Fields Institute Communications
Volume 11, 1997

Some Comments on Nonlinear Time Series Analysis

Howell Tong
Institute of Mathematics and Statistics
University of Kent
Canterbury, Kent CT2 7NF UK

Abstract. Through two simple examples, we reveal some basic similarities as well as basic differences between two recent approaches to defining initial-value sensitivity for a stochastic dynamical system. We also make some comments relevant to the practice of prediction.

1 Introduction

Recent developments in nonlinear time series have substantially widened the scope of traditional time series analysis. They have deepened our understanding of prediction and the role of initial values. This paper makes a few comments and suggests a couple of issues which require further research.

2 Initial-Value Sensitivity for a Stochastic Dynamical System

There are at least two different approaches to defining initial-value sensitivity for a stochastic dynamical system. (For references see, e.g., Tong [1995]). The earlier approach starts with a deterministic dynamical system typified by

$$X_t = F(X_{t-1}), \tag{2.1}$$

where X_t denotes a state vector in R^d and $t \geq 1$. For simplicity of discussion, let $d = 1$, which will suffice for our purpose of raising the forthcoming issues. If the system is ergodic, then it is well known that the so-called *Lyapunov exponent* λ defined by $\lim_{t\to\infty} t^{-1} \sum_{m=0}^{t-1} \ln |\dot{F}(X_m)|$, or equivalently $E \ln |\dot{F}(X)|$ by ergodicity, plays a central role in quantifying the initial-value sensitivity. Here, the expectation is taken with respect to the invariant measure associated with the ergodic deterministic dynamical system. If the time series $\{X_t, -\infty < t < \infty\}$ is generated by an ergodic stochastic dynamical system of the form

$$X_t = F(X_{t-1}) + \epsilon_t, \tag{2.2}$$

where $\{\epsilon_t, t \geq 1\}$ is a (dynamic) noise sequence taking the form of a sequence of independent identically distributed random variables, then it would seem quite natural to adopt essentially the same Lyapunov exponent as above to quantify the initial-value sensitivity of the stochastic dynamical system (2.2), which is obtained

1991 *Mathematics Subject Classification.* Primary: 62M20, 62M10; Secondary: 58F13.

by "clothing" its obvious deterministic counterpart (2.1), also called the *skeleton*. The only modification required is to take the expectation with respect to the invariant measure of the stochastic dynamical system instead, assuming that the latter is ergodic. Specifically, for almost all $\omega \in \Omega$, where Ω is the underlying probability space, we get

$$\lambda(\omega) = \lim_{t\to\infty} t^{-1} \sum_{m=0}^{t-1} \ln|\dot{F}(X_m(\omega))|, \tag{2.3}$$

which, by ergodicity, yields a Lyapunov exponent independent of ω, namely

$$\lambda = E \ln|\dot{F}(X)|, \tag{2.4}$$

where the expectation is taken with respect to the invariant measure associated with the stochastic dynamical system (2.2). Note that in this approach, the same $\epsilon_t(\omega)$ is used when we study the effect of perturbing $X_0(\omega)$ to, say, $X_0(\omega) + \delta$. For future reference, we call this the *identical noise realization approach*. K. S. Chan in his contribution to the discussion of Tong [1995] has generalized (2.3) to

$$\lambda^*(\omega) = \lim_{t\to\infty} t^{-1} \ln|\partial X_t(\omega)/\partial X_0(\omega)|. \tag{2.5}$$

As pointed out by Chan (*op. cit.*), one advantage of this generalization is that it can be applied to the more general stochastic dynamical system of the form

$$X_t = F(X_{t-1}, \epsilon_t), \tag{2.6}$$

where the dynamic noise need not be additive. Another advantage is that the new definition is invariant under a 1-1 co-ordinate transformation of the state. We shall adopt the generalized version in this paper. Note that it is obviously still in keeping with the identical noise realization scenario.

