

- [4] Ragueneau, F., Lancelot, C., Egorov, V., Vervlimmeren, J., Cociasu, A., Deliat, G., Krastev, A., Daoud, N., Rousseau, V., Popovitchev, V., Brion, N., Popa, L. & Cauwet, G., Biogeochemical transformations of inorganic nutrients in the mixing zone between the Danube river and the north western Black Sea *Estuarine Coastal & Shelf Sci.*, **54**, 321-336, 2002
- [5] Menendez, M., Hernandez, O. & Comin, F.A., Spatial distribution and ecophysiological characteristics of macrophytes in a mediterranean coastal lagoon *Estuarine Coastal & Shelf Sci.*, **55**, 403-413, 2002.
- [6] Andrews, J.E., Brimblecombe, P., Jickells, T.D., Liss, P.S. & Reid, B.J., *An Introduction to Environmental Chemistry*, 2nd Ed, Blackwells, Oxford, 296pp., 2004.
- [7] Towner, J.V., Contaminants in estuarine sediments (Chapter 40). *Coastal, Estuarial and Harbour Engineer's Reference Book*, ed. M.B. Abbott & W.A. Price, Chapman & Hall, London, 585-596, 1993.
- [8] Law, C.S., Rees, A.P. & Owens, N.J.P., Temporal variability of denitrification in estuarine sediments *Estuarine Coastal & Shelf Sci.*, **33**, 37-56, 1991.
- [9] Malcolm, S.J., Sivyer, D.B., Trimmer, M. & Nedwell, D.B., *Study to establish rates of flux of nitrogen and phosphorus in Langstone and Chichester Harbours*, CEFAS report to the Environment Agency, 1996
- [10] Dyer K.R., *Estuaries: a physical introduction, 2nd edition*, John Wiley & Sons Ltd, Chichester, 156pp., 1997.
- [11] Mitchell, S.B., Burgess, H.M. & Pope, D.J., Effect on estuarine fine sediment transport of intermittent pump discharge at Pagham Harbour, West Sussex, *Journal of the Chartered Institute of Water and Environmental Management*, In press
- [12] Hydes, D., Nutrients in the Solent. *Proceedings in Marine Science 1: Solent Science - A review*, ed. M. Collins and K. Ansell, Elsevier Science B.V. Amsterdam; London. 135-148, 2000.

A stochastic-dynamic model to predict fecal coliforms at the mouth of the Añasco River

N. D. Ramirez-Beltran¹, F. Gilbes² & J. M. Castro³

¹*Department of Industrial Engineering, University of Puerto Rico, Mayaguez, Puerto Rico*

²*Department of Geology, University of Puerto Rico, Mayaguez, Puerto Rico*

³*Department of Electrical and Computer Engineering, University of Puerto Rico, Mayaguez, Puerto Rico*

Abstract

This research effort attempts to predict one year ahead the concentration of fecal coliforms at the mouth of the Añasco River, located in Puerto Rico. One of the most efficient techniques to represent stochastic processes is time series modeling. These models decompose the process into three major components: trend, seasonality and stochastic components. Unfortunately, time series models require observations at equal time intervals. Since the water quality data are not given at regular intervals, an adaptive estimation technique is proposed to estimate the missing values and, therefore, to generate an approximated time series at equal time intervals, to be able to study trend, seasonality, and stochastic components of fecal coliforms. Water quality data were collected from three water quality stations, located on the Añasco River. Historical data from 1973 to 2000 were used to model fecal coliforms at each of the water quality stations of the Añasco River. Time series models were identified at each station and were used to predict for one year the concentration of fecal coliforms at each station. A spatial interpolation algorithm was used to estimate the fecal coliforms at the mouth of the river.

Keywords: fecal coliforms, time series models, sequential quadratic programming, maximum likelihood estimation, Kriging algorithm, missing values.



1 Introduction

Environmental pollution is an important problem that humans are facing. The increasing population along with agricultural and industrial activities has generated more scenarios for water contamination. The Añasco River basin, located on the west coast of Puerto Rico is not an exception. The Environmental Quality Board operates the water quality-monitoring network for Puerto Rico. The surface-monitoring program consists of about 76 water stations located in 22 rivers (U.S. Geological Survey [12]). The current monitoring program performs water sampling on a quarterly basis. The presence of high concentrations of fecal coliforms and fecal streptococcal bacteria are one of the major pollution problems in Puerto Rico. This research attempts to develop a statistical model to predict the concentration of fecal coliforms at the mouth of the Añasco River.

