Prediction of 0-1-events for short- and long-memory time series

Jan Beran

Abstract. The problem of predicting 0-1-events is considered under general conditions, including stationary processes with short and long memory as well as processes with changing distribution patterns. Nonparametric estimates of the probability function and prediction intervals are obtained.

Keywords: 0-1-events, long-range dependence, short-range dependence, antipersistence, kernel smoothing, bandwidth, prediction

1. The general problem

In time series applications, the main concern is sometimes to predict whether a certain event will occur or not. For instance, in finance, a decision may be based on the probability that a stock price stays within certain bounds; in meteorology, we may want to know whether certain disastrous weather conditions are likely to occur or not etc. This motivates the following problem: Let $X_t (t \in N)$ be a stochastic process on a probability space $(\Omega, \mathcal{A}, P)$ where $\Omega \subset R^N$ is a subspace of real valued functions on N and $\mathcal{A}$ is a suitable $\sigma-$algebra. For a fixed $k \in N_+$, and $\tau_1, ..., \tau_k \in N$, let $A_{\tau_1,...,\tau_k}(i) \in \mathcal{A}$ be such that

$$A_{\tau_1,...,\tau_k}(i) = \{\omega : (X_{i+\tau_1}, X_{i+\tau_2}, ..., X_{i+\tau_k}) \in B\},$$

for some $B \subset R^k$ and

$$p = p(i; \tau_1, ..., \tau_k) = P(A_{\tau_1,...,\tau_k}(i))$$

the probability of this event. The general question is now: Given observations $X_1, ..., X_n$, how can we estimate p, without making too strong assumptions on the unknown underlying probability distribution P.

2. Specific assumptions

The following assumptions will be used: Let

$$Y_i = 1\{A_{\tau_1,...,\tau_k}(i)\} \tag{1}$$

and

(2) $$p = p(t_i) = P(Y_i = 1).$$

The process Y_i is assumed to have the following properties:

- **(A1)**

(3) $$Z_i = \frac{Y_i - p(t_i)}{\sqrt{p(t_i)(1-p(t_i))}}$$

  is a second order stationary process with autocovariances $\gamma(k)$ and spectral density $f(\lambda) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \exp(ik\lambda)\gamma(k)$.
- **(A2)** The spectral density is continuous in $[-\pi, 0) \cup (0, \pi]$ and at the origin we have

(4) $$f(\lambda) \sim c_f |\lambda|^{-2d} \ (|\lambda| \to 0)$$

  for a constant $c_f > 0$ and $d \in (-\frac{1}{2}, \frac{1}{2})$, where " $\sim$ " means that the ratio of the left and right hand side converges to one.
- **(A3)** $p \in C^2[0,1]$
- **(A4)** $\sup_{0<x<1} \max_{j=0,1,2} |p^{(j)}(x)| \leq C_1 < \infty$ where $p^{(j)}$ denotes the j'th derivative of p.
- **(A5)** $|p^{''}(x) - p^{''}(y)| \leq C_2 \cdot |x-y|^\beta$ for all $x, y \in [0,1]$, constants $C_1, C_2 < \infty$, and some $\beta \in (2,3]$.
- **(A6)** For a given $\Delta \in (0, \frac{1}{2})$, $\sup_{t \in [\Delta, 1-\Delta]} |p^{(l+1)}(t)| > 0$ for at least one $l \in \{0,1\}$ and $p^{(l)}$ achieves an absolute maximum or minimum in $[\Delta, 1-\Delta]$.

Remarks:

1. Since Y_i is a 0-1-process, we have $var(Y_i) = p(t_i)(1-p(t_i))$ so that $Y_i - p(t_i)$ can not be stationary. Therefore, the standardized process Z_i is considered.
2. Z_i can be second order stationary even if neither the X_i nor $X_i - E(X_i)$ are stationary. For instance, let X_i be iid with fixed $\alpha-$quantile q_α but arbitrary distributions F_i that differ, for instance, in their variance. Then X_i is not second order stationary, in contrast to the 0-1-process $Y_i = 1\{A_\tau(i)\}$ with $A_\tau(i) = \{X_{i+\tau-1} > q_\alpha\}$.
3. Three cases can be distinguished (see e.g. Beran 1994 and references therein):
   (a) Short memory: $d = 0$, f is continuous in the whole interval $[-\pi, \pi]$ and $0 \neq \sum \gamma(k) < \infty$;
   (b) Long memory: $d > 0$, f is infinite at zero and $\sum \gamma(k) = \infty$;
   (c) Antipersistence: $d < 0$, $f(0) = 0$ and $\sum \gamma(k) = 0$.
4. The assumptions include in particular the special case where the original process itself consists of 0-1-variables, i.e. where $Y_i = X_i$.

