Occurrence of Bacteria With Visual Beach Pollution

P. M. Jellett: Charles Sturt University
Wagga Wagga NSW Australia
pjellett@csu.edu.au

Abstract Faecal bacterial counts were performed on water samples from Sydney beaches. In addition, beach pollution was visually assessed by beach inspectors on a daily basis using a 5-point ratings scale at the same beaches. The bacterial counts were summed over all beaches to produce an irregularly spaced daily total count. The surveyed visual pollution scores were averaged over all beaches to produce a daily time series of visual pollution for Sydney beaches. The total bacterial counts were modelled as Normal random variables via a generalised linear model in which the independent variables entailed a transfer funcion of the average daily visual pollution scores. The use of a generalised linear model with a logarithmic link function permitted the untransformed bacterial counts to be modelled without the more commonly performed explicit logarithmic transformation. The model explained approximately three quarters of the variability in the raw bacterial counts. Earlier work (Jellett 1996) reported a relationship between the visual pollution scores and wind, rain, ocean current and temperature, establishing the usefullness of the visual ratings as a measure of pollution. The current work establishes a link between visually assessed pollution and the occurrence of faecal bacteria in the water. The paper also describes the method which was employed to simultaneously estimate the transfer function parameters for filtering the visual ratings together with the generalised linear model.

1. INTRODUCTION

Bacterial counts for faecal coliforms and faecal streptococci were performed on water samples which were collected approximately every second day from 34 Sydney beaches. In addition, beach pollution was assessed visually by beach inspectors on a daily basis using a 5-point ratings scale at the same beaches. The time period of the modelled data covered 1 July 1991 to 30 April 1992 with the bacterial counts as the dependent variable and the visual pollution scores as the independent variable. The bacterial counts were summed over all beaches to produce an irregularly spaced daily total count with almost half the record missing. The surveyed visual pollution scores were averaged over all beaches to produce a daily time series of visual pollution for Sydney beaches with no missing values (Figure 1). In 1996 Jellett reported a relationship between the visual pollution scores and wind, rain, ocean current and temperature, establishing the usefullness of the visual ratings as a measure of pollution. The current work establishes a link between visually assessed pollution and the occurrence of faecal bacteria in the water which was sampled at the beaches (Figure 2).

2. RESULTS

The estimated model equations are

$$y_t = \exp(c \quad + \quad x_t) \quad + \quad e_t \tag{1}$$

$$x_t = a_1 x_{t-1} + b_0 v_t + b_3 v_{t-3} \tag{2}$$

where y_t are the total observed bacterial counts and v_t are the observed average daily coastal visual pollution ratings (Figure 1). The a, b and c terms in equations (1) and (2) were estimated and are given in Table 1 while e_t are the model residuals. The linear transfer function output, x_t, was computed from equation (2). Figure 2 shows the fitted model from the right hand side of equation (1) together with the observed total bacterial counts. Figures 3 and 4 summarise the effect of the transfer function equation (2) in the model. Figure 5 shows the model residuals, e_t, from equation (1).

Figure 3 shows hypothetical bacterial counts calculated from equations (1) and (2) corresponding to one unit of average visual pollution on day 1 and zero on other days. The effect of such a unit will be to multiply bacterial levels to produce the levels shown day by day in Figure 3. Thus bacterial levels will peak on the same day and then rapidly settle back to previous levels. Likewise Figure 4 shows hypothetical bacterial counts calculated from equations (1) and (2) but corresponding to one unit of visual pollution on day 1 and succeeding days and zero prior to day 1. Figure 4 shows that, after an initial peak, bacterial levels will settle down to a steady factor of 3.3 increase compared with day zero.

The model r^2 value was 71%. Thus the correlation between the total bacterial counts and the average daily coastal visual ratings was 0.84. This is remarkably high given the small number of estimated parameters, the large number of missing values and the fact that the visual pollution scores amounted to a daily survey which was completed by separate beach inspectors. The result suggests a common source for both the visual pollution and the bulk of the bacterial pollution. Earlier work suggested that the source is largely from the sea since the visual pollution was found to be associated with onshore winds and currents (Jellett 1996) as well as rainfall. Rainfall could produce an increase through direct urban runoff or through the urban drainage/sewerage system as well as the sea.