Rodriguez and Muñoz [11] studied the most important water quality parameters of 10 rivers in Puerto Rico. They used the Kruskal-Wallis test to detect seasonal behavior in water quality parameters. They found that 16% of the analyzed data sets showed seasonal components. Darken et al. [7] show that although water quality data is not given at equal time interval the data resembles a strong serial correlation and consequently tests of significance such as seasonal Kendall test are misleading. They show that the seasonal Kendall analysis does not hold for even small departures from independency. Lohre et al. [8] use seasonal long-memory time series model to represent the water flow of the Rhine River. The memory parameter was estimated using the log periodogram regression for every seasonal frequency. Some researchers instead of using time series models they have identified external variables that exhibit some relationships with fecal coliforms behavior. For instance, Youn-Joo et al. [13] identified a relationship between fecal coliform and the amount of gasoline sold, which was related to recreational boating activity, and the resuspension of *Escherichia coli*.

A robust method to model fecal coliform is proposed here and consists of reconstructing the time series and properly model trend, seasonality, and stochastic components of the fecal coliform process.

2 Data

Water quality data were obtained from the Water Resource Data for Puerto Rico and Virgin Islands (USGS [12]). Data were collected on a bimonthly basis from 1953 to 1996 and after the 1996 observations have been collected on a quarterly basis. Although, water quality data exist since 1953, information for coliforms is only available since 1973. Therefore, this study includes only those 28 years of information (1973 to 2000). The Añasco River has three water quality stations. The first station is located in Lares, the second one in San Sebastian, and the third one is located in Añasco. The code number and location of each station are given in Table 1. The mouth of the Añasco River is located at latitude 18.2667° N and longitude 67.1883° W (see Figure 1). The question mark located in the map indicates the place where the fecal coliforms will be estimated.

Table 1: Location of water quality stations.

Station name	Code	Data since	Location	
			Latitude	Longitude
Añasco	50146000	1960	18.2667° N	67.1347° W
San Sebastian	50144000	1963	18.2847° N	67.0514° W
Lares	50143000	1959	18.2572° N	66.9167° W

The proposed methodology includes five major tasks: (1) a rationale of the proposed algorithm is introduced; (2) an adaptive estimation procedure is developed to reconstruct the monthly time series; (3) seasonal time series models are identified to model the behavior of fecal coliforms at each water quality station; (4) seasonal time series models are used to predict the fecal coliforms at each station; and (5) a spatial interpolation procedure is used to estimate the concentration of fecal coliforms at the mouth of the Añasco River.

3 Methodology

3.1 Rationale of the proposed algorithm

An efficient technique to represent stochastic processes includes time series models. These models decompose the process into three major components: trend, seasonality and stochastic component. Unfortunately, time series models require observations at equal time intervals. Since the water quality data are not given at regular time intervals, an adaptive estimation technique is proposed to estimate the missing values, and to generate an approximated time series at equal time intervals to study trend, seasonality, and stochastic components of fecal coliforms.

The challenge of our research is to develop a sequence of values at equal time intervals given a set of observations obtained at unequal time intervals. The rationale of the proposed algorithm is based on the continuous time representation of a dynamic system. The theoretical representation of a dynamic system is used to generate observations at equal and at irregular time intervals. Time series models are identified with data at equal points in time and are used to predict the unknown values, and these results are compared with the theoretical values. Deviations from theoretical values are used to assess the precision of the suggested algorithm.

