SERIE RESEARCH MEMORANDA

STOCHASTIC NONLINEARITY

A Firm Basis for the Flexible Functional Form

by

Aart F. de Vos
Ivo J. Steyn

Research Memorandum 1990-13

March 1990

vrije Universiteit amsterdam
Faculteit der Economische Wetenschappen en Econometrie
AMSTERDAM
STOCHASTIC NONLINEARITY
A FIRM BASIS FOR THE FLEXIBLE FUNCTIONAL FORM

Aart F. de Vos
Ivo J. Steyn

ABSTRACT
This paper deals with a basic problem: the relation between one explanatory variable $x$ and one dependent variable $y$:

$y_i = g(x_i) + u_i = g_i + u_i$ with $u_i \sim \text{NID}(0, \sigma^2_u)$

We propose a novel solution by specifying for $g(x)$ a parsimonious, flexible statistical model that corresponds with prior ideas about the nature of the uncertainty in specifying a function, permits rigorous statistical treatment and leads to robust, smooth estimates of $g$. The basis for this model is a stochastic specification in continuous $x$, by analogy to models specified in continuous time. In our preferred model, the slope of $g(x)$ follows a continuous random walk. The model may be estimated with a Kalman Filter. As a byproduct, a useful way to initialise the Kalman Filter is presented.


Free University,
Dept. of Economics,
De Boelelaan 1105,
1081 HV Amsterdam,
The Netherlands.
1. Introduction

This paper deals with a basic problem: the relation between one explanatory variable x and one dependent variable y:

\[ y_i = g(x_i) + u_i = g + u_i \quad \text{with} \quad u_i \overset{i.i.d.}{\sim} NID(0,\sigma_u^2) \]

Even within this simple context the specification of g is the subject of much discussion. The standard solution is the use of deterministic functions, but these are considered to be too restrictive in many applications. Alternatives like spline functions and kernel estimators are more flexible but use many parameters and entail messy statistical problems.

We propose a novel solution by specifying for g a parsimonious, flexible statistical model that corresponds in our view with prior ideas about the nature of the uncertainty in specifying g, permits rigorous statistical treatment and leads to robust, smooth estimates of g.

The general idea is to use a stochastic functional form, more specifically the form specified by

\[ \frac{d}{dx} \left[ \frac{dg(x)}{dx} \right] = ds(x) = \xi(dx) \quad \text{with} \quad \xi(dx) \overset{i.i.d.}{\sim} NID(0,\sigma^2 dx) \]

a notation known from continuous stochastic processes as reviewed in Bergstrom[1984], where \( \xi(dx) \) is a random measure, implying that for ordered values of x:

\[ \int_{x(i-1)}^{x(i)} \xi(dr) = \varepsilon_i \quad \text{with} \quad \varepsilon_i \overset{i.i.d.}{\sim} NID(0,\sigma^2(x_i-x_{i-1})) \quad \text{for all} \quad i \]

so the slope s(x) is a Brownian motion in x, its changes between two successive values x are independent and have a variance proportional to the

---

1 Many thanks are due to Herman Bierens and Ben Hanzon for stimulating discussions, and to Harro Merkus for providing superb computational assistance.
distance between the $x$ values. The implication of (2) for the curve $g(x)$ is that it has an unknown but small curvature at every point, while larger curvatures may occur but are unlikely, and equally unlikely at every point.

Models containing (2) as a special case were analyzed by Wymer[1972], but only in the situation where $x$ represents time and the observations are equidistant. In econometrics (Bergstrom[1984]), and also in system theory (Jazwinsky[1970]) continuous stochastic processes are almost exclusively treated in terms of time.

As continuous time processes are unfamiliar to most statisticians, and higher order processes like (2) even more so, we will give in section 2 a simple, rather informal, justification of the model following from (1) and (2). This derivation is based on the discrete analog of (2):

(4) $(1-L)^2 g(t) = c(t)$ or $\Delta^2 g_t = c_t$

This model is more familiar, it has been used by Gersch & Kitagawa[1983] for trend components in a structural time series model. Structural models, like (4) with superimposed errors, may be estimated by the Kalman filter, which also provides estimates of the components. Estimated trends that are specified as (4) are generally very smooth.

Assuming that (4) holds on intervals that are much shorter than the observation period, an analogous model in continuous time results, somewhat different from (4), but still implying smooth trends. This model can also be estimated using the Kalman filter. Moreover this model may be used when observations are not equidistant. And the same model may be used for functional forms, where $x$ values generally are not equidistant.

The idea to use stochastic functional forms instead of deterministic functions, splines or kernel estimators has a quite successful predecessor in time series analysis. Trend specifications based on deterministic functions, splines and other ad hoc procedures have been crowded out by stochastic difference equations: the ARIMA models. (see Box & Jenkins[1970], section A5.3.3, for a comparison between ARIMA models and an ad hoc procedure)

Next, structural time series models as advocated by Harvey[1984] have shown how fruitful it is to consider data as the result of the sum of different stochastic processes especially in combination with the recursive estimation
and smoothing procedures going back to Kalman[1960]. Application of the same ideas to functional forms thus may be seen as a logical step in a historical trend.