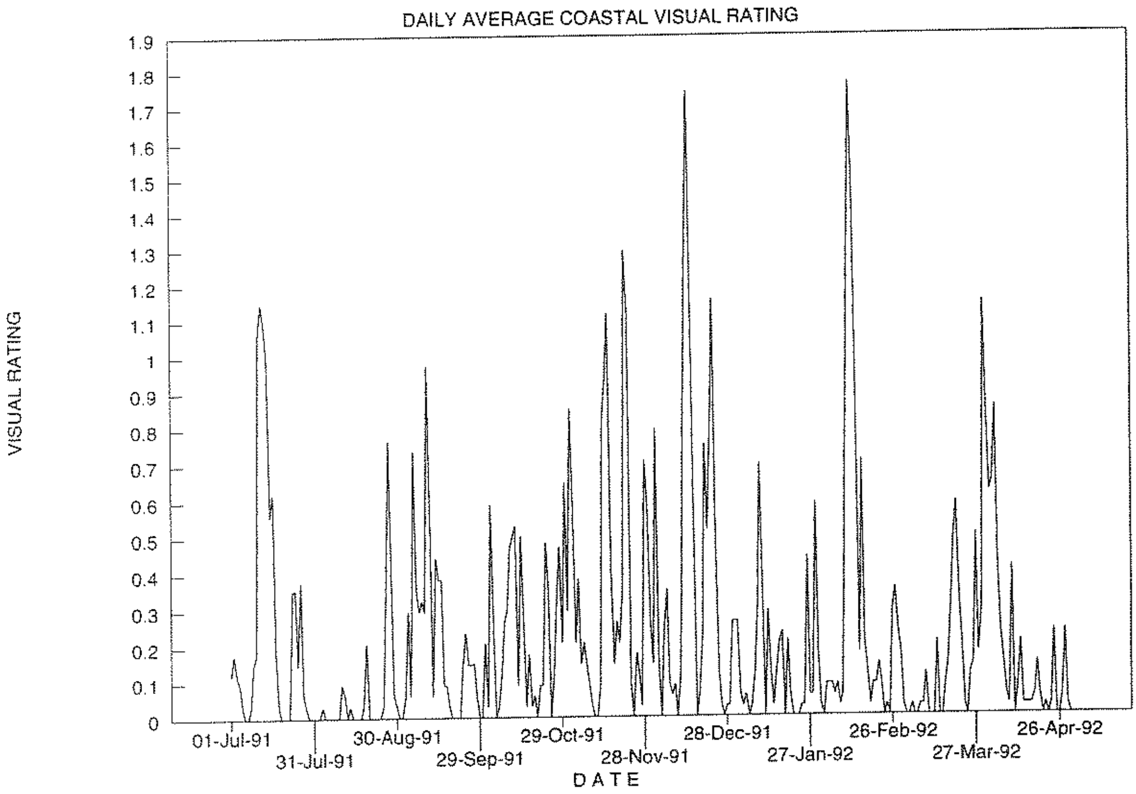

Figure 1: Daily Coastal Average Visual Rating

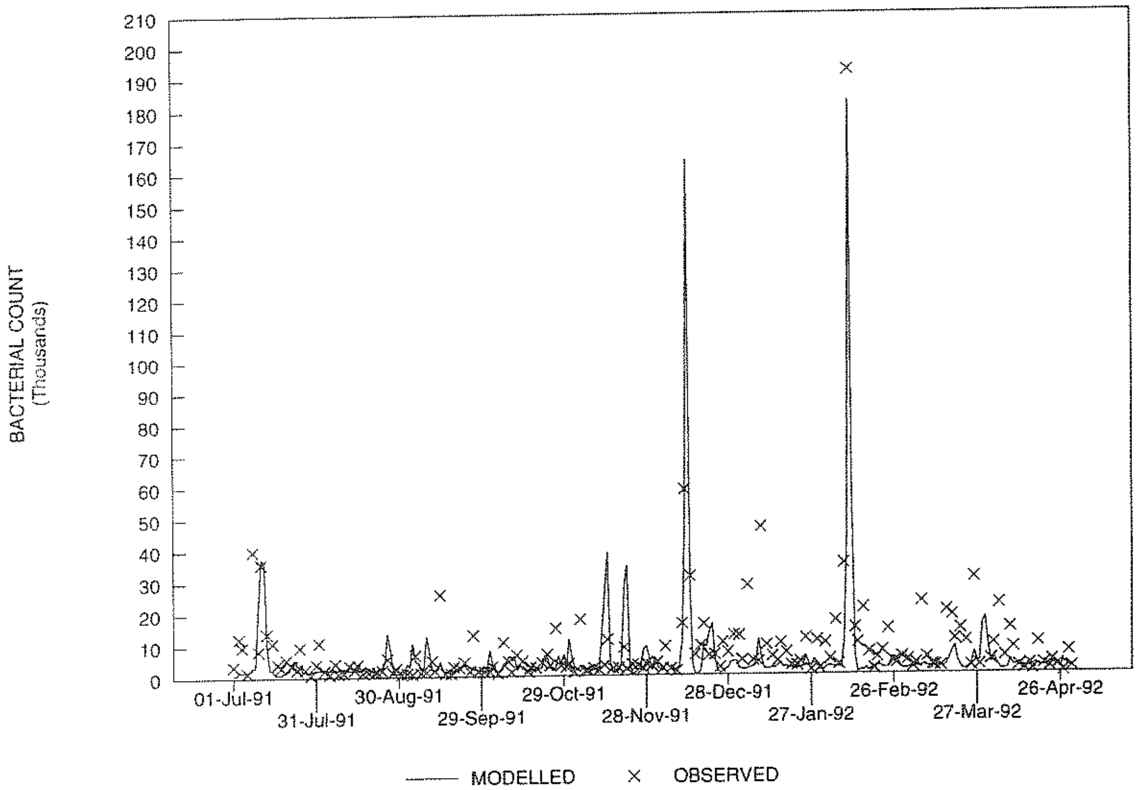

Figure 2: Observed and Modelled Bacterial Counts

Table 1: Fitted Model (R-Squared=71%)

Model Term	Parameter Estimate	Standard Error	Description
c	7.71224	0.21738	Constant
b_0	2.12453	0.23109	Lag 0 Visual Rating
b_3	-1.21813	0.43255	Lag 3 Visual Rating
a_1	0.24040	0.09879	Lag1 Filtered Visual Rating

3. PARAMETER ESTIMATION

Equation (1) is nonlinear and is of a kind which is common in the modelling of discrete counts (Bishop et al 1980). The GLIM language was used (Nelder and Wedderburn 1972). Equation (2) is a transfer function and is common in time series models (Box and Jenkins 1976, Brockwell and Davis 1991. Young 1984). Thus the parameter estimation method below combined two methodologies, generalised linear models and transfer function models. The method was also used by Jellett (1996).

The data covered 305 days but only 170 bacterial counts were observed. having been measured roughly every second day. Equation (2). being a recursive difference equation. relies on an unbroken data record for its computation. This was not a problem since there were no missing values in the visual pollution data v_t. The GLIM language. which was used to estimate the model, does not easily support recursive calculations down a column of numbers. Besides combining methodologies, another useful feature of the method which is described below is that it does not involve recursive evaluation of x_t in equation (2).