Suppose that the concentration of a chemical pollutant in a river is changing with time, and the associated dynamic system is known. Assume that the concentration of a pollutant is completely described by a second order differential equation with nonlinear trend and seasonal component. River pollutants usually exhibit the second order of dependency as it is shown in the empirical equations presented in section 3.3. The seasonal component intends to represent the influence of the climate conditions, and the nonlinear trend reveals the human influences. The influence of climate conditions is also exhibited in the empirical equations presented in section 3.3. Thus, the pollution concentration throughout time may be represented by:

$$\beta_1 \frac{d^2 y}{dt^2} + \beta_2 \frac{dy}{dt} + \beta_3 y = \alpha_1 \sin(\alpha_2(t-1)) + \alpha_3 \cos(\alpha_2(t-1)) + \alpha_4 + \alpha_5 t + \alpha_6 t^2 \quad (1)$$

where y is the chemical concentration of the pollutant at time t . The constants β 's, and α 's are selected in such a way that the solution of the differential equation is easily obtained with the desired seasonality and trend components. Thus, the alphas and betas can be postulated as follows:

$$\beta_1 = 1, \quad \beta_2 = \frac{3}{4}, \quad \beta_3 = \frac{1}{8}, \quad \alpha_1 = \left(\frac{1}{8} - \frac{4\pi}{a_2^2}\right)a_1, \quad \alpha_2 = \frac{2\pi}{a_2}, \quad \alpha_3 = \frac{3\pi}{2a_2}a_1,$$

$$\alpha_4 = \frac{3}{4}a_3 + 2a_4, \quad \alpha_5 = \frac{a_3}{8} + \frac{3}{2}a_4, \quad \alpha_6 = \frac{a_4}{8}, \quad a_1 = 5, \quad a_2 = 12, \quad a_3 = \frac{1}{5}, \text{ and}$$

$$\beta_1 = 1 \quad a_4 = \frac{1}{100}.$$

It can be shown that the general solution of the above differential equation is:

$$y = c_1 e^{-0.25t} + c_2 e^{-0.5t} + a_1 \sin\left(\frac{2\pi}{a_2}(t-1)\right) + a_3 t + a_4 t^2 \quad (2)$$

where c_1 and c_2 are constants that define a particular solution. Suppose that the concentration of the pollutant, and the rate of change at time zero are $f(0) = 100$, and $f'(0) = 10$, respectively; and $f(t) = y$. Thus, the particular solution of the system is given by the following equation:

$$y = 225.12e^{-0.25t} - 127.62e^{-0.5t} + 5 \sin\left(\frac{\pi}{6}(t-1)\right) + \frac{t}{5} + \frac{t^2}{100} \quad (3)$$

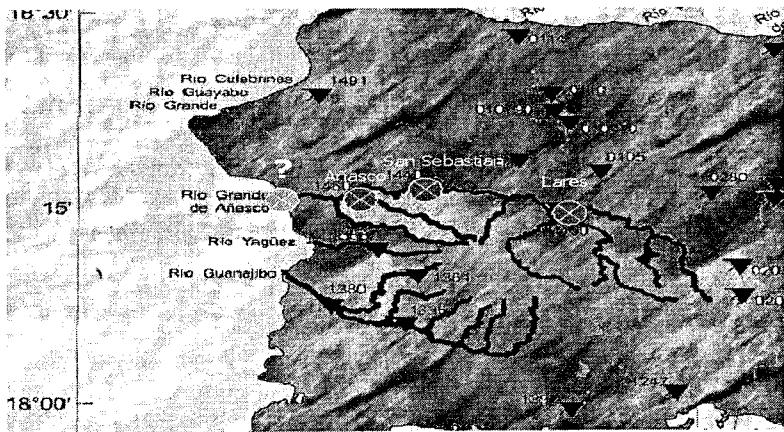


Figure 1: Location of water quality stations.

Eqn (3) represents the system pollution concentration at any time. This equation was used to generate a sequence of pollution concentration at equal

time intervals for $t = 1, 2, \dots, 100$. These data sets were divided into two parts each one with 50 values. The first 50 values were used to identify a time series model. For the last 50 values, thirty percent of the data was randomly eliminated. This random elimination was implemented to generate a sequence with different time intervals. Thus, the generated data have 15 missing values in its second part.

A seasonal autoregressive integrated moving average (SARIMA) model was identified using the first 50 values and it has the following form (Brockwell and Davis [5]):

$$(1-B)(1-B^{12})(1-\phi_1 B - \phi_2 B^2 - \phi_3 B^3)y_t = (1 + \theta_1 B + \theta_2 B^2)a_t,$$

where y_t represents the pollution concentration at time t , a_t is a sequence of random noise at time t , B is the back-shift operator, ϕ 's, and θ 's are parameters of the time series model. Estimates of these parameters are shown in Table 2.

Table 2: Estimation of parameters.

