Statistical aspects of analysing effects of rainfall on water quality

Penev, Spiridon
The University of New South Wales, Department of Statistics
ANZAC Pde
Sydney, 2052 NSW, Australia
E-mail: S.Penev@unsw.edu.au

Leonte, Daniela
National Industrial Chemicals Notification and Assessment Scheme
Level 7, 260 Elizabeth Street, Surry Hills
Sydney, 2010 NSW, Australia
E-mail: Daniela.Leonte@nicnas.gov.au

Lazarov, Zdravetz
Boronia Capital Pty Ltd
12 Holtermann Street, Crows Nest
Sydney, 2065 NSW, Australia
E-mail: Zdravetz.Lazarov@boroniacapital.com.au

Water quality can be substantially affected during a rainfall day. It is also possible that persistent rainfall would have an impact on the long-term water quality trends. In other words, a statistically significant trend in some water quality variables might disappear when the rainfall is accounted for, while for other variables the reverse could occur.

The standard approach in the water quality literature for accounting for the impact of the rainfall on the water quality is to use an ad-hoc semi-parametric approach based on LOESS (synonym LOWESS), that is, to use a locally weighted scatterplot smoothing. A detailed description of this method is given in [4]. Under the LOESS approach, the rainfall data is used as an input to derive a new time series from the original water quality series. The former time series is assumed to be free from any rainfall influence but this assumption obviously is too strong. Also, this approach does not allow to quantify the influence of rainfall, if present, or to obtain some of its qualitative properties. In particular, testing for rainfall effect on water quality variables is often of interest.

Other simplistic alternatives take the arithmetic average of the past values of rainfall and use them as an exogenous regressor in a statistical trend model. In this way, a test for the presence of rainfall effect can be easily performed, and its impact on water quality quantified. However, the dynamic effect of rainfall can not be observed in this simplistic alternative. It is indeed possible, as we will demonstrate below, that rainfall on past days can affect the present water quality in certain way, and this effect should be accounted for. For example, one reasonable assumption would be that the rainfall on the current day has the highest influence on water quality, and this influence decreases for the rainfall on previous days. It is also of interest to see if more complex patterns of influence could occur, and to get an idea of their shape(s).

1 The approach

Our approach to investigate these potentially complex patterns of rainfall influence on water quality is through the use of a flexible statistical model called Mixed Data Sampling (MIDAS) regression (see [1] for a review of the method).

Typically, data collection for water quality variables happens in fortnightly or even monthly

intervals. On the other hand, rainfall data is collected daily. Even without any missing data, a five years collection of data would give a rise of just about 110 fortnightly of 60 monthly observations. These are just a very small portion of the rainfall measurements for one year only. The relatively small number of observations of the water quality variables necessitates the use of parsimonious time series models. The advantage of using the MIDAS regression is precisely in the possibility to achieve parsimonious model description. When using MIDAS regression, with only three parameters, different shapes, including humped shapes, for the weight coefficients of the lags in the rain impact can be modelled.

1.1 Least squares

In practice, missing data appears routinely for time series of water quality variables. As a result, not all fortnightly data records will be available. Substituting missing values in short time series-type data is a risky operation and should be avoided if possible. We describe two procedures to deal with this issue. The first one, based on least squares ideas, is described first. Let $X_1, X_2, \ldots, X_M$ be the time series of all monitoring records of a specific water quality variable (when the value is missing it is coded with a missing value code, otherwise it is a positive reading). These are the records of the response variable in the model. Water quality variables of interest are: Aluminum Total , Manganese Total, Iron Total, Nitorgen Oxidised etc. Since, when not missing, the monitoring records are positive variables, we consider taking the logarithm of the original time series which brings the transformed variables closer to normally distributed ones. This transformation is also helpful in achieving stationarity. In addition, the logarithmic transformations helps us to avoid the need to put restrictions on the model's residuals in order to guarantee positivity of the dependent variable. Let $d_1, d_2, \ldots, d_M$ be a sequence of days when the data records are taken. We denote by N the number of non-zero values (i.e., actual observations) given by the sub-sequence $X_{t_1}, X_{t_2}, \ldots, X_{t_N}$ for the set of indices $t_1 < t_2 < \ldots < t_N$.

