An Analysis of a Bivariate Time Series in Which the Components Are Sampled at Different Instants

A la recherche du temps perdus

David R. Brillinger

Abstract

It is desired to express the relationship between the components of a bivariate time series. What is unusual is that the components are observed at different times and that the observation times are irregularly distributed. The problem of different sampling times is dealt with by interpolating the values of the dependent series to the times of the independent. This is followed by nonparametric regression to estimate the relationship. The research is motivated by data collected at a station along the Solimões River in central Brazil and also at a second station along a branch of the Solimões. Of interest to geographers is the possible change in the proportion of the Solimões waters entering the branch. This is because an increased flow of Solimões water into the branch might lead to the branch's widening and becoming the main stream. This could have substantial environmental effects.

Keywords. Amazonia; irregular sampling times; marked point process; nonparametric regression; river discharge; time series.

1 Introduction

Constance van Eeden has worked in many areas of statistics. Perhaps the problem that she has studied that is closest to the work of this paper is that of density estimation via kernels, e.g. [24]. Her work on that problem, like on many others, has been via very careful analysis.

The genesis of our work is that H. O'R. Sternberg, Professor of Geography at Berkeley, visited with a problem and a data set. In statistical terms the problem concerned the development of an instantaneous relationship between the components of a bivariate time series. The difficulty was that the components were sampled at

different instants. Further the spacings of the time points were irregular. The series were river flow rates measured at two places of a river system in Central Brazil.

According to Brazilian usage, the name Amazonas is applied to the Amazon river below the mouth of the Rio Negro. Upstream from the city of Manaus the main stem is known as the Solimões. Characteristically, the waters of the Solimões-Amazonas deposit, fork and come together, embracing islands approximately lenticular in shape. Strung along the river for many hundreds of kilometers, these islands split the stream bed into a master channel and one or more side channels, called *paranás*. One such channel exists just upstream from the mouth of the Rio Negro, where the Solimões (Amazon) sends off a branch, the Paraná do Careiro, that rejoins the trunk stream about 40 km. downvalley. Figure 1 shows various features. The large bright spot midway up the figure is the city of Manaus. The light line heading across the figure from the left to the right is the Solimões. The concern is the split of the Solimoes just to the right of Manaus. The large dark river coming towards Manaus from the upper right is the Rio Negro. This image was taken from the NASA web site eosweb.larc.nasa.gov.

A reason for the study is that concern has been expressed that an increased flow of Solimões water into the branch might lead to the widening of the Careiro and even to its eventual usurpation of the trunk stream. Such a process is of scientific and socio-economic interest, since it would destroy valuable floodplain land that supports a significant farm population and tens of thousands of cattle. In the 1950s, the matter drew the attention of Professor Sternberg, [19, 20]. In 1963 he coordinated a joint project of the US Geologic Survey, the University of Brazil and the Brazilian Navy, with the objective of carrying out discharge measurements in the Brazilian Amazon, [17]. Following this initial work, the Brazilian government embarked upon a program of systematic discharge measurements in Amazonia. This supplied the data of the work. A collaborative paper planned with Professor Sternberg will highlight the geomorphological framework of the problem, and further discuss analytical procedures applied to the issue, [21].

Specific questions that arise are: Does the proportion of the discharge that enters the Careiro depend on the discharge level of the Solimões? Is the relationship between the Careiro and the Solimões changing with time? Is the flow of the Solimões increasing?

Seeking answers to these questions leads to some interesting statistical problems: 1. the series are sampled at different times, 2. the series are sampled irregularly, 3. the relationship is possibly nonlinear and 4. the need for some theoretical properties of the proposed solutions. To study the questions there appears a desire for nonparametric estimates together with estimates of the associated uncertainty and further an assessment of model fit. The emphasis of this paper is on the statistical methods rather than the geographical interpretation of the empirical results obtained. More careful checking of assumptions and serious evaluation of uncertainties is needed before embarking on interpretations, [21].

Figure 1: NASA image of Central Brazil. The large bright spot towards the middle is the city of Manaus.

The sections of the paper are: Introduction, The data, Analytic formulations, Results, Goodness of fit of the models, Some extensions, and Discussion and summary. Lastly there is an Appendix laying out some analytic details and prooofs.

2 The data

The rivers' flow rates are measured upriver on the Solimões at a station near the town of Manacapuru. They are also measured at a station on the Paraná do Careiro. These stations are approximately 90 km. apart. Figure 1 provides a NASA image of the region on July 23, 2000.

The data available are for the years 1977 through 1998. Almost invariably they are

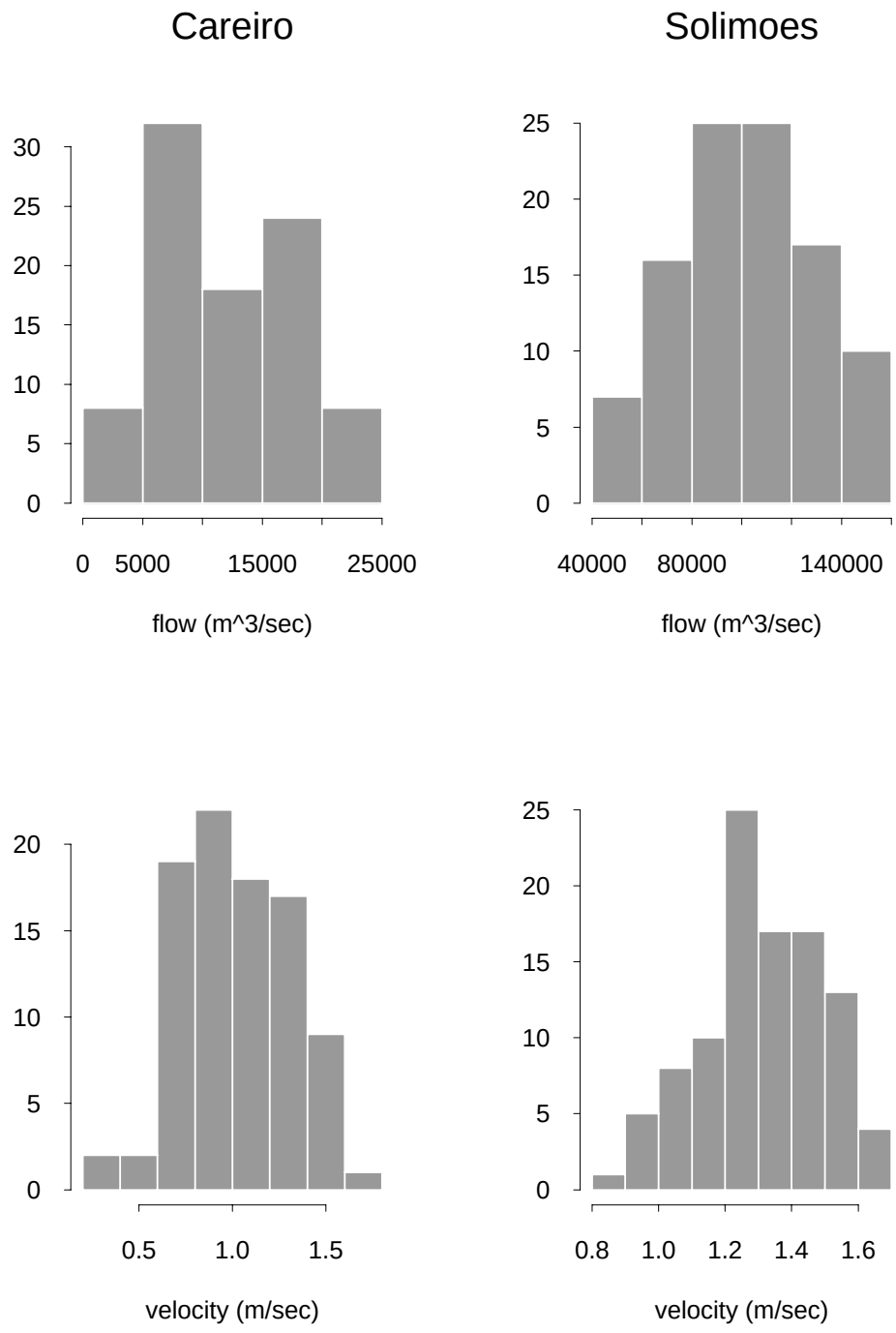


Figure 2: Histograms of the discharge rates and velocities for the two rivers.