Parameters	Estimate	Std. Error	p-value
ϕ_1	2.15568	0.1016	0.0000
ϕ_2	-1.62219	0.2076	0.0000
ϕ_3	0.398399	0.1174	0.0019
θ_1	-1.20785	0.0401	0.0000
θ_2	-0.802352	0.1428	0.0000

The estimation of the missing values consists on removing the trend component first and the missing values are obtained in such a way that the autocorrelation function (ACF) with and without missing values remain about the same. Thus, the missing values are found in such a way that the following difference is minimized:

$$v = \sum_{k=1}^M (\omega_k - \varpi_k)^2$$

where ϖ_k and ω_k are the sample autocorrelation function of data with and without missing values, respectively, M is the number of lags to be studied. The estimation strategy includes two major steps: estimation of an initial point, and deriving the final estimation. 1) Estimation of the initial point consists of finding a value that minimizes v and ϖ_k is computed by using the time series with missing values and at the end of the series an estimated value of a missing value is included. It should be noted that the estimated value of a missing value is the one that minimizes v . Once the first missing value is computed, it is assumed that the selected value is a known observation. This process is repeated over and over until the initial point is completed. 2) The final estimation is obtained by using a multivariate optimization technique such as the simplex search, which

available at Matlab software. The optimization search starts at the identified initial point and the routine provides the optimal estimates of missing values. These estimates guaranty that the sample ACF when there are no missing values is similar to the sample ACF when estimates of missing values are included. The suggested algorithm was implemented in the simulated exercise and provides a $\nu = 9.37$.

The new time series is known as the reconstructed time series. Figure 2 shows the comparison between the original and the reconstructed time series, an original value is represented by a dot and a predicted value by the symbol "+". This figure shows that there is almost no difference between the observed and the predicted values.

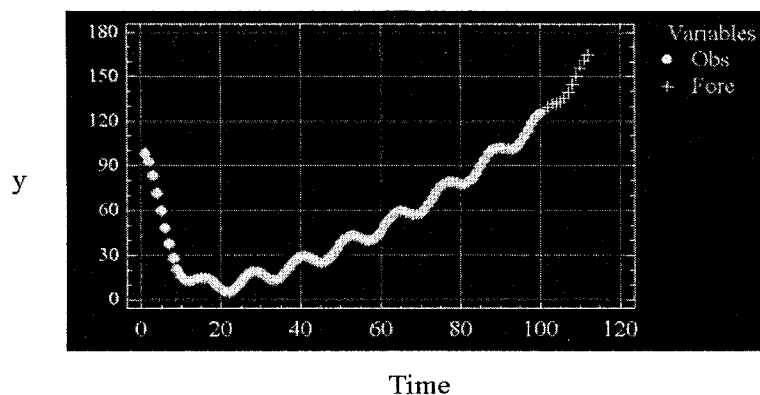


Figure 2: Observed vs. predicted values.

Figure 2 shows that the adaptive algorithm approximately reconstructs the original time series. This methodology was implemented to reconstruct the time series of fecal coliforms at the Añasco River and it is described in the next section.

3.2 Reconstruction of time series.

The U.S. Geological Survey developed a special research project and collected water quality data at the San Sebastian station on a monthly basis from October 1974 to October 1981. These seven years of monthly and consecutive observations provide sufficient information to expand the series from 7 to 28 years.

The seven years of reconstructed data at the San Sebastian station were used to develop seven years of monthly data for the remaining stations. This task was accomplished by using regression method to perform spatial interpolation of fecal coliforms among the stations of the Añasco River. A regression equation for fecal coliforms was identified between San Sebastian and Añasco stations as well as between the San Sebastian and Lares stations. Regression models were

derived after considering observations at the same point in time but at different locations. The regression models identified are the following:

Model 1, between San Sebastian and Añasco:

$$y_{1,i} = 1/(b_0 - b_1 y_{2,i}) + e_{1,i} \quad (1)$$

where $y_{1,i}$, and $y_{2,i}$ represent the natural logarithm of fecal coliform at the Añasco and San Sebastian stations at the i^{th} month, respectively; $e_{1,i}$ represents random noise, and b 's are the parameters of the model. The natural logarithm was applied to mitigate the deviation from normal distribution and also stabilize the variance of the process.