3. Estimation of p

Under the assumptions given above, the estimation problem consists of estimating a smooth function $p(t)$, where $0 \leq p \leq 1$. If the distribution of the process X_i is known, except for a finite dimensional parameter vector θ, then the optimal method is to estimate θ from the original observations X_i (for instance by maximum likelihood) and set $\widehat{p}(t) = p(t;\widehat{\theta})$. Here, we address the problem of estimating p, when only the assumptions given in the previous section are known. Note that these are assumptions on the process Y_i - no knowledge about the distribution of the original process X_i is needed. Thus, we consider estimation of $p(t)$ ($t \in [0,1]$) where $p(t_i) = E(Y_i), Y_i \in \{0,1\}$and

$$Y_i = p(t_i) + \sqrt{p(t_i)(1-p(t_i))}Z_i \tag{5}$$

where Z_i is a stationary zero mean process as defined in (A1). We will consider kernel estimation of p : Let $K : [-1,1] \to R_+$ be a positive symmetric function with support $[-1,1]$ and $b > 0$ a bandwidth, then we define

$$\widehat{p}(t;b) = \frac{1}{nb}\sum_{i=1}^{n} K(\frac{t-t_i}{b})Y_i. \tag{6}$$

The general problem of estimating a smooth function μ from data of the form

$$Y_i = \mu(t_i) + Z_i \tag{7}$$

has been considered by various authors for the case where the error process Z_i is stationary with (i) short-range dependence (see e.g. Chiu, 1989; Altman, 1990; Hall and Hart, 1990; Herrmann, Gasser and Kneip, 1992) or (ii) long-range dependence, i.e. $0 < \alpha < 1$ (see e.g. Hall and Hart, 1990; Csörgö and Mielniczuk, 1995; Ray and Tsay, 1997) or (iii) antipersistence (Beran and Feng 2002a). Beran and Feng (2002a,b,c) consider the more general case where it is not known a priori whether Z_i is stationary (including antipersistence as well as short- and long-range dependence) or nonstationary.

The essential question to be solved is how to choose the bandwidth b optimally. Note that, in contrast to the usual setup, for 0-1-processes the variance of the error process is related to the mean function and the mean function is bounded from below and above. One may therefore either estimate p itself, under the constraint $0 \leq p \leq 1$ or one may instead estimate a suitable transformation of p. Obvious transformations are, for instance, the logistic transformation $g(p) = \log[p/(1-p)]$ or the variance stabilizing transformation $g(p) = \arcsin\sqrt{p}$. Asymptotically, the choice of g does not influence bandwidth selection, if the criterion is the mean squared error. This follows from standard arguments: Assume that $\widehat{p}(t;b_n)$ ($n \in N$) is a (weakly) consistent sequence of estimates of $p(t)$. Then $g(\widehat{p}) = g(p) + g^{'}(p)(\widehat{p}-p) + o_p(\widehat{p}-p)$ so that (under suitable regularity conditions on the sequence $\widehat{p}$) we have $MSE(g(\widehat{p})) = E\{[g(\widehat{p}) - g(p)]^2\} = [g^{'}(p)]^2 MSE(\widehat{p}) + r$ where r is of smaller order than $MSE(\widehat{p})$. Since $g^{'}(p)$ is a constant, independent of b, the bandwidth minizing $MSE(g(\widehat{p}))$ is asymptotically the same as the one minimizing $MSE(\widehat{p})$. In the following, we thus use the mean squared error of $\widehat{p}$ as a criterion for choosing b.