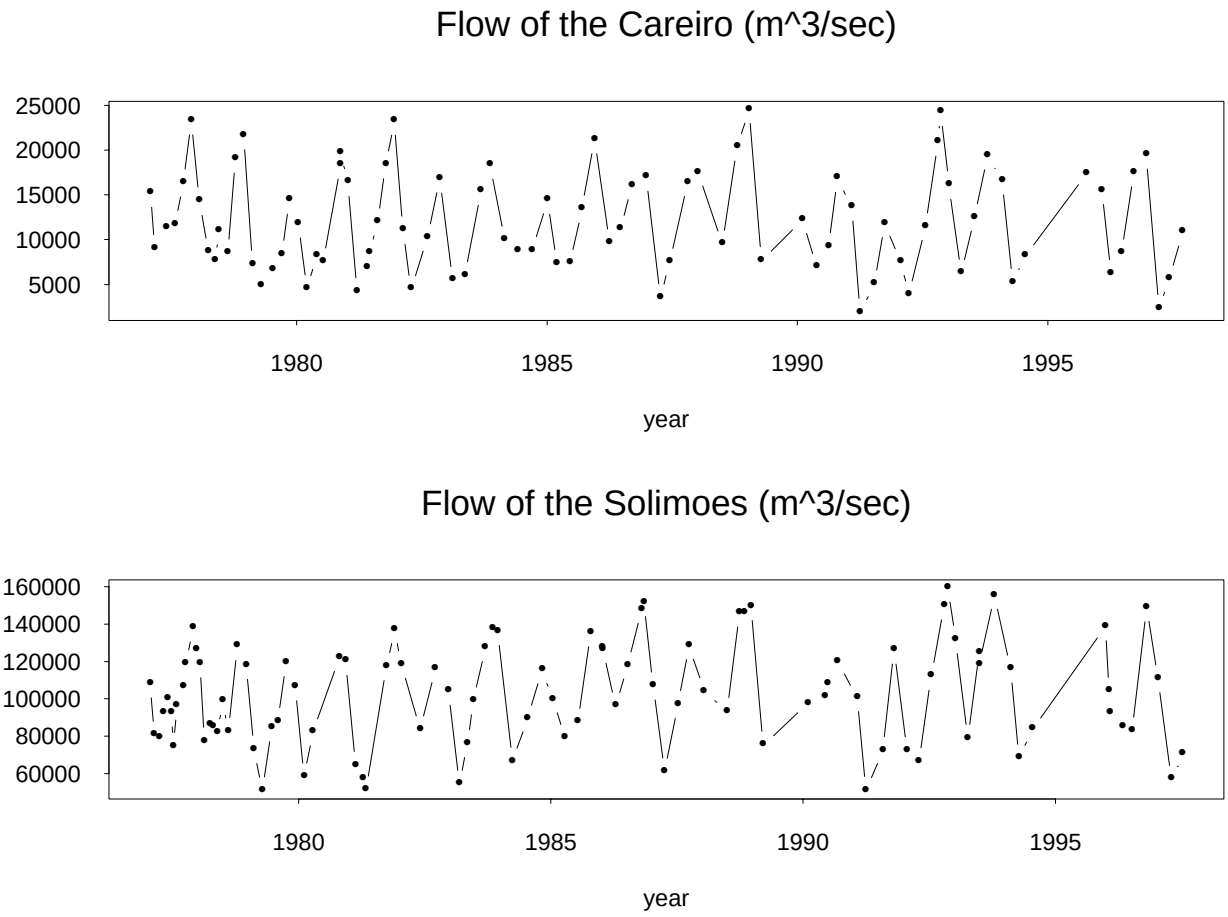
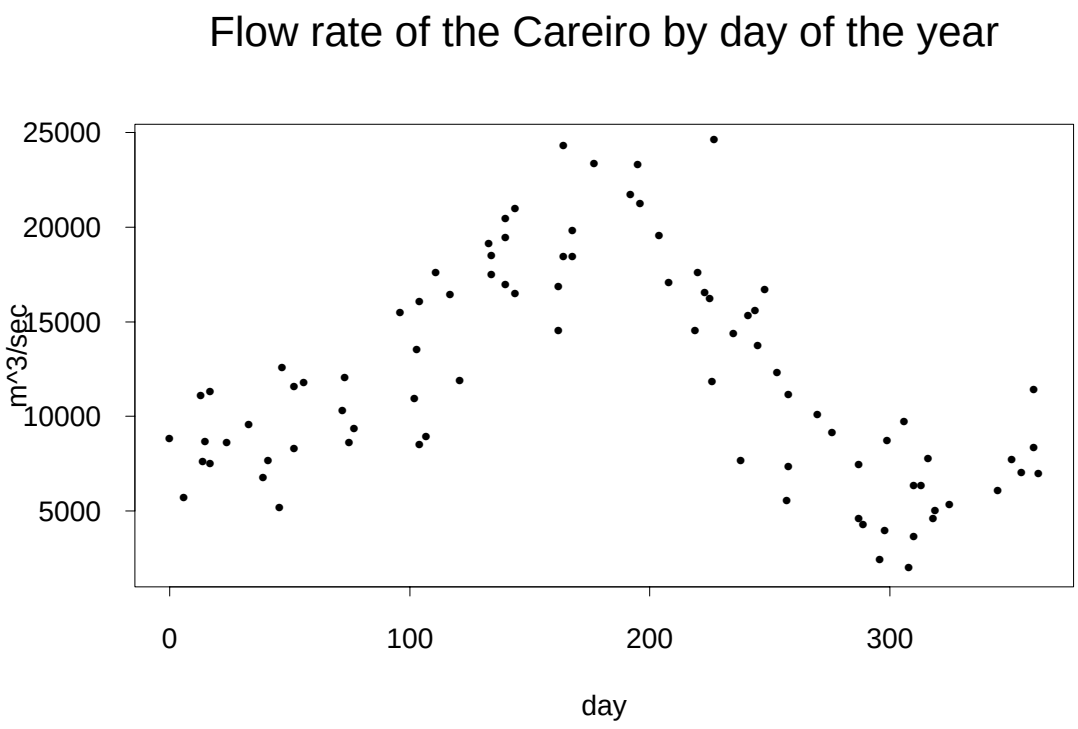


Figure 3: The discharges of the Careiro and the Solimões Rivers in cubic-meters/second. The points indicate the dates of available measurements.



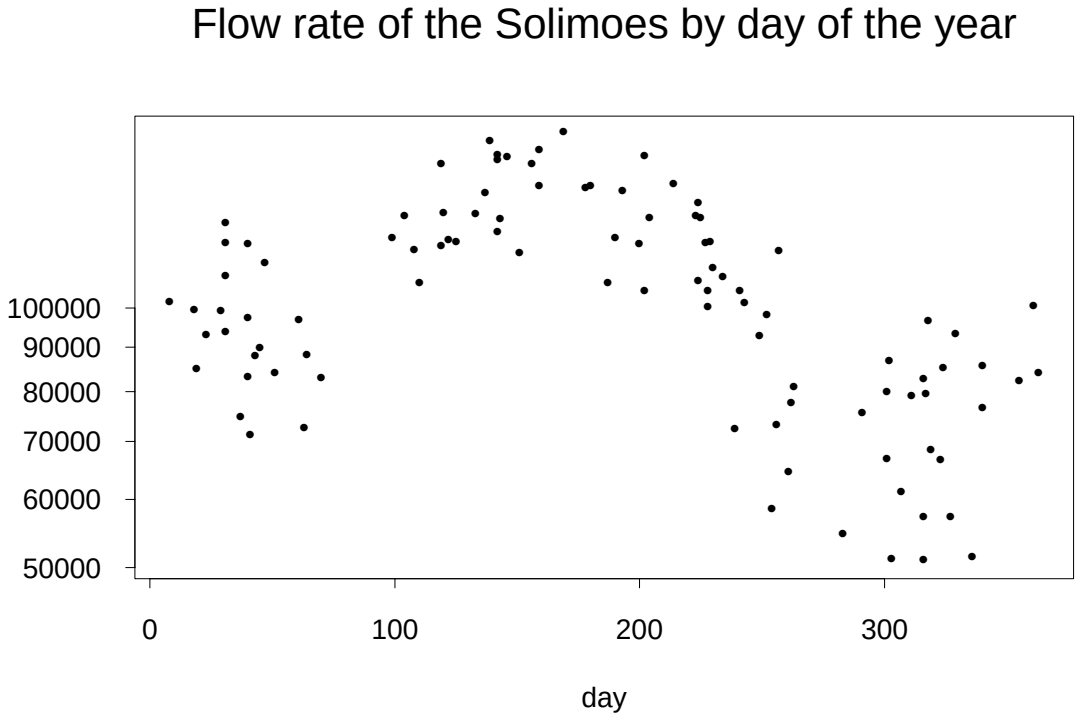


Figure 4: Observed flow rates of the two rivers plotted by day of the year.