The structure of this paper is as follows. Section 2 shows in an informal way how, when x is time, (2) may be transformed to a statistical model for trends. It is a self-contained derivation, starting with the familiar expression (4), avoiding the use of random measures. Section 3 shows how such models look in the case where x is not time, and how they may be estimated using the Kalman Filter. Section 4 deals with statistical inference: interpolation and extrapolation of g, s, and y based on the model specified by (1) and (2), while Section 5 gives some examples. In section 6 our model is compared with spline and kernel estimators, while section 7 summarizes and speculates about possible extensions.

2. A smooth trend in discrete and continuous time

To reiterate, model (4) has been used by Gersch & Kitagawa to represent smooth trend curves underlying some time series. Examples of resulting estimates are also given in De Vos[1987]. In this section we show what happens when this model is specified for time intervals that are much smaller than the observation period. To proceed, we first rewrite (4) as :

\[(1-L)g_t = s_{t-1}\]
\[(1-L)s_t = e_t\]

where \(s_t\) is the slope of the trend curve at time \(t\).

Next we may wonder how, \(g_t\) and \(s_t\) evolve during \(k\) time periods. (5a)-(5b) implies :

\[
g_{t+k} = g_t + ks_t + \sum_{j=t+1}^{t+k} e_j \quad (k>1)
\]

\[
s_{t+k} = s_t + \sum_{i=t+1}^{t+k} e_i \quad (k>1)
\]

We may rewrite this as :

\[
E[\eta_{t+k}^2] = \sigma_\epsilon^2 k(k-1)(2k-1)/6 = \sigma_\epsilon^2 k^3/3
\]

\[
E[\zeta_{t+k}] = \sigma_\epsilon^2 k
\]

\[
E[\zeta_{t+k} \eta_{t+k}] = \sigma_\epsilon^2 k(k-1)/2 = \sigma_\epsilon^2 k^2/2
\]
For large \( k \) the approximations are relatively accurate. The idea behind a specification in continuous time is to consider the evolution of \( g \) and \( s \) between two successive data points as the result of a large number - say \( k \) - of stochastic changes, which justifies the use of the approximations. This makes sense if it is irrelevant how large \( k \) is, and this appears to be the case.

To see this, suppose another metric such that \( b \) instead of \( k \) steps divide two data points, with disturbances having a different variance, \( \sigma^2 \). The slope in the new metric, \( s^*_t \), is different: \( s^*_t = ks_t/b \). Substitution of these values in (7a) and (7b), using the approximations, learns that exactly the same model for the evolution of \( g_t \) between two data points results if \( \sigma^2 = \sigma^2 k^3/b^3 \). Thus, as \( \sigma^2 \) is a parameter, one is free to choose the metric, as long as one keeps in mind that the approximations only make sense in very fine metrics. In Appendix A it is shown that the approximations are the exact result of a derivation in continuous time.

The step to continuous time is now a small one. We simply allow \( k \) to take any value, and slightly rewrite the model as:

\[
\begin{align*}
(8a) \quad g_t &= g_{t-k} + ks_{t-k} + \eta_t \quad E[\eta_t^2] = \sigma^2 k^3/3 \\
(8b) \quad s_t &= s_{t-k} + \zeta_t \quad E[\zeta_t^2] = \sigma^2 k \\
\end{align*}
\]

In a discrete time context the obvious metric is \( k=1 \). Note that after this substitution (8a)-(8b) is not the same model as (5a)-(5b). It is thus an alternative model for smooth trend curves. More generally (8a)-(8b) implies for equidistant observations at distance \( k \):

\[
(9) \quad (1-L)g_t = k\zeta_t + \eta_t - \eta_{t-k} = \xi_t + \xi_{t-k}
\]

an ARIMA(0,2,1) process. The parameters \( \theta \) and \( \text{var}(\xi) \) follow by equating the variance and autocovariance of the latter two expressions:

\[
\begin{align*}
(1+\theta^2)\sigma^2 &= k^2E[\zeta_t^2] + 2E[\eta_t^2] - 2kE[\zeta_t\eta_t] = 2\sigma^2 k^3/3 \\
\phi \sigma^2 &= kE[\zeta_t\eta_t] - E[\eta_t^2] = \sigma^2 k^3/2 \\
\end{align*}
\]

The invertible root of the MA part thus follows from \((1+\theta^2)/\theta = 4\), so \( \theta = 2-\sqrt{3} \). The fact that this does not depend on \( k \) confirms that this model does not depend on the metric chosen.

Summarizing, model (8a)-(8b) gives rise to the following specification for smooth trend curves:
(10) \[(1-L)^2g_t = \xi_t + (2-\sqrt{3})\xi_{t-1}\]
a result similar to one from the formal theory of stochastic processes in continuous time, derived by Wymer[1972]. The trend specification (10) has the advantage that it is essentially independent of the observation frequency. This should hold in view of (9), and in Appendix B it is proved that (10) is the only ARIMA(0,2,1) model with this property.

