed similarly for a potential break to the right of $T_1$.

It is a common practice to replace $\zeta_2$ and similar statistics defined via the $R^2$ by F–type transformations. Such transformations achieve distributions with better invariance properties with respect to $T$. They also achieve a somehow more convenient 'resolution', whereas $\zeta_2$ has most of its mass very close to 1. Because the focus of this work is not so much on distributional properties but rather on decision bounds that vary with $T$ anyway, I refrain from F–transformations, particularly as the bounded range of $\zeta_2$, the interval (0,1], is quite convenient.

Because the process C is really a second-order integrated process, an interesting alternative statistic to $\zeta_2$ would be the common Dickey-Fuller statistic calculated on the differenced series. However, it is assumed that the decision maker views this process as an SBDT process and does not have full information about its structure. If the decision maker has such full information, she would be better off anyway by calculating the computationally costly likelihood ratios.

Other possible alternative statistics that have been used in the literature would be the average ratio, suggested in the form of average F by HANSEN (1992), and the random-walk coefficients test considered by LEYBOURNE AND MCCABE (1989). Because the coefficients of process C actually follow random walks, the latter statistic may be particularly appealing. However, the low probability of breaks may impair its discriminatory power.

Figure 3 shows frequency histograms of the finite-sample distribution of $\zeta_2$ for the processes A–C. Because the minimum sample size $\min(T_1, T_2 - T_1, \ldots, T - T_K)$ with $K = [(T + 49)/100]$ increases toward $\infty$ as $T \to \infty$, the condenser $\zeta_2$ converges to 1 for large $T$ if the applied non-breaking model is correct. Processes A and B are correctly specified, hence for these two processes the distribution $\zeta_2$ should be concentrated close to 1. Process C is not correctly specified and the asymptotic properties of $\zeta_2$ are rather complicated. Figure 3 shows that the mode of $\zeta_2$ is in the range 0.8–0.9 for the processes A and B, with much mass concentrated close to the ideal value of 1.0. For process C, the distribution is rather flat with a mode at 0.75, hence the discriminatory power of $\zeta_2$ appears to be unsatisfactory. However, due to the prior distribution on $\lambda$, many C trajectories are not revealing in the sense that the breaking process is never activated. No test can work miracles in this situation.

An analysis of the bivariate distribution of $(\zeta_1, \zeta_2)$ confirms this interpretation (not shown). For processes A and B, visual inspection can hardly reject independence of the two condensers. For process C, the density shows a curious accumulation of mass for very negative values of $\zeta_1$ and large values of $\zeta_2$. This area corresponds to small draws of $\lambda$ and to trajectories without revealing shape. The effects persist for larger $T$.

## 5 Loss functions

A decision problem is defined by two main constituent components, the prior setup and the aim of the exercise. The prior setup is summarized in the frame — the considered probability model, the way it is split into classes or hypotheses of interest, the prior distributions over the classes and within the classes. The aim of the exercise is expressed via a loss criterion that defines the loss or disutility suffered if incorrect or simply bad decisions are adopted. I consider two types of loss criteria.

Firstly, I consider *discrete quadratic loss*, which is a counterpart to the continuous quadratic loss used in classical statistical estimation. With discrete loss, the theoretical disutility depends on the selected decision and on the true model only, whereas other characteristics of the trajectories, expressible through primary continuous parameters, play a role only for calculating loss expectations. Bayesians argue sometimes that discrete loss functions rarely express the true cost of decision problems (cf. Poirier, 1995). However, if one does not happen to know the true cost function, they are convenient approximations and their most simple version, the 0-1 loss, is commonly seen in Bayesian applications and leads directly to an evaluation of posterior odds.

Secondly, I assume that the researcher is interested not so much in identifying the true data-generating mechanism but rather views modeling as an intermediate aim in obtaining a good prediction of future values of the variable of interest $X$. A way to express this loss concept is by the squared prediction error for a one-step prediction. Notice that the two criteria are not equivalent. Firstly, using an incorrect model may help in obtaining a better prediction than using the correct model if parameters have to be estimated. Hence, loss is not trivially zero if the 'correct' decision is taken. Secondly, prediction loss depends on characteristics of the process trajectories, which are expressed by primary parameters. Hence, loss is not constant within the hypothesis classes.

## 5.1 Discrete quadratic loss

In estimating a continuous parameter $\theta$ by a point estimate $\hat{\theta}$, the quadratic loss function

$$l_2(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

is commonly used, either explicitly or implicitly. In classical inference, this loss function is at the basis of the traditional evaluation of the efficiency of estimators by their variance. In Bayesian inference, $l_2$ is used explicitly and motivates point estimation by the posterior mean. The concept is also easily adopted if the parameter space is ordered and discrete, e.g. $\Theta = \mathsf{N}$, where the integer that is closest to the posterior mean constitutes a reasonable point estimate. The same concept is also useful if the rank of an $n$–dimensional matrix represents the parameter of interest, as in cointegration analysis (see Kunst, 1996).

If, as in the present case, the discrete parameter of interest is an element of an unordered set $(A, B, C)$, one may either order the set by an appropriate function $\{0, 1, 2\} \rightarrow \Xi$ and thus express directly that the SBDT process C is 'further away' from the unit-root process A than the trend-stationary B, or code the parameter values by an inherent presence or absence of $K$ features within $\{0, 1\}^K$. Here, there are two characteristic features ($K = 2$), the presence of a unit root and the presence of change in the trend part of the generating mechanism, which suggests the following correspondence

$$(1, 0) \mapsto A$$

$$\begin{aligned} (0,0) &\mapsto B \\ (0,1) &\mapsto C \end{aligned} \tag{21}$$

Conforming with the view that trend shocks and regular innovations have different economic interpretation, I do not code process C as (2,1), although this may be justified on statistical grounds. Obviously, $\{0,1\}^2$ is not exhausted, as breaks and unit roots do not co-exist in the frame. Adopting the original $l_2$ function on the ordered set does not seem to be advisable, as such a loss function would correspond to *linear* and not to *quadratic* weighting of misclassifications. I therefore rather adopt the double-quadratic loss

$$l_{d2}(\kappa, \hat{\kappa}) = \left( \sum_{i=1}^{K} |\kappa_i - \hat{\kappa}_i| \right)^2$$

which is called 'double-quadratic', as the absolute value is here equivalent to squares. The discrete parameter is naturally expressed as an element of $\{0,1\}^K$ as $\kappa = (\kappa_1, \ldots, \kappa_K)'$.