made on different days for the two stations because they are made by a single vessel which has to journey the 90km. distance between the stations. For the Solimões there are 100 measurements in all while for the Careiro there are 90.

The basic data are the measured river discharges, discharge being defined as: the area of the section across the river where the measurements are made multiplied by the water's velocity. For these rivers the values are high going along with the fact that the Amazon outputs a substantial proportion of the world's fresh water. Figure 2 provides histograms of the flow rates and the rivers' velocities. The rivers are seen to flow at around 1 cubic-meter/sec. The discharge for the Solimões is quite a bit higher than that of the Careiro while the velocities are comparable. The discharges will fluctuate with the heights of the rivers. Time series plots of the data sets are provided in Figure 3. One notes the irregularity of the measurement dates.

Figure 4 stacks the flow data for the whole period by day of the year. An annual effect becomes apparent with the flows being high in July-August and low at October-November. One again notes the irregularities of the measurement dates.

Figure 5 provides the cumulative counts of measurements made as a function of date. Such a plot is useful for examining the stationarity of the measuring process, specifically in the stationary case the points plotted will fluctuate around a straight line. One sees for example that for the Solimões measurements were being made at a higher rate at the beginning of the period of data collection. There are two noteably flat stretches corresponding to measurements not being made often.

3 Analytic formulations

Let $Y(t)$ denote the discharge rate of the Careiro at time t and $X(t)$ that of the Solimões. Being downriver the Careiro will be considered the dependent series and the Solimões the explanatory. The model that will be considered is:

$$Y(t) \;=\; g(X(t)) \;+\; E(t), \;\; t \;=\; 0, \pm 1, \pm 2, ... \tag{1}$$

with the error series E having mean 0. One might ask for example is the function g linear? Is its derivative increasing? Is it changing with time?

Turning to statistical formulations of the problems one can start by thinking about the classical situation where the data available are $(X(t), Y(t)),\ t = 0, ..., T-1$, i.e. the observations of the two components are made at the same times and are equi-spaced.

For this model and smooth $g(.)$ one could compute a kernel estimate of $g(.)$. With the weight function $w_T(.)$ such an estimate has the form

$$\hat{g}(x) \;=\; \sum_t Y(t)\ w_T(x - X(t)) / \sum_t w_T(x - X(t))$$

at the point x. If the noise series, $\{E(t)\}$ is white with variance σ^2 the estimate's

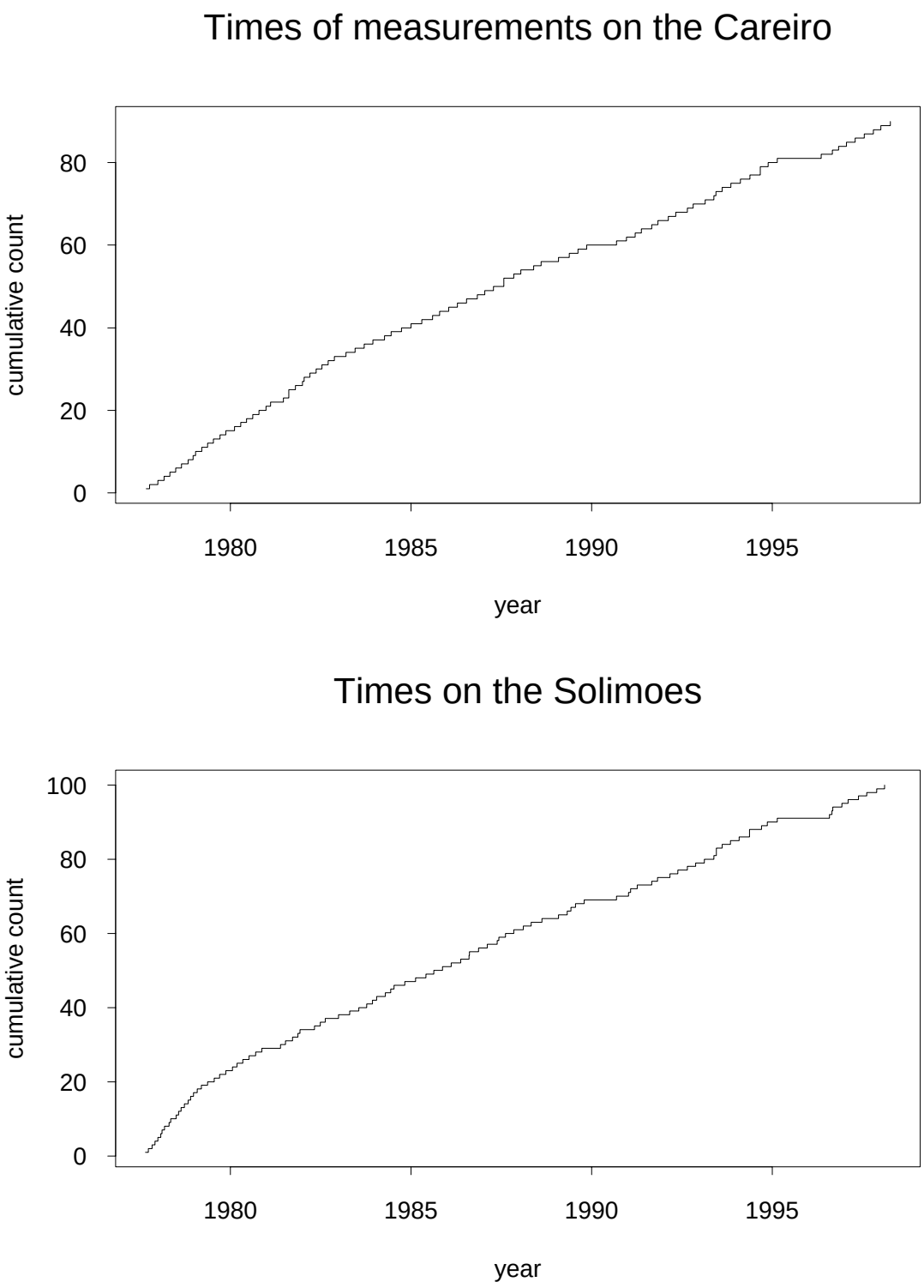


Figure 5: The cumulative counts of measurements made for each river.

variance is

$$var\{\ \hat{g}(x)|X\}\ =\ \sigma^2 \sum_t w_T(x - X(t))^2/(\sum_t w_T(x - X(t))^2$$

If the variance of $E(t)$ depends on the level of X it may be estimated for $X\ =\ x$ by

$$\hat{\sigma}^2(x)\ =\ \sum_t (Y(t) - \hat{g}(x))^2\ w_T(x - X(t))/\sum_t w_T(x - X(t)) \tag{2}$$

see Hardle [13]. This reference also provides a substantial review of nonparametric regression in the case that the noise values are independent.

Alternately to estimate the function g one might employ a local linear estimate such as produced by the Splus function loess(.) [11]. References concerned with nonparametric regression in the presence of time series errors include: [1], [15], [23], [14], [22] and [18].

