Quaderni di Statistica
Vol. 5, 2003

Kernel smoothing for the analysis of climatic data

Marcella Niglio
Di.S.E.S., Università di Salerno
E-mail: mniglio@unisa.it

Cira Perna
Di.S.E.S., Università di Salerno
E-mail: perna@unisa.it

Summary: Climatic temperature time series show in their behaviour a strong and regular seasonality which can be differently analysed. When the series are examined using kernel regression, the classical approaches for the bandwidth selection fail involving undersmooth or oversmooth of data. The aim of the paper is to describe and discuss different methodologies for the bandwidth selection which take into account the correlation structure of the errors. The approaches, based on corrected versions of the Generalized Cross Validation criterion, have been used to analyze two climatic data collected on the South of Italy from January 1960 to December 2000.

Keywords: Temperature data, Nonparametric regression, Kernel estimator, Bandwidth selection, Cross validation.

1. Introduction

The analysis of climatic temperature data has a relevant role in the study of environmental fenomena. The interest of meteorologists often relies on the estimation of missing values which frequently affect temperature and on the generation of predictions usually based on quite complicated models.

These two aspects imply a preliminary study of data and a careful research of their behaviour.

In this paper we refer to climatic time series decomposed using a signal plus noise model:

$$Y_t = m(t) + \epsilon_t \qquad t = 1, 2, \ldots, n \tag{1}$$

where $m(t)$ is the regular component and ϵ_t are zero-mean errors often coming from a stationary correlated process.

We focus on the estimation of the smooth deterministic $m(t)$ function when the structure of data is not known a priori and nonparametric techniques are used. In particular, we refer to kernel estimators which imply the selection of a smoothing parameter called *bandwidth*.

The classical approach for the bandwidth selection, which is based on the assumption of independent errors, fails involving an undersmooth or oversmooth of data when a positive or a negative correlation is respectively recognized and so reducing, in the former case, the estimation of $m(t)$ to an interpolation of observations (see Hart (1991), Herrmann *et al.* (1992) among the others).

In order to limit this effect, different bandwidth selection procedures have been proposed in the literature. Some of them are based on heavy assumptions on the correlation structure of the errors (see Chiu (1989), Hart (1991, 1994)) so limiting their use. Definitely model free approaches (Altman (1990), Chou and Marron (1995), Hall *et al.* (1995)) appear more suitable when no information is available on the errors structure.

The aim of the paper is to describe and discuss different methodologies for the bandwidth selection which are particularly useful in the context of climatic data. The approaches, based on corrected versions of the Generalized Cross Validation criterion, have the advantage to be very general avoiding assumptions on the errors correlation structure.

The paper is organized as follows. In Section 2 the theoretic results of the applied methodologies are presented focusing on the bandwidth selection. In Section 3 the proposed different procedures are applied to the analysis of climatic time series related to the mean and maximum temperature recorded in Scafati (Salerno), from January 1960 to December 2000. Some concluding remarks are given in the final section.

2. The methodology

Let $m(t) = s(t/n)$, where $s(\cdot)$ is a real smooth function defined on $[0, 1]$. The estimation of $s(\cdot)$ is carried out by using the Priestley and Chao kernel (Priestley and Chao, 1972) defined as:

$$s(x) = (nh)^{-1} \sum_{t=1}^{n} K\left(\frac{x - t/n}{h}\right) Y_t \tag{2}$$

where h is the bandwidth parameter and the kernel $K(\cdot)$ is a symmetric probability density function.

One of the main problem related to the estimation of $s(x)$ is the choice of the bandwidth. When a cross validation (CV henceforth) criterion is selected, the optimal bandwidth, h, is taken as the minimizer of the mean square error:

$$CV(h) = \frac{1}{n} \sum_{t=1}^{n} [Y_t - \hat{s}_{-t}(t/n)]^2 \tag{3}$$

where $\hat{s}_{-t}(t/n)$ is the kernel estimator of Y_t, obtained by omitting the t-th observation $(t/n, Y_t)$.

Unfortunately, this approach is not suitable when the errors are correlated. This is due to the fact that this bandwidth selection method associates all the structure of data to the mean function even incorporating the error correlation. When no assumption is given on the structure of the mean and of the error components, model (1) is unidentifiable and so the estimate of the signal cannot be separated from the noise.