Model 2, between San Sebastian and Lares:

$$y_{3,i} = c_0 + c_1 y_{2,i} + c_2 z_{1,i} + c_3 z_{2,i} + c_4 y_{2,i}^{0.1} + e_{3,i} \quad (2)$$

where $y_{3,i}$, and $y_{2,i}$ represent the natural logarithm of fecal coliforms at the Lares and San Sebastian stations at the i^{th} month, respectively; $e_{3,i}$ represents random noise, $z_{1,i}$ and $z_{2,i}$ are the natural logarithm of the discharge and natural logarithm of turbidity at the San Sebastian station at the i^{th} month respectively; and c 's are the parameters of the regression model.

The regression models were used to estimate the monthly data for the period from October 1974 to October 1981. The seven years of data for Lares and Añasco stations were expanded to 28 years by using the adaptive estimation algorithm. Figure 3 shows the reconstructed data for the San Sebastian station

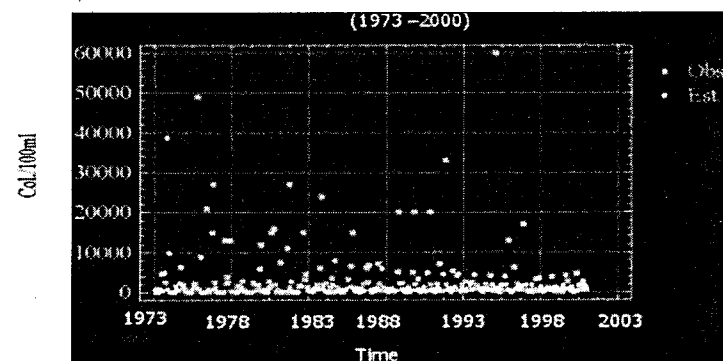


Figure 3: Reconstructed time series at San Sebastian station.

3.3 Time series models

The stochastic behavior of the reconstructed monthly time series data was modeled by a SARIMA model. The trend and the seasonal components are removed by differencing the original process. The stochastic component is usually modeled by autoregressive and moving average structures. Autoregressive structure means that the present values of the process can be expressed by its past values, and the moving average structure expresses the current output of the process by a linear combination of its past errors. Thus, the moving average structure operates as a feedback learning process, i.e., the model improves its estimation by regressing onto its past errors. Essentially, the SARIMA model is a linear difference equation with constant coefficients, which are typically estimated by using a nonlinear regression method or by using the innovation algorithm (Brockwell, and Davis [5, 6]; Box and Jenkins [3]).

A SARIMA model was identified at each one of the reconstructed time series. The logarithm of the reconstructed time series at San Sebastian exhibits a strong seasonal component and therefore a 12-order difference was applied to remove the seasonal effect.

After removing the seasonal component, the process exhibits a stationary behavior and therefore a stochastic model was identified at each one of the time series. The identified models are shown below.

At Añasco: $(1 - \phi_1 B)(1 - B^{12})y_{1,t} = \theta_0 + (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta B^{12})a_{1,t}$ (8)

At San Sebastian: $(1 - \phi_1 B - \phi_2 B^2)(1 - B^{12})y_{2,t} = (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta B^{12})a_{2,t}$ (9)

At Lares: $(1 - B^{12})y_{3,t} = (1 + \Theta B^{12})a_{3,t}$ (10)

where $y_{i,t}$ represents the natural logarithm of the fecal coliform at the i^{th} station and at the t^{th} month, $a_{i,t}$ is sequence of random noise for the i^{th} station and at the t^{th} month, B is the back-shift operator, and ϕ 's, θ 's, and Θ are the parameters of the time series models. Results of the estimation procedure were omitted because of space limitations. The selected parameters are statistically significant and also the correlation on residuals resembles an independent sequence of random values. The p-values indicated that the selected coefficients are statistically significant.

3.4 Prediction of fecal coliform

The identified time series models, eqns (8) to (10), were applied to perform predictions for year 2001 based on information up to the year 2000. SARIMA models provide the capability of predicting one year ahead, with reasonable skillfulness.