The case of concern in this paper involves the data values

$$(\{X(\sigma_j)\},\ \{Y(\tau_k)\}),\ j = 1, ..., J_T;\ k = 1, ..., K_T$$

where T the observation interval is $[0, T)$ and where there are J_T observations at the times $\{\sigma_j\}$ for the Solimões and K_T at the times $\{\tau_k\}$ for the Careiro.

The observations are not in immediate correspondence, but many of the σ_j and τ_k are close. (See Figure 6 below which graphs the Careiro times relative to those of the Solimões.) This occurrence suggests that in the particular situation at hand one might be able to obtain useful a estimate of g.

Two approaches come to mind directly. One could interpolate the $Y(\tau_k)$ values to obtain estimates $\tilde{Y}(\sigma_j)$ and then work with the values $(X(\sigma_j), \tilde{Y}(\sigma_j)),\ j = 1, ..., J_T$ or one could interpolate the $X(\sigma_j)$ to obtain estimates $\tilde{X}(\tau_k)$ and then work with the values $(\tilde{X}(\tau_k), Y(\tau_k)),\ k = 1, ..., K_T$. The first approach appears the simpler and will be the one pursued. The second will be discussed later and some references provided.

Consider the model (1). The interpolated values can be written:

$$\tilde{Y}(t)\ =\ \sum_k\ B_{k,1}(t/T)Y(\tau_{k+1}))$$

with the $\{B_{k,1}\}$ B-splines of order 1, see the Appendix, [2], [12], [16] for details of these. Next

$$\tilde{Y}(\sigma_j)\ =\ Y(\sigma_j)\ +\ noise_1$$

and from (1)

$$Y(\sigma_j)\ =\ g(X(\sigma_j))\ +\ noise_2$$

leading to

$$\tilde{Y}(\sigma_j)\ =\ g(X(\sigma_j))\ +\ noise_2\ +\ noise_1 \tag{3}$$

Because of the additivity of the overall error term, equation (3) has the form of the nonparametric regression model, with the exception that perhaps the errors are

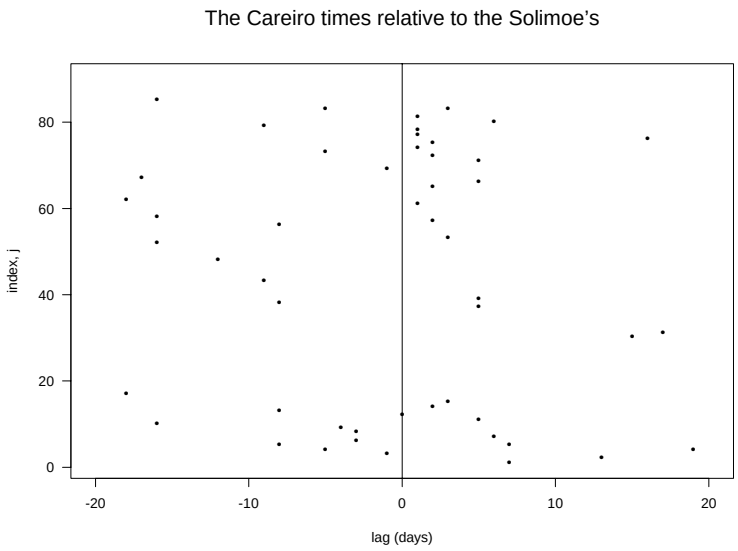


Figure 6: A plot of the points $(j,\ \tau_k - \sigma_j),\ j = 1, ..., J$

autocorrelated and heteroscedastic. The interpolation proposed involves only the $\{Y(\tau_k)\}$ values, i.e. not the $\{X(\sigma_j)\}$. The kernel estimate of $g(x)$ is now,

$$\hat{g}(x) \ = \ \sum_j \tilde{Y}(\sigma_j) w_T(x - X(\sigma_j)) / \sum_j w_T(x - X(\sigma_j)) \tag{4}$$

with $w_T(x) \ = \ {b_T}^{-1} w({b_T}^{-1} x)$ the kernel function. However the reasonableness of standard errors provided by naive programs is suspect their being based on an assumption of independent noise errors in (3).

In assessing the properties of the estimate it is assumed that the Careiro flow does not change rapidly, see Assumption b) in the Appendix. Also as it takes some time for the waters to flow from Manacapuru to Careiro time delay δ is included in the model writing $Y(t + \delta)$ instead of $Y(t)$.

Some large sample properties of the estimate, e.g. the variance, are developed in the Appendix. The estimate is asymptotically unbiased, consistent and normal.

4 Results

Already Figures 2-6 have been introduced. Figure 6 provides information on the relative timings of the Solimões and Careiro measurements. The center vertical line corresponds to the times of Solimões measurements. Spread on either side of the line are the nearby Careiro times. Careiro measurement times are seen to follow Solimões

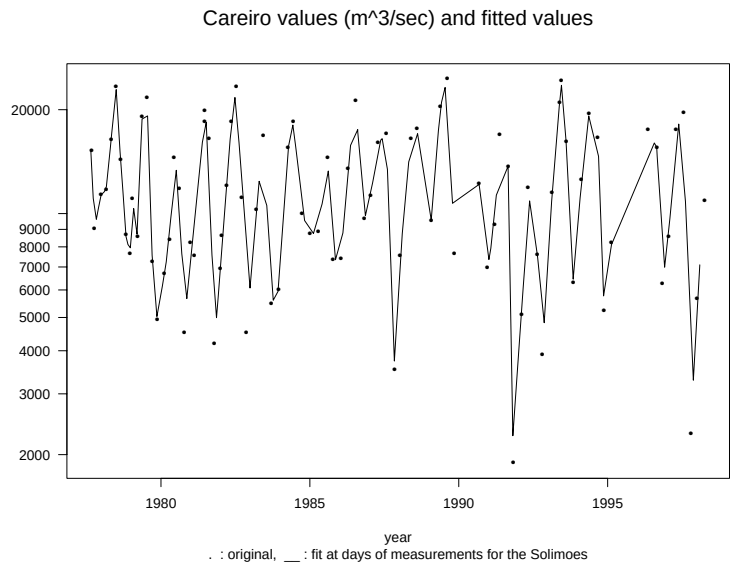


Figure 7: The points provide the positions $(\tau_k, Y(\tau_k))$, i.e. the actual data for the Careiro. The lines join the positions $(\sigma_j, \hat{Y}(\sigma_j))$, i.e. the interpolated values.

to an extent and many occur within 7 days. There are enough near 0 that it seems that reasonable estimation of g is possible. This type of rastor plot appears in [8].

We proceed following the scheme proposed in the previous section. The data available are denoted $\{Y(\tau_k)\}$ for the Careiro and $\{X(\sigma_j)\}$ for the Solimões. Being downriver from Manacapuro the Careiro series is viewed as the dependent one.

To begin a linear interpolation spline is run through the Careiro data [2], [12]. Specifically the missing Careiro discharge values were estimated by linear spline interpolation, i.e. a curve is passed through the points $(\tau_k, Y(\tau_k))$ and consists of straight lines between the points. This has the advantage that if a σ_j is actually a τ_k then the observation $Y(\tau_k)$ itself is used, i.e. no smoothing is carried out. Nearby times will have nearby values. The linear spline was employed to avoid oscillations between the data points, but higher-order splines may be considered.

In the computations the logarithms of the discharges are taken as the basic values for analysis since the variance of the additive noise appeared more nearly constant when logs were employed. Henceforth X and Y will refer to the logarithms of the basic data. Figure 7 provides the results of the interpolation of the Y values. The original data values $(\tau_k, Y(\tau_k))$ are plotted as points. The interpolated values are joined by lines. The fitted values do not appear inappropriate.

Next the values $Y(\sigma_j + \hat{\delta})$ are estimated where the σ_j are the available times for the Solimões and $\hat{\delta}$ is an estimate of the time that it takes the water to flow between the two stations. The results of the interpolation are denoted $\tilde{Y}(\sigma_j + \hat{\delta}),\ j = 1, ..., J$.

The value $\hat{\delta} = .80$ days employed is obtained using a distance of 90 km. and a speed of 1.3m/sec. (See Figure 2.)