A recent alternative is to take full account of the fact that the states are now random variables and study the effect of perturbing the initial value on the distributions (or some summary characteristics) of the later states. Now, let $\{X_t\}$ denote a *general* real-valued stochastic process indexed by $t \in N$ and with finite variance. Assume that the conditional density of X_m given $X_0 = x$ exists and let us denote it by $g_m(.|x)$. We suppose that $g_m(.|x)$ is sufficiently smooth in x. For two neighbouring initial values x, $x + \delta \in R$, after time $m \geq 1$, the *divergence* of the conditional distribution of X_m may be defined once the choice of a distance function over the space of distributions is made. The choice is clearly non-unique. For example, if the negative Kullback-Leibler mutual information is used, then we may define the divergence as

$$K_m(x;\delta) = \int \{g_m(z|x+\delta) - g_m(z|x)\} \ln\{g_m(z|x+\delta)/g_m(z|x)\} dz. \tag{2.7}$$

Note that stationarity is not required for this equation. Now, for small δ, we can expand the right hand side using Taylor's series about x. This gives the approximation

$$K_m(x;\delta) = \delta^2 I_{1,m}(x) + o(\delta^2), \tag{2.8}$$

where

$$I_{1,m}(x) = \int \{\partial g_m(z|x)/\partial x\}^2 / g_m(z|x) dz. \tag{2.9}$$

If the L_2 norm is used, then we may define another divergence as

$$D_m(x;\delta) = \int \{g_m(z|x+\delta) - g_m(z|x)\}^2 dz. \tag{2.10}$$

It follows from the Taylor's expansion that

$$D_m(x;\delta) = \delta^2 I_{2,m}(x) + o(\delta^2), \tag{2.11}$$

where

$$I_{2,m}(x) = \int \{\partial g_m(z|x)/\partial x\}^2 dz. \tag{2.12}$$

Sometimes it is convenient to specialize the above setup to just the conditional mean function when the latter exists. Consider model (2.2) but now we weaken the assumption of independence and identical distribution of the dynamic noise ϵ_t to $E(\epsilon_t|X_k, k<t) = 0$. Let us refer to this model as model (2.2*). Let $F_m(x) = E[X_m|X_0 = x]$, $x \in R$ and $m \geq 1$. For $\delta \in R$,

$$F_m(x+\delta) - F_m(x) = \dot{F}_m(x)\delta + o(|\delta|), \tag{2.13}$$

where $\dot{F}_m(x)$ denotes $dF_m(x)/dx$. For model (2.2*), $F_1(x) = F(x)$ and

$$\begin{aligned} F_m(x) &= E\{F(X_{m-1})|X_0 = x\} \\ &= E\{F(F(X_{m-2}) + \epsilon_{m-1})|X_0 = x\} \\ &= E\{F((\cdots(F(x)+\epsilon_1)+\cdots)+\epsilon_{m-1})|X_0 = x\}. \end{aligned} \tag{2.14}$$

By the chain rule, differentiation of the right hand side of the above equation gives

$$\dot{F}_m(x) = E\{\prod_{k=1}^{m} \dot{F}(X_{k-1})|X_0 = x\}, \tag{2.15}$$

where we have assumed that the differentiation under the integral sign is justified. We may interpret equations (2.13) and (2.15) as stochastic generalizations of their obvious deterministic counterparts.

Given the existence of different approaches, it is natural to enquire what different aspects of initial-value sensitivity they reflect. The following very simple examples are quite instructive.

Example 2.1 Consider a simple linear first order autoregressive process

$$X_m = \epsilon_m + \alpha X_{m-1}, \tag{2.16}$$

where $\{\epsilon_j\}$ is a sequence of independent random variables each distributed as $\mathcal{N}(0,1)$. Clearly,

$$X_m = \epsilon_m + \alpha\epsilon_{m-1} + ... + \alpha^{m-1}\epsilon_1 + \alpha^m X_0. \tag{2.17}$$