Hart (1991) demonstrates that, under well defined conditions on the kernel function and on the stationarity of the errors, when data are positively correlated, the CV criterion in (3) will choose h such that the kernel estimate very nearly interpolates the data. To overcome this limit, he suggests appropriate differencing of data and the use of the spectral density of ϵ_i to estimate the covariance $c(k) = cov(\epsilon_i, \epsilon_{i+k})$, $|k| = 1, 2, \ldots$, which affects the bandwidth selection.

This procedure is not completely free from parametric assumptions on the error process which are instead absent in the procedure proposed in Chou and Marron (1991). It is based on the estimate of h, an adjustment of the CV criterion called modified cross validation (MCV) which is a $leave-(2\ell+1)-out$ version of the criterion in (3). In particular, in order to remove the short range dependence among data, the MCV is based on the minimization of (3) such that $\hat{s}_{-t}(t/n)$ is a $leave-(2\ell+1)-out$ estimator of Y_t obtained removing the observations $(t/n, Y_t)$, $-\ell \leq t \leq \ell$.

Asymptotic results on the behavior of h_{MCV} are also shown under well defined assumptions and the convergence rate is highlighted to be of the same order than the one obtained with independent observations.

The problem which clearly arises in this approach is the selection of ℓ which has to increase as the dependence among data becomes longer.

A further approach for the estimation of h, which is going to be used in the following, has been proposed in Altman (1990). It has a relevant use when time series data show heavy regularities in their behaviour.

The procedure makes use of the kernel estimator (2) for $s(\cdot)$ and assumes that the errors ϵ_t come from a weak stationary process with covariance function:

$$E[\epsilon_t, \epsilon_{t+k}] = \sigma^2 \rho_n(|k|) \qquad t = 1, \ldots, n$$

with σ^2 the error variance and $\rho_n(|k|)$ the correlation function such that:

$$\lim_{n\to\infty} \sum_{k=1}^{n} |\rho_n(k)| < \infty \qquad\qquad \lim_{n\to\infty} \sum_{k=1}^{n} k|\rho_n(k)| = 0$$

where first condition ensures that sufficiently far observations are uncorrelated.

The generalized cross validation (GCV) criterion (Craven and Wahba, 1979), generally employed for the estimation of h, can be affected from the errors structure. In this case two different procedures can be used to correct the selection criterion.

The first one, called *direct*, allows a correction of the GCV to reduce the biasedness induced by the correlation.

In particular, given the square residuals in (3):

$$r^2(t, h, n) = [Y_t - \hat{s}_{-t}(t/n)]^2 \qquad\qquad t = 1, 2, \ldots, n$$

the *direct* method corrects the GCV such that the loss function is constructed using the following square residuals:

$$r^2_{GCV,D}(t,h,n) = \frac{r^2(t,h,n)}{1 - n^{-1}tr(\mathbf{W}_h\mathbf{R}_n)^2} \tag{4}$$

where $\mathbf{W}_h$ is the square matrix $(n \times n)$ of the kernel weights, $tr(\cdot)$ is the matrix trace and $\mathbf{R}_n$ is the correlation matrix.

The second *indirect* procedure transforms the residuals to limit their linear relation. In this case $r(t,h,n)$ becomes:

$$\mathbf{R}_n^{-1/2} \cdot r(\cdot,h,n) = r_{\rho^{-1}}(\cdot,h,n)$$

and so the GCV criterion is defined as:

$$r^2_{GCV,I}(t,h,n) = \frac{r^2_{\rho^{-1}}(t,h,n)}{[1 - n^{-1}tr(\mathbf{W}_h)]^2} \tag{5}$$

To estimate $\mathbf{R}_n$ in (4) and (5), it can be used an estimator based on the method of moments whose consistency is shown, under some regularity conditions on the errors, in Altman (1990). The procedure implies a preliminary estimate of $m(\cdot)$ in (1) and the use of its residuals to estimate $\mathbf{R}_n$.

The theoretic assumptions of this procedure can be recognized in the generating process of climatic climatic data where the seasonal component prevails on the other components of the observed data (Altman, 1991; Bowman and Azzalini, 1997).