The model considered has now become:

$$\tilde{Y}(\sigma_j + \hat{\delta}) \;=\; g(X(\sigma_j)) \;+\; noise \tag{5}$$

and it is to be noted that the *noise* is additive.

Figure 8 plots the points $(X(\sigma_j), \tilde{Y}(\sigma_j + \hat{\delta})), j = 1, ..., J$ and as a smooth curve the estimate of g computed via a kernel smoother using a gaussian kernel. The curve is approximately linear. The dashed lines are approximate 95% marginal confidence limits based on the variance formula (9) of the Appendix. Some 6 or 7 of the 87 points lie outside the 95% bounds.

The kernel smoother has the disadvantage of difficulties at boundaries, but it has the advantage of simplicity in the derivation of analytic results.

5 Goodness of fit of the model

The confidence limits of Figure 8 are basic to drawing inferences. The assumptions of independence and constant variance involved need to be considered.

First consider the assumption of constant variance. The top panel of Figure 9 plots the residuals $\hat{E}_k \;=\; Y(\tau_k) - \hat{g}(\tilde{X}(\tau_k - \hat{\delta})$. One notes some narrowing on the righthand side, but perhaps not enough to invalidate the use of an assumption of heteroscedasticity.

Next the assumption of independence is examined. A convenient way to look for time series autocorrelation is to examine a periodogram. Periodograms can highlight a broad variety of departures from white noise, see the discussion in the Appendix. Here the form

$$\frac{1}{2\pi T}|\sum_k \hat{E}_k \; exp\{-i\lambda\tau_k\}|^2, \;\; -\infty < \lambda < \infty$$

is computed. Inferences can be made based on approximating the distribution by a multiple of a chi-squared. This is discussed in the Appendix.

The results are presented in Figure 9. There are so many points in the periodogram plot because the sampling interval employed in its computation was 1 day. One notes 5.03% of the periodogram ordinates lying outside the 95% confidence limits in the botton panel of Figure 9. One has no strong evidence for a departure from approximately white noise errors.

In summary the results of the studies presented in the two panels of Figure 9 suggest that assumptions leading to the uncertainty limits of Figure 8 may not be inappropriate.

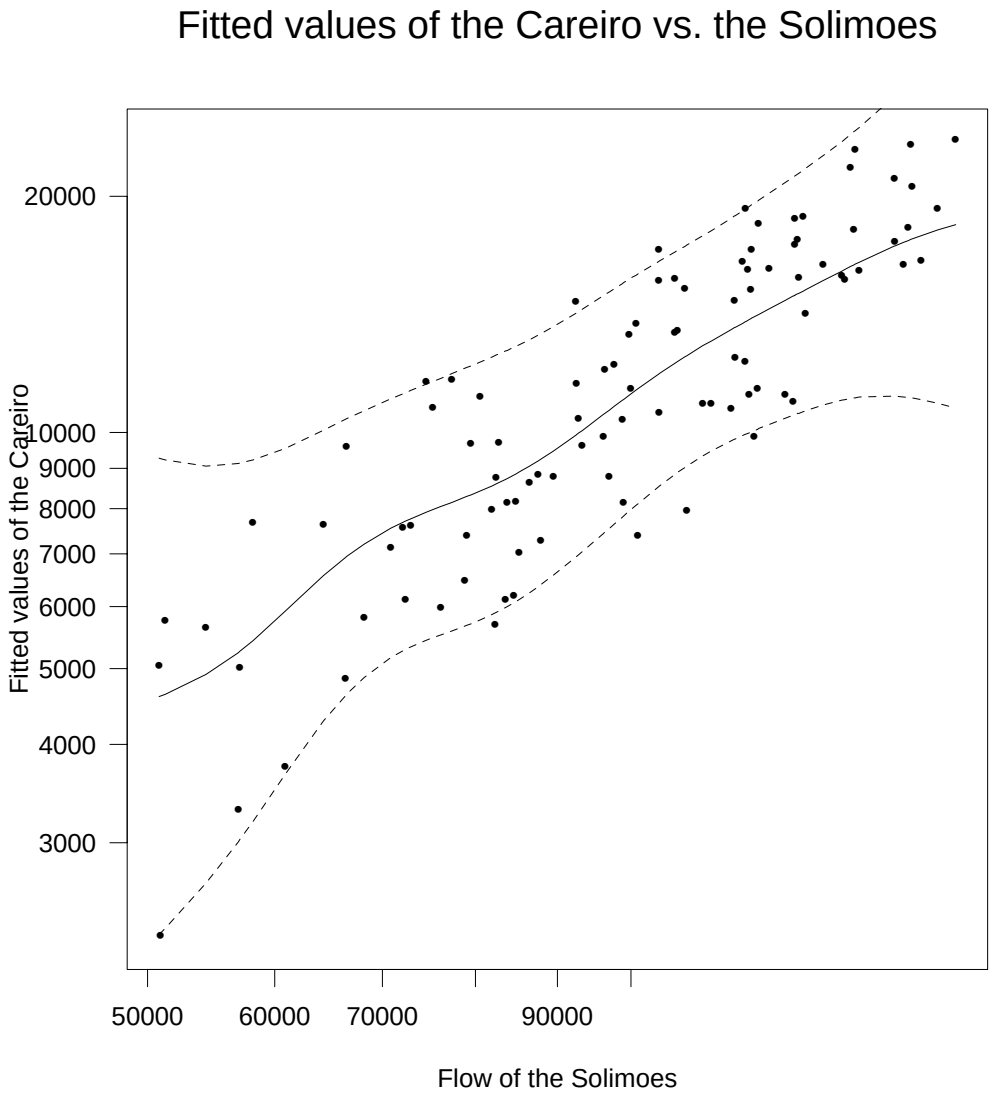


Figure 8: The points are $(X(\sigma_j, \tilde{Y}(\sigma_j))$, the interpolated Careiro observations plotted against the actual Solimões values. The dashed lines are approximate marginal 95% confidence limits derived using expression (9).

