Yadolah Dodge · Joe Whittaker
Editors

Computational Statistics

Volume 1

Proceedings of the
10th Symposium on
Computational Statistics

COMPSTAT
Neuchâtel, Switzerland, August 1992

With 102 Figures

Physica-Verlag
A Springer-Verlag Company

Forecasting Using a Semiparametric Model
R. Cao, W. González-Manteiga, J.M. Prada-Sánchez, I. García-Jurado and M. Febrero-Bande 327

Computing ARMA Models with MATLAB
J.F. Emmenegger 331

Analyzing Treatment Effects - The WAMASTEX Approach to Paired Sample Data
K.A. Froeschl, W. Dorda and W. Grossmann 337

FORTUNE: Improving Forecasts by Tuning the Forecasting Process
T. Kötter and A. Benner 343

A Comparative Study of Outlier Detection and Missing Value Estimation Methods Applied to Time Series Transport Data
E.J. Redfern, S.M. Watson, S.D. Clark and M.R. Tight 349

Linking Judgements to Forecasting Models
M. Talbot, T. Koetter, A. Benner, G. Ricken, A. Prenneis, J. Medding, R. Bohan, G. Hawkins, F. Kelly and Z. Luo 355

Numerical Computation of Exact Distributions for First Order Stochastic Difference Equations
A.K. Tsui 359

VII. NONLINEAR REGRESSION

Estimation of Radioimmunoassay Data Using Robust Nonlinear Regression Methods
H.P. Altenburg 367

An Artificial Intelligence Approach for Modeling in Nonlinear Regression Parametric Models
N. Caouder 373

Accurate Multivariable Numerical Derivatives for Inference in Nonlinear Regression
R. Gonin 379

Advantages of the Approximative Interpretation - An Algorithm and Program for Finding Solutions of Large Non-Linear Problems
B. Jasiński 385

Providing for the Analysis of Generalized Additive Models within a System already Capable of Generalized Linear and Nonlinear Regression
P.W. Lane and T.J. Hastie 391

A Note on Local Sensitivity of Regression Estimates
H. Nyquist 397

Parallel Model Analysis with Factorial Parameter Structure
G.J.S. Ross 403

VIII. ROBUSTNESS AND SMOOTHING TECHNIQUES

Time-Efficient Algorithms for Two Highly Robust Estimators of Scale
C. Croux and P.J. Rousseeuw 411

Universal Consistency of Partitioning-Estimates of a Regression Function for Randomly Missing Data
A. Carbonez, L. Györfi and E.C. van der Meulen 429

The Use of Slices in the LMS and the Method of Density Slices: Foundation and Comparison
G. Antille and H. El May 441

On Some Statistical Properties of Bézier Curves
A. Blejec 446

TRADE Regression
J.B. Dijkstra 453

A Review on Smoothing Methods for the Estimation of the Hazard Rate Based on Kernel Functions
O. Gefeller and P. Michels 459

Departures from Assumptions in Two-Phase Regression
H.J. Kim 465

An Analysis of the Least Median of Squares Regression Problem
N. Krivulin 471

Nonparametric Regression Methods for Smoothing Curves
M.A.A. Moussa 477

An Inference Procedure for the Minimum Sum of Absolute Errors Regression
S.C. Narula, G. Stangenhaus and P. Ferreira 481

An Operator Method for Backfitting with Smoothing Splines in Additive Models
M.G. Schimek, H. Stettner and J. Haberl 487

Sensitivity Analysis in Structural Equation Models
Y. Tanaka, S. Watadani and K. Inoue 493

Methods for Robust Non-Linear Regression
B.C. van Zomeren 499

IX. INDUSTRIAL APPLICATIONS: PHARMACEUTICS AND QUALITY CONTROL

Statistical Thinking in Total Quality Management: Implications for Improvement of Quality - with Ideas for Software Development
T.J. Boardman and E.C. Boardman 505

Statistical Computation in the Pharmaceutical Industry
A. Racine-Poon 523

Stochastic Simulations of Population Variability in Pharmacokinetics
S. Guzy, C.A. Hunt and R.D. MacGregor 533

Computational Aspects in Uncertainty Analyses of Physiologically-Based Pharmacokinetic Models
L. Edler 539

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Jour. Roy. Statist. Soc., Series B*, 39(1):1-38.

Harvey, A. (1989). *Forecastin, structural time series models and the Kalman filter*. Cambridge Univ. Press, Cambridge.