Therefore , given $X_0 = x$, we have $X_m \sim \mathcal{N}(\alpha^m x, \sigma_m^2)$, where

$$\sigma_m^2 = \left(\frac{1-\alpha^{2m}}{1-\alpha^2}\right), \tag{2.18}$$

for $\alpha \neq 1$, and

$$\sigma_m^2 = m, \tag{2.19}$$

for $\alpha = 1$. Let us consider the conditional distribution approach first since it does not require stationarity. Now, simple calculations yield that

$$I_{1,m}(x) = \alpha^{2m}/\sigma_m^2. \tag{2.20}$$

Equation (2.20) shows a sensitivity measure which differs from the classical Lyapunov exponent in that it incorporates directly the effect of the dynamic noise in the form of a diffusion term σ_m^2 in order to adjust the impact of the disturbance δ on the drift term. Substituting the value of σ_m^2 from (2.18) in (2.20), we get

$$I_{1,m}(x) = \alpha^{2m}(1-\alpha^2)/(1-\alpha^{2m}). \tag{2.21}$$

To investigate the asymptotic behaviour of $I_{1,m}$, we consider three cases separately as follows.

(i) $|\alpha| < 1$: In this case, $I_{1,m}(x) \to 0$ as $m \to \infty$. This mimics the behaviour of the globally stable skeleton, i.e., the case with the dynamic noise switched off. Thus, even after clothing the skeleton remains initial-value insensitive.

(ii) $|\alpha| > 1$: In this case, $I_{1,m}(x) \to (\alpha^2 - 1)$ as $m \to \infty$. It is interesting to note that the limit is positive but *finite*. Thus, the stochastic model is sensitive to initial values; the sensitivity is clearly induced by the instability of the skeleton. (Recall that stationarity is not required in the definition of $K_m(x;\delta)$.)

(iii) $|\alpha| = 1$: This has points of contact with the well-known *unit-root* model in econometrics. In this case, $I_{1,m}(x) = m^{-1} \to 0$ as $m \to \infty$.

Next, let us turn to the conditional mean specialization and the identical noise realization approach. For these, we need to impose the stationarity condition $|\alpha| < 1$. Plainly for the conditional mean specialization, $|\dot{F}_m(x)| = |\alpha|^m$ for each m. Therefore

$$|F_m(x+\delta) - F_m(x)| = |\alpha|^m\delta + o(|\delta|). \tag{2.22}$$

For the identical noise approach, we have

$$|\partial X_m(\omega)/\partial X_0(\omega)| \equiv |\alpha|^m, \tag{2.23}$$

in agreement with the conditional mean specialization. In fact, all the above results are broadly in agreement if we note that under stationarity $|I_{1,m}|^{\frac{1}{2m}} \to |\alpha|$.

Example 2.2 Consider the following simple model which concentrates on the diffusion term

$$X_t = \epsilon_t X_{t-1}, \tag{2.24}$$

where $\epsilon_t \sim \mathcal{N}(0,1)$ and ϵ_t is independent of $X_s, s < t$. Clearly $\{X_t\}$ is a sequence of uncorrelated random variables, i.e., a white noise sequence.

First, the conditional mean specialization has nothing to say on the initial-value sensitivity because $E[X_m|X_0 = x] \equiv 0$, all m.

For the identical noise realization approach, simple calculation gives

$$\lambda^* = E[\ln|\epsilon_1|] = (2/\pi)^{\frac{1}{2}} \int_0^\infty (\ln x)\exp(-x^2/2)dx = -(\gamma + \ln 2)/2, \tag{2.25}$$

where γ denotes Euler's constant. Thus,

$$\lambda^* \approx -0.635 < 0.$$

Apparently the identical noise realization approach attends to characteristics of the random system other than the conditional mean. Equations (2.4) and (2.5) seem to suggest that it is unlikely that the approach attends to any particular (conditional or unconditional) moments associated with the joint distributions of $\{X_t\}$. On noting that $X_m = X_0 \prod_{j=1}^m \epsilon_j$, clearly any perturbation of X_0 will be successively shrunk by $\epsilon_j, j = 1, .., m$. The identical noise realization approach