3.1 Iterated Regression On Lagged Fitted Values

Firstly consider the estimation of the linear transfer function model. equations (3) and (2),

$$y_t = c + x_t + e_t \tag{3}$$

The method involves determining the *a*, *b* and *c* parameter estimates in these model equations by carrying out an iterative sequence of multiple linear regressions of a dependent variable. y_t, on the terms on the right hand side of equation (4). Equation (4) is a linear regression to determine the parameter estimates with the superscript, *i*. Equation (5) was then calculated with parameter estimates from the i-th regression. Quantities with the superscript, *i-1*, are from the preceding iteration.

$$y_t = c^i + a_1^i x_{t-1}^{i-1} + b_0^i v_t + b_3^i v_{t-3} + e_t \tag{4}$$

$$x_t^i = a_1^i x_{t-1}^{i-1} + b_0^i v_t + b_3^i v_{t-3} \tag{5}$$

When the parameter estimates and model outputs have converged.

$$x_t^i = x_t^{i-1} \tag{6}$$

and equation (5) will be the same as the recursive difference equation, (2). To initialise equations (4) and (5),

$$x_t^0 = 0 \tag{7}$$

$$x_t^0 = y_t \tag{8}$$

equation (7) was used. If this is not acceptable to a regression package then equation (8) would be satisfactory. Equation (8) could not be used here because of the presence of missing values. At each iteration the vector x_t^i could be viewed as the output of a discrete vector transfer function which, as the parameter estimates converge, has constant inputs on the right hand side of equation (5). The vector transfer function output in equation (5) has the superscript, *i*, as the discrete sequencing variable while *t* runs over elements of the vector as a dummy variable. Hence each element of the output vector will converge to a steady state, provided *a* is less than one in absolute value.

3.2 Combination with Generalised Linear Model

The scheme of section 3.1 will lead to estimates for transfer function models, for example, equations (3) and (2). The method is easily adaptable to cover multiple transfer function models (Jellett 1996) or time-varying regressions (Jellett 1997). Because each iteration consists of a multiple linear regression, it is a simple matter to replace that step with another type of model which utilises a dependent variable and independent variables. This is the case with generalised linear models of which equation (1) is an example. Thus all that needed to be changed in the GLIM language was to specify equation (9) as the model form in order to estimate equations (1) and (2) rather than the multiple linear regression equation (4). Equation (5) remains unchanged.

$$y_t = \exp(c^i + a_1^i x_{t-1}^{i-1} + b_0^i v_t + b_3^i v_{t-3}) + e_t \tag{9}$$

3.3 Statistical Properties of The Estimates

Study of the nonlinear equations which are satisfied by the parameter estimates at convergence of equations (4) and (5) reveal that the method of section 3.1 is an example of an iterative instrumental variable estimator. (Young 1984), and will have satisfactory bias and efficiency properties following from the orthogonality of x_t and e_t at each iteration of the regression model (4). The method of computing both the estimates and the model output appears to be new. The estimates produced by equations (4) and (5) are related to those for simplified refined instrumental variables, except that the model outputs are not pre-filtered here and hence are not the derivatives of the model residuals with respect to the *a*-parameter as is the case with simplified refined instrumental variables which are optimal in the sense that they produce least squares estimates conditional upon start-up values for time one. The method of section 3.1 is distinguished by its extreme simplicity by comparison with other time series estimation methods and permits time series models to be fitted in any software environment where it is possible to fit a sequence of regressions. The method even permits combination with ARMA model residuals (Jellett 1996).