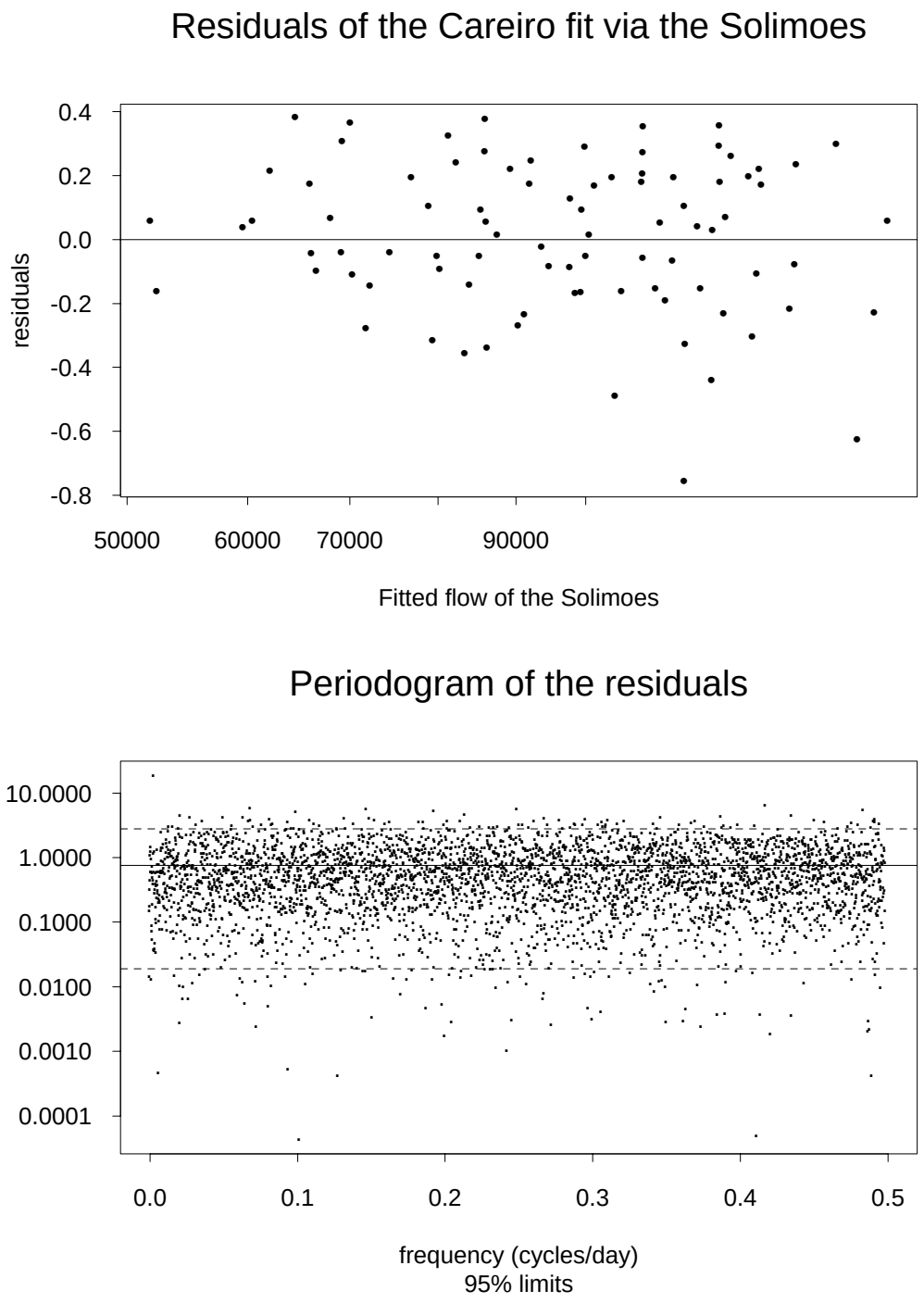


Figure 9: The top panel provides the residuals of the Careiro data as explained by the interpolated Solimões values. The bottom panel provides the periodogram. The horizontal line in it is the average of all the periodogram values. The dashed lines in the lower panel provide approximate 95% marginal confidence limits.

6 Discussion and summary

The series $\{X(t\})$ and $\{Y(t)\}$ were sampled irregularly and at different times, yet it has proved possible to proceed to study their relationship via smoothing and Fourier analysis. To begin some descriptive statistics were presented, then modelling was carried out. An approximate linear dependence of the Careiro flow upon that of the Solimões flow was found.

Contributions of the work include assessing the whiteness of a sampled time series via a periodogram and derivation of some properties of a nonparametric estimate involving interpolation of the dependent variate's values.

The method of predicting the $Y(\sigma_j)$ could perhaps be improved, e.g. by including a seasonal effect explanatory series. There are other methods of constructing approximate confidence intervals such as the bootstrap. Also the problem of the estimation of the smoothing parameter, b_T, might be considered.

As mentioned earlier an alternate approach to dealing with the irregularity of observation times is to interpolate the $\{X(\sigma_j\})$ to the $\{\tau_k\}$ time points. This may be implimented by an interpolation spline as was done to obtain the $\{\tilde{Y}(\sigma_j)\}$. Supposing the values $(\tilde{X}(\tau_k), Y(\tau_k))$ are then employed in a nonparametric regression analysis the circumstance could be viewed as an errors in variables problem. One difference from the classical situation though is that here one has some control over the bias and the error variance through choice of the interpolation/smoothing procedure. There is a literature on this situation, [3] and [10]. There is also a related literature on endogeneity in nonparametric models [5]. The approach actually implimented is simpler because of the additivity of the noise, the fact that one may argue conditionally on the X-values and because it seemed allowable to treat the error values as white noise. It remains to be seen which approach leads to better estimates and this surely depends on the values of J_T and K_T, which series can be better interpolated amongst other things. It may be that interpolating each series to the same equi-spaced grid is to be preferred.

In connection with the interpretation of the results of the data analyses it needs to be remembered that the work is preliminary. It will be further developed in [21].

Acknowledgements

The work of this paper was supported by NSF Grants DMS 97-04739, DMS 99-71309 and DMS 02-03921. As indicated at the outset, the problem and the data were brought to this writer's attention by Professor H. O'R. Sternberg of the Department of Geography, University of California, Berkeley. I thank him profusely for the interactions on this topic and on so many others. Professor Sternberg obtained the data from: CPRM (Companhia de Pesquisas de Recursos Minerais), ANEEL (Agência Nacional de Energia Elétrica), DNAEE (Dep. Nacional de Aguas e Energia Eléctrica) through the courtesy of Eduardo de Freitas Madeira and Fernando Pereira

de Carvalho.

Jim Powell made some helpful comments and provided a preprint of [5]. Phil Spector helped with some Splus coding. The Referees made comments leading to improvements in the paper. One suggested the consideration of the log transform. Michael Last carried out a literature search on nonlinear models involving measurement errors. I thank them all.

Constance van Eeden has done a great deal for statistics in Canada. It is indeed an honour to know her and to contribute to her *Festschrift*. (I hope that she won't be put off by my not so "very careful analysis".)

David R. Brillinger, Department of Statistics, University of California, Berkeley, CA 94720-3860, `brill@stat.berkeley.edu`.

References

[1] N. S. Altman. Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, 85:749–759, 1990.

[2] K. E. Atkinson. *An Introduction to Numerical Analysis.* J. Wiley, New York, 1978.

[3] S. M. Berry, R. J. Carroll, and D. Ruppert. Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.*, 97:160–169, 2002.

[4] P. Bloomfield. Spectral analysis with randomly missing observations. *J. Royal Statist. Soc.*, 32:369–380, 1970.

[5] R. Blundell and J. L. Powell. Endogeneity in nonparametric and semiparametric regression models. In L.Hansen and S. Turnovsky, editors, *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Econometric Society Monograph Series. Cambridge University Press, 2002.

[6] D. R. Brillinger. *Time Series: Data Analysis and Theory.* Holt-Rinehart, New York, 1975.

[7] D. R. Brillinger. A continuous form of post-stratification. *Ann. Inst. Stat. Math.*, 31:271–277, 1979.

[8] D. R. Brillinger. Nerve cell spike train data analysis: A progression of technique. *J. Amer. Statist. Assoc.*, 87:260–272, 1992.

[9] D. R. Brillinger. Examining an irregularly sampled time series for whiteness. *Resenhas*, 4:423–431, 2000.

[10] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement Error in Nonlinear Models.* Chapman and Hall, London, 1995.

[11] W. S. Cleveland, E. Grosse, and W. M. Shyu. Local regression models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, pages 309–376, Pacific Grove, California, 1992. Wadsworth.

[12] C. de Boor. *A Practical Guide to Splines*. Springer, New York, 1978.

[13] W. Hardle. *Smoothing: With Implementation in S*. Springer, New York, 1991.

[14] W. Hardle and P-D. Tuan. Some theory on M-smoothing of time series. *J. Time Series Analysis*, 7:191–204, 1986.

[15] J. D. Hart. Kernel regression estimation with time series errors. *J. Royal Statist. Soc.*, 53:173–187, 1991.

[16] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[17] R. E. Oltman, H. O'R. Sternberg, Ames F. C., and L.C. Davis, Jr. *Amazon River Investigations, Reconnaissance Measurements of July 1963*, volume 486 of *Geological Survey Circular*. U.S. Department of the Interior, Washington, 1964.

[18] P. M. Robinson. Large-sample inference for nonparametric regression with dependent errors. Technical Report EM/97/336, London School of Economics, London, 1997.

[19] H. O'R. Sternberg. Radiocarbon dating as applied to a problem of Amazonian morphology. In *Comptes Rendus XVIIIe Congrès International de Géographie, Rio de Janeiro - 1956*, volume 2, pages 399–424, Rio de Janeiro, 1960. Comité National du Brésil.

[20] H. O'R. Sternberg. *A água e o Homem na Vârzea do Careiro (2nd ed.)*, volume 1, 2 of *Coleção Friedrich Katzer*. CNPq Museu Paraense Emílio Goeldi, Belém, Pará, 1998.

[21] H. O'R. Sternberg and D. R. Brillinger. Streamflow at branching channels; geomorphic and statistical issues from a case study in the Brazilian Amazon. *In preparation*, 2002.