quantifies the shrinkage factor (i.e., $\exp\{\lambda^*\}$) to be roughly 0.5 for the case with standard Gaussian noise. Clearly, λ^* depends on the noise distribution. For example, for $\epsilon_1 \sim U(-\sqrt{3}, \sqrt{3})$, we have $\exp\{\lambda^*\} \approx 0.2$. Obviously, the assumption of identical noise realization yields for each m, $\partial X_t(\omega)/\partial X_0(\omega) = \prod_{j=1}^m \epsilon_j(\omega)$, in which $X_0(\omega)$ is absent. Suppose we remove this assumption. Let $X_0 = x_0$. Then $X_m(\omega) = x_0 \prod_{j=1}^m \epsilon_j(\omega)$. After perturbing x_0 to $x_0 + \delta$ and employing a *different* noise realization, say $\{\epsilon_j(\omega^*), j = 1, .., m\}$, we get $X_m(\omega^*) = (x_0 + \delta) \prod_{j=1}^m \epsilon_j(\omega^*)$. We note that in contrast to the above, x_0 will now be present in $X_m(\omega^*) - X_m(\omega)$ at least for each finite $m > 0$.

For the conditional distribution approach, simple calculation yields

$$\sqrt{I_{1,1}(x_0)} = \sqrt{2}/|x_0|.$$

The result suggests that any perturbation to the initial value x_0 is amplified or reduced accordingly as $|x_0|$ is smaller than or larger than $\sqrt{2}$. Note that conditional on x_0, $X_1 \sim \mathcal{N}(0, x_0^2)$. It is intuitively obvious that the discrepancy between say $\mathcal{N}(0, 1000)$ and $\mathcal{N}(0, 1000.5)$ is insignificant when compared with that between say $\mathcal{N}(0, 0)$ (a Dirac delta function) and $\mathcal{N}(0, 0.5)$. For $I_{1,m}(x_0), m > 1$, it seems decidedly difficult to obtain a closed form expression. However, the following approximate result holds for large m (mostly due to Professor K. S. Chan–private communication). First, we note that $X_m/X_0 = M_m S_m$, where $M_m = |\epsilon_m||\epsilon_{m-1}|...|\epsilon_1|$ and $S_m = \text{sign}(\epsilon_m \epsilon_{m-1} ... \epsilon_1)$. Second, we note that for large m, we may appeal to the Central Limit Theorem and deduce that $\ln M_m$ is approximately $\mathcal{N}(m\mu, m\sigma^2)$, where $\mu = E(\ln|\epsilon_1|)$ and $\sigma^2 = \text{Var}(\ln|\epsilon_1|)$. Finally, we deduce that

$$\sqrt{I_{1,m}(x_0)} \approx \{\sqrt{m}\sigma|x_0|\}^{-1}.$$

It is interesting to note the continuing presence of $|x_0|$ in the denominator. Further, it is clear that $I_{1,m}(x_0)$ tends to 0 as m tends to ∞ for all $x_0 \neq 0$.

Clearly the identical noise realization approach and the conditional distribution approach have overlapping as well as non-overlapping targets. The above simple examples have revealed some of their basic similarities as well as basic differences but a more systematic study is clearly necessary.

3 Miscellaneous Remarks on Prediction

(i) Do not be fooled by exaggerated claims

Prediction is a risky business. Perhaps this is one of the many reasons why it features so significantly in almost every ancient civilization and continues to occupy such a premier position in modern times. As far as gazing beyond the crystal balls and the oracle bones is concerned, fashions have come and gone even within the statistical community and now and then exaggerated claims have been made if only just to provide the occasional illusory excitement. In this respect, measures such as $F_m(x), I_{1,m}(x), I_{2,m}(x)$ and others with their nonparametric estimates may be seen as timely moderators. It is now well known (e.g., Tong [1995]) that the predictability of a time series is reflected by the above measures which typically depend on x, the present state, and m, sometimes called the *predictive horizon.*

