JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 8(1), 15–32

SUR Models Applied to an Environmental Situation With Missing Data and Censored Values

ROSS SPARKS[†] ross.sparks@csiro.au
CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 1670, Australia

Abstract. This paper develops methodology for predicting faecal coliform values in waterways when some of the data are missing, and some of the data are left-censored. Such predictions are important in predicting when bathing is safe in specific areas of Sydney harbor. The approach taken in the paper makes use of spatial information to improve these predictions.

Keywords: EM-algorithm, Generalized least squares, Left-censoring, Parameter estimation, Simulation and Spatio-temporal modelling.

1. Introduction

The Seemingly Unrelated Regression (SUR) model is well-known in the econometric literature (Zellner [19], Srivastava and Giles [17], Greene [4]), but is less known else where. Its benefits have been explored by several authors (Zellner [20], Kmenta and Gilbert [6], Revankar [13] & [14], Mehta and Swamy [10], Dwivedi and Srivastava [3], Maeshiro [9], Blatten and George [1], Kurata and Kariya [7]). More recently the SUR model is being applied in agricultural economics (O'Donnell, Shumway and Ball [11], Wilde, McNamara, and Ranney [18]), aquaculture economics (Samonte-Tan and Davis [15]), and forestry for modelling tree growth (Hasenauer, Monserud and Gregoire [5]). Its application in the natural sciences is likely to increase once scientists in these disciplines are exposed to its potential. Liu and Wang [8] demonstrated to natural scientists the superiority of SUR model over the unrelated model that uses ordinary least squares to estimate parameters.

This paper illustrates how the SUR model could be used to fit spatio-temporal data in an environmental setting. It starts by introducing the SUR model and presents some of its properties when covariances are known.

[†] Requests for reprints should be sent to Ross Sparks, CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 1670, Australia.

It then examines how regression parameters will be estimated if covariances are unknown. A simulation study is used to compare SUR model with unrelated regression models.

Missing data is fairly common in the applications cited in this paper. Thus the paper indicates how the EM algorithm (Dempster, Laird, and Rubin [2]) can be used to give Maximum Likelihood Estimates of regression and dispersion parameters. A simulated examples illustrate the advantage of using all the data to estimate parameters when some of the responses are missing.

The applications considered also have left-censored values. This paper illustrates how the EM algorithm is used to overcome this incomplete data problem. A simulation study demonstrates the methodology. Followed by a detailed consideration of an example of application.

2. Introduction to the SUR Model

Assume that you have p response variables each with n observations denoted by vectors $y_1, y_2, \cdots, y_p$ with associated explanatory variables $X_1, X_2, \cdot\cdot\cdot, X_p$, respectively. One way of fitting these models is to treat them as unrelated multiple regression models of the form of

$$y_j = X_j \beta_j + e_j \tag{1}$$

where β_j is a vector of unknown regression parameters and e_j is a vector of random errors with each element having variance σ_j^2, for $j = 1, 2, \cdot\cdot\cdot, p$. Let

$$X = \begin{bmatrix} X_1 & 0 & . & . & 0 \\ 0 & X_2 & . & . & 0 \\ . & . & . & & . \\ . & . & & . & . \\ 0 & 0 & . & . & X_p \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_p \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{bmatrix}, e = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_p \end{bmatrix} \tag{2}$$

$E(e_i e_j^{'}) = \sigma_{ij} I$ where $\sigma_{ij} \neq 0$, $E(ee') = \Sigma = I \otimes \Delta$, where $\Delta = \{\sigma_{ij}\}$ is the covariance matrix capturing the variances and