An important advantage of the specification is that it also may be used when observations are not equidistant. Choosing the metric in which time is measured, (8a)-(8b) may be written as:

\[
\begin{align*}
(11a) \quad g_{t1} &= g_{t0} + (t_1 - t_0)s_{t0} + \eta_{t1} \\
E[\eta^2_{t1}] &= \sigma^2(t_1 - t_0)^3/3 \\
(11b) \quad s_{t1} &= s_{t0} + \zeta_{t1} \\
E[\zeta^2_{t1}] &= \sigma^2(t_1 - t_0) \\
E[\eta_{t1}\zeta_{t1}] &= \sigma^2(t_1 - t_0)^2/2 
\end{align*}
\]

This specification, together with a measurement equation (and in realistic time series models with other components like AR and a seasonal term):

\[
\begin{align*}
(11c) \quad y_{t1} &= g_{t1} + u_{t1} \\
E[u^2_{t1}] &= \sigma^2 \quad \forall \, l
\end{align*}
\]

may be cast directly in the Kalman filter form, as is shown in the next section for the more general case where \( x \) is not necessarily time.

3. A fundamentally flexible functional form

With the yoke of equidistance shed, we may leave the realm of time series, and write the model following from (2) as:

\[
\begin{align*}
(11a') \quad g_x &= g_{x_{l-1}} + (x_l - x_{l-1})s_x + \eta_x \\
E[\eta^2_x] &= \sigma^2(x_l - x_{l-1})^3/3 \\
(11b') \quad s_x &= s_x + \zeta_x \\
E[\zeta^2_x] &= \sigma^2(x_l - x_{l-1}) \\
E[\eta_x\zeta_x] &= \sigma^2(x_l - x_{l-1})^2/2 
\end{align*}
\]

the only stipulation being that the \( x_l \) are ordered observations of some indexing variable. Adding the disturbance term from (1) and dropping the awkward double indexing we arrive at the model we shall employ from now on:

\[
\begin{align*}
(12a) \quad y_1 &= g_1 + u_1 \\
E[u^2_1] &= \sigma^2 \\
(12b) \quad g_i &= g_{i-1} + (x_i - x_{i-1})s_{i-1} + \eta_i \\
E[\eta^2_i] &= \sigma^2(h(x_i - x_{i-1})^3/3 \\
(12c) \quad s_i &= s_{i-1} + \zeta_i \\
E[\zeta^2_i] &= \sigma^2 h(x_i - x_{i-1}) \\
E[\eta_i\zeta_i] &= \sigma^2 h(x_i - x_{i-1})^2/2 
\end{align*}
\]
A few words of explanation. h is the ratio between the systemic variance and the measurement noise variance ($\sigma_e^2/\sigma^2$). All disturbance terms are assumed to have expectation 0 and to be uncorrelated apart from the specified correlation between $\eta_k$ and $\xi_j$ if $k=l$.

Model (12) describes a stochastic functional relation between an explanatory variable $x$ and a dependent variable $y$, just like model (11) describes a series with a smooth time trend and disturbances: the relation between the indexing variable and the dependent variable is assumed to be locally linear, with nonzero curvatures less likely as they grow larger. In the time series case, (11) produces smooth trend curves; in the functional form case (12) produces smooth graphs in $xy$-space.

Obviously, (12a)-(12c) is just one member of a large class of models. We might also have taken an I(3) discrete model as our starting point, and ended up with a model where the relation between $x$ and $y$ is locally parabolic. However, (12) is already a very flexible generalization of the regression model, as will become clear from the examples, and we believe that it will be the most important possibility for practical purposes.

The estimation of (12) proceeds just like that of its time series ancestor: with the Kalman Filter. The model can be written in State Space form:

\begin{align}
(13a) \quad y_i &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} g_i \\ s_i \end{bmatrix} + u_i = Z \alpha_i + u_i \\
(13b) \quad \alpha_i &= \begin{bmatrix} 1 & x_i-x_{i-1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} g_{i-1} \\ s_{i-1} \end{bmatrix} + \begin{bmatrix} \eta_i \\ \xi_i \end{bmatrix} = T \alpha_{i-1} + \omega_i \\
\end{align}

where the form of $Q_i$ follows from (12a)-(12c). Maximum likelihood estimation by the Kalman Filter is straightforward, and only one parameter needs to be estimated numerically: $h$, the signal/noise ratio. After initialization (see Appendix C for an initialization method which does not involve numerical difficulties) the recursive filter gives, for any value of $h$, the "prediction errors" and their variances, with $\sigma^2$ as a scaling factor. In itself this use of the prediction error decomposition in functional forms is remarkable: it shows how the two sources of uncertainty are combined in a natural way. The maximum likelihood estimate of $\sigma^2$ conditional upon $h$ is standard. Alternatively Bayesian methods may be used; the posterior distribution of $h$
and $\sigma^2$ then allows interesting unconditional inference. Tests of the interesting hypothesis $h=0$, which implies the linear regression model, may be based on the likelihood ratio $^2$, information criteria like those of Akaike or Schwarz, or Bayesian posterior odds.

For given parameters (we will employ the maximum likelihood estimates) standard smoothing algorithms (see Appendix) provide estimates of $g_i (= g(x_i))$ and its slope $s_i$ at every point $x_i$ conditional upon all information. Also the variances and covariances result. These provide the information for further inference, which is covered in the next section.

We can summarize the course and results of the Kalman Filter estimation as follows:
1. Compute numerical maximum likelihood estimates of $h$ and $\sigma^2$
2. Run data through a final Kalman Filter with $h = \hat{h}$.
3. Smooth the estimates.