We take a subset of indices $f_1 < f_2 < \ldots < f_P, (P \leq N)$ from the set of indices $t_1 < t_2 < \ldots < t_N$ such that $Y_{f_i} = \log(X_{t_j})$ for some $X_{f_i} = X_{t_j} > 0$ and $t_j = t_{j-1} + 1$. In other words, we choose all recorded observations for which there exists another recorded observation which is taken a fortnight apart.

$$Y_{f_i}, i = 1, 2, \ldots, P \tag{1}$$

as the output data.

It is a typical feature of the water quality data that it is autocorrelated. It has been observed in the literature that most of the time it is a short-term autocorrelation (see for example, [5], p. 95), which can be captured using the lagged value of the dependent variable (an AR(p) term), and possibly an MA(q)-term) and small values of p and q. The goal is to model the dynamics with an ARMA(p, q)-type model with a relatively small values of p and q. Taking into account these remarks, we consider one of the possible (and simplest) models that can be entertained when the small number of available observations and the parsimony are to be observed: a starting model for water quality trends can be given by the system of linear regressions of the form:

$$Y_{f_i} = \alpha + \beta_1 f_i + \beta_2 \sin(2\pi d_{f_i}/365) + \beta_3 \cos(2\pi d_{f_i}/365) + \beta_4 Y_{f_{i-1}} + \epsilon_{f_i}, i = 1, 2, \ldots, P. \tag{2}$$

Here $\alpha, \beta_i, i = 1, 2, 3, 4$ are unknown parameters and $\epsilon_{f_i}, i = 1, 2, \ldots, P$ are the error terms. In this simple model, the significance and the sign of the coefficient β_1 indicate the presence and direction of a trend. We use a simple linear specification for the trend to minimize the number of parameters in the model. The regressors $\sin(2\pi d_{f_i}/365)$ and $\beta_3 \cos(2\pi d_{fi}/365)$ account for the intra-yearly seasonality usually present in water quality data. We add the lagged value of the depended variable $Y_{f_{t-1}}$ to control

for presence of autocorellation. Finally, $\epsilon_{f_i}, i = 1, 2, \ldots, P$ represents a sequence of independently distributed error terms.

Model (2) can be estimated by Ordinary Least Squares (OLS), the advantage being robustness and simplicity. However, cases are included for analysis only when a neigbouring observation recorded a fortnight apart exists and the procedure is not efficient. It still could be used in its own right when there are no many missing values or it could be used to generate initial estimators for a more sophisticated iterative Full Maximum Likelihood procedure. The latter uses all available non-missing data points by taking into account the time gap between them. To describe the procedure we make the additional assumption of normality of the errors. Next we describe this procedure.

1.2 Full Maximum Likelihood

When estimating the model (2) we restrict our attention only to those observations for which there exists an observation in the previous sampling interval (say monthly or fortnightly, for routinely collected water quality samples). In this section we develop a Full Maximum Likelihood (FML) estimation (under normality assumption for the errors) which allows us to impute missing data.

To fix ideas, suppose for a moment that instead of the given time series of monitoring records $X_1, X_2, \ldots, X_M$ we had the time series $X_1^*, X_2^*, \ldots, X_M^*$ where no observation was missing (i.e., a positive value $X_i^*, i = 1, 2, \ldots, M$ each time was recorded). Let $Y_1^*, Y_2^*, \ldots, Y_M^*$ be the corresponding log-transformed values. Similarly to (2) we assume that the dynamics of the time series has the linear form

$$Y_i^* = \alpha + \beta_1 i + \beta_2 \sin(2\pi d_i/365) + \beta_3 \cos(2\pi d_i/365) + \beta_4 Y_{i-1}^* + \eta_i, i = 1, 2, \ldots, M, \tag{3}$$

where the error terms $\eta_i, i = 2, 3, \ldots, M$ are independent identically distributed normal zero mean variables with variance σ^2. If we had observed the whole sequence $Y_1^*, Y_2^*, \ldots, Y_M^*$ we could have estimated the model (3) by least squares even without normality assumption on η_i. Now the presence of missing data means that only a subset $Y_1, Y_2, \ldots, Y_N$ of the time series is observed. To estimate the model (3) in such situations, we derive relations between adjacent recorded variables by recursively applying the relation (3).