(ii) Never take comparative results on trust

Nobody can deny the value of a comparative study of different prediction methods. However, for such a study to have any scientific value, we maintain that it should be conducted on *genuine* post-sample data in the sense that the human operators really have no access to the post-sample data in any way at the time when the predictions are made. Despite the obvious truth of this statement, we observe that there are far more "fake" post-sample data predictions than genuine ones reported in the literature! The most recent large-scale example of genuine post-sample prediction seems to be the "Sante Fe Time Series Prediction and Analysis Competition, 1991" (results summarized in Weigend and Gershenfeld [1994]). Earlier large-scale examples were Makridakis and Hibon [1979] and Makridakis et al. [1984]. However, even in the above comparative studies, the "true answers" were actually known to the organisers (but not to the competitors). This state of affairs is not surprising in view of the necessary time constraint on the organisation. The situation may be compared with the 8-year waiting time for Ghaddar and Tong [1981], who in 1980 submitted their genuine out-of-sample predictions for the annual sunspot numbers for the years 1980 to 1987 inclusively. (For further discussion see Tong [1990], esp. Section 7.3.2.) They had to wait for almost 8 years before they knew if they had done a reasonable job! Of course, this was only a small scale exercise involving only one time series, albeit over a long time on a human scale.

Our experience convinces us that we must maintain a healthy scepticism when presented with comparative results purporting to demonstrate the superiority of a new method and yet only "historical" data are used.

We are of the opinion that when a method does show genuinely promising performance, it is most important to understand why, because no method will produce promising predictions all the time under all conditions. For example, Table 4 of Yao and Tong [1995, p. 84] shows that we should expect poorer prediction performance for the sunspot numbers over the "peak" years. Therefore, any claim of excellent performance over the "peak" years would be suspect.

Another example is the use of the neural network approach reported in Weigend and Gershenfeld [1994]. For the out-of-sample prediction of the laser data of the competition, the neural network approach produces the best entry as well as (almost) the worst entry. As neural networks tend to involve vast numbers of parameters to be tuned, different tunings can produce vastly different qualities of prediction. There is clearly a need for some systematic guidance in the tuning of the parameters.

(iii) Multiple-step models vs single-step models

Suppose we assume that model (2.2) is the true model. (This assumption could well be discovered to be invalid at a later stage leading to the problem of mis-specification.) Then the m-step ahead least-squares prediction of X_{t+m} based on observations up to now is simply the conditional mean of X_{t+m} given X_t. We can then use either the Chapman-Kolmogorov relation or the Monte Carlo method to evaluate it. (See, e.g., Tong [1995]). It is important to note that if F is *nonlinear* then

$$E[X_{t+2}|X_t] = E[F(X_{t+1})|X_t] \tag{3.1}$$

$$\neq F(E[X_{t+1}|X_t])\,. \tag{3.2}$$

In other words, it is unwise to replace X_{t+1} by its 1-step ahead least-squares prediction obtained at time t in order to obtain the 2-step ahead least-squares prediction. The argument obviously applies to higher steps. The nonlinearity of F simply kills the commutativity of F and the conditional expectation operator. It is amazing how often people fall into this *commutativity-trap* and some of these are supposed to be nonlinear time series specialists! Linear prejudice dies hard.

Of course, we know that in practice F is rarely known and we replace it by its estimate, say $\hat{F}_1$. Typically, $\hat{F}_1$ is obtained as a result of minimizing an appropriate functional of the 1-step ahead predictors, say $\{X_{t+1} - \hat{F}_1(X_t),\ t = 1, \ldots, N-1\}$. Clearly $\hat{F}_1$ is no longer the true model even if it is so asymptotically. Another complication which often exists in practice arises because we may be using a *mis-specified* model. For example, we may be using the model class (2.2) whilst the true model class is (2.6). In this case, it is pertinent to ask the question: if we want to predict m-steps ahead, is it better to first of all fit a multiple-step model of the form

$$X_{t+m} = F_m(X_t) + \epsilon_{t+m}\ , \tag{3.3}$$

by minimizing an appropriate functional of the m-step ahead prediction errors, say $\{X_{t+m} - \hat{F}_m(X_t),\ t = 1, \ldots, N-m\}$, and then use $\hat{F}_m(X_t)$ as the m-step least-squares predictor of X_{t+m}? The comparison here is with the use of $\hat{F}_1$ of the single-step model coupled with the Chapman-Kolmogorov relation to produce an m-step least-squares prediction . (Of course, we will not commit the commutativity sin!) As far as we know, there is inadequate research on this question for the *nonlinear* case. However, R. L. Smith has presented some recent results at the Workshop reported in this volume. For the purely deterministic dynamical system, see, e.g., Farmer and Sidorowich [1987]. We conjecture that there is no uniform domination between the two approaches. In a sense the issue has to do with the robustness of $\hat{F}_1$.