4. Interpolation and extrapolation

Inference from functional forms is much more challenging than inference in the time domain. The estimates of a historical trend are of dubious importance, forecasts are always extrapolations. Forecasts of $y$ conditional upon $x$ in functional forms may concern interpolation and extrapolation at both sides. Also, direct inference on $g(x)$ or $s(x)$ may be important, e.g. when $g(x)$ stands for the effects of drugs on health or the effect of inflation on economic growth. All these types of inference may be based on the results of the Kalman filter, which are smoothed estimates of level and slope of $g(x_i)$ for all $i$, and their variances and covariances.

To address the problem of interpolation first, assume we wish to estimate the curve between two points $x_0$ and $x_1$. We have estimates of $g_0$, $s_0$, $g_1$ and $s_1$.

Consider the simultaneous distribution of $g_{\lambda} = g(x_{\lambda})$, $g_1$ and $s_1$, conditional on $g_0$ and $s_0$, where $0 \leq \lambda \leq x_1 - x_0$. Let $\Delta x = x_1 - x_0$. We have:

---

2 The models are nested, and, due to our initialization procedure (outlined in Appendix C) the likelihood and parameter estimates from a regression, and those from the Kalman Filter coincide.
(14) \[ g_\lambda = g_0 + \lambda s_0 + \eta_\lambda \]
\[ s_\lambda = s_0 + \zeta_\lambda \]

var(\eta_\lambda) = \sigma^2h\lambda^3/3

var(\zeta_\lambda) = \sigma^2h\lambda

cov(\eta_\lambda, \zeta_\lambda) = \sigma^2h\lambda^2/2

(15) \[ g_1 = g_\lambda + (\Delta x - \lambda)s_\lambda + \eta_{\lambda 1} \]
\[ s_1 = s_\lambda + \zeta_{\lambda 1} \]

var(\eta_{\lambda 1}) = \sigma^2h(\Delta x - \lambda)^3/3

var(\zeta_{\lambda 1}) = \sigma^2h(\Delta x - \lambda)

cov(\eta_{\lambda 1}, \zeta_{\lambda 1}) = \sigma^2h(\Delta x - \lambda)^2/2

By substitution we obtain :

(15) \[ g_1 = g_0 + \Delta x s_0 + (\Delta x - \lambda )\zeta_\lambda + \eta_{\lambda 1} + \eta_\lambda \]

which leads to :

(16) \[ \begin{bmatrix} g_1 \\ s_1 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} g_0 + \lambda s_0 \\ s_0 \end{bmatrix}, \sigma^2h \begin{bmatrix} \frac{1}{3}\lambda^3 + \frac{1}{2}\lambda^2(\Delta x - \lambda) & \frac{1}{2}\lambda^2 \\ \frac{1}{3}(\Delta x)^3 & \frac{1}{2}(\Delta x)^2 \end{bmatrix} \right) \]

so

(17) \[ \mathbb{E}[g_1 | g_0, s_1, s_0] = g_0 + \lambda s_0 + \left[ \frac{1}{3}\lambda^3 + \frac{1}{2}\lambda^2(\Delta x - \lambda) \right] \left[ \frac{(\Delta x)^3}{3} (\Delta x)^2/2 \Delta x \right]^{-1} \begin{bmatrix} g_1 - g_0 - \Delta x s_0 \\ s_1 - s_0 \end{bmatrix} = \]
\[ g_0 (2\lambda + \Delta x)(\lambda - \Delta x)^{2/3} + \]
\[ g_1 \lambda^2(2(\Delta x - \lambda) + \Delta x) / (\Delta x)^3 + \]
\[ s_0 \lambda(\lambda - \Delta x)^{2/3} + \]
\[ s_1 \lambda^2(\lambda - \Delta x) / (\Delta x)^2 \]

(18) \[ \text{var}[g_1 | g_0, s_1, s_0] / (\sigma^2h) = \]
\[ \lambda^3/3 - \left[ \frac{1}{3}\lambda^3 + \frac{1}{2}\lambda^2(\Delta x - \lambda) \right] \left[ \frac{(\Delta x)^3}{3} (\Delta x)^2/2 \Delta x \right]^{-1} \begin{bmatrix} \frac{1}{3}\lambda^3 + \frac{1}{2}\lambda^2(\Delta x - \lambda) \\ \frac{1}{2}\lambda^2 \end{bmatrix} = \]
\[ \frac{1}{3}\lambda^3(\Delta x - \lambda)^3 / (\Delta x)^3 \]

Note that the conditional mean is just a weighted average of the four items of information, as expected, and that the conditional variance is 0 for \( \lambda = 0 \) and \( \lambda = \Delta x \), and symmetric around \( \lambda = \Delta x / 2 \).