Notice that exactly the same loss function evolves either from ordering the set of secondary parameters and applying $l_2$ directly or from interpreting process C as a second-order integrated process and considering the problem of estimating a discrete parameter in $\{0,1,2\}$ that corresponds to the number of unit roots at 1. The loss penalizes a misclassification of $A$ as $C$ and vice versa more heavily than other misclassifications. In the absence of sample information, this results in a unique optimal decision: all variables are classified as $B$. Although the loss is perfectly symmetric and does not depend on the primary parameters of the individual processes, the risk to be minimized

$$R(\hat{\kappa}|X) = E(l_{d2}(\kappa, \hat{\kappa})|X) = \sum_{\kappa \in \Xi} l_{d2}(\kappa, \hat{\kappa}) P(\kappa|X)$$

is not constant in $\theta|\kappa$. Condensing the information via the statistics $\zeta = (\zeta_1, \zeta_2)'$ leads to the modified problem of minimizing

$$R(\hat{\kappa}|\zeta) = E(l_{d2}(\kappa, \hat{\kappa})|\zeta) = \sum_{\kappa \in \Xi} l_{d2}(\kappa, \hat{\kappa}) P(\kappa|\zeta)$$

in $\hat{\kappa}$ as a function of $\zeta$. It is the expression $P(\kappa|\zeta)$ that creates problems and that can usually only be approached by numerical integration or by simulation. If the condenser statistic has an accessible small-sample distribution, it may be easier to access the risk via Bayes' theorem

$$P(\kappa|\zeta) = P(\zeta|\kappa) / \sum_{\xi \in \Xi} P(\zeta|\xi)$$

which expression is particularly simple due to the uniform prior distribution over $\kappa$. However, in the present case the small-sample distribution is not accessible and hence this transformation offers no simplification. A substantial simplification can be achieved if the areas where a certain decision is optimal are bounded by lines in the $(\zeta_1, \ldots, \zeta_m)$ space that are parallel to axes. Then, a grid search over potential critical values $(\bar{\zeta}_1, \ldots, \bar{\zeta}_m)$ results in an optimal decision configuration. This procedure is particularly attractive if $m > 2$ but $m$ is also not too large, as then the grid search becomes time-consuming. If optimal bounds are not parallel to axes, this method may still result in a readily applicable low-risk and consistent, albeit not globally optimal, technique. For applications of such *decision bounds*, see also KUNST (1996).

## 5.2 Prediction error loss

The specification of a loss function in econometric model selection is difficult, as the ultimate aim of modeling in economics is often unclear. Many models are used to corroborate economic theories or to assess the explanatory power of rival approaches. Summarizing such aims as 'academic discussion', technical loss functions such as the squared loss may be appropriate for allowing an *a priori* unbiased discussion.

However, some models are constructed for a well specified purpose. The most common purposes in economics are policy analysis and forecasting. In policy analysis, an evaluation of a social welfare indicator may define a loss function, similar to monetary costs and profits in business analysis. In forecasting, prediction accuracy is the most natural target. Hence, in a forecasting framework the specification of loss follows from the specification of a measure of predictive accuracy.

Suppose the forecaster considers a finite set of models $\mathcal{M} = \{M_1, \ldots, M_m\}$. Each model $M_i$ can be expressed as a stochastic process connected with a vector of primary parameters $\theta_i$ and these may partially be identified from observing a single process trajectory. The forecaster observes a portion of such a trajectory $X_1^T = (x_1, \ldots, x_T)$ and approximates the primary parameter $\theta_i$ by an estimate $\hat{\theta}_i$. Acting as if $M_i$ were the correct model, this estimation can be conducted over the whole range of secondary parameters $\Xi = \{1, \ldots, m\}$. The procedure yields $m$ different predictors for the next value $x_{T+1}$, which may be denoted by $G(x_{T+1}|X_1^T, M_i)$. Typically, unless evaluation of the conditional expectation is numerically complicated, these forecasts will be determined by conditional expectation, hence $G(x_{T+1}|X_1^T, M_i) = E(x_{T+1}|X_1^T, M_i)$. Nonlinearities in the models or asymmetries in the forecaster's subjective loss function may suggest replacing conditional expectation by an approximation or by altogether different concepts, such as the median of the predictive distribution. Typically, the information in the trajectory cannot be condensed in the estimate $\hat{\theta}_i$ and $G(x_{T+1}|X_1^T, M_i)$ cannot be written as $G(x_{T+1}|\hat{\theta}_i)$.

Suppose then that the forecaster observes $x_{T+1}$. She will tend to prefer the model $M_i$ where $G(x_{T+1}|X_1^T, M_i)$ is 'closest' to $x_{T+1}$ and tend to dislike the model where that distance is large. Hence, the forecaster's loss is conveniently modeled as a monotonous function of $\left\|x_{T+1} - G(x_{T+1}|X_1^T, M_i)\right\|$. Absolute values or squares define such norms on $\mathsf{R}$ and also on $\mathsf{R}^n$. Here, all processes are real-valued and squared loss is used.

For processes A and B, the forecaster is assumed to use the conditionally correct parametric model structure. For process C, the forecaster is assumed to take the coefficient parameters from the segment of the trajectory between the identified break point and the end of the sample. This procedure may bias the results against model C. However, as in the construction of $\zeta_2$, it is assumed that modelers that use structural breaks are not fully aware of the underlying break-generating mechanism.

# 6 Results

## 6.1 Results for the technical loss function

Figures 4a–c are based on 270,000 replications of the frame and hence on approximately 90,000 replications of each of the processes A–C. With $T =$ 50, 100, 150, the statistic $\zeta_2$ is calculated for just one possible break point. Based on the simulated distribution of $(\zeta_1, \zeta_2)$, the minimum expected loss conditional on the observed $(\zeta_1, \zeta_2)$ is evaluated, which splits the area of admissible values into three regions.

The histograms in Figures 2–3 show that values outside the area $\{(\zeta_1, \zeta_2) \in (-9, 4) \times (0.3, 0.95)\}$ are very rare for $T = 100$. In the absence of sample information, the outlined decision procedure always prefers the process B, hence the strip along the upper margin of the graphs possibly just belongs to the B area because of small density values for *all* processes. The *glacis* that separates the A and the C areas has a different reason. A and C are about equally likely there but misclassification bears a higher risk than opting for the (rather unlikely) B.

For $T = 100$, the unit-root process A is preferred in a rather narrow zone similar to $\{(\zeta_1, \zeta_2) \in (-3.2, -0.8) \times (0.75, 0.95)\}$. The procedure is two-sided: $\zeta_1 < -3.2$ is taken as indicating rejection of the unit-root hypothesis in favor of trend stationarity, whereas $\zeta_1 > -0.8$ indicates structural breaks that generate apparently explosive behavior. Apart from the glacis and the thinly covered strip $\{\zeta_2 \in (0.95, 1.0)\}$, the breaking process C is preferred if $\zeta_2 < 0.75$ or if $\zeta_1 > 0$. The wide area of preference for C represents different manifestations of breaking. Breaking may generate seemingly explosive and also seemingly trend-stationary trajectories, depending on the frequency and size of the breaks.