3. The analysis of temperature data

We consider two climatic time series related to the mean and the maximum temperature recorded in Scafati (Salerno, Italy) from January 1960 to December 2000. The original data, collected by the *'Istituto Sperimentale per il Tabacco'* are ten-days spaced for a total of 1476 observations. In particular, the *mean temperature* (M) is the average among the maximum and the minimum temperatures over ten-days whereas the *maximum* series (MX) is the maximum temperature observed in each ten-days over the period under study.

In order to have monthly data, the ten-days observations related to each month of a single series have been aggregated computing their mean so obtaining two new series of length 492.

The time plot of the monthly series are shown in Figure 1 where the main aspect which comes out is the strong seasonality whose behaviour can be approximated using a deterministic sinusoidal function over the period under analysis. In particular, when model (1) is used, it is fitted by the $m(\cdot)$ function whereas the analysis of ϵ_t need further investigation.

A simple differencing of data which allows to easily remove the seasonal component, $Y_{t,12} = Y_t - Y_{t-12}$, outlines that the transformed data $Y_{t,12}$ (for $t = 13, \ldots, 492$) show an autocorrelation structure which is significant at low order autocorrelation lags as presented in Figure 2.

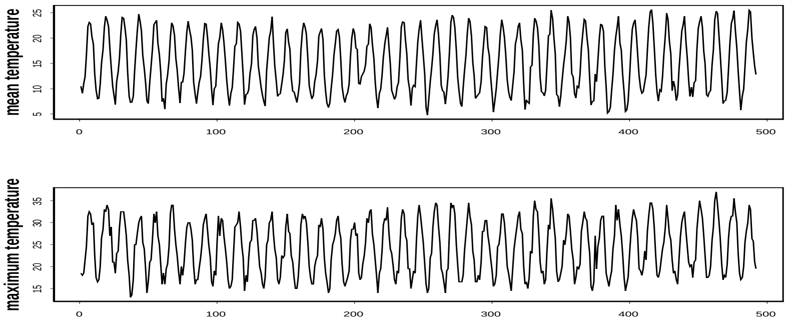

Figure 1: *Mean and maximum monthly climatic temperature data collected in Scafati (Italy) from 01.1960 to 12.2000*

The plots of $Y_{t,12}$ for the two series (Figure 3) graphically show the stationarity in mean of the differenced data and so further highlighting that the data under analysis satisfy the theoretical conditions of the procedure discussed in the previous section.

The use of a non-parametric approach to study the data in Figure 1 implies the selection of a bandwidth parameter. In this context a GCV criterion based on the assumption of independent data clearly fails carrying out the undersmooth of data.

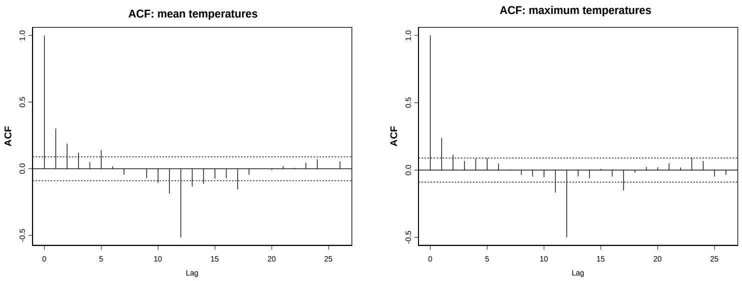

Figure 2: *Correlograms of the monthly Mean and Maximum temperature data*

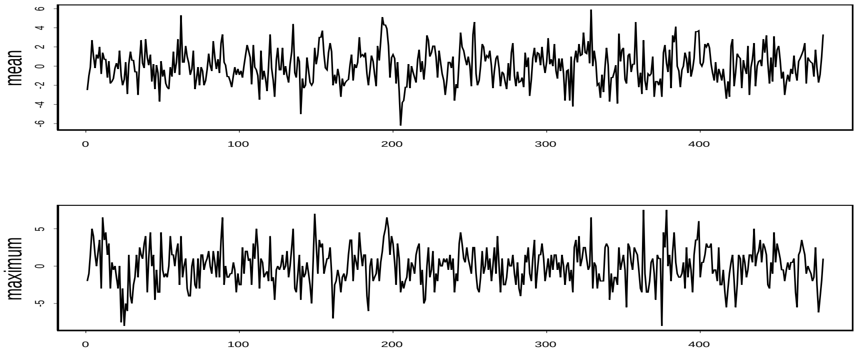

Figure 3: *Plots of the Mean and Maximum differenced time series*

Different results are instead obtained by using the procedure in Section 2. It is based on three main steps:

1. estimate the preliminary $m(\cdot)$ choosing h such that the data are overmoothed;

2. estimate the correlation matrix $\mathbf{R}_n$ using the residuals obtained in Step 1.

3. select h such that the criterion in (3) is minimized using the square

loss functions (4) or (5) when the *direct* or *indirect* method is respectively preferred.

The two temperature time series have been studied following the previous steps and fixing, for the preliminary estimate of $m(\cdot)$, $h = 3$.

The search for the selection of h through the *direct* and *indirect* method, has been carried out, over a grid of 100 points, on the interval $[0.5, 3]$ whereas a wider interval has been considered for the GCV criterion based on the hypothesis of independence.

In Table 1 we report the selected bandwidths obtained using the classical GCV criterion and its direct and indirect versions. In Figure 4 the values of the two GCV criteria over the grid search are shown for series M (the results of series MX, not reported here, replicate those of series M).

Table 1. Bandwidths selected for the two series under analysis using three GCV criteria

	Mean series (M)	Maximum series (MX)
GCV-INDEPENDENT	0.213	0.213
GCV-DIRECT	1.409	1.434
GCV-INDIRECT	2.394	2.071

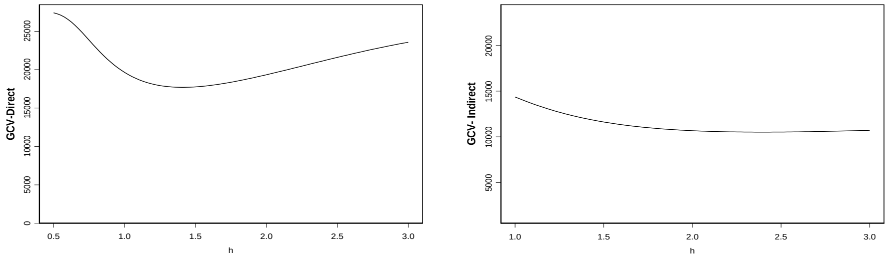

Figure 4: GCV_D *and* GCV_I *plots of series M over the grid search*

Table 1 and Figure 4 clearly show that the bandwidths selected taking into account the dependent structure of data allow to estimate a mean function which is smoother than that obtained when the correlation is neglected.

This is further shown in Figure 5 and in Figure 6 where the observed time series M and MX are respectively compared to that fitted using model (1) whose bandwidth h is selected with the three method under analysis.

In order to present more clear plots, the representations are focused on the time interval 1970-1980.

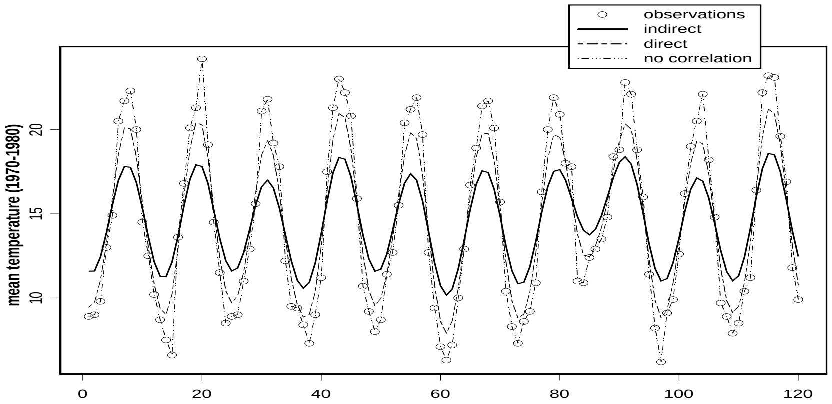

Figure 5: *Time plot of the Mean temperature data from 01.1970 to 12.1980 with the smoothing curves obtained using GCV-indipendent, GCV-direct and GCV-indirect.*

Both figures show the interpolation carried out from model (1) when h is chosen assuming the independence of the observations whereas more smoother curves are obtained using the other two bandwidth selection procedures.