4. Asymptotically optimal bandwidth choice

In this section, asymptotic expressions for the mean squared error and the asymptotically optimal bandwidth are given. Using the notations $I(p^{''}) = \int_{\Delta}^{1-\Delta}[p^{''}(t)]^2 dt$ and $I(K) = \int_{-1}^{1} x^2 K(x)dx$, the following results can be derived in a similar way as in Beran and Feng (2002a) by taking into account the heteroskedasticity factor proportional $w(t) = p(t)(1-p(t))$:

Theorem 1. *Let $b_n > 0$ be a sequence of bandwidths such that $b_n \to 0$ and $nb_n \to \infty$, then we have*

(i): *Bias:*

$$E[\widehat{p}(t) - p(t)] = b_n^2 \frac{p^{''}(t)I(K)}{2} + o(b_n^2) \tag{8}$$

uniformly in $\Delta < t < 1-\Delta$;

(ii): *Variance:*

$$(nb_n)^{1-2d}\mathrm{var}(\widehat{p}(t)) = w(t)V(\theta) + o(1) \tag{9}$$

uniformly in $\Delta < t < 1-\Delta$ where $0 < V(\theta) < \infty$ is a constant;

(iii): *IMSE: The integrated mean squared error in $[\Delta, 1-\Delta]$ is given by*

$$\int_{\Delta}^{1-\Delta} E\{[\widehat{p}(t) - p(t)]^2\}dt = IMSE_{asympt}(n, b_n) + o(\max(b_n^4, (nb_n)^{2d-1}))$$

$$= b_n^4 \frac{I(g^{''})I^2(K)}{4} + (nb_n)^{2d-1}V(\theta)\int_{\Delta}^{1-\Delta} w(t)dt + o(\max(b_n^4, (nb_n)^{2d-1})) \tag{10}$$

(v): *Optimal bandwidth: The bandwidth that minimizes the asymptotic IMSE is given by*

$$b_{opt} = C_{opt}\ n^{(2d-1)/(5-2d)} \tag{11}$$

where

$$C_{opt} = C_{opt}(\theta) = [\frac{(1-2d)V(\theta)\int_{\Delta}^{1-\Delta} w(t)dt}{I(g^{''})I^2(K)}]^{1/(5-2d)}. \tag{12}$$

Similar results can be obtained for kernel estimates of derivatives of p. For instance, the second derivative can be estimated by $\widehat{p}^{''}(t) = n^{-1}b^{-3}\sum K((t_j - t)/b)Y_j$ where K is a symmetric kernel such that $\int K(x)dx = 0$ and $\int K(x)x^2dx = 2$. The optimal bandwidth for estimating the second derivative is of the order $O(n^{(2d-1)/(9-2d)})$. The asymptotic expression $V(\theta)$ can be given explicitly for $d = 0$ and $d > 0$:

$$V(\theta) = 2\pi c_f \int_{-1}^{1} K^2(x)dx,\ (d = 0), \tag{13}$$

$$V(\theta) = 2c_f\Gamma(1-2d)\sin\pi d \int_{-1}^{1}\int_{-1}^{1} K(x)K(y)|x-y|^{2d-1}dxdy,\ (d>0). \tag{14}$$

For $d < 0$, a general simple formula for V does not seem to be available., except in special cases. For the box-kernel, we obtain (see Beran and Feng 2002a)

Corollary 1. *Let* $K(x) = \frac{1}{2}1\{x \in [-1,1]\}$. *Define*

$$\nu(d) = \frac{2^{2d}\Gamma(1-2d)\sin(\pi d)}{d(2d+1)} \tag{15}$$

with $\nu(0) = \lim_{d\to 0}\ \nu(d) = \pi$. *Then, under the assumptions of Theorem 1, we have*

(i): *Bias:*

$$E[\widehat{p}(t) - p(t)] = b_n^2 \frac{p''(t)}{6} + o(b_n^2); \tag{16}$$

(ii): *Variance:*

$$\mathrm{var}(\widehat{p}(t)) = (nb_n)^{2d-1}\nu(d)c_f w(t) + o((nb_n)^{2d-1}); \tag{17}$$

(iii): *IMSE:*

$$\int_{\Delta}^{1-\Delta} E\{[\widehat{p}(t) - p(t)]^2\}dt = b_n^4 \frac{I(p'')}{36} + (nb_n)^{2d-1}\nu(d)c_f W$$

$$+o(\max(b_n^4, (nb_n)^{2d-1})) \tag{18}$$

where $W = \int_{\Delta}^{1-\Delta} w(t)dt$.