[22] L. Tran, G. Roussas, S. Yakowitz, and B. T. Van. Fixed-design regression for linear time series. *Ann. Statist.*, 24:975–991, 1996.

[23] Y. K. Truong. Nonparametric curve estimation with time series errors. *J. Stat. Planning Inference*, 28:167–183, 1991.

[24] C. van Eeden. Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Ann. Inst. Stat. Math.*, 37:461–472, 1985.

Appendix

Throughout the Appendix the arguments are conditional on X as is usual in statistical inference.

The theorems do not present best possible results rather the material has been set up to suggest possible routes for later work. In the data analyses linear splines have been employed, and the analytic results developed. Further the error process E has been assumed to have cumulants of all orders in order to obtain asymptotic normaility directly. Lastly the case of a kernel smoother is the one studied. This was done in the interests of simpler proofs. The results are preliminary, proofs are sketched and assumptions have been set down rather casually. It is possible to bound the errors of all the approximations however.

Consistency and asymptotic normality.

The model considered is

$$Y(t) \;=\; g(X(t)) \;+\; E(t),\; t = 0, ..., T-1$$

with the distinction that the data values available are

$$X(\sigma_j),\; j = 1, ..., J_T;\; Y(\tau_k)),\; k = 1, ..., K_T$$

with $0 \;\leq\; \sigma_k,\; \tau_k \;<\; T$. The results are developed for the case of a kernel smoother.
Notations.

The spline of order 1 based on the points $\{(\tau_k, Y(\tau_k))\}$ is denoted $\tilde{Y}(t)$. It may be written

$$\tilde{Y}(t) \;=\; \sum_k Y(\tau_{k+1}) B_{k,1}(t/T)$$

where the k-th B-spline function of degree 0 with knots τ_k is given by

$$B_{k,1}(t/T) \;=\; (t - \tau_k)/(\tau_{k+1} - \tau_k) \quad for \quad \tau_k \leq t < \tau_{k+1}$$

$$=\; (\tau_{k+2} - t)/(\tau_{k+2} - \tau_{k+1}) \quad for \quad \tau_{k+1} \leq t < \tau_{k+2}$$

and equals 0 otherwise.

The kernel smooth studied to estimate $g(x_0)$ is

$$\hat{g}(x_0) \;=\; \sum_j \; w_T(x_0 - X(\sigma_j))\tilde{Y}(\sigma_j) \;/\; \sum_j \; w_T(x_0 - X(\sigma_j)) \tag{6}$$

In this definition

$$w_T(x) \;=\; b_T^{-1} w(b_T^{-1} x)$$

with $b_T \;>\; 0$ a binwidth.

To develop asymptotic approximations the following forms are used

$$g(X(t)) \;=\; H_T(t) \;=\; h(t/T),\; X(t) \;=\; x(t/T)$$

For a a function on [0,1]

$$||a|| \;=\; \sup_u |a(u)|$$

Assumptions. Basic ones include

a) The kernel w is non-negative, of finite support, continuous and $\int w(u)du = 1$.

b) The functions h and x are in $C^1[0,1]$. The equation $x(u) = x_0$ has a finite number of solutions.

c) The sequence $\{\tau_k\}$ is strictly increasing. There exist $F_M(u)$ and 1-1 $F_N(v)$ such that

$$\hat{F}_M(u) = \frac{1}{J_T}\#\{\sigma_j/T \le u; j = 1, ..., J_T\} \to F_M(u)$$

and

$$\hat{F}_N(v) = \frac{1}{K_T}\#\{\tau_k/T \le v; k = 1, ..., K_T\} \to F_N(v)$$

weakly as $T \to \infty$. The functions F_M and F_N have densities f_M and f_N.

d) The process E is zero mean white noise and has cumulants of all orders.

Results.

Lemma 1. Under Assumption b)

$$\tilde{H}_T(t) = H_T(t) + O(\Delta_T^2 T^{-2}||h^{(2)}||) \tag{7}$$

as $T \to \infty$ where $\Delta_T = max\{\tau_{k+1} - \tau_k,\ k = 1, ..., K_T\}$ and h' is the derivative of h. The error term is uniform in t.

Proof. Since h is in $C^1[0,1]$, $\tilde{h}$ the spline of order 1 passing through the points $(v_k, h_k = h(v_k))$, satisfies

$$|h(v) - \tilde{h}(v)| \le \frac{1}{8}\Delta^2 \sup_v |h^{(2)}(v)|$$

where $\Delta = max\{v_{k+1} - v_k\}$, [12]. The result of the Lemma follows taking $v = t/T$, $v_k = \tau_k/T$ and $H_T(t) = h(t/T)$.

Theorem 1. Under the Assumptions

$$E\{\hat{g}(x)\} = g(x) + O(b_T) + O(\Delta_T^2 T^{-2}) \tag{8}$$

Proof. To begin because the spline is linear in the data values

$$\tilde{Y}(t) = \tilde{H}_T(t) + \tilde{E}(t)$$

From the result (7) and the fact that the series E has mean 0

$$E\{\sum_j \tilde{Y}(\sigma_j) w_T(x - X(\sigma_j))\} \ / \ \sum_j w_T(x - X(\sigma_j))$$

$$= \sum_j (g(X(\sigma_j)) + O(\Delta_T T^{-1})) w_T(x - X(\sigma_j)) \ / \ \sum_j w_T(x - X(\sigma_j))$$

From the non-negativity, the finite support of w and the boundedness of the derivative of h

$$g(X(\sigma_j)) = g(x) + O(b_T)$$

uniformly giving the result.

Corollary. Under the Assumptions and if $\Delta_T/T,\ b_T \ \rightarrow\ 0$ as $T \ \rightarrow \ \infty$ the estimate $\hat{g}(x)$ is asymptotically unbiased.

Theorem 2. Under the Assumptions with the $E(\tau_k)$ independent and of variance σ^2

$$var\{\hat{g}(x_0)\} \ = \ \sigma^2 \sum_k \ [\sum_j \ w_T(x_0 - X(\sigma_j))B_{k,1}(\frac{\sigma_j}{T})]^2 \ / \ (\sum_j \ w_T(x_0 - X(\sigma_j)))^2 \quad (9)$$

Proof. By construction

$$\tilde{E}(t) \ = \ \sum_k \ E(\tau_{k+1})B_{k,1}(\frac{t}{T})$$

Now

$$\sum_j \tilde{E}(\sigma_j)w_T(x - X(\sigma_j)) \ = \ \sum_k [\sum_j B_{k,1}(\frac{\sigma_j}{T})w_T(x - X(\sigma_j)]E(\tau_{k+1}) \qquad (10)$$

and the result is immediate since the error series, E, is white noise with variance σ^2.

The form (9) may be contrasted with that were the values $X(\sigma_j), Y(\sigma_j)),\ j = 1, ..., J_T$ available. The variance then would be

$$\sigma^2 \sum_j w_T(x_0 - X(\sigma_j))^2 \ / \ (\sum_j w_T(x_0 - X(\sigma_j)))^2$$

Corollary. With $\hat{E}(\tau_k) \ = \ Y(\tau_k) - \hat{g}(\tilde{X}(\tau_k))$ the error variance σ^2 may be estimated by

$$\sum_k \hat{E}(\tau_k)^2 \ / \ K_T$$

The asymptotic behavior of the variance may be investigated further. In the proofs below certain sums will be replaced by integrals. The accuracy of these approximations may be bounded using Lemma 4 below. Basically one will want $\sup_u|\#\{\sigma_j/T \ \leq \ u\}/J_T - F_M(u)|$ and $\sup_v|\#\{\tau_k/T \ \leq \ v\}/K_T - F_N(v)|$ to become small at a fast enough rate as $J_T,\ K_T \rightarrow \infty$.

Consider the variance at $X(t) \ = \ x_0$.

Lemma 2. Under asumptions like those already indicated plus $b_T \rightarrow 0$ and $J_T \rightarrow \infty$ as $T \rightarrow \infty$ the denominator of (6) is asymptotically

$$(J_T \ \sum\nolimits' f_M(x^{-1}(x_0))/|x'(x^{-1}(x_0))|)^2$$

where the sum, $\sum'$ is over the solutions of $x(u) \ = \ x_0$.