More specifically, let us pick a pair of neighbouring observations Y_{j-1} and Y_j, some $j = 2, 3, \ldots, N$, remembering that $Y_j = \log(X_{t_j}), Y_{j-1} = \log(X_{t_{j-1}})$. Setting $s_j = t_j - t_{j-1}$ and

$$\mu_i = \alpha + \beta_1 i + \beta_2 \sin(2\pi d_i/365) + \beta_3 \cos(2\pi d_i/365)$$

for $i = 2, 3, \ldots, M$ we have

$$Y_j = \log(X_{t_j}) = \mu_{t_j} + \beta_4 \log(X_{t_j-1}^*) + \eta_{t_j} \tag{4}$$

Applying (3) to the lagged term $\log(X_{t_j-1}^*)$ it follows

$$Y_j = \mu_{t_j} + \beta_4 \log(X_{t_j-1}^*) + \eta_{t_j} = \mu_{t_j} + \beta_4 \mu_{t_j-1} + \beta_4^2 \log(X_{t_j-2}^*) + \beta_4 \eta_{t_j-1} + \eta_{t_j}.$$

Continuing applying recursively Equation (4) in a similar way, we end up with

$$Y_j = \sum_{k=0}^{s_j-1} \beta_4^k \mu_{t_j-k} + \beta_4^{s_j} Y_{j-1} + \zeta_{t_j}, j = 2, 3, \ldots, N. \tag{5}$$

where $\zeta_{t_j} \sim N(0, \sigma^2 \frac{1-\beta^{2s_j}}{1-\beta^2})$.

Working in a similar fashion, we can derive similar relationships for all pairs $Y_j, Y_{j-1}, j = 2, 3, \ldots, N$. Hence we can write down the Likelihood of the observed data and can estimate the parameters by using Maximum Likelihood. It is known from the general Likelihood Theory (e.g., [2])

that the maximum likelihood estimate of the parameter vector $(\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \sigma^2)'$ is asymptotically normal, asymptotically unbiased and its variance-covariance matrix can be evaluated by the inverse of the Fisher information matrix.

We have implemented an iterative procedure using SAS for fitting this model. The initial values for the iterative procedure can be obtained by using the Least Squares method of the previous subsection. The program reports the parameter estimates, their standard errors and the associated p-values. Among other things, this allows us to gauge the magnitude and statistical significance of the water quality trend.

1.3 Rainfall influence

In this section we describe how to incorporate the impact of rainfall on the water quality variable. For simplicity of the presentation let us assume that there are no missing observations. For each $i = 1, 2, \ldots, M$ we denote by $V_{i,1}$ the rainfall on the current day. Similarly, for $k = 2, 3, \ldots, l$, denote the rainfall from the previous days, up to $l-1$ before, by $V_{i,k}$. We note that the rainfall variables will be often be equal to zero, and also can take widely varying positive values ranging from 0.5 to 200-300mm. With this in mind, we introduced a new scaled variable $r_{i,k} = \log(1 + V_{i,k})$ instead of $V_{i,k}$.