(iv): *Optimal bandwidth:*

$$b_{opt} = C_{opt}\ n^{(2d-1)/(5-2d)} \tag{19}$$

with

$$C_{opt} = [\frac{9(1-2d)\nu(d)c_f W}{I(g'')}]^{1/(5-2d)} \tag{20}$$

5. Data driven bandwidth choice

An iterative algorithm for choosing the bandwidth for a modell with a smooth trend function and a stationary or nonstationary error processes Z_i is defined in Beran and Feng (2002a,b). The error process is modelled by a (possibly integrated) Gaussian fractional ARIMA process (Granger and Joyeux 1980, Hosking 1981). Beran and Feng prove convergence of the algorithm and provide finite sample modifications to improve its performance for short series. Convergence of the algorithm relies on consistency of the estimate of the spectral distribution f. For a 0-1-process Y_i, the spectral distribution function can be estimated consistently by the Gaussian maximum likelihood estimate for FARIMA-processes if f is indeed identical with the spectral density of a fractional ARIMA process. We thus assume the following additional assumption **(A7)**:

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \left| \frac{\psi(e^{i\lambda})}{\phi(e^{i\lambda})} \right|^2 |1 - e^{i\lambda}|^{1-2d} \tag{21}$$

for some $-\frac{1}{2} < d < \frac{1}{2}$. Here, $\phi(x)$ and $\psi(x)$ are polynomials of finite orders m_1 and m_2 respectively with roots outside the unit circle.

A suitable modification of the algorithm in Beran and Feng (2002a,b,c) can now be defined. (Note that in Beran and Feng (2002c), m_1 is set equal to zero.) The main steps of the algorithm are as follows:

Algorithm:

Step 1: Set $j = 1$, define a maximal autoregressive order M and an initial bandwidth b_o, and carry out Steps 2 to 5 for each $m_2 \in \{0, 1, ..., M\}$.

Step 2: Estimate p and Z_i using an the bandwidth b_{j-1};

Step 3: Estimate f by maximum likelihood (Beran 1995).

Step 4: Given the estimated spectral density f, calculate a new optimal bandwidth b_j; set $j = j + 1$ and a new optimal bandwidth for estimating p''.

Step 5: Stop, if the change in the bandwidth does not exceed a certain bound. Otherwise go to Step 2.

Step 6: Select the solution that minimizes a consistent model choice criterion such as the BIC (Schwarz 1978, Beran et al. 1998).

6. Testing and Prediction

An approximate pointwise test for testing the null hypothesis $H_o : p(t) \equiv const$, can be obtained by defining the rejection region $|\widehat{p}(t)| > c_{\frac{\alpha}{2}} \hat{v}$ where $\hat{v}$ is equal to the square root of $(nb_n)^{2\widehat{d}-1} \nu(\widehat{d}) c_f \widehat{p}(t)(1 - \widehat{p}(t))$ and $c_{\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})-$quantile of the standard normal distribution. Note that, alternatively, we may test the null hypothesis $H_o : p^{'}(t) \equiv 0$.

Prediction of Y_{n+k} from $Y_1, ..., Y_n$ reduces to predicting the succes probability $p(1 + k/n)$. Beran and Ocker (1999) propose a prediction method in the context of general nonparametric trend functions that is based on Taylor expansion and optimal linear prediction of the stochastic component. This approach can, in principle, be carried over to forecasts of $p(1 + k/n)$. Note, however, that this may lead to values outside of the interval $[0, 1]$. As an alternative, one may extrapolate a suitable transformation of p. More specifically, let $g : [0, 1] \to R$ be a one-to-one monotonic function such that $\lim_{x\to\infty} g(x) = \infty$ and $\lim_{x\to-\infty} g(x) = -\infty$. Then $g(p(1 + k/n))$ may be approximated by $g(1) + g^{'}(1)\frac{k}{n}$. The predicted value of p is $\widehat{p}(1 + k/n) = g^{-1}[g(1) + g^{'}(1)\frac{k}{n}]$. In the context of quantile estimation for certain long-memory processes, this approach is used, for instance, in Ghosh and Draghicescu (2001).