Proof. Consider

$$\frac{1}{J_T} \ \sum_j w_T(x_0 - x(\sigma_j/T))$$

$$\begin{aligned} &= \int w_T(x_0 - x(u)) d\hat{F}_M(u) \\ &\approx \int b_T^{-1} w(b_T^{-1}(x_0 - x(u)) f_M(u) du \\ &\approx \sum{}' f_M(x^{-1}(x_0) \frac{1}{|x'(x^{-1}(x_0))|} \int w(u) du \end{aligned}$$

Bounds on the error of approximating the sum by an integral may be found using Lemma 7.

Lemma 3. Under the assumptions of Lemma 2 plus $K_T \to \infty$ the numerator of (6) is asymptotically

$$J_T^2 \frac{1}{b_T K_T} \sum{}' f_N(x^{-1}(x_0)^{-1} f_M(x^{-1}(x_0))^2 \frac{1}{|x'(x^{-1}(x_0)|} \int w(u)^2 du$$

Proof. Consider

$$\sum_k [\sum_j B_{k,1}(\frac{\sigma_j}{T}) w_T(x_0 - X(\sigma_j))]^2 \approx J_T^2 \sum_k [\int B_{k,1}(u) w_T(x_0 - x(u)) f_M(u) du]^2$$

Because $2TB_{k,1}(u)/(\tau_{k+2} - \tau_k)$ acts like a Dirac delta function centered at τ_{k+1}/T for a continuous function a

$$\int B_{k,1}(u) a(u) du \approx a(\tau_{k+1}/T)(\tau_{k+2} - \tau_k)/2T$$

With the approximation

$$\frac{\tau_{k+2} - \tau_k}{2T} \approx \frac{1}{K_T f_N(\tau_k/K_T)}$$

the expression being studied is approximately

$$\frac{J_T^2}{K_T} \sum_k \frac{(\tau_{k+2} - \tau_k)^2}{4T^2} w_T(x_0 - x(\tau_{k+1}/T))^2 f_M(\tau_{k+1}/T)^2 \approx \frac{J_T^2}{K_T} \int \frac{1}{f_M(v)^2} w_T(x_0 - x(v))^2 f_M(v)^2 f_N(v) dv$$

With the change of variable $s = x(v)$ this becomes

$$\frac{J_T^2}{K_T} \int w_T(x_0 - s)^2 f_M(x^{-1}(s))^2 f_N(x^{-1}(s))^{-1} \frac{1}{|x'(x^{-1}(s))|} ds$$

giving the indicated result as $T \to \infty$.

Corollary 1. The variance of $\hat{g}(x_0)$ is asymptotically

$$\frac{\sigma^2}{b_T K_T} \sum{}' f_N(x^{-1}(x_0))^{-1} f_M(x^{-1}(x_0))^2 \frac{1}{|x'(x^{-1}(x_0)|} \int w(u)^2 du \,/\, (\sum{}' f_M(x^{-1}(x_0))/|x'(x^{-1}(x_0))|)^2$$

Corollary 2. Assuming in addition that $b_T K_T \to \infty$ the estimate is consistent.

The result simplifies when $f_N \equiv f_M$.

Theorem 3. Assuming that $E(t)$ has finite cumulants, $\kappa_m,\ m = 1, 2, ...$

$$(\hat{g}(x) \ - \ E\{\hat{g}(x)\}) \ / \ \sqrt{var\{\hat{g}(x)\}}$$

is asymptotically normal with mean 0 and variance 1.

Proof. The cumulant of order m of (10) is

$$\kappa_m \ \sum_k [\sum_j B_{k,1}(\frac{\sigma_j}{T}) w_T(x_0 - X(\sigma_j))]^m$$

Arguing as above one sees that this is asymptotically

$$\frac{\kappa_m}{b_T^{m-1} K_T^{m-1}} \sum{}' f_N(x^{-1}(x_0))^{1-m} f_M(x^{-1}(x_0))^m \frac{1}{|x'(x^{-1}(x_0)))|} \int w(u)^m du$$

The standardized cumulants of order $m > 2$ are seen to tend to 0 as $T \to \infty$. The normal distribution being determined by its moments, the Theorem follows.

Sampled time series.

Consideration next turns to the logic lying behind the model assessment via the periodogram of the residuals.

A sampled time series may be represented as $\{R(t) \ = \ M(t)X(t),\ t = 0, \pm 1, \pm 2, ...\}$ where $\{X(t),\ t = 0, \pm 1, \pm 2, ...\}$ is the series and $\{M(t),\ t = 0, \pm 1, \pm 2, ...\}$ is a 0-1 valued series representing the sampling times, $\{\sigma_j\}$. M takes on the value 1 when the X observation is present and 0 otherwise. The non-zero values of R are the $\{X(\sigma_j)\}$. The series R reflects the statistical properties of the series X and M. For example if X and M are mutually independent and stationary with power spectra $f_{XX}(\lambda),\ f_{MM}(\lambda)$ then the process R has power spectrum

$$f_{RR}(\lambda) \ = \ c_M^2 f_{XX}(\lambda) \ + \ \int_{-\pi}^{\pi} f_{MM}(\lambda - \alpha) f_{XX}(\alpha) d\alpha$$

assuming further that X has mean 0 and that M has mean c_M. This expression was given in [4] and follows from Example 2.10.4 in [6]. From it one sees that if X is white noise then the spectrum of R is constant. The heuristics of this result are clear: if the series is white then the sampled values are a separate selection and themselves white. One has a means of assessing whether a sampled time series is white. (These remarks correct some incorrect discussion in [9].)

The spectrum of the series R may be estimated by smoothing the periodogram

$$\frac{1}{2\pi T} |\sum_0^{T-1} \ e^{-i\lambda t} R(t)|^2 \ = \ \frac{1}{2\pi T} |\sum_1^{K_T} \ e^{-i\lambda \tau_k} X(\tau_k)|^2$$

for example. In many situations when L_T periodogram ordinates are averaged $\hat{f}_{RR}(\lambda)/f_{RR}(\lambda)$ is distributed approximately as $\chi^2_{2L_T}/2L_T$. When $L_T \rightarrow \infty$ as $T \rightarrow \infty$ the asymptotic distribution of $log(\hat{f}_{RR}(\lambda)/f_{RR}(\lambda))$ is normal with mean 0 and variance $1/L_T$.

For examples of the assumptions leading to such results see e.g. [6]. Such results are often used to set confidence limits for a spectrum estimate. They may also be used to test whether $f_{RR}(\lambda) = \gamma^2$, e.g. by constructing confidence bounds in a plot of the estimated spectrum $\hat{f}_{RR}(\lambda)$. Such a test will be consistent at frequency λ as $L_T \rightarrow \infty$ for

$$\sqrt{L_T}\ log(\hat{f}_{RR}(\lambda)/\gamma^2) =$$
$$\sqrt{L_T}\ log(\hat{f}_{RR}(\lambda)/f_{RR}(\lambda) + \sqrt{L_T}\ log(f_{RR}(\lambda)/\gamma^2)$$

The first term here is asymptotically standard normal while when $f_{RR}(\lambda) \neq \gamma^2$ the absolute value of the second grows to infinity as $L_T \rightarrow \infty$.

Figure 9 presents the spectral results taking $R = \hat{E}_k = Y_k - \hat{g}_k$. In Figure 9 the baseline plotted is an estimate of γ^2, specifically the average of all the periodogram ordinates. It tends to $\int f_{RR}(\lambda)d\lambda$ and has variance $O(1/T)$, i.e. its variability may be neglected.

Lemma 4. Let $h(u)$ be a function with variation $V(h)$. Let $F(u)$ be a distribution function on $(-\infty, \infty)$. Given $u_1,\ u_2,\ ...,\ u_J$ let

$$\hat{F}(u) = \#\{j|u_j \leq u\}/J$$

and let

$$d_J = \sup_u |\hat{F}(u) - F(u)|$$

Then

$$|\frac{1}{J}\sum_j h(u_j) - \int h(u)du| \leq V(h)d_J$$

Proof. See [7].