From (17)-(18) the conditional distribution of \( g_\lambda \) may be written as
\( g_\lambda \sim N(c_\lambda^a, \sigma_\lambda^2) \)

with \( \sigma_\lambda^2 \) the expression from (18), \( c_\lambda \) a 4-vector with as its elements the four polynomials from (17), and \( a = [g_0^1, g_0^1, g_0^1, s_0^1]' \). In reality however, we work with the smoothed estimate of \( a \), derived from the Kalman Filter. Since

\[
f(g_\lambda, g_0^1, s_0^1, s_0 | \text{data}) = f(g_\lambda | g_0^1, g_0^1, s_0^1, s_0) \cdot f(g_0^1, g_0^1, s_0 | \text{data})
\]

we use \( f(g_\lambda | g_0^1, g_0^1, s_0^1, s_0) = f(g_\lambda | g_0^1, g_0^1, s_0^1, s_0) \cdot f(g_0^1, g_0^1, s_0 | \text{data}) \) to get

\[
f(g_\lambda, g_0^1, g_0^1, s_0^1, s_0 | \text{data}) = f(g_\lambda | g_0^1, g_0^1, s_0^1, s_0) \cdot f(g_0^1, g_0^1, s_0 | \text{data})
\]

The distribution of \( g_\lambda \) given the data can then be found by integrating \( g_1^0, g_0^1, s_1^0, s_0^1 \) out of the above distribution. The end result is that:

\[
g_\lambda | \text{data} \sim N(c_\lambda^\hat{a}, \sigma_\lambda^2 + c_\lambda^TP_\lambda a_\lambda)
\]

with \( \hat{a} \) the smoothed estimate of \( a \) and \( P_\lambda \) its covariance matrix. Note that we do not only require the variance of the smoothed estimates, but also the covariance of each pair of successive state vectors. These are obtained by defining a new state vector consisting of the old state vector from (13a)-(13b), and that vector from the previous observation and using this augmented state during the Kalman Filter calculations. During the smoothing phase we then get smoothed estimates of the states, the variance matrices and the covariance of the smoothed estimates of each pair of successive states, just enough to calculate (19). An alternative procedure, which does not involve doubling the dimension of the State vector is described in de Vos & Merkus[1990].

By comparison, extrapolation is far simpler. Let \( x_n \) be the largest observation on \( x \) we had. To extrapolate the curve to values of \( x \) larger than \( x_n \), we note that:

\[
g_x = g_n + (x-x_n)s_n + \eta_x \quad \text{var}\{\eta_x\} = \sigma_x^2 h(x-x_n)^3/3
\]

By a similar argument to that used for interpolation, we get:

\[
E[g_x] = \hat{g}_n + (x-x_n)\hat{s}_n
\]

\[
\text{var}\{g_x\} = \frac{1}{3}\sigma_x^2 h(x-x_n)^3 + \text{var}\{\hat{g}_n\} + (x-x_n)^2\text{var}\{\hat{s}_n\} + 2(x-x_n)\text{cov}(\hat{g}_n, \hat{s}_n)
\]

where the variance and covariance terms are again those of the smoothed estimates of \( g_n \) and \( s_n \). Note that the variance increases as the third power of \( x \). Extrapolation to the left of \( x_n \) proceeds in a similar fashion.
It is worth noting that the expressions for the variance in (19) and (21b) are overly optimistic. Our inference is conditional on the maximum likelihood estimate for \( h \). Bayesian methods may be used to produce unconditional inference, or a procedure pointed out by Hamilton[1986] in a non-Bayesian framework may be used.

5. Some illustrative examples

In this Section we will show how model (12a)-(12c) performs in a number of examples. To give a first impression we generated 3 x 100 data points with (12a)-(12c) for given \( x \), using parameters \( h=100 \) and \( \sigma^2=0.0001 \). Fig.1a shows the first batch of generated data points, the true underlying curve \( g(x) \) (anchored by '+' symbols) and the estimated curve. The parameter estimates were \( \hat{h}=236 \) and \( \hat{\sigma}^2=0.00056 \). The outermost curves are the boundaries of a 95% confidence interval for inter- and extrapolated data (\( y_1 \)). In fig.1b, the true curve \( g(x) \) is shown again, together with the estimated curve and confidence intervals for that curve. Finally, in fig.1c the true slope \( s(x) \), the estimated slope \( \hat{s}(x) \) and its confidence intervals are shown. Note that the slope of the apparently smooth curve \( g(x) \) is not smooth itself: it follows a random walk. The estimated slope, by comparison, is far smoother.

In example 1, the curve that underlied the data could be estimated with surprising success. By contrast, in example 2 \( g(x) \), though simulated with the same parameters, happens to show a pattern that is much more obscured by measurement noise. As a result, the estimated curve is a simplified version of the true curve: more simply couldn't be extracted from the polluted data. Still, Figs. 2a-2c, similar to figs. 1a-1c in subject, do not reveal great shortcomings in the inferences to be made. Note particularly how cautious the inferences are in extrapolations and even in interpolations near the boundaries of the observations. Figs.3a-3c show a more volatile curve underlying the data being reconstructed with impressive accuracy.

Cautiousness in the form of wide intervals might also be expected between data points that are wide apart (see the formulae in section 4) but this effect is negligible in the first three examples. In our last example, we generated a dataset with a fairly dense cluster of \( x \)-values and one "\( x \)-outlier". Fig.4 shows the dramatic effects on the reliability of interpolation when the data points are far apart.
6. A comparison with kernel and spline estimators

Drawing smooth lines that claim to represent a continuous relation between y and x, observed with disturbances, has been the subject of many articles.

Spline estimators, dating back to Whittaker(1923), describe g(x) as a set of cubic polynomials, linked at "knots" in such a way that the first two derivatives are continuous. For each interval between knots one free parameter results. The main problem with this setup is its hybrid nature: deterministic functions, changing at points to other deterministic functions cannot be a good description of uncertainty about the course of a function as long as these knots have no natural interpretation. Poirier (1973) mitigates this problem by formulating probabilities that points are knots, but this complicates the analysis severely without offering a new interpretation of the nature of the knots.