Prediction made at time n with h steps ahead is computed by using the innovation algorithm (Brockwell and Davis [5]).

$$\hat{y}_{n+h} = \sum_{j=1}^N \phi_j^* \hat{y}_{n+h-j} + \sum_{j=h}^{q+sQ} \theta_{n+h-1,j} (y_{n+h-j} - \hat{y}_{n+h-j}) \quad (11)$$

where \hat{y}_{n+h} is the optimal linear estimator of y_{n+h} made at time n with a lead time h , y_t is the natural logarithm of fecal coliforms at time t and at a specific water quality station, $N = s^D + sP + p + d$, and ϕ_j^* are the coefficients of the autoregressive structure defined by the following polynomial:

$$(1 - B^s)^p (1 - B)^d (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_Q B^{Qs}) = 0 \quad (12)$$

s is the seasonal order, D is the order of difference to remove seasonal component, and d is the order of difference to remove trend component, p , P and Q are the order of autoregressive and moving average components, ϕ 's and Φ 's are the autoregressive coefficients. $\theta_{n,j}$ are the innovation coefficients that can be estimated once the autoregressive and moving average parameter of the model are known.

3.5 Spatial interpolation

The reconstructed time series for each station were organized for every 6 months and the Kriging algorithm was used to estimate the concentration of fecal coliforms at the mouth of the Añasco River. The Kriging algorithm performs spatial and optimal interpolation considering the geographical location of each station, and concentration of fecal coliforms. The Kriging interpolation is optimal in the sense that it minimizes the estimation of variance (Bras and Kafritas [4]; Matheron [9]). Since the Kriging algorithm is applied at different points in time, the resulting scheme will be a spatial and temporal interpolation algorithm. The expected values of fecal coliforms at the mouth of the Añasco River are shown in Figure 4.

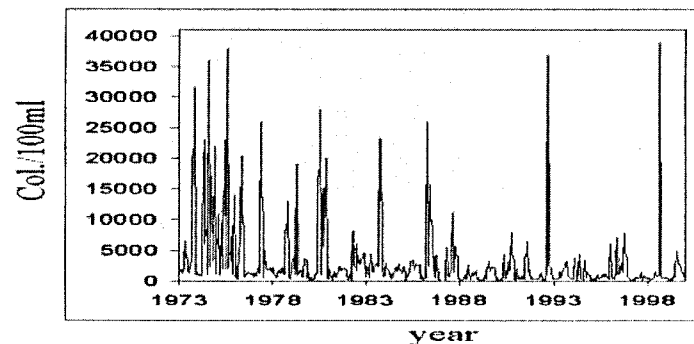


Figure 4: Expected value of fecal coliform at the mouth of the Añasco River

The predicted values from SARIMA models were used as the inputs to the spatial interpolation algorithm to predict fecal coliforms at the mouth of the Añasco River. The predictions of fecal coliforms at the mouth of the Añasco River for year 2001 are shown in Figure 5.

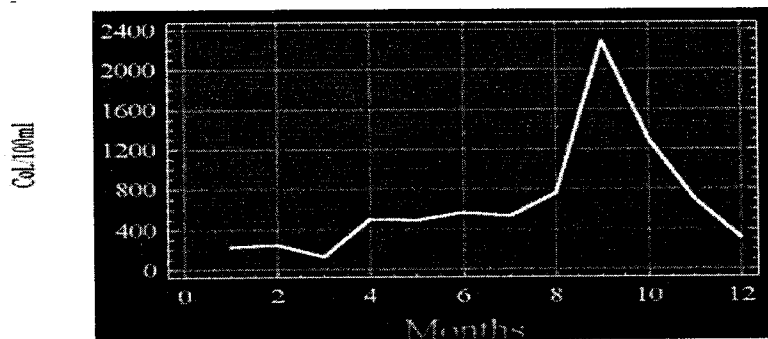


Figure 5: Prediction for 2001 of fecal coliform at the mouth of the Añasco River.

4 Conclusions

A method is proposed to estimate the missing values of time series data. The method consists on identifying a time series model based on available consecutive observations. The identified model is used to predict the missing values and the predicted missing values are assumed to be real observations. This process is repeated until the underlying time series is complete. Cross-validation results show that the proposed method is a potential tool to estimate the missing values of an autocorrelated time series.

It has been shown that SARIMA models properly represent the concentration of fecal coliforms at each station of the Añasco River. The identified SARIMA models were used to predict one-year into the future of fecal coliforms at each of the water quality stations. The Kriging algorithm was adopted to estimate fecal coliforms at the mouth of the Añasco River.