Harrison, P.J. and Stevens, C.F. (1976). "Bayesian Forecasting" *Journal of the Roy. Statist. Soc., Series B*, 38:205-247.

Martin, R.D. and Yohai, V.J. (1992). "Highly robust estimation of autoregressions". Technical Report, Dept. of Statistics, Univ. of Washington, Seattle, WA.

Masreliez, C.J. (1975). "Approximate non-Gaussian filtering with linear state and observation relations" *IEEE Trans. Automat. Control*, 20:107-110.

Masreliez, C.J. and Martin, R.D. (1977). "Robust Bayesian estimation for the linear model and robustifying the Kalman filter" *IEEE Trans. on Automatic Control*, AC-22:361-371.

Shumway, R.H. and Stoffer, D.S. (1991). "Dynamic linear models with switchin." *Jour. of the Amer. Statist. Assoc.*, 86:763-769.

Smith, A. and West, M. (1983). "Monitoring renal transplants: an application of the Multiprocess Kalman filter." *Biometrics*, 39:867-878.

West, M. and Harrison, P.J. (1986). "Monitoring and adaptation in Bayesian forecasting models." *Jour. of the Amer. Statist. Assoc.* 81:741-750.

West, M. and Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.

Forecasting Using a Semiparametric Model

R. Cao, W. González-Manteiga, J.M. Prada-Sánchez, I. García-Jurado and M. Febrero-Bande

Department of Statistics and Operations Research. University of Santiago de Compostela. 15771 Santiago de Compostela. Spain

Abstract

In this paper we present a forecasting system that has been used to control the contamination in the surroundings of a Power Station in Northwestern Spain. The system provides forecastings of the immision levels every five minutes as well as confidence intervals for such levels.

1. INTRODUCTION

The concentration of SO_2 in the soil, known as immision, is a variable associated to the contamination produced by coal combustion. As it is a good indicator of the alterations on the atmospheric environment caused by the activity of a (coal) power station, in most developed countries there are official rules that this kind of industries must regard to control such a variable. In Spain, for instance, the mean of immision levels recorded in the proximities of a power station in the last two hours should not exceed certain values. It is a remarkable fact that since the actions to reduce the future immision levels are taken until that reduction is achieved some time goes by. Consequently, in a Spanish power station, in order to regard the laws (optimizing the productiveness of the plant), it would be very helpful to have a prediction device to estimate future immision levels.

In this work we describe the forecasting system we have developed for the Power Station of As Pontes, placed in the Northwest of Spain. In its proximities there are six tracking stations, provided with automatic analysts, which record immision levels and transmit them to the central laboratory of the plant every ten seconds. Every five minutes these data are averaged and the resulting value (together with the other 23 analogously obtained in the last two hours) is used to produce the *bihourly mean*. These bihourly means are the values to be controlled according to the Spanish laws. With the resources available in the Power Station of As Pontes, it takes about half an hour since we decide to intervene in the combustion process to reduce the immision level, changing the type of coal, until such a reduction is achieved. Hence, considering that we have a datum every five minutes, each time we receive a new observation we need to predict the immision level six times ahead and, according to that prediction, we must decide whether to intervene or not. Of course, it is very important that the forecasting system produces the prediction as fast as possible. Otherwise, we would have to predict seven or more instants ahead and, then, we would loose accuracy in the forecasting.

In the following section we introduce the semiparametric model which is the basis of our forecasting system. In the third section we describe how we obtain the point forecasts and how we construct confidence intervals using Box-Jenkins (see Wei, 1990) and bootstrap (following

Thombs and Schucany, 1990) methods. In the fourth section we show that our system works properly comparing the forecastings it produced and the data which later were observed.

2. THE MODEL

After verifying the inefficacy of the ARIMA models to solve our problem (because they produce unstable predictions when certain values are overcome) and the inefficacy of the nonparametric models (because they can underestimate high immission levels), we have designed a mechanism (based on a semiparametric model) which uses selectively information from the past to predict the future.

Such a mechanism utilizes two sources of information. First, it considers a matrix containing historical immission data. The matrix is formed by a big number of arrays which have the same structure as the forecasting (in our case, $(\ldots,X_{t-1},X_t,X_{t+6})$). They are composed by a regressor component (in this case $(\ldots,X_{t-1},X_t)$) and a response component (X_{t+6} in our problem). Each of the arrays is assigned to a certain box of the matrix, according to the value of its response component. The matrix is dynamically and selectively actualized in such a way that every time a new array enters it (a certain box of it), the oldest array of the box leaves it.