A straightforward and flexible approach is to introduce the rainfall variables $r_{i,k}$ as exogenous regressors in the specification (3) is via

$$Y_i = \alpha + \beta_1 i + \beta_2 \sin(2\pi d_i/365) + \beta_3 \cos(2\pi d_i/365) + \beta_4 Y_{i-1} + \gamma_1 r_{i,1} + \gamma_2 r_{i,2} + \ldots + \gamma_l r_{i,l} + \epsilon_i, \tag{6}$$

$i = 1, 2, \ldots, M$. Here d_i denotes the day record and ϵ_i is the error term. The coefficients $\gamma_1, \gamma_2, \ldots, \gamma_l$ account for the impact of rainfall in the previous $(l-1)$ days, as well as the current day. For brevity, we will refer to the function $f(k) = \gamma_k$ as the *rainfall impact function.* The specification (6) is flexible yet can be impractical for inference purposes for large l especially since M is usually small. The alternative to set all γ_i to be equal (which is equivalent to prefiltering) does not allow to accommodate complex patterns of rainfall influences. Another alternative is to use the Koyck parameterization with two parameters δ and λ to set $\gamma_1 = \delta, \gamma_2 = \delta\lambda, \ldots, \gamma_l = \delta\lambda^{l-1}$. Other parameterizations from the Distributed Lag Models literature are surveyed in ([3]) but their common characteristic is that they allow for decreasing weights only. This is too restrictive for our models. Indeed, the full impact of a heavy rainfall might be felt a few days after the event. Here we propose to model the impact of rainfall on water quality via MIDAS regression [1]. The main advantage is that it is parsimonious (it uses only three parameters to model the weights). It also allows for very flexible shapes of the function form of coefficient weights, including a decreasing or humped shape. The latter form of the coefficient weights allow incorporating delayed peak in the rainfall impact on water quality. More specifically, we set

$$\gamma_k = \delta B(\frac{k}{l}; \theta_1, \theta_2). \tag{7}$$

Here $B(x; \theta_1, \theta_2) = \frac{x^{\theta_1 - 1}(1-x)^{\theta_2 - 1}\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}$. The Beta function's coefficients θ_1 and θ_2 are always positive hence whether there will be a statistically significant impact of the rainfall on water quality is determined by the significance of the coefficient δ. High significance (low p-values) of θ_1 and θ_2 is yet helpful to access model adequacy. We have implemented an iterative SAS procedure to fit this model.

2 Implementation and results

In this section we estimate the trend model (6) with the MIDAS specification of the weights given by (7) for a few water quality variables at a specific catchment monitoring site. The dataset contains

measurements on water quality variables from the Shoalhaven supply system and catchments in the Sydney region.