Kernel estimators are much more popular nowadays. They go back to Rosenblatt(1956), Parzen(1962), Nadaraya(1964) and Watson(1964). Kernel estimators relate the estimate of g(x) to a set of surrounding observations of y. The kernel describes the functional form of this relation, and most methods also contain an extra degree of freedom of description, the "band width", which regulates how many points are considered. Härdle(1988) gives an up to date treatment with hundreds of references. His subject is "nonparametric regression", a name induced by the fact that the methods described do not correspond to any known parametric model. In fact, the models are highly parametric, but only implicitly.

Kernel estimators give results similar to ours, but:

- there is no satisfying treatment of points near the endpoints of the observations, and of extrapolations. The choice of a specific method generally requires subjective assessment of the results and is thus only partly empirical.

- probability statements on forecasts of y for given x may at best be derived from asymptotic distributions or by using confidence bands of which the only known property is that they cover the true curve at least with probability 1-α.

- Bayesian methods, answering the more relevant question "what probability statements may be made about the curve and forecasts, given the data?" cannot be used as the likelihood is not known.
testing of various methods against each other is not possible on generally accepted grounds.

- it is very difficult to understand how the methods incorporate prior ideas about smoothness of the underlying curve.

Most of these problems also apply to spline functions, but in a more indirect way. Our method, as opposed to these features in reverse order:

- starts from the notion that an unknown smooth curve may bend at each point in a nondeterministic way.
- uses only one parameter to describe the degree of smoothness.
- allows testing of competing specifications with likelihood-based and Bayesian tests.
- is suited for Bayesian estimation, though specifying priors for the hyperparameter is far from trivial.
- provides uniform confidence intervals, either conditional upon one parameter (if the maximum likelihood estimate is used) or conditional upon a possibly noninformative prior distribution for the smoothness parameter (if Bayesian intervals are used)
- requires little subjective judgement due to the parsimonious parameterization. Moreover the subjective judgement may be formulated a priori, instead of during the inspection of the resulting pictures (in the Bayesian view a mortal sin).
- treats interpolation and extrapolation of the function, its derivative and new data in one, coherent, way.

Whether these arguments convince depends on the statistical paradigms one believes in. Our basic attitude is a Bayesian one. From a Bayesian viewpoint the only thing that matters is whether one believes the model to be a good representation of prior knowledge. In combination with the possible extensions treated in the next section, we think any reasonable type of prior knowledge may be incorporated into our model. These extensions also provide the context for studies within the maximum likelihood paradigm. Asymptotic properties or simulation studies cannot add anything within the Bayesian paradigm. That many points, reasonably distributed on a restricted interval for $x$, will in general lead to effective inference on $g(x)$ is obvious, and illustrated by our examples. These illustrate as well that some
cases are more difficult than others (example 2). The rate of convergence of
the estimated function towards the real function - much studied by
frequentists in the context of kernel estimation - is of little help in such a
situation with a limited amount of data. Simulation studies may only reveal
what comes out of the procedures for given $g(x)$, and not what may be said
about $g(x)$ for given data.
It will take many studies in which real data are used to make forecasts to
reveal which procedure is best in a more objective sense. That may take some
time, until then we will have to make do with subjective arguments.

7. Conclusions, extensions, and further perspectives.
We have shown that the relation between one dependent and one independent
variable may be described in a very flexible way by assuming that the slope of
the functional relationship follows a Brownian motion. The resulting model may
be estimated by using the Kalman filter and all Bayesian and likelihood-based
types of statistical inference are possible in this context.
A straightforward extension results by considering a pre-transformation of the
explanatory variable. Taking logarithms, or using the Box-Cox transformation
(involving one new parameter) of $x$ before applying our algorithm changes the
model. The likelihoods of the different models may be compared directly. This
option thus drastically extends the class of accessible functional forms.
Transformations of $y$ may also be considered, and compared using the Jacobian
of the transformation.
Other possibly important generalizations require further research. In the
first place, a different stochastic model could be formulated for $g(x)$. An
interesting and easily implemented example is the addition of a constant to
the equation governing the evolution of the slope. This leads to a class of
locally parabolic functions. Another idea is the use of the third derivative
rather than the second as the stochastic basis for the model. Such simple
alterations imply wholly different assumptions about the underlying curve
$g(x)$. A random walk for the second derivative probably defines a class of
functions that is much too volatile, but combined with prior information an
attractive class may result. A third idea to generalise the specification of
$g(x)$ is to specify the smoothness parameter $h$ as a random walk, allowing
ARCH-like changes in volatility.
In the second place assumptions on the nature of the measurement error may be changed. The use of distributions with fat tails for instance may meet the presence of outliers. Unfortunately the standard Kalman filter does no longer apply in that case. This considerably complicates matters. Another useful feature may, if the data concern time processes, be the addition of correlation structures in time for the measurement error. This too complicates matters; the strictly recursive relations of the Kalman Filter no longer apply. Finally, generalization to more explanatory variables seems highly desirable. This is much more difficult as it is no longer possible to order the observations. So far, this extension has not proven amenable to a tractable estimation method.