For the linear case, there is some evidence which suggests that the multiple-step model may produce better prediction than the single-step model in *some* mis-specified situations. We summarize the findings of Tiao and Tsay [1994]. Specifically, let B denote the back-shift operator, namely $BX_t = X_{t-1}$. Suppose that the *true* model is

$$(1 - B)^d X_t = \epsilon_t\ , \tag{3.4}$$

where $\{\epsilon_t\}$ is a sequence of independent identically distributed random variables each with a $\mathcal{N}(0,1)$ distribution, and $-0.5 < d < 0.5$. (It is known that models of this form have long-memory in the sense that the autocorrelation function is not absolutely summable.) Suppose that we do not know what the true model is and use instead a mixed autoregressive/moving average model to fit the data generated by (3.4); the *mis-specified* model takes the form

$$(1 - \beta B)X_t = (1 - \eta B)b_t\ , \tag{3.5}$$

with $|\beta| < 1$ and $|\eta| < 1$. We refer to this model by the usual acronym ARMA (1,1). For this ARMA (1,1) model, the standard m-step ahead least-squares predictor, $\hat{X}_t(m)$, and its corresponding error, $\hat{e}_t(m)$, are

$$\hat{X}_t(m) = \begin{cases} \beta X_t - \eta b_t & \text{for } m = 1\ , \\ \beta^{m-1}\hat{X}_t(1) & \text{for } m > 1\ , \end{cases} \tag{3.6}$$

and

$$\hat{e}_t(m) = \begin{cases} X_{t+1} - \hat{X}_t(1) & \text{for } m = 1\,, \\ X_{t+m} - \beta^{m-1}(\beta X_t - \eta b_t) & \text{for } m > 1\,. \end{cases} \tag{3.7}$$

Clearly, minimizing $Var[\hat{e}_t(m)]$ would produce different "least-squares estimates" of β and η for different m. (In practice, we would minimize the sample version of $Var[\hat{e}_t(m)]$.) Let $\beta(m)$ and $\eta(m)$ denote the minimizers of $Var[\hat{e}_t(m)]$ with respect to β and η. Then the multiple-step model would substitute $\beta(m)$ and $\eta(m)$ for β and η respectively in equation (3.5) to yield the m-step prediction. On the other hand, the single-step model would substitute $\beta(1)$ and $\eta(1)$ for β and η respectively in equation (3.5) to yield the m-step prediction. Tiao and Tsay [1994] have given Table 1 below to compare their prediction performance.

Table 1

Comparison of prediction performance by a multiple-step model and a single-step model. Min(m) denotes the theoretical variance of the prediction error when the true model is used. (This represents the minimum achievable value of the prediction variance for the different m.) Mℓ(m) denotes the ratio of the prediction variance for the multiple-step model related to Min(m). (The ratio is always greater than 1; the smaller the ratio the better the performance of the multiple-step model.) Similarly, Sℓ(m) denotes the variance ratio when the single-step model is used.

d	m	Min(m)	M$\ell(m)$	$\beta(m)$	$\eta(m)$	S$\ell(m)$
.45	1	1.000	1.044	.965	.554	1.000
	2	1.203	1.045	.982	.737	1.019
	3	1.309	1.043	.988	.813	1.041
	4	1.380	1.042	.991	.857	1.059
	5	1.433	1.040	.993	.885	1.073
	6	1.475	1.040	.994	.904	1.086
	7	1.509	1.039	.995	.917	1.098
	8	1.538	1.038	.996	.928	1.110
	9	1.564	1.037	.996	.936	1.121
	10	1.586	1.037	.997	.942	1.133
	15	1.670	1.034	.998	.961	1.193
	20	1.726	1.033	.998	.971	1.258
	50	1.896	1.028	.999	.989	1.572
	100	2.013	1.024	≈ 1	.994	1.716
	200	2.122	1.024	≈ 1	.996	1.674