In particular the direct and the indirect methods show more similar results with the MX series (Figure 6) whereas in Figure 5 the estimation of h with the GCV-indirect method implies a higher smoothing of data with respect to the direct one.

This difference can be due to the nature of data. In fact, M is obtained as average of the minimum and the maximum temperatures and so this

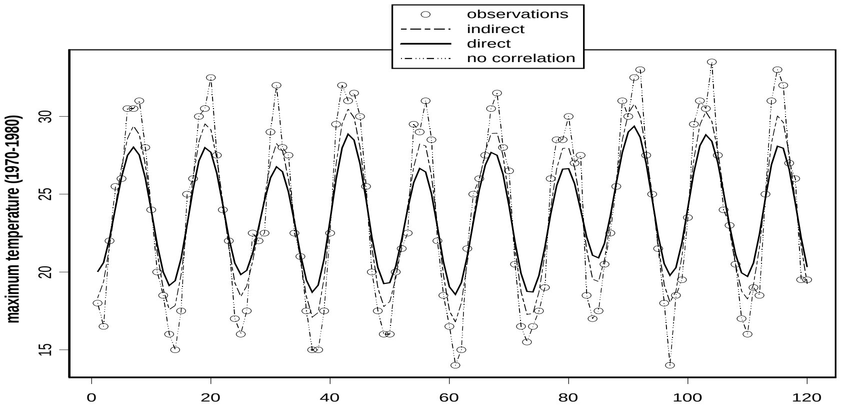

Figure 6: *Time plot of the Maximum temperature data from 01.1970 to 12.1980 with the smoothing curves obtained using GCV-indipendent, GCV-direct and GCV-indirect.*

preliminary treatment of data should modify the dependence structure of data which instead has been left unchanged in the MX time series.

4. Concluding remarks

The bandwidth selection in kernel regression with dependent errors is tackled when climatic temperature data are under analysis. In particular it is shown that, when the dependence of data is neglected, the GCV bandwidth selection procedure implies an undersmooth of data which leads to the interpolation of the observations.

In this context different approaches have been proposed in literature. Among them the Altman (1990) procedure has been selected for the analysis of two climatic temperature time series.

The results obtained show that when the selection criterion is suitably corrected for the dependence among data, a smoother curve which fits the data is obtained. This allows to reduce the variability with respect to the case when the dependence is neglected.

Acknowledgments. We thank Dr Salvatore Ascione (*Istituto Sperimentale per il Tabacco*, Scafati, Italy) who kindly gave us the analysed data allowing the realization of the present paper.

References

Altman N.S. (1990), Kernel smoothing of data with correlated errors, *Journal of the American Statistical Association*, 85, 749-759.

Bowman A.W., Azzalini A. (1997), *Applied smoothing techniques for data analysis*, Clarendon Press, Oxford.

Chiu S.T. (1989), Bandwidth selection for kernel estimate with correlated noise, *Statistics & Probability Letters*, 8, 347-354.

Chou C.K., Marron J.S. (1991), Comparison of two bandwidth selectors with dependent errors, *The Annals of Statistics*, 19, 1906-1918.

Craven P., Wahba G. (1979), Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized crossvalidation, *Numerische Mathematik*, 31, 377-403.

Hall P., Lahiri S.N., Polzehl J. (1995), On bandwidth choice in nonparametric regression with both short and long range dependent errors, *The Annals of Statistics*, 23, 1921-1936.

Hart J.D. (1991), Kernel regression estimation with time series errors, *Journal of the Royal Statistical Society (B)*, 53, 173-187.

Hart J.D. (1994), Automated kernel smoothing of dependent data by using time series cross-validation, *Journal of he Royal Statistical Society (B)*, 56, 529-542.

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

Priestley M.B., Chao M.T. (1972), Non-parametric function fitting, *Journal of the Royal Statistical Society (B)*, 34, 385-392.