For $T = 150$, the area where the unit-root process A is preferred is extended leftward and is approximately bounded by $\zeta_1 = -3.6$. Simultaneously, the $\zeta_2$ boundary creeps upward such that process C is generally preferred if $\zeta_2 < 0.83$. For even larger sample size, the leftward movement of the A area continues whereas the additional permitted breaks slow down the upward movement of the horizontal separating line between the B and C areas.

For $T = 50$, the A preference area shrinks to a small island centered around $(\zeta_1, \zeta_2) = (-1.8, 0.72)$. A rather ragged boundary curve separates the large B preference area from the C preference area, and the *glacis* is particularly spacious.

Figure 4d uses the sample size $T = 200$. Simulations calculate the statistic $\zeta_2$ under the assumption of two potential breaks and are more time consuming than for smaller $T$. Hence, the number of replications was set at only 70,000. The additional break point decreases $\zeta_2$, whereas the left bound of the A area continues its westward drift. Also notice that the *glacis* shrinks considerably as compared to smaller $T$.

Table 1 shows how many of the generated processes are classified correctly if the decisions represented in Figure 4 are followed literally with a numerical precision grid of 0.1 for $\zeta_1$ and 0.02 for $\zeta_2$. The most frequent events of misclassification are breaking process C trajectories and unit-root process A trajectories classified as stemming from the trend-stationary process B. The decision that is optimal in the absence of information, i.e. B, still dominates at larger sample sizes.

## 6.2 Results for the prediction loss function

Figure 5b shows the decision map for the prediction loss function based on the same frame as before, with $T = 100$, and 120,000 replications. The shape of the map is remarkable. It must be noted, though, that drawing the map turned out to be very difficult. For most trajectories and at most locations of $(\zeta_1, \zeta_2)$ the forecasting performance of all three prediction models is nearly equivalent. Therefore, a rougher grid was used than for the technical loss function and the boundary curves may be less reliable.

The area with preference for process A is surprisingly large. It contains a large portion of the process A area attached to the technical loss function. For large $\zeta_2$, its left boundary is around $-3.7$ which is a marked shift to the left from the technical A area. Its right bound is approximately $\zeta_2 = -1.0$ and is largely *independent* of $\zeta_1$. The area stretches far into the southwest corner and includes many cases where the *true* process is the breaking process C. This means that one should use the integrated model for forecasting if the true data-generating mechanism is supposed to be integrated but also if it is supposed to be a structural-breaks model, as long as the Dickey-Fuller statistic does not indicate 'explosive' behavior.

The B preference area is unconnected. One of its parts consists of the northwest corner, which is roughly equivalent to the technical preference area without the *glacis*, though its right boundary has shifted leftward, as was already noted. The other part is situated to the right of the A area and overlaps with a part of the *glacis* in Figure 4a. Hence, trend-stationary models should be used for forecasting either when the data-generating mechanism is actually trend-stationary or when slightly explosive behavior in the trajectory is observed without too much evidence on a structural break.

The C preference area is comparatively small. It contains those trajectories where explosive behavior *and* structural breaks are indicated jointly by the traditional interpretation of the test statistics. Only in these cases does it make sense to use a parameter estimate that has been based solely on the last part of the sample.

Figure 5c shows a comparable decision map for $T = 150$. The boundaries between the A and B areas in the northwest corner are at similar locations as for the technical loss function. The northeast portion of the B area seems to grow but gains over the C model are very small there and the boundary of this enclave is rather uncertain. Figure 5b relies on 180,000 replications of the frame.

Figure 5a gives the decision map for $T = 50$ based on 270,000 replications of the frame. Although the A preference area is smaller than for larger $T$, it is much larger than for the technical loss function shown in Figure 4a. In many empirically relevant cases, it pays to use the unit-root assumption for prediction, even though one should be rather uncertain whether the process actually *has* a unit root. Process C loses some ground relative to technical loss to the benefit of process B but the difference between the two is rather small over large areas such that these boundaries may shift if the number of replications is extended further. In contrast, process A dominates by a wider margin in the central area.

Figure 5d relies on $\zeta_2$ with two possible breaks, $T = 200$, and 70,000 replications of the frame. The upper bound of the C area moves south but the loss of C decisions in the north-east corner is only slightly larger than that of B decisions and this boundary is not very precise.

Because the validity of the maps in Figure 5 is so uncertain, the numerical evaluation in Table 2 relies on the original optimal decision with respect to prediction loss and a grid with resolution 0.333 for $\zeta_1$ and 0.02 for $\zeta_2$. Also the double-quadratic risk was evaluated at the solution where prediction risk attains its minimum. These values are clearly much higher than for the decision contours shown in Figure 4. The reason for this increase is the high percentage of C trajectories that are now classified as A, which is advisable in order to obtain a nice prediction but reduces the probability of correct classification *per se*. The fact that technical risk increases from $T = 150$ to $T = 200$ probably has no significance and may be due to the reduced number of replications for the latter experiment.

It may seem surprising that prediction risk is an increasing function of $T$. This effect has two main reasons. Firstly, whereas larger-sample forecasts rely on better parameter estimates for processes A and B, this is not necessarily so for process C. For process C, estimates are taken from the time range between the most significant break and the end of the sample whereas true slopes possibly change more often in this range for higher $T$. It was outlined in Section 4.3 how the potential number of breaks increases with $T$ and this problem is overcome as $T \to \infty$. Secondly, all processes are non-stationary and are started in 0 at $t = 0$. With larger $T$, the absolute values of the realizations typically increase and this yields a higher expected prediction loss.

## 6.3 An empirical illustration

The OECD publishes national accounts data for its member countries on a quarterly basis. As a convenient example for the application of the decision procedure, I used the OECD series on gross domestic product (GDP) in logarithms, a summary measure of economic activity that was also in the focus of previous studies on unit roots and structural breaks.

A first inspection reduced the number of usable country data to 17, among which 12 had at least 80 observations, so that the decision map for $T = 100$ may be applied to them, at least approximately. For the remaining 5 countries, the map for $T = 50$ was used with a similar result but I will not report on that experiment in detail. Out of the 12 longer series, some were long enough to match the map for $T = 150$ but I cut the series from the beginning in order to maximize the number of observations in the cross-country comparison. In detail, the countries analyzed are: Australia, Canada, Denmark, Finland, France, Italy, Japan, the Netherlands, Spain, Switzerland, the United Kingdom, the United States of America. This sample represents an interesting cross-section of small and large economies across the globe.