d	m	Min(m)	M$\ell(m)$	$\beta(m)$	$\eta(m)$	S$\ell(m)$
.25	1	1.000	1.011	.835	.609	1.000
	2	1.063	1.009	.915	.784	1.005
	3	1.087	1.007	.944	.854	1.009
	4	1.101	1.006	.959	.890	1.011
	5	1.110	1.006	.967	.912	1.013
	6	1.116	1.005	.973	.927	1.015
	10	1.131	1.004	.984	.956	1.021
	15	1.141	1.003	.989	.971	1.024
	20	1.146	1.003	.992	.978	1.024
	50	1.159	1.002	.997	.991	1.017
	100	1.165	1.001	.998	.996	1.012
	200	1.170	1.001	.999	.997	1.008

The table shows that (i) for d near to 0.5, the single-step model leads to poor multiple-step ahead predictions and a substantial gain can result using the multiple-step model, and (ii) more importantly, the penalty of mis-specifying the model is not serious if the multiple-step model is used.

(iv) To transform or not to transform?

Critics of instantaneous transformation of the data prior to modelling have argued that it is difficult to undo the transformation if we want to use the model fitted to the transformed data to predict the future values of the untransformed data, unless the transformation is linear. It has been shown that whilst it is true that a naive back-transformation will lead to bias in the prediction, the bias can be removed substantially by introducing a bias correction. (See, e.g., Tong [1990], Section 6.2.4.) Specifically, let $y = f(z)$ denote the 1-1 smooth transformation from z to y. Let g denote the inverse of f. Let $\hat{Z}_t(m)$ and $\hat{Y}_t(m)$ denote the least-squares m-step ahead prediction at time t of Z_{t+m} and Y_{t+m} respectively. Then we may correct the bias by observing that

$$\hat{Z}_t(m) \approx g(\hat{Y}_t(m)) + \tfrac{1}{2}g''(\hat{Y}_t(m))Var[Y_{t+m}|Y_tY_{t-1}, \ldots,]. \tag{3.8}$$

(Cf. for $Z = e^Y$ where $Y \sim \mathcal{N}(\mu, \sigma^2), EZ = \exp\{\mu + \frac{1}{2}\sigma^2\}$.) For an example of a successful application, see, e.g., Tong [1990], p.422. Therefore, we would argue that prediction consideration alone does not constitute valid grounds for objection to instantaneous transformations.

(v) Nearly the same goodness of fit need not mean nearly the same prediction

It is recognised that sometimes different classes of models can give almost equally good fit to the same data set. However, it might not be as widely recognised that their predictions can sometimes be dramatically different. In fact, models which fit the same data set equally well could, in some circumstances, lead to dramatically different predictions. It is perhaps instructive to think of the fitting model as part of the initial condition of a *super dynamical system*, whose output is the prediction; this super dynamical system may be sensitively dependent on its initial condition including the fitting model.

Cox and Medley [1989] give a particularly striking illustration in the short-term prediction of the AIDS epidemic in the United Kingdom, which we summarize here.

As the issue is quite complex, we shall have to leave out some of the details. The basic model used is one of a series of point events occurring in continuous time in a Poisson process of rate $\lambda(t)$, to be called the *incidence function*. The point events represent patients newly diagnosed as having AIDS. Once diagnosed, the case is notified to the Communicable Disease Surveillance Centre (CDSC) of the Public Health Laboratories of the U.K. with a time delay (of between a few weeks and, in extreme circumstances, two or more years), called the notification delay. Cox and Medley [1989] used the AIDS data supplied by the CDSC for all cases up to 1 July 1988. These data give the calendar months of diagnosis and of the arrival of the report to CDSC. The purpose is to predict the actual (not just the notified) number of AIDS patients at any one time.