7. Data examples

The method introduced here can be used to explore various linear and nonlinear properties of time series. This is illustrated by the following application to daily values of three stock market indices between january 1, 1992 and november 10, 1995. The indices are: FTSE 100 (figure 1), CAC (figure 2) and the Swiss Bank Corporation Index (figure 3). We consider the event

$$A(i) = \{\omega : X_{i+20} > X_i \text{ and } \min_{s=1,\ldots,20} X_{i+s} > 0.9X_i\}. \tag{22}$$

The event $A(i)$ means that in one month (20 work days) the index will be higher than the initial value X_i and during this one-month period it never drops below 90% of X_i. The estimated probability functions $p(t_i) = P(A(i))$ are displayed in figures 1c, 2c and 3c respectively. The shaded areas correspond to time stretches where p is significantly different from a constant. (fully shaded area for "high level" and area shaded with lines for "low level"). Although the critical limits to test for non-constant $p(t)$ are pointwise limits only, the similar patterns for all three series support the conjecture that p is not constant. In particular, there is a period (around observation 400) where p is considerably higher for all three series.

Acknowledgment

This research was supported in part by the Center of Finance and Econometrics, University of Konstanz, Germany. I would also like to thank Dr. Dirk Ocker (Swiss Union of Raiffeisenbanks) and Dr. Elke M. Hennig (Citibank, Frankfurt) for providing me with the data. Thanks go also to a referee for comments that helped to improve the presentation of the results.

References

: Altman, N.S. 1990. Kernel smoothing of data with correlated errors. J. Am. Statist. Assoc. 85, 749-759.

: Beran, J. 1994. Statistics for long-memory processes. Chapman and Hall, New York.

: Beran, J. 1995. Maximum likelihood estimation of the differencing parameter for invertible short- and long-memory ARIMA models. J. Roy. Statist. Soc. B 57, No. 4, 695-672.

: Beran, J., Bhansali, R.J. and Ocker, D. 1998. On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes. Biometrika Vol. 85, No. 4, 921-934.

: Beran, J., Feng, Y. (2002a) 'SEMIFAR models - a semiparametric framework for modelling trends, long-range dependence and nonstationarity'. Computational Statistics & Data Analysis} (in press).

: Beran, J. and Feng, Y. (2002b) 'Data driven bandwidth choice for SEMIFAR models'. Journal of Computational and Graphical Statistics} (in press).

: Beran, J. and Feng, Y. (2002c) 'Local polynomial fitting with long memory, short memory and antipersistent errors'. Annals of Statistical Mathematics (in press).

: Beran, J. and Ocker, D. (1999) SEMIFAR forecasts, with applications to foreign exchange rates'. Journal of Statistical Planning and Inference}, 80, 137-153.

: Chiu, S.T. 1989. Bandwidth selection for kernel estimates with correlated noise. Statist. Probab. Lett. 8, 347-354.

: Csörgö, S. and Mielniczuk, J. 1995. Nonparametric regression under long-range dependent normal errors. Ann. Statist. 23, 1000-1014.

: Ghosh, S. and Draghicescu, D. (2001) Predicting the distribution function for long-memory processes. International Journal of Forecasting (in press).

: Granger, C.W.J., Joyeux, R. 1980. An introduction to long-range time series models and fractional differencing. J. Time Series Anal. 1 15-30.

: Hall, P. and Hart, J. 1990. Nonparametric regression with long-range dependence. Stoch. Processes Appl. 36, 339-351.

: Herrmann, E., Gasser, T. and Kneip, A. 1992. Choice of bandwidth for kernel regression when residuals are correlated. Biometrika 79, 783-795.

: Hosking, J.R.M. 1981. Fractional differencing. Biometrika 68 165-176.

: Ray, B.K. and Tsay, R.S. 1997. Bandwidth selection for kernel regression with long-range dependence. Biometrika 84, 791–802.

: Schwarz, G. 1978. Estimating the dimension of a model. Ann. Statist. 6, 461-464.