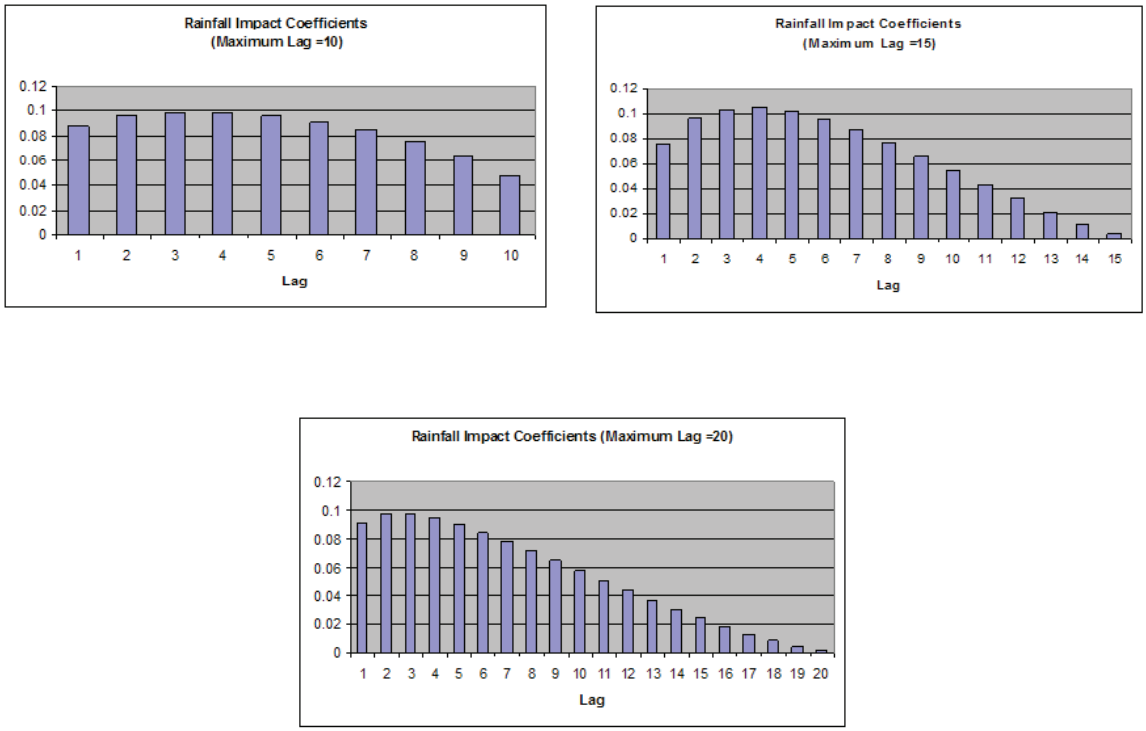

Figure 1: *Rainfall Impact Coefficients for Aluminium Total.*

We briefly highlight our empirical analysis by considering a catchment site coded as E822 and three water quality variables, namely Aluminium Total, Turbidity Field and Manganese Total. The lag parameter l is specified beforehand, and three different lag sizes of 10, 15 and 20 days were used. Our main point of interest is whether the trend coefficient changes value and statistical significance substantially when accounting for the rainfall. Of interest for this paper is also the actual sign and shape of the rainfall impact function. We used the Full Maximum Likelihood Estimation for Aluminium Total, without and with a rainfall impact function. We observed that the coefficient δ in front of the Beta function is highly significant and positive ($0.0873, s.e. = 0.0132, P \approx 0$ when $l = 10; 0.0629, s.e. = 0.0073, P \approx 0$ when $l = 15$; $0.0524, s.e. = 0.0074, P \approx 0$ when $l = 15$) indicating a significant rain impact. The trend coefficient in the Full Maximum Likelihood model (with rain impact excluded) is also highly significant and negative ($-0.00318, s.e. = 0.00136, P \approx 0.019$). In this case, including the rain impact did not change the sign, size and significance of the trend ($-0.00374, e.s. = 0.0103, P \approx 0.00026$ when $l = 10$ and similar values when $l = 15, l = 20$). The graph (1) of the lag coefficients for $l = 10, 15, 20$ suggest a slightly hump shaped rainfall impact function. This delayed impact reflects the time needed for certain chemical reactions to take place during a rainfall.

Next we present a similar analysis for Turbidity Field at the same catchment site. This time the trend coefficient in the model with rainfall excluded was non-significant ($-0.0019, s.e. = 0.00125, P = 0.129$) but became slightly significant when rainfall's influence is included ($-0.0025, s.e. = 0.00106, P \approx 0.0018$ for $l = 10$, with similar values for $l = 15, 20$). The coefficient δ is highly significant and positive ($0.0726, s.e. = 0.0018, P \approx 0$ for $l = 10$ with similar values for $l = 15, 20$). This time, not surprisingly (see Figure (2)), the rainfall impact function coefficients exhibit a different pattern as the impact of