The second source of information are the last 72 observations (corresponding to the last six hours) of the series. We call it the *active series* (and denote it by X_t, $t=1,\ldots,72$).

Three effects determine X_t: a historically far one (or general trend of the prediction), which will be estimated using the historical matrix, a historically recent one, which depends on the last values observed of the variable and on errors corresponding to the recent past (it will be estimated using the active series) and the one due to the error committed in the instant under consideration. Then, we propose the following model of prediction:

$$\Phi_p(B)\nabla^d Z_t = \Psi_q(B)a_t,$$

where $\Phi_p(B)$ is a stationary AR operator (order p), $\Psi_q(B)$ an invertible MA operator (order q), ∇^d a difference operator (order d), a_t a random noise (which we suppose to be of mean 0 and variance σ^2) and $Z_t = X_t - E(X_t/X_{t-6},X_{t-7},\ldots)$ the series resulting of subtracting the general trend (expressed by the conditioned expectation) to the active series.

3. FORECASTING

We use the Nadaraya-Watson estimator with two explicative variables, cross-validation bandwidth h and gaussian kernel K, to approximate the general trend,

$$\hat{E}(X_t/X_{t-6},X_{t-7}) = \frac{\sum_{\tau\in HM} X_{\tau+6}K\left(\frac{1}{h}((X_{t-6},X_{t-7})-(X_\tau,X_{\tau-1}))\right)}{\sum_{\tau\in HM} K\left(\frac{1}{h}((X_{t-6},X_{t-7})-(X_\tau,X_{\tau-1}))\right)}$$

with $t=8,\ldots,78$ (HM denotes the historical matrix). Then, we model (using ARIMA methodology and the routine FTCP of the IMSL library) the series $\hat{Z}_t$, $t=8,\ldots,72$, resulting of subtracting the estimated trend to the active series, to obtain the corresponding prediction $\hat{Z}_{78}$. We consider three prediction rules: a point one,

$$\hat{E}(X_{78}/X_{72},X_{71}) + \hat{Z}_{78},$$

and two confidence intervals of level α: the classical one, obtained with Box-Jenkins methodology, given by

$$\hat{E}(X_{78}/X_{72},X_{71}) + \hat{Z}_{78} \pm z_{\alpha/2}\left(\hat{\sigma}^2\sum_{j=0}^{6-1}\hat{\pi}_j^2\right)^{1/2}$$

where $\hat{\sigma}^2$ is the classical estimated variance of the white noise and the $\hat{\pi}_j$ weights are calculated from the relation

$$\Pi(B) = \frac{\Psi_q(B)}{\Phi_p(B)(1-B)^d}$$

(see Wei 1990, p. 91), and the bootstrap one, obtained by adding to the nonparametric estimated trend, the quantiles corresponding to Z^*_{78}, bootstrap distribution of Z_{78} (following the ideas proposed in Thombs and Schucany, 1990, that we have generalized for ARI models).

4. RESULTS

In Figure 1 we show the predictions (always six instants ahead) using the semiparametric model (dotted line) and the series observed (solid line) along the selected immision episode.

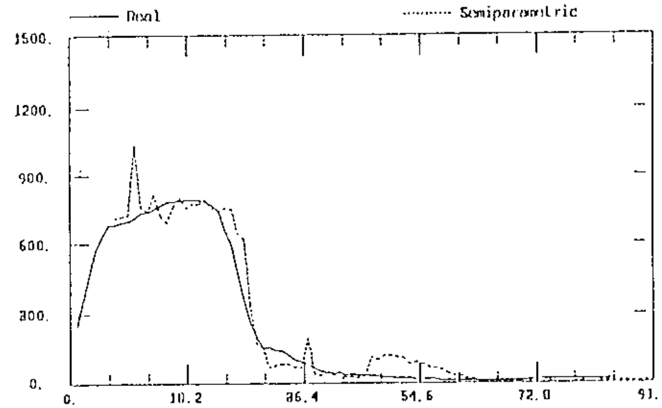

Figure 1. *Prediction using a semiparametric model.*

In Figures 2 and 3 we present the Box-Jenkins and bootstrap confidence intervals respectively ($\alpha = 0.05$), obtained in every instant as described in section 3. We can clearly see that the bootstrap intervals are better than the classical ones in this immision episode.