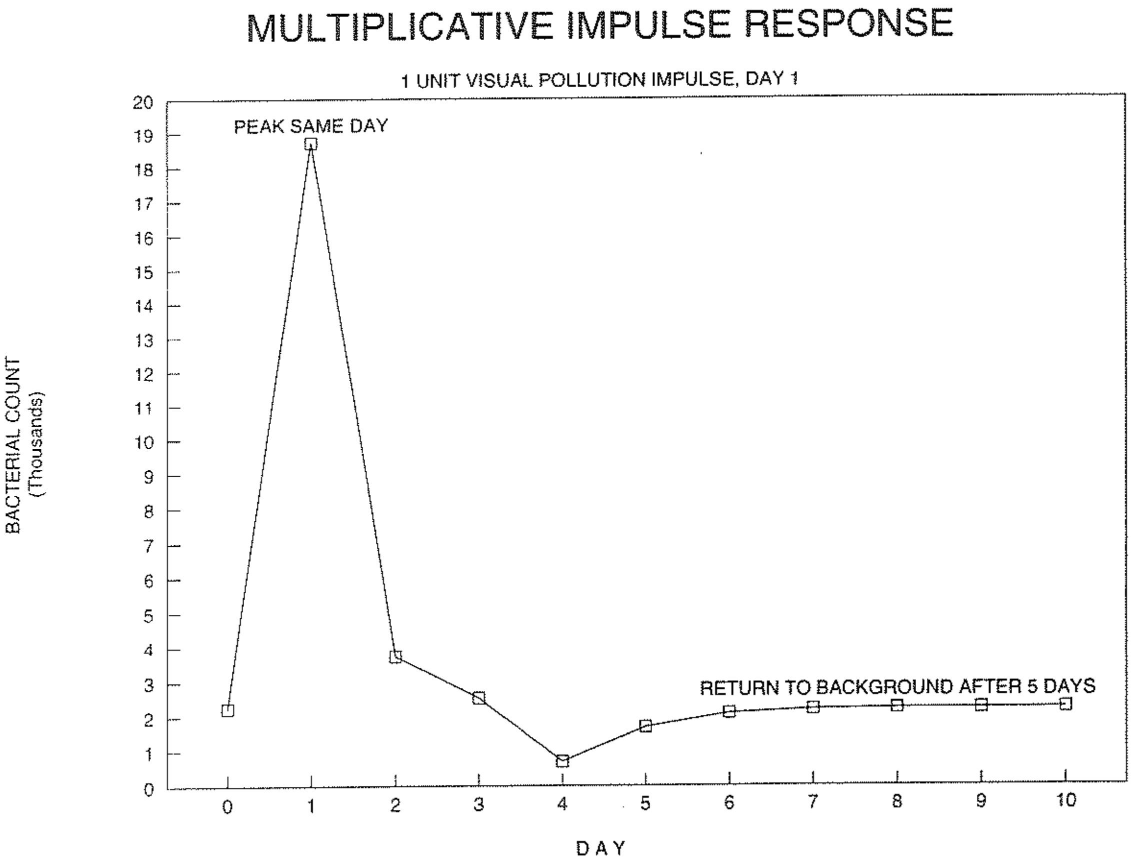

Figure 3: Impulse Response

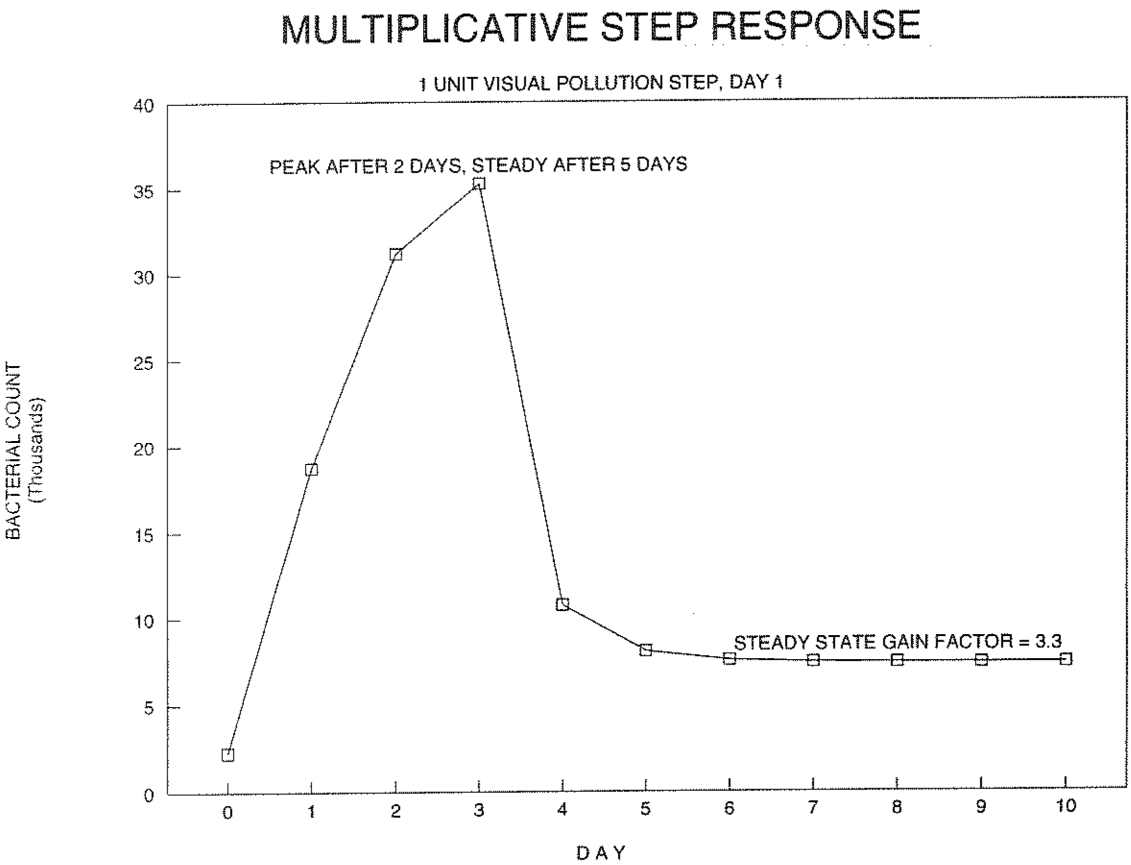

Figure 4: Step Response

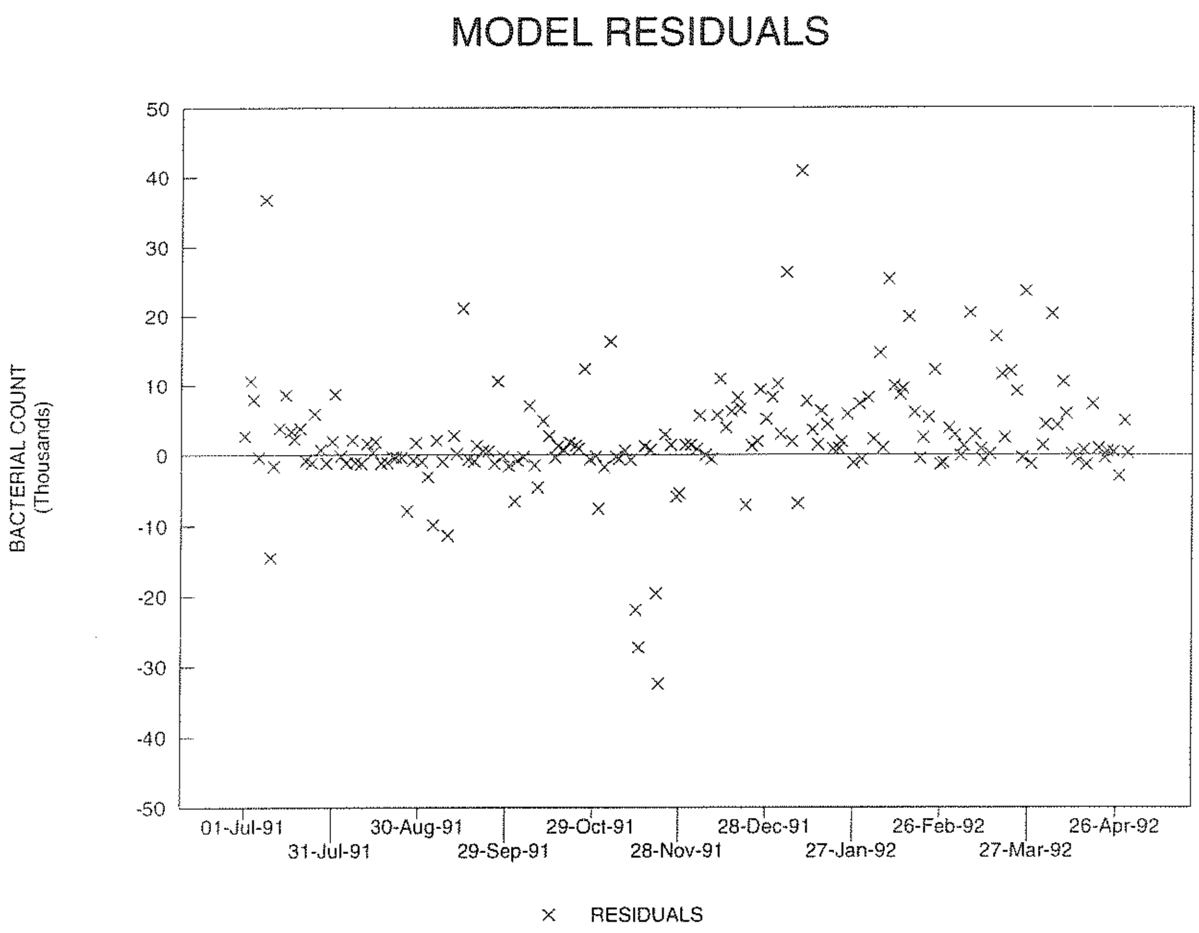

Figure 5: Model Residuals

4. CHARACTERISATION OF ESTIMATES

Consider a first order transfer function and white noise model where y_t and u_t are observed and a, b and x_t are to be estimated according to equations (10) and (11), leading to model residuals, e_t.

$$y_t = x_t + e_t \tag{10}$$

$$x_t = a x_{t-1} + b u_t \tag{11}$$

4.1 Simplified Refined Instrumental Variables

Write pre-filters as

$$Y_t = y_t + a Y_{t-1} \tag{12}$$

$$U_t = u_t + a U_{t-1} \tag{13}$$

$$X_t = x_t + a X_{t-1} \tag{14}$$

Then the model equations (10) and (11) become

$$Y_t = a Y_{t-1} + b U_t + e_t \tag{15}$$

$$X_t = a X_{t-1} + b U_t \tag{16}$$

Now partially differentiate the model residuals, e_t, which are the same in equations (10) and (15), with respect to a and b. Superscripts denote partial differentiation in equations (17) and (18).

$$e_t^a = -X_{t-1} \tag{17}$$

$$e_t^b = -U_t \tag{18}$$