For each country, the information condenser statistics $\zeta_1$ and $\zeta_2$ were calculated from the GDP time series and the resulting values were inserted into the decision contour map. The result is shown in Figure 6a–b for the double-square loss function and the prediction loss function. Noting that the series for the United Kingdom (UK) and Finland (FIN) are different as they are not calculated on a seasonally adjusted basis, the unanimity of the evidence is strikingly in favor of the unit-root model A. Only Italy (I) yields evidence on structural breaks that are also clearly visible from a time-series plot. Spain (E) reverts to a linear trend in longer swings and hence is classified as a deterministic-trend model B. Japan (JAP) and the Netherlands (NL) are located in the glacis of Figure 5a. Their classification is very uncertain, and hence they are classified as deterministic-trend models B in order to minimize expected loss. Japan may also be better forecasted by using that model but for the Dutch series it may be more advisable to use a unit-root model for prediction.

Time-series plots of the two outliers UK and FIN show dominant cyclical seasonal patterns. Only for these two cases did some diagnostic statistics, that I calculated routinely in the auxiliary regression for $\zeta_1$, point to a possible gain in information by using an extended frame, for example by using more lags in this regression. The regular seasonal patterns are interpreted as structural breaks and the series are classified as breaking (C). In contrast, if the aim is prediction, usage of a unit-root model is more advisable.

This exercise is meant as an illustration of the procedure only. In order to obtain further conclusions, one may for example seasonally adjust the outlying series, extend the frame to allow for seasonal unit roots, investigate the influence of seasonal adjustment on classification etc. It is obvious, however, that, firstly, the outcome is rather similar across countries and, secondly, the maps can be consulted quickly without any further computational requirement than some linear regressions. The breaking model, that appears to be attractive to some researchers, is not supported as a good descriptive model and even less as a good prediction model.

# 7 Summary and conclusion

To my knowledge, this study is the first attempt at comparing technical loss and prediction loss with respect to model selection in the three-model set of unit-roots models, trend-stationary models, and structural-breaks models, although these models are commonly seen as alternative ways of formalizing trending behavior in economic time series. It is obvious that much more detailed work has to be done before the results can be viewed as a definitive guideline for empirical economists and economic forecasters. The most interesting extensions of the approach should include extensive Monte Carlo to investigate the reaction of decision contours as functions of $T$ and discrete lag-order priors with infinite support. Sensitivity studies could handle non-normal errors and variations of the functional form in prediction loss. It may also be interesting to investigate the robustness of the results to an increase of the forecast horizon. Some rudimentary experiments showed little reaction of the decision contours but more research is needed to permit a definite conclusion.

On the whole, the results confirm the fact that the true model — I define a model as a parameterized collection of processes, not as one specific fixed-parameter process — is not necessarily the best forecasting model, due to sampling variation in its parameter estimates. Both faces of this simple truth are important to empirical researchers. Neither does the true or 'valid' model guarantee optimal forecasting nor must a good forecasting model fulfill any criteria of statistical validity besides small prediction loss. A cursory glance at the current forecasting literature reveals a widespread ignorance of this simple truth.

# References

ANDREWS, D.W.K., and PLOBERGER, W. (1994), 'Optimal Tests when a Nuisance Parameter is Present Only under the Alternative', *Econometrica* **62**, 1383–1414.

BANERJEE, A., DOLADO, J., GALBRAITH, J.W., and HENDRY, D.F. (1993), *Co-integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data.* Oxford University Press.

BOX, G.E.P., JENKINS, G., and REINSEL, G.C. (1994), *Time Series Analysis. Forecasting and Control.* 3rd edition, Prentice-Hall.

CHEN, C., and TIAO, G.C. (1990) 'Random Level-Shift Time Series Models, ARIMA Approximations, and Level-Shift Detection'. *Journal of Business & Economic Statistics* **8**, 83–97.

CHOW, G.C. (1960) 'Tests of equality between sets of coefficients in two linear regressions'. *Econometrica* **28**, 591–605.

CHRISTIANO, L.J. (1992) 'Searching for a Break in GNP'. *Journal of Business & Economic Statistics* **10**, 237–250.

DAVIES, R.B. (1977) 'Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative'. *Biometrika* **64**, 247–254.

—— (1987) 'Hypothesis Testing when a Nuisance Parameter is Present Only under the Alternative'. *Biometrika* **74**, 33–43.

DICKEY, D.A., and FULLER, W.A. (1979) 'Distribution of the Estimators for Autoregressive Time Series with a Unit Root'. *Journal of the American Statistical Association* **74**, 427–431.

FULLER, W.A. (1996) *The Statistical Analysis of Time Series.* Wiley.

GOURIEROUX, C., AND MONFORT, A. (1995) *Statistics and Econometric Models.* Cambridge University Press.

HACKL, P., and WESTLUND, A.H. (1991) (ed.) *Economic Structural Change: Analysis and Forecasting.* Springer-Verlag.

HANSEN, B.E. (1992) 'Tests for Parameter Instability in Regressions with I(1) Processes'. *Journal of Business & Economic Statistics* **10**, 321–335.

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

HATANAKA, M. (1996) *Time-Series-Based Econometrics: Unit Roots and Co-Integration.* Oxford University Press.

HENDRY, D.F. (1995) *Dynamic Econometrics.* Oxford University Press.

KUNST, R.M. (1996) 'Estimating Discrete Parameters: An Application to Cointegration and Unit Roots'. *Österreichische Zeitschrift für Statistik* **25**, 7-32.

LEYBOURNE, S.J., and MCCABE, B.P.M. (1989) 'On the distribution of some test statistics for coefficient constancy'. *Biometrika* **76**, 169–178.

NELSON, C.R., and PLOSSER, C.I. (1982) 'Trends and random walks in macroeconomic time series: some evidence and implications'. *Journal of Monetary Economics* **10**, 139–162.

PERRON, P. (1989) 'The Great Crash, the Oil Shock and the Unit Root Hypothesis'. *Econometrica* **57**, 1361–1402.

POIRIER, D.J. (1995) *Intermediate Statistics and Econometrics.* MIT Press.

QUANDT, R.E. (1960) 'Tests of the Hypothesis That a Linear Regression System Obeys Two Separate Regimes'. *Journal of the American Statistical Association* **55**, 324–330.

SPANOS, A. (1986) *Statistical foundations of econometric modelling.* Cambridge University Press.

STOCK, J.H. (1994) 'Deciding between I(1) and I(0)'. *Journal of Econometrics* **63**, 105–131.

WEST, M., and HARRISON, J. (1989) *Bayesian Forecasting & Dynamic Models.* Springer-Verlag.

WHITE, H. (1996) *Estimation, Inference and Specification Analysis.* Cambridge University Press.