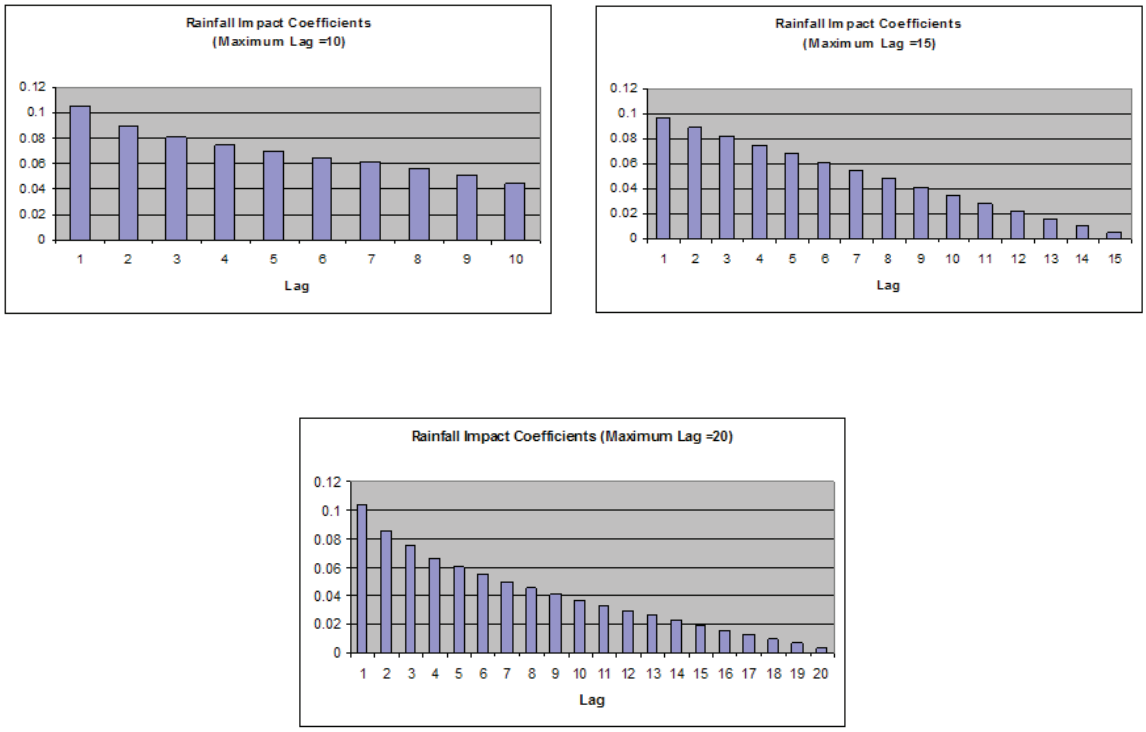

Figure 2: *Rainfall Impact for Turbidity Field.*

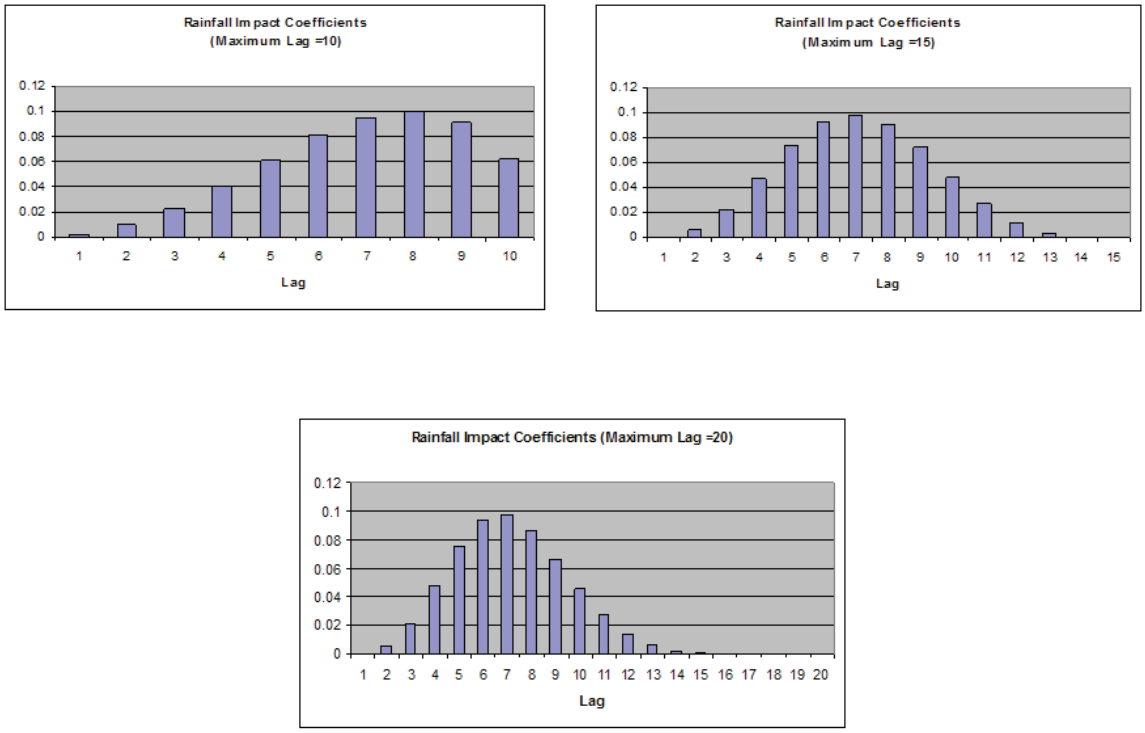

Figure 3: *Rainfall Impact Coefficients for Manganese Total.*

rainfall is at its maximum on the rainfall day and then gradually and smoothly decreases.

Our final example is for Manganese Total (see Figure (3)). Again, there is no substantial change in the trend's coefficient size and significance with rainfall included or excluded. Now, δ in front of the Beta function is negative this time and is again highly significant. The rain impact function is markedly different, being pronouncedly hump shaped, reaching a maximum about 6-8 days after rainfall and approaching zero after about 14 days.

In conclusion, rainfall may have little or no significant effect on the general trend, however in all cases it has a more complicated impact on the water quality compared to what is reported in the current literature. In particular, the hump shaped or decreasing rainfall impact function indicates that models that use arithmetic averages of the rainfall could be grossly miss-specified.

3 Acknowledgements

The authors would like to thank Sydney Catchment Authority (SCA), who provided funding for the work reported here through a collaborative research grant. The work of Dr Zdravetz Lazarov was sponsored by the grant. We thank Dr Rob Mann who managed the grant on behalf of SCA.

References

[1] Ghysels, E., Sinko, A., and Valkanov, R. (2007) MIDAS Regressions: Further Results and New Directions. Econometric Reviews 26 (1), 53–90.

[2] Greene, W. (2000) Econometric Analysis, 4th Edition. Prentice Hall.

[3] Gujarati, D. (2003) Basic Econometrics, 4th Edition, McGraw Hill.

[4] Helsel, D.R. and Hirsch, M., (2000) Statistical Methods in Water Resources, Studies in Environmental Science, 49. Elsevier, New York.