Department of Mathematics and Statistics, University of Konstanz, Universitätsstr. 10, Postfach 5560, 78457 Konstanz, Germany

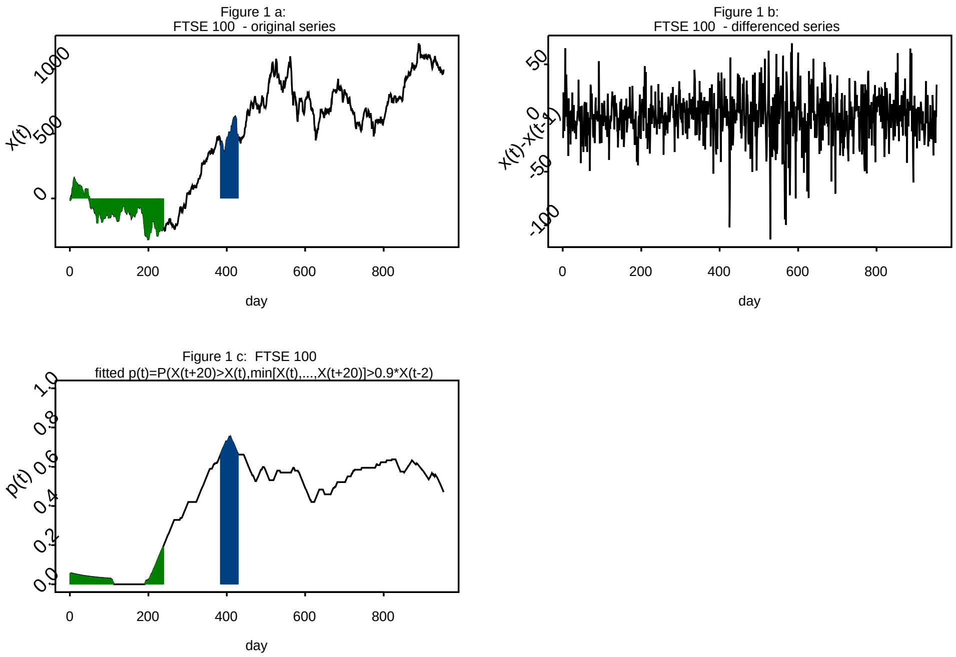

Figure 1. FTSE 100 between january 1, 1992 and november 10, 1995 - original series (figure 1a), differenced series (figure 1b) and estimated probability function $p(i/n) = P(A(i))$ where $A(i) = \{\omega : X_{i+20} > X_i \text{ and } \min_{s=1,\dots,20} X_{i+s} > 0.9X_i\}$. Periods with significant departures from $H_o : p \equiv const$ are shaded with lines (for p below critical bound) and fully shaded (for p above critical bound) respectively.