ZIVOT, E., and ANDREWS, D.W.K. (1992) 'Further Evidence of the Great Crash, the Oil Price Shock and the Unit Root Hypothesis'. *Journal of Business & Economic Statistics* **10**, 251–270.

# Tables

TABLE 1. Classification frequencies based on minimizing the double-squared technical loss function.

| identified | generated model | | | expected risk |
|---|---|---|---|---|
| model | A | B | C | |
| $T = 50$ | | | | 0.5808 |
| A | 0.138 | 0.034 | 0.031 | |
| B | 0.843 | 0.959 | 0.658 | |
| C | 0.019 | 0.007 | 0.311 | |
| $T = 100$ | | | | 0.3873 |
| A | 0.686 | 0.127 | 0.093 | |
| B | 0.296 | 0.865 | 0.284 | |
| C | 0.018 | 0.008 | 0.623 | |
| $T = 150$ | | | | 0.2610 |
| A | 0.816 | 0.133 | 0.076 | |
| B | 0.172 | 0.860 | 0.185 | |
| C | 0.012 | 0.006 | 0.739 | |
| $T = 200$ | | | | 0.1754 |
| A | 0.894 | 0.090 | 0.047 | |
| B | 0.096 | 0.907 | 0.113 | |
| C | 0.009 | 0.002 | 0.841 | |

Note: Probabilities have been normalized conditional on the generated model classes.

TABLE 2. Classification frequencies based on minimizing the prediction loss function.

| identified | generated model | | | expected risk | technical risk |
|---|---|---|---|---|---|
| model | A | B | C | | |
| $T = 50$ | | | | 1.7624 | 0.8671 |
| A | 0.809 | 0.326 | 0.404 | | |
| B | 0.176 | 0.663 | 0.413 | | |
| C | 0.015 | 0.010 | 0.183 | | |
| $T = 100$ | | | | 1.8245 | 0.7622 |
| A | 0.682 | 0.183 | 0.330 | | |
| B | 0.295 | 0.805 | 0.388 | | |
| C | 0.022 | 0.012 | 0.283 | | |
| $T = 150$ | | | | 1.9241 | 0.6122 |
| A | 0.744 | 0.125 | 0.282 | | |
| B | 0.244 | 0.869 | 0.336 | | |
| C | 0.013 | 0.006 | 0.381 | | |
| $T = 200$ | | | | 1.9906 | 0.6421 |
| A | 0.836 | 0.139 | 0.311 | | |
| B | 0.145 | 0.846 | 0.306 | | |
| C | 0.019 | 0.015 | 0.383 | | |

Note: Probabilities have been normalized conditional on the generated model classes.

# Figures

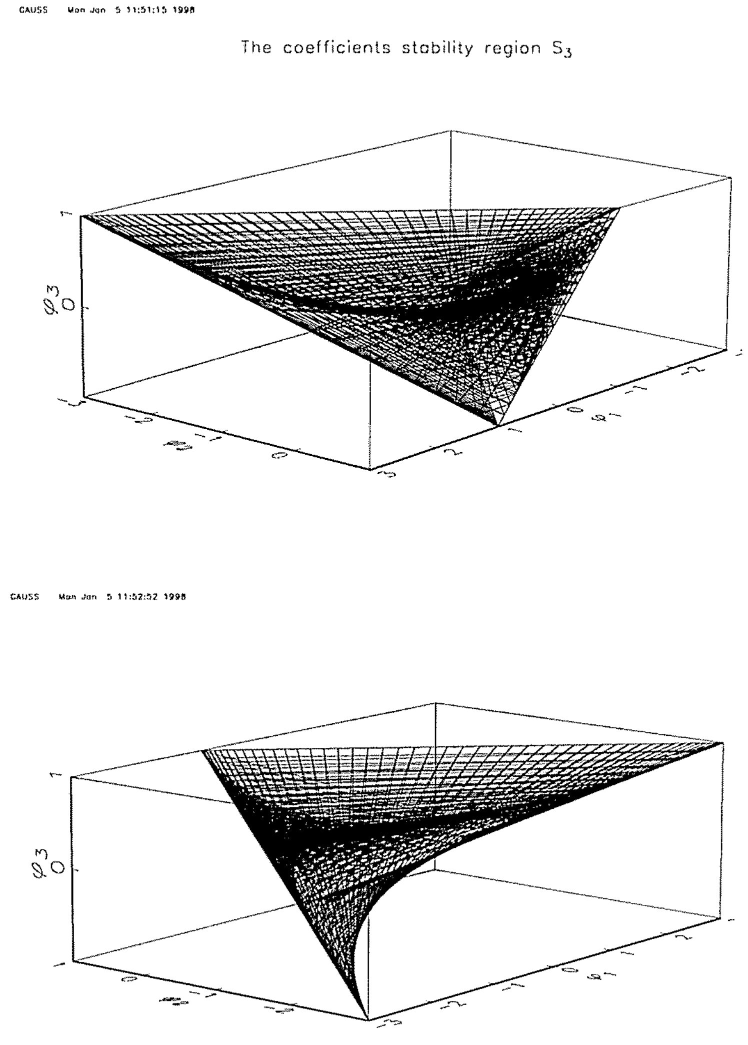

FIGURE 1. The stability region for the coefficients in a third-order autoregression.