If the historical trend is any indication, these problems will be solved in time, although, as shown in Section 5, one must be careful with extrapolations.
APPENDIX A

**Derivation of the recursive relations in continuous x**

In continuous x (time in the literature) the random measure $\xi(dx)$ may be defined (see Bergstrom[1984]) such that $\int_0^x \xi(du)$ is normally distributed with

$$E[\xi(du)] = 0 \quad \text{and} \quad E[\int_0^x \xi(du)\xi(dv)] = \sigma^2(x-x_0)$$

Thus the model (2), which may be written

$$d(g'(x)) = \xi(dx)$$

implies

$$g'(x) = g'(x_0) + \int_0^x \xi(du) = g'(x_0) + \zeta$$

$$g(x) = g(x_0) + \int_0^x g'(x) = g(x_0) + (x-x_0)g'(x_0) + \int_0^x \xi(dw)du$$

Thus $g(x) = g(x_0) + (x-x_0)g'(x_0) + \eta$

with

$$\text{Var}(\zeta) = \sigma^2(x-x_0)$$

$$\text{Var}(\eta) = \int_0^x \int_0^x \int_0^z \xi(dw)\xi(dv)dzdu$$

$$= \sigma^2 \int_0^x \int_0^z (\text{min}(u,z)-x_0)dzdu = 2\sigma^2 \int_0^x \int_0^z (u-x_0)dzdu$$

$$= \sigma^2 \int_0^x (x-x_0)^2dz = \sigma^2(x-x_0)^3/3$$

$$\text{Cov}(\zeta,\eta) = \int_0^x \int_0^x \int_0^z \xi(dw)\xi(dz)du$$

$$= \sigma^2 \int_0^x (u-x_0)du = \sigma^2(x-x_0)^2/2$$

These are the expressions used in model (8) where $x=t$ and in model (12).
APPENDIX B

The ARIMA(0,2,1) model with MA-parameter $2-\sqrt{3}$ is the only one of its class to remain unchanged when the observation frequency is halved.

Proof

An ARIMA(0,2,1) model is:

$$(1-L)^2y_t = \zeta_t + \phi \zeta_{t-1}$$

When the observation frequency is halved, we must consider:

$$(1-L)^2y_t = (1+L)^2(1-L)^2y_t = (1+L)^2(1+\phi L)\zeta_t = w_t$$

Now

$$E[w_t^2] = [1 + (2+\phi)^2 + (1+2\phi)^2 + \phi^2]\sigma^2$$

$$E[w_t w_{t-2}] = [1 + 2\phi + \phi(2+\phi)]\sigma^2$$

For this to be the same model in halved observation frequency as the original ARIMA(0,2,1) model,

$$E[w_t^2]/E[w_t w_{t-2}] = (1+\phi^2)/\phi$$

Substituting gives:

$$(1-2\phi+\phi^2)(1-4\phi+\phi^2) = 0$$

so $\phi = 2-\sqrt{3}$ is the only root corresponding to an invertible process.

APPENDIX C

Initialization of the Kalman Filter

The following question is central to our method of initialization:

Does the State $\alpha_1$ have a well-defined distribution conditional on $y_1 .. y_1$?

As an example, assume we have one observation. Since

$$y_1 = Z\alpha_1 + u_1$$

$\alpha_1$ will only have a well-defined distribution if $Z$ is of full column rank.

After 2 observations, we can summarize our information about $\alpha_2$ as follows:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} ZT_2^{-1} \\ Z \end{bmatrix} \alpha_2 + \begin{bmatrix} u_1 - ZT_2^{-1} \omega_2 \\ u_2 \end{bmatrix} = \mathbb{E}_2 \alpha_2 + \psi_2$$

where we have used the fact that $\alpha_2 = T_2 \alpha_1 + \omega_2$. If the matrix $\mathbb{E}_2$ is of full column rank, then we can write:
The conditional distribution of $\alpha_2$, given $y_1$ and $y_2$, follows from this:

$$
\alpha_2 \sim N\left( \left( \Sigma_z \Sigma_z \right)^{-1} \Sigma_z \left[ \begin{array}{c} y_1 \\ y_2 \\ y_3 \\ Z \\ Z \end{array} \right], \left( \Sigma_z \Sigma_z \right)^{-1} \Sigma_z \Sigma_z \left( \Sigma_z \Sigma_z \right)^{-1} \right)
$$

where $\Sigma_z$ is the variance matrix of $\psi_z$. These expressions simplify considerably when $\Sigma_z$ is not only of full column rank, but regular.

If $\Sigma_z$ still isn't of full column rank, we involve yet a third observation, and write:

$$
\left[ \begin{array}{c} y_1 \\ y_2 \\ y_3 \\ Z \\ Z \\ Z \end{array} \right] = \left[ \begin{array}{c} Z^T \omega_3 \\ Z^T \omega_3 \\ Z^T \omega_3 \\ Z \end{array} \right] \alpha_3 + \left[ \begin{array}{c} u_1 - Z^T \omega_3 \\ u_2 - Z^T \omega_3 \\ u_3 \end{array} \right] = \Sigma_z \alpha_3 + \psi_3
$$

If $\Sigma_z$ is of full column rank, then we proceed as above, etc.