Cox and Medley [1989] have considered three different parametric forms of $\lambda(t)$, each coupled to a notification delay distribution which takes the form of a mixture of two first order Gamma distributions. The three parametric forms of $\lambda(t)$ are

$$\lambda_1(t) = \exp(\rho_1 + \rho_2 t - \rho_3 t^2), \tag{3.9}$$

$$\lambda_2(t) = \rho_3/[1 + \exp(\rho_1 - \rho_2 t)], \tag{3.10}$$

and

$$\lambda_3(t) = (\rho_1 + \rho_2 t)/[1 + \rho_3 \rho_1 \exp(-\rho_4 t)]. \tag{3.11}$$

The notification delay distribution is of the form

$$\int [\theta_1 \theta_2^2 x e^{-\theta_2 x} + (1 - \theta_1)\theta_3^2 x e^{-\theta_3 x}] dx, \tag{3.12}$$

where the integral is over the appropriate month.

Table 2

Maximum log likelihood estimates and maximized log likelihoods, ℓ, for different incidence functions and notification delay

ρ_1	ρ_2	ρ_3	ρ_4	θ_1	θ_2	θ_2	ℓ
$\lambda_1(t)$							
−2.57	1.55	0.05	–	0.57	10.65	1.52	5353.58
$\lambda_2(t)$							
8.22	0.97	1519	–	0.57	11.36	1.49	5348.09
$\lambda_3(t)$							
70.99	162.06	13.57	0.82	0.57	11.35	1.52	5350.80

Table 2 shows that over the period of the data there is no great difference between the $\lambda(t)$'s, and the maximized log likelihoods are quite close.

Cox and Medley [1989] have shown that despite the similar goodness of fit of the three models the differences between their predictions are major, even after a year or so. In fact, differences are qualitatively very significant. To proceed further it is wise that subject matter considerations be included in order to throw some light on the more plausible model and hence prediction of what is clearly an extremely pressing problem in human terms. Cox and Medley [1989] have invoked epidemiological theory to suggest that $\lambda_3(t)$ is the most plausible. Note that a standard black-box modelling approach based on the principle of parsimony, e.g., the use of model selection criteria (such as Akaike's information criterion, Schwartz's

information criterion, Rissanen's minimizing description-length criterion, etc.) is here superceded by a substantive approach.

Acknowledgments

This paper was partially supported by the European Community grant ERB CHRX-CT 940693. I thank Professor Kung-sik Chan for discussion.

References

Cox, D. R. and Medley, G. F. [1989] *A process of events with notification delay and the forecasting of AIDS*, Philos. Trans. Roy. Soc. Lond., **B**, **325**, 135–45.

Farmer, J. D. and Sidorowich, J. J. [1987], *Predicting chaotic time series*, Phys. Rev. Lett., **59**, 845.

Ghaddar, D. K, and Tong, H. [1981], *Data transformation and self-exciting threshold autoregression*, Appl. Statist., **30**, 238.48.

Makridakis, S. and Hibon, M. [1979], *Accuracy of forecasting: an empirical investigation (with discussion)*, J. Roy. Statist. Soc., **A142**, 97–145.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. [1984] *The Forecasting Accuracy of Major Time Series Methods*. John Wiley, New York.

Tiao, G. C. and Tsay, R. S. [1994], *Some advances in nonlinear and adaptive modeling in time series analysis*, J. Forecasting, **13**, 109–31.

Tong, H. [1990], *Nonlinear Time Series*, Oxford: Oxford Univ. Press.

Tong, H. [1995], *A personal overview of non-linear time series analysis from a chaos perspective (with discussion)*, Scand. J. Statist., **22**, 399–445.

Weigend, A. S. and Gershenfeld, N. A. [1994], *Time series prediction: forecasting the future and understanding the past*, Proceeding volume XV, Santa Fe Inst. Studies in the Sciences of Complexity, New York: Addison-Wesley.

Yao, Q. and Tong, H. [1995], *On prediction and chaos in stochastic systems*, in Chaos and Forecasting–Proceedings of the Royal Society Discussion Meeting, (H. Tong, ed.) Singapore: World Scientific.