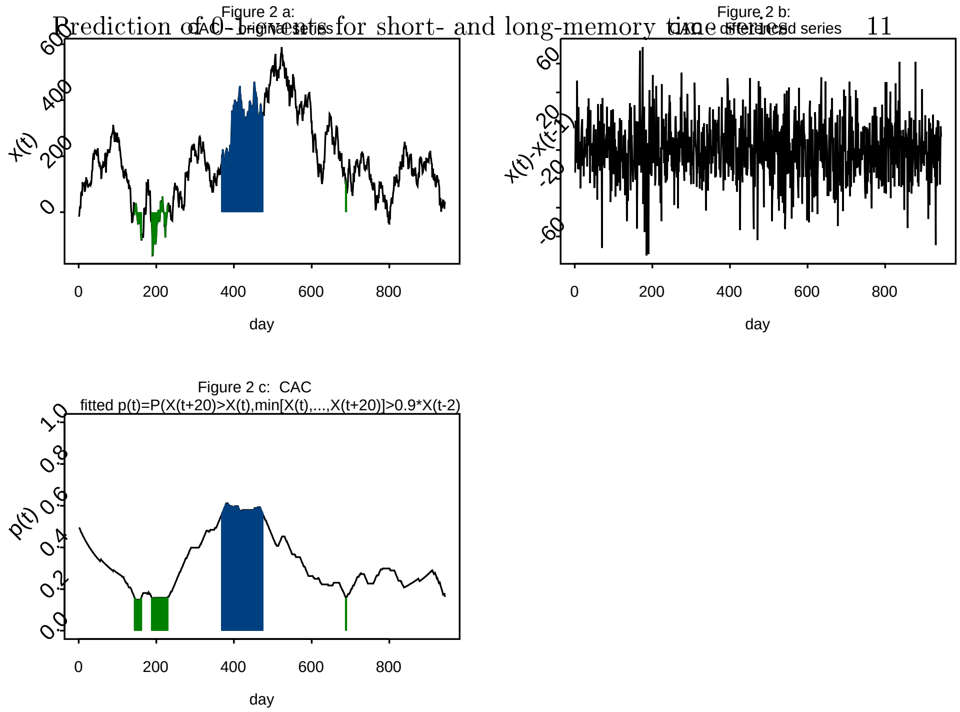

FIGURE 2. CAC between january 1, 1992 and november 10, 1995 - original series (figure 1a), differenced series (figure 1b) and estimated probability function $p(i/n) = P(A(i))$ where $A(i) = \{\omega : X_{i+20} > X_i$ and $\min_{s=1,\ldots,20} X_{i+s} > 0.9X_i\}$. Periods with significant departures from $H_o : p \equiv const$ are shaded with lines (for p below critical bound) and fully shaded (for p above critical bound) respectively.

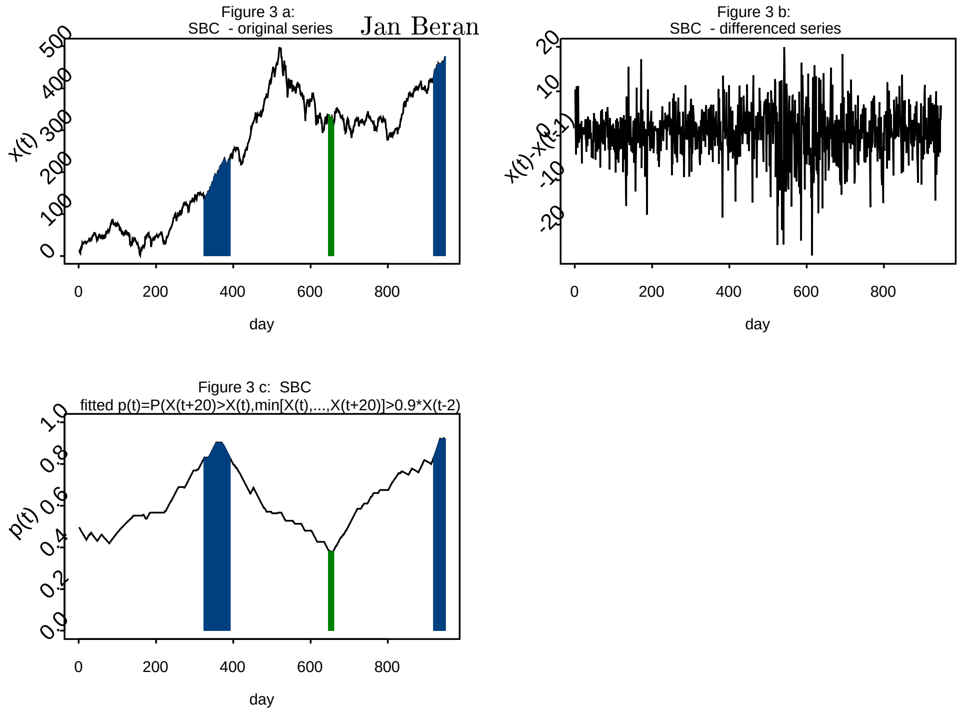

FIGURE 3. SBC between january 1, 1992 and november 10, 1995 - original series (figure 1a), differenced series (figure 1b) and estimated probability function $p(i/n) = P(A(i))$ where $A(i) = \{\omega : X_{i+20} > X_i \text{ and } \min_{s=1,\ldots,20} X_{i+s} > 0.9X_i\}$. Periods with significant departures from $H_o : p \equiv const$ are shaded with lines (for p below critical bound) and fully shaded (for p above critical bound) respectively.