In this way we proceed to increase the number of observations on which we condition, until, for some $i$, $\Sigma_i$ is of full column rank. We can then calculate a well-defined distribution of $\alpha_i$, conditional on $y_1, y_i, y_2$. It is interesting to note that, if the State vector is of dimension $k$, then we will need exactly $k$ observations to get a $\Sigma_k$ of full rank if and only if the model satisfies a condition known as observability in the systems theoretical literature. In general, $i \geq k$.

In our case, 2 observations will generally suffice, but only if $x_1 \neq x_2$. In that case, we can write:

$$
\left[ \begin{array}{c} y_1 \\ y_2 \\ y_3 \\ Z \\ Z \end{array} \right] = \left[ \begin{array}{c} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{array} \right] \left[ \begin{array}{c} g_2 \\ s_2 \\ s_2 \\ g_2 \\ g_2 \end{array} \right] + \left[ \begin{array}{c} u_1 \\ u_2 \\ u_2 \\ u_1 \\ u_1 \end{array} \right] \text{ with } \omega_2 = \left[ \begin{array}{c} \eta_2 \\ \zeta_2 \end{array} \right]
$$

from which follows:

$$
\left[ \begin{array}{c} g_2 \\ s_2 \\ s_2 \\ g_2 \\ g_2 \end{array} \right] = \Delta X \left[ \begin{array}{c} 0 \\ 0 \\ 1 \\ -1 \\ -1 \end{array} \right] \left[ \begin{array}{c} y_1 \\ y_2 \\ y_3 \\ Z \\ Z \end{array} \right] = \left[ \begin{array}{c} 0 \\ 0 \\ 1 \\ -1 \\ -1 \end{array} \right] \left[ \begin{array}{c} u_1 \\ u_2 \\ u_2 \\ u_1 \\ u_1 \end{array} \right] \Delta X \left[ \begin{array}{c} \eta_2 \\ \zeta_2 \end{array} \right]
$$

Or, put more simply:

$$
\left[ \begin{array}{c} g_2 \\ s_2 \\ s_2 \\ g_2 \\ g_2 \end{array} \right] = \left[ \begin{array}{c} y_2 \\ y_2 \\ y_2 \\ y_2 \\ y_2 \\ y_2 \end{array} \right] - \frac{1}{\Delta X} \left[ \begin{array}{c} \Delta X u_2 \\ \Delta X u_2 \\ \Delta X u_2 \\ \Delta X u_2 \\ \Delta X u_2 \end{array} \right]
$$

As initialization we therefore use the following:
This method avoids all numerical and methodological problems involved with so-called diffuse initial conditions (see Steyn[1989]). An equivalent alternative would be to start with a precision matrix $0$ and use a precision filter until the precision matrix becomes invertible, as advocated by Maybeck[1979,1982].
A time-series of Total Accounts for the Netherlands 1978-1984

Self-Service Activities and Legal or Illegal Market Services

The Monetary Order

Price Rigidities and Rationing

A Simple Throughput Bound for Large Closed Queuing Networks with Finite Capacities

Analytic Error Bounds for Approximations of Queuing Networks with an Application to Alternate Routing

Selfexciting Counting Process Systems with Finite State Space

Rational Expectations and New Classical Macroeconomics

De Nederlandse Boekenmarkt tussen Stabiliteit en Verandering

Traditional "Social Security Systems" and Socio-economic Processes of Changes: The Chase of Swaziland: opportunities for research

"Stop = Recirculate" for Exponential Product Form Queuing Networks with Departure Blocking

Modelbouw en matigingsbeleid in Nederland

Simple performance estimates and error bounds for slotted Aloha loss systems

Sugar Crisis, a Comparison of two Peripheral Economies

Consistent Diffuse Initial Conditions in the Kalman Filter

Als Estimation of Parameters in a State Space Model

Dynamic Interrelationships between Spot Prices of some Agricultural Commodities on Related Markets

Zoeken naar nieuwe medewerkers

Integration of Environmental Externalities in International Commodity Agreements

Micro Labour Demand Functions with Heterogeneous Output for Dutch Housing-Construction

An empirical Analysis of Employers' Search

Product Forms for Queuing Networks with State Dependent Multiple Job Transitions

On "stop = repeat" Servicing for Non-Exponential Queuing Networks with Blocking

An Empirical General Equilibrium Model for the Spatial Interactions of Supply, Demand and Choice

Kansen en risico's - Over de fundamenten van statistische uitspraken door accountants

A Note on Extended Uniformization for Non-Exponential Stochastic Networks

Cortadores de Cosecha en Tres Regiones Cafetaleras en Nicaragua (1960-81)

Exact Solutions for Central Service Systems With Breakdowns

Product Forms for Queuing With Limited Clusters

Tijdsindeling van huishoudelijke activiteiten in relatie tot kenmerken van huishoudens

An Empirical Analysis of Vacancy Durations and Vacancy Flow: Cyclical Variation and Job Requirements

A Simple Performability Estimate for Jackson Networks with an Unreliable Output Channel

Price Adjustment in a Two-Country Model

An Equivalence of Communication Protocols for Interconnection Networks

Micro-Foundations of Money and Finance

The Importance of Bias-Terms for Error Bounds and Comparison Results

On Regression Sampling in Statistical Auditing: Bad Answers to the Wrong Questions?

Company Strategy and the (Re)Design of Industrial Relations, some case studies in the Netherlands