[5] Ward, R., Loftis, J. and McBride, G. (1990) Design of water quality monitoring systems. New York, Van Nostrand Reinhold.

RÉSUMÉ (ABSTRACT)

We discuss novel statistical methods in analysing trends in water quality. Analysis of these trends is a very important activity in catchment management. Such analysis deals with large and complex data sets of different classes of variables like water quality variables, hydrological, meteorological variables and others. Distinguishing features of water quality data set records include: irregularity of the days when observations were taken, presence of multiple observations on a single day, changing detection limits, different frequencies of data collection etc. Non-normality of many of the variables, presence of missing data and presence of seasonal patterns must also be accounted for in the analysis.

We concentrate on analysing the effect of rainfall on trends in water quality variables. We utilise a flexible model called Mixed Data Sampling (MIDAS). The mixed data sampling arises because of the mixed frequency in the data collection: typically, water quality variables are sampled fortnightly, whereas the rain data is sampled daily. Rainfall can have an impact on the quality on the day of the measurement of the water quality variable or via its weighted influence from previous days. The advantage of using MIDAS regression is in the simple, flexible and parsimonious modelling of the influence of the rain on trends in water quality variables. Only three additional parameters are used to fit a variety of monotone and humped influence shapes for the weight coefficients of the lags in the rain impact. We discuss the model, its implementation on a water quality data set, and some outcomes to justify its benefits.