$$\sum_{t=2}^{n} Y_t X_{t-1} = a \sum_{t=2}^{n} Y_{t-1} X_{t-1} + b \sum_{t=2}^{n} U_t X_{t-1} \tag{19}$$

$$\sum_{t=2}^{n} Y_t U_t = a \sum_{t=2}^{n} Y_{t-1} U_t + b \sum_{t=2}^{n} U_t^2 \tag{20}$$

The simplified refined instrumental variable method determines a and b iteratively from equations (19) and (20) so that the model residuals, e_t, are orthogonal to the prefiltered outputs and inputs, X_{t-1} and U_t, respectively, used as instrumental variables and so that the model equations, (15) and (16) are satisfied. But in view of equations (17) and (18) these estimates will minimise the sum of squared residuals and so will be the same as those computed from a nonlinear least squares optimisation of equations (10) and (11), conditional on the start-up value for x_1.

4.2 Iterated Regression on Lagged Fitted Values

The Simplified Refined Instrumental Variable Method is due to Young, and is a straightforward application of the

Refined Instrumental Variable Method (eg Young 1984) in which the errors are white noise. If, in the Simplified Refined Instrumental Variable Method, equations (19) and (20) are replaced by equations (21) and (22) respectively

$$\sum_{t=2}^{n} Y_t x_{t-1} = a \sum_{t=2}^{n} Y_{t-1} x_{t-1} + b \sum_{t=2}^{n} U_t x_{t-1} \tag{21}$$

$$\sum_{t=2}^{n} Y_t u_t = a \sum_{t=2}^{n} Y_{t-1} u_t + b \sum_{t=2}^{n} U_t u_t \tag{22}$$

$$\sum_{t=2}^{n} y_t x_{t-1} = a \sum_{t=2}^{n} x_{t-1}^2 + b \sum_{t=2}^{n} u_t x_{t-1} \tag{23}$$

$$\sum_{t=2}^{n} y_t u_t = a \sum_{t=2}^{n} x_{t-1} u_t + b \sum_{t=2}^{n} u_t^2 \tag{24}$$

then alternative estimates will be obtained in which the model residuals, e_t, are orthogonal to the unfiltered outputs and inputs, x_{t-1} and u_t, respectively, used as instrumental variables. The rapid convergence properties of the algorithm seem to be retained and these orthogonalities are sufficient to define the estimates which are the same as those for estimates obtained from *Iterated Regression On Lagged Fitted Values*, equations (23) and (24). Thus only the methods of computation of the estimates vary between the equations which are satisfied at convergence, (21,22) and (23,24).

The orthogonalities mentioned above will lead to asymptotic unbiasedness and efficiency, though the estimator will be slightly sub-optimal in the least squares sense.

The reasons for use of the method of equations (23) and (24), as applied in equations (4) and (5) and extended in equations (9) and (5) are

- Simplicity of concept and application.
- Natural extension to multiple transfer functions with ARMA noise.
- Easy combination with generalised linear models or time-varying regressions.
- No use of recursive calculations down a column of numbers. Only a shift or lagging of one column with respect to others is required.

The model below illustrates the use of the method of iterated regression on lagged fitted values and residuals for a multiple transfer function model where y_t, u_t and v_t are observed and the noise, E_t, follows an ARMA(1,1) model.

$$y_t = x_t + z_t + E_t$$

$$x_t = a x_{t-1} + b u_t$$

$$z_t = c z_{t-1} + d v_t$$

$$E_t = \phi E_{t-1} + \theta e_{t-1} + e_t$$

The next equation indicates a multiple linear regression for the i-th iteration with succeeding equations updating the columns to be used in the next iteration of the regression fit.

$$y_t = a^i x_{t-1}^{i-1} + b^i u_t + c^i z_{t-1}^{i-1} + d^i v_t + \phi^i E_{t-1}^{i-1} + \theta^i e_{t-1}^{i-1} + e_t^i$$

$$x_t^i = a^i x_{t-1}^{i-1} + b^i u_t$$

$$z_t^i = c^i z_{t-1}^{i-1} + d^i v_t$$

$$E_t^i = y_t - x_t^i - z_t^i$$

$$e_t^i = E_t^i - \phi^i E_{t-1}^{i-1} - \theta^i e_{t-1}^{i-1}$$

5. CONCLUSION

The model results suggest a common source for both the visual pollution and the bulk of the bacterial pollution. Earlier work suggested that the source is largely from the sea since the visual pollution was found to be associated with onshore winds and currents (Jellett 1996) as well as rainfall. Rainfall could produce an increase through direct urban runoff or through the urban drainage/sewerage system as well as the sea.

A new iterative instrumental variable estimator and computational method was described for time series models. The method facilitates the fitting of transfer functions in combination with other models such as generalised linear models or time-varying parameters. The method also provides a simple approach to computing estimates for multiple input, single output transfer functions (Jellett 1996). The method also permits time series models to be fitted using software tools which support only regression without recursive calculations.

6. ACKNOWLEDGEMENT

The New South Wales Environment Protection Agency provided the data for this paper.

7. REFERENCES

Bishop, Y. M. M., S. E. Fienberg and P. W. Holland, Discrete multivariate analysis, *MIT Press*, 1980.

Box, G.E.P., and G.M. Jenkins, Time series analysis: forecasting and control, *Holden-Day*, 1976.

Brockwell, P. J. and R. A. Davis, Time series: theory and methods, *Springer-Verlag*, 1991.

Jellett, P.M., Time series models for beach pollution, *Environmental Software*, Vol 11, 25-33, 1996.

Jellett, P.M., A seasonal model for general use, *MODSIM97 Proceedings*, University of Tasmania, Hobart, Modelling and Simulation Society of Australia, Inc, 1997.

Nelder, J. A. and R. W. M. Wedderburn, Generalised Linear Models, *Journal of the Royal statistical Society*, A, 135, 370-384, 1972.

The GLIM language, *Numeral Algorithms Group*, Wilkinson House, Jordan Hill Road, Oxford, United Kingdom, OX2 8DR.

Young, P., Recursive estimation and time-series analysis, *Springer-Verlag*, 1984.