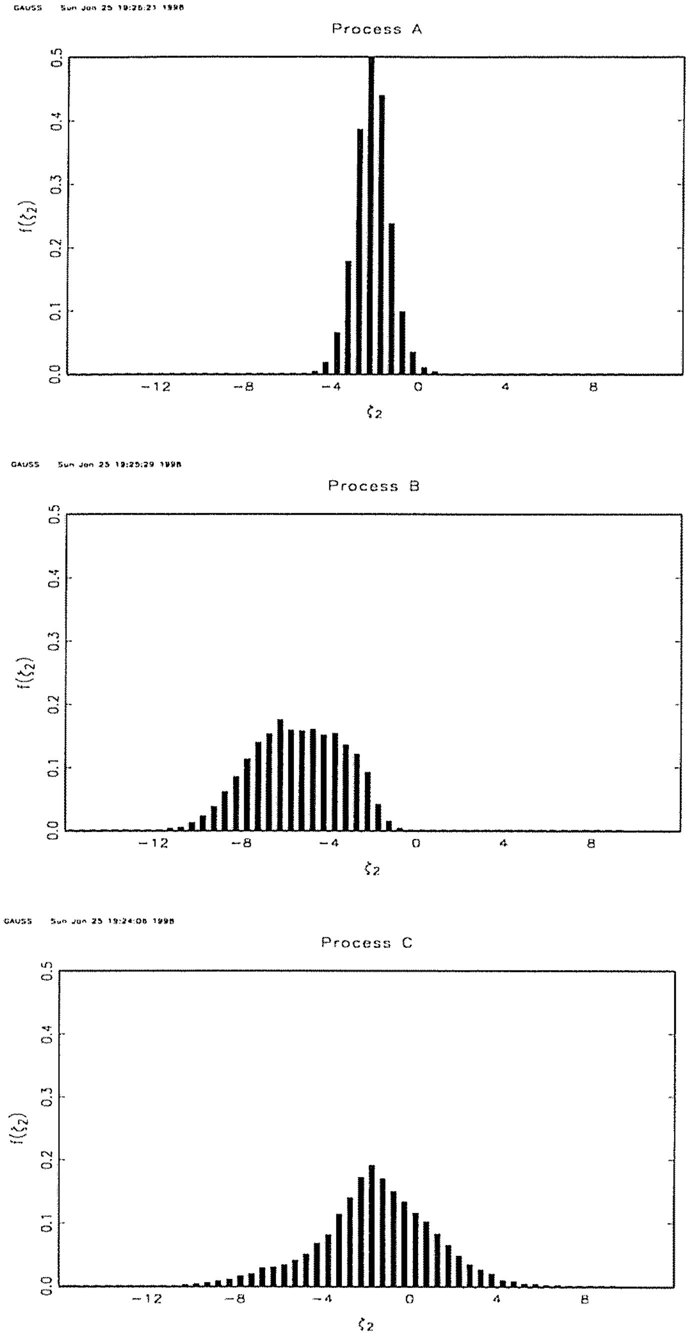

FIGURE 2. Frequency histograms of the probability distribution of $\zeta_1$ for $T = 100$.

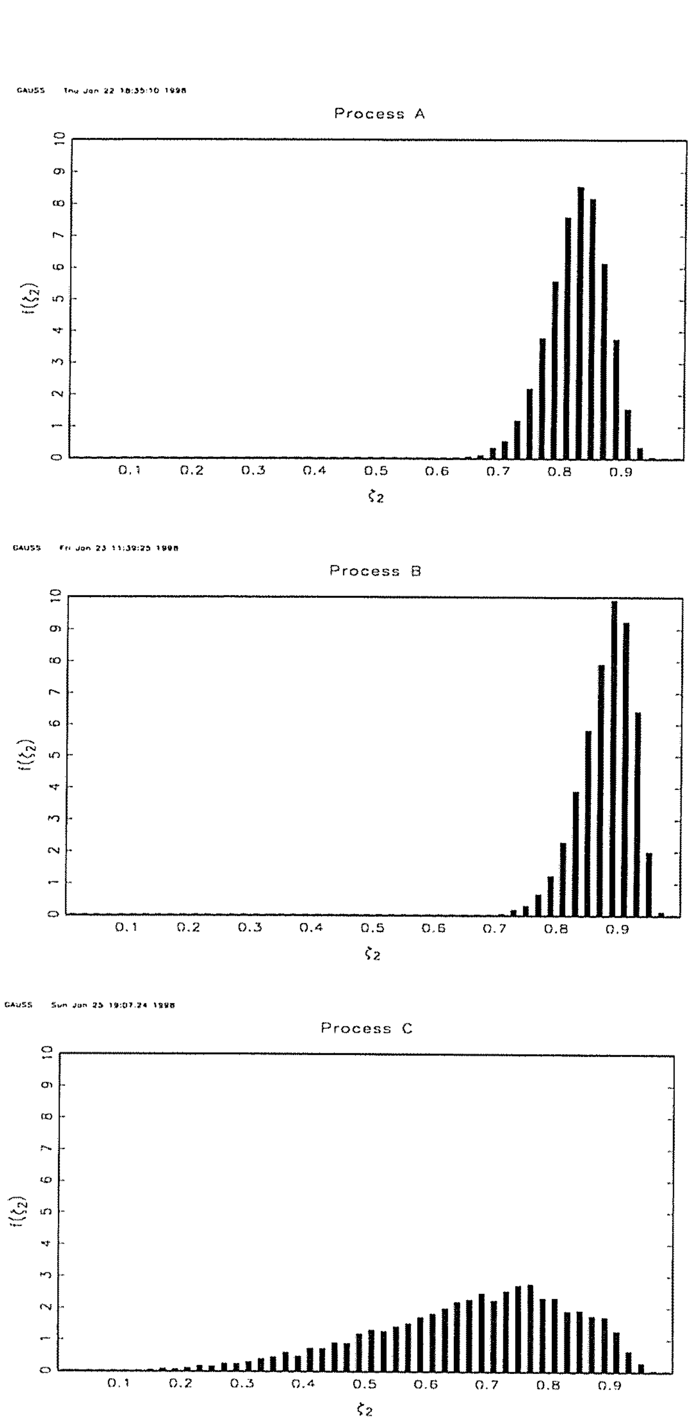

FIGURE 3. Frequency histograms of the probability distribution of $\zeta_2$ for $T = 100$.

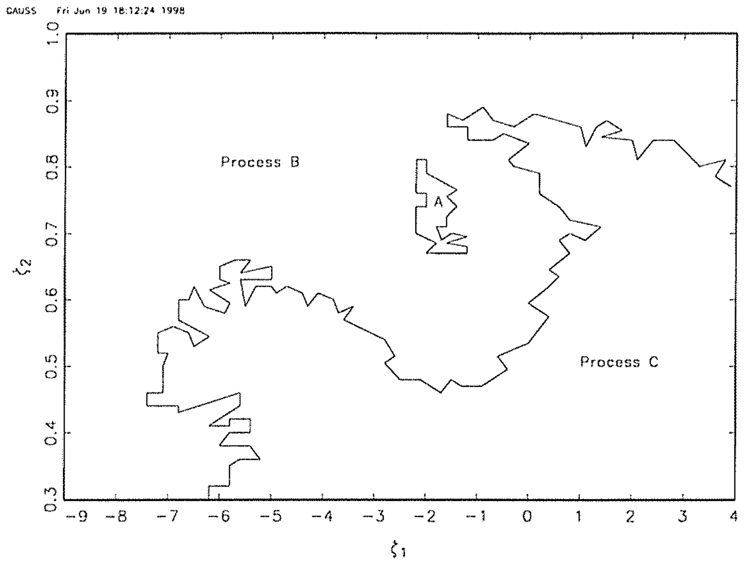

FIGURE 4A. Optimal decision for double-squared loss and $T = 50$.