The parameter estimation of the dynamic probability model is obtained by using the maximum likelihood method. In this case, obtaining the likelihood estimates requires solving a constrained nonlinear optimization problem. The SQP algorithm was successfully implemented to solve the underlying optimization problem. Convergence of the nonlinear optimization algorithm was successfully achieved by using a mathematical transformation of the observed fecal coliforms and also by using the outputs of a regression model to initialize the SQP algorithm.

Acknowledgments

This research was supported by the "Puerto Rico Water and Sewer Authority through the Institute of Water Resources of the University of Puerto Rico.

References

- [1] Bazaraa, M.S. Sherali, H.D., and Shetty, C.M. (1993). *Nonlinear Programming Theory and Algorithms*. Second Ed., John Wiley & Sons, New York.
- [2] Bickel, P.J., and Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Inc. Oakland, CA.
- [3] Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis Forecasting and Control*, Holden-Day Oakland, California.
- [4] Bras, R. and J. Kafritas, 1981: *The Practice of Kriging*. Cambridge Massachusetts, MIT Press.
- [5] Brockwell, P.J., and Davis, R.A., (1996). *Introduction to Time Series and Forecasting*, Springer-Verlag New York Inc.
- [6] Brockwell, P.J., and Davis, R.A., (2002). *Introduction to Time Series and Forecasting*, 2nd Ed. Springer-Verlag New York Inc.
- [7] Darken, P.F., Zipper, C.E., Holtzman, G.I., and Smith, E.P. (2002), *Series correlation in water quality variables: estimation and implications for trend analysis*, *Water Resources Research*, Vol. 38, No. 7, pp 22-1, 22-7.
- [8] Lohre, M., Sibbertsen, P., and Konning, T., (2003). *Modeling water flow of the Rhine River using seasonal long memory*. *Water Resources Research*, Vol. 38, No. 7, pp 22-1, 22-7.
- [9] Matheron, G., (1971). *The Theory of Regionalized Variables and its Applications*: Centre de Geostatistique, Fontainebleau, France.
- [10] Ramirez-Beltran, N.D., and T. Sastri. (1997). *Transient Detection with an Application to a Chemical Process*. *Computers & Industrial Engineering*, Vol. 32, No. 4, pp 891-908.
- [11] Rodriguez, A., and Muñoz, R. (1990). *Analysis of Surface-Water Quality Data for Puerto Rico*, Technical Report to US Department of the Interior.
- [12] US Geological Survey (2000). *Water Resources Data for Puerto Rico and Virgin Islands*, Commonwealth of Puerto Rico and Virgin Islands.
- [13] Youn-Joo Y.-J. Ana, Donald H. D.H. Kampbellb, G. Peter Breidenbach (2002). *Escherichia coli and total coliforms in water and sediments at lal marinas*. *Environmental Pollution*, Vol. 120, No. 3, pp. 771-778,

FIFTH INTERNATIONAL CONFERENCE ON
ENVIRONMENTAL PROBLEMS IN COASTAL REGIONS
INCORPORATING OIL SPILL STUDIES

COASTAL ENVIRONMENT V

CONFERENCE CHAIRMEN

J.M. Saval Perez
University of Alicante, Spain

C.A. Brebbia
Wessex Institute of Technology, UK

L. Garcia Andion
University of Alicante, Spain

INTERNATIONAL SCIENTIFIC ADVISORY COMMITTEE

J Antunes do Carmo
I M Banat
R Cocci-Grifoni
B. Fasino
S Mitchell
J Nouwen
S Palmieri
M Tzatzanis

LOCAL ORGANISING COMMITTEE

S Campos Ferrara
F Sanchez Amillategui
A Trapote Jaume

Organised by:

Wessex Institute of Technology, UK
and
The University of Alicante, Spain

N. 11 2000 D. Ramirez

**Coastal
Environment V**
incorporating
Oil Spill Studies

Editors:

C.A. Brebbia
Wessex Institute of Technology, UK

J.M. Saval Perez
University of Alicante, Spain

L. Garcia Andion
University of Alicante, Spain

Y. Villacampa
University of Alicante, Spain

WITPRESS Southampton, Boston