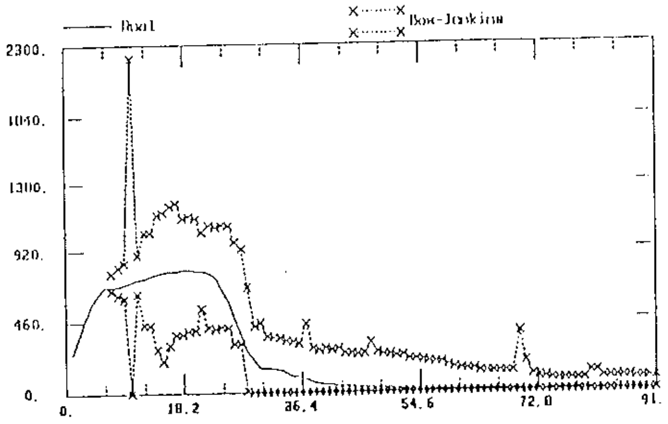

Figure 2. *Classical confidence intervals.*

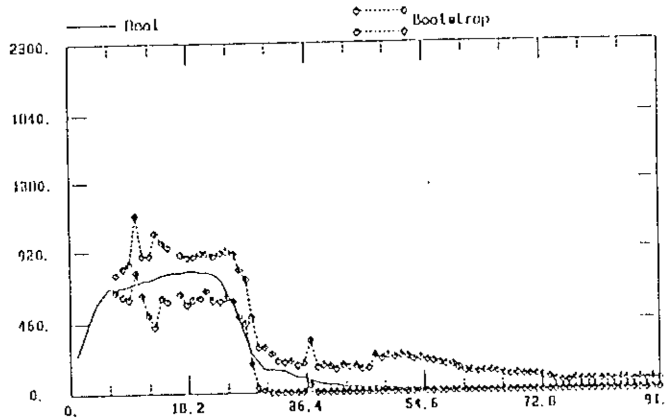

Figure 3. *Bootstrap confidence intervals.*

5 REFERENCES

[1] Thombs, L.A. and Schucany, W.R. (1990). Bootstrap Prediction Intervals for Auto-regression. Journal of American Statistical Association 85, 486-492.

[2] Wei, W. (1990). Time Series Analysis. Addison Wesley.

Computing ARMA Models with MATLAB

J.-F. Emmenegger, University of Fribourg, Switzerland

Abstract

By now, ARMA modeling (cf. Box and Jenkins [1]) is a well known method to analyse time series. Application of the method of Box and Jenkins, starting with the choice of the suitable software on the available computer-environment and ending up with the presentation of final results and conclusions is a hard process of data analysis, decision making, computation, and edition.

The widespread MATLAB software package contains a System Identification Toolbox [5] edited by L. Ljung [4] which is specially constructed to solve system identification problems. As time series can be considered as a subclass of system identification problems, the MATLAB environment has been chosen for time series analysis. The latest version of July 1991 includes the necessary routines to build seasonal ARMA models, with respect to the rule of parsimony.

At the basis of the present analysis, there is a sample–path of a time series of monthly electrical energy distribution data, concerning a well determined area of Western Switzerland. The data used for determination of a seasonal ARMA model covers the period from 1950 to 1988. The recently available data for the three year period from 1989 to 1991 will later be used for model evaluation and prediction.

The aim of this paper is to show, that the MATLAB software package is a suitable tool for time series analysis, involving estimation of seasonal or non-seasonal ARMA models.

Computational and programming work has been carried out on a Macintosh SE/30 PC with the appropriate version of MATLAB [5].

1. Construction of a seasonal ARMA model

The sample-path of monthly measured energy distribution data, measured in MKWh, with $n = 468$ values $x_t := x(t)$; $t = 1,..,n$, providing an underlying seasonal ARIMA process $\{X_t\}$, ([1], p. 85) is constructed. Analysis, according to the three phases of the method of Box and Jenkins: *Identification, Estimation, Diagnostic Checking* ([1], p. 171), is performed. The seasonal period is $s = 12$. The *Identification phase* contains *preliminary transformations* which mainly are:

- logarithms of the sample values (elimination of exponential trend): $y_t = \log(x_t)$
- Given the sample-path y_t and the backshift operator U for $k \in \mathbb{N}$: $U^k X_t = X_{t-k}$,