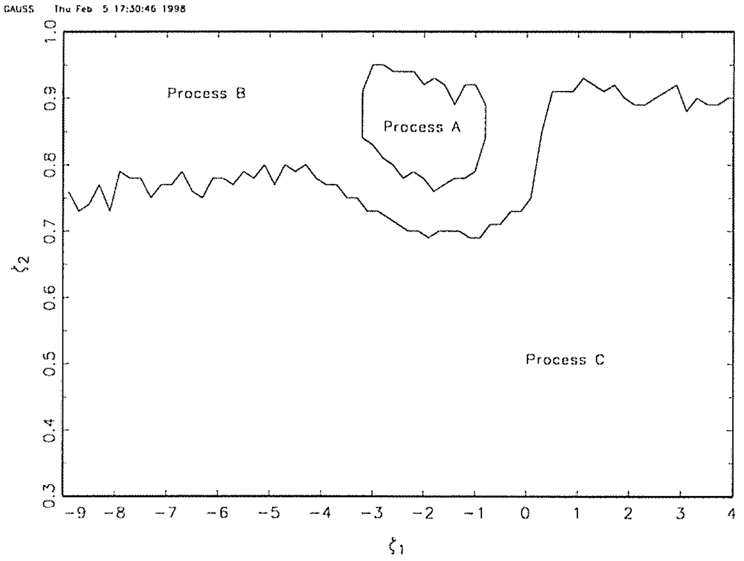

FIGURE 4B. Optimal decision for double-squared loss and $T = 100$.

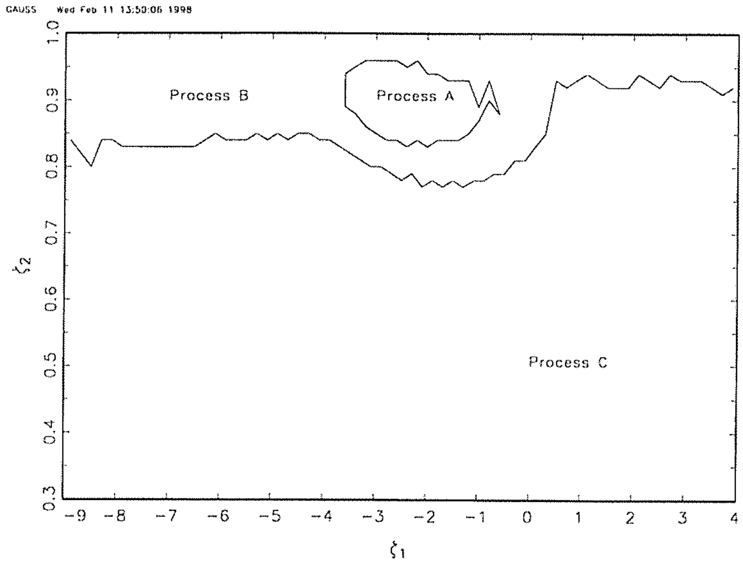

FIGURE 4C. Optimal decision for double-squared loss and $T = 150$.

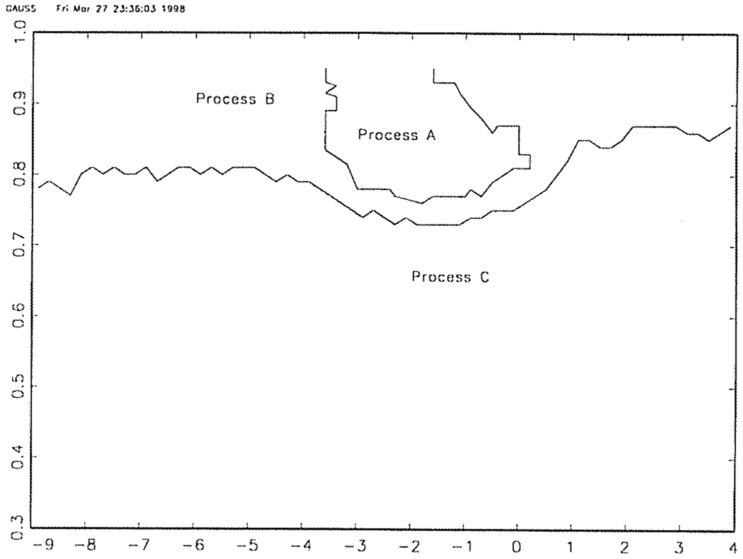

FIGURE 4D. Optimal decision for double-squared loss and $T = 200$.

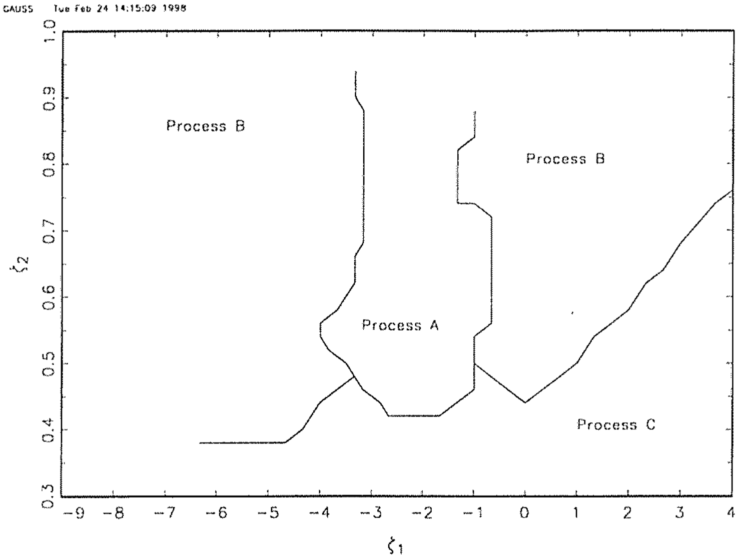

FIGURE 5A. Optimal decision for prediction loss and $T = 50$.

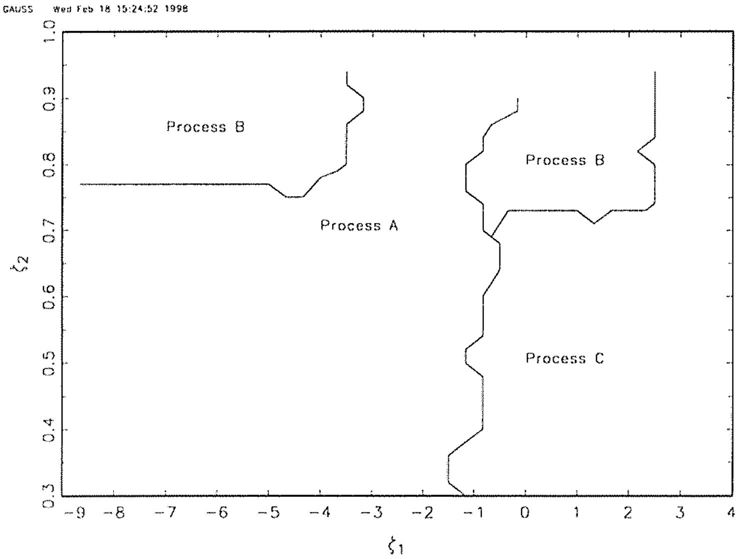

FIGURE 5B. Optimal decision for prediction loss and $T = 100$.

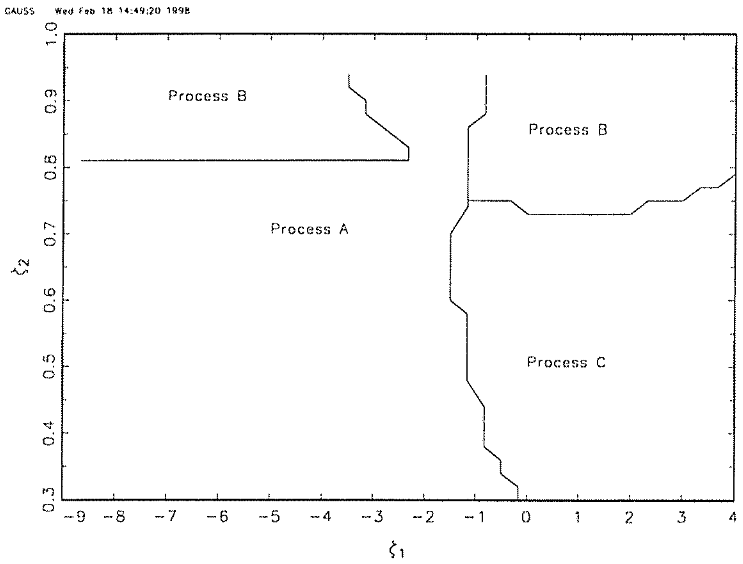

FIGURE 5C. Optimal decision for prediction loss and $T = 150$.

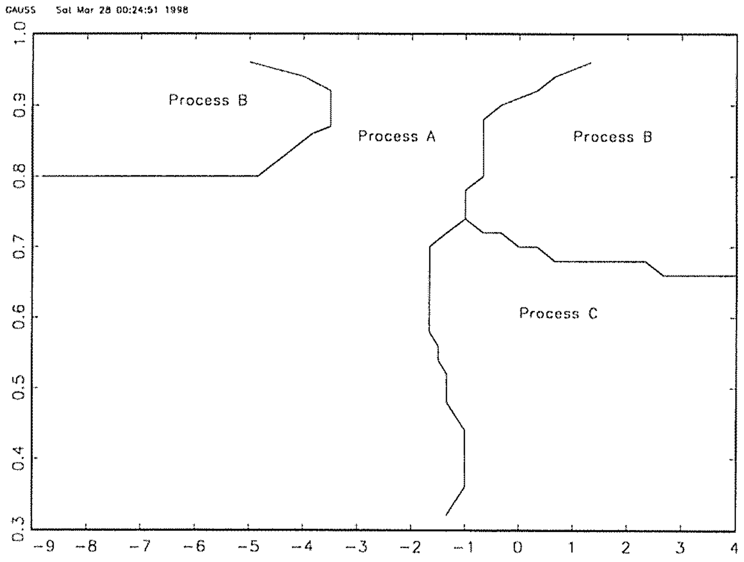

FIGURE 5D. Optimal decision for prediction loss and $T = 200$.

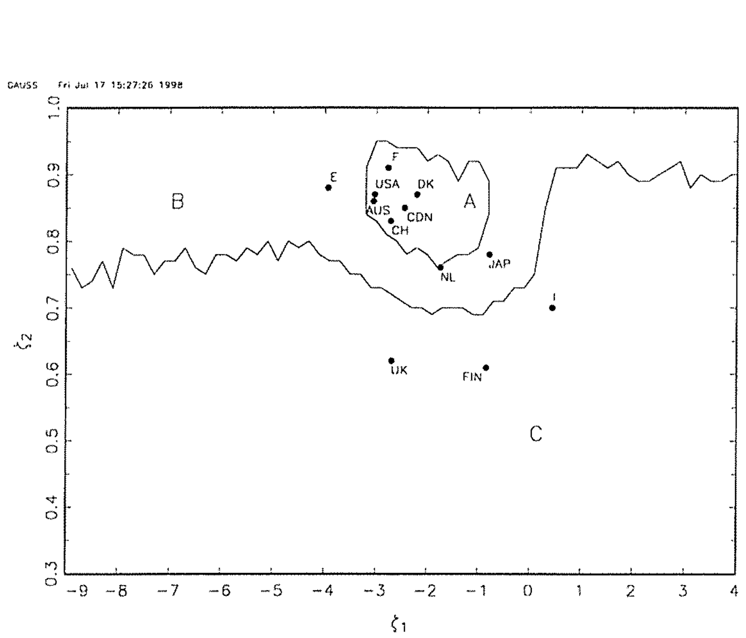

FIGURE 6A. Optimal decision for GDP series from 12 OECD countries and double-squared loss.

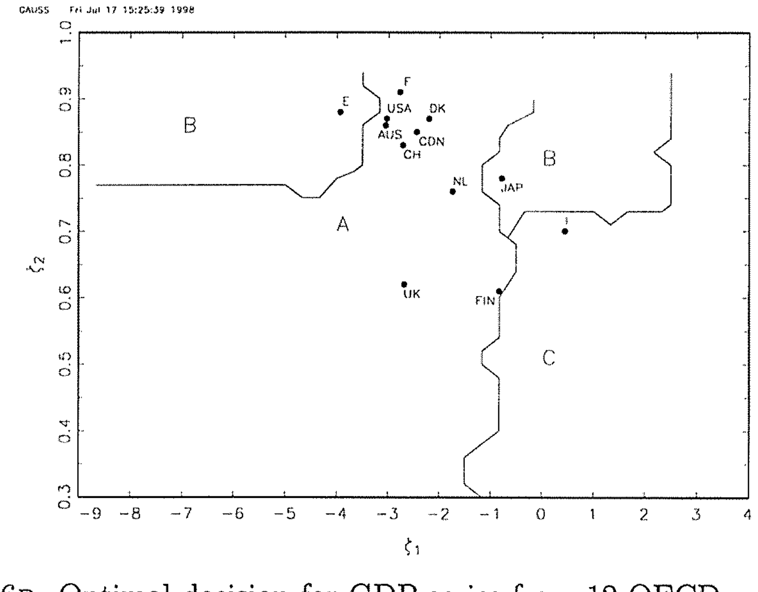

FIGURE 6B. Optimal decision for GDP series from 12 OECD countries and prediction loss.

**Institut für Höhere Studien**
**Institute for Advanced Studies**
Stumpergasse 56
A-1060 Vienna
Austria

Phone: +43-1-599 91-145
Fax: +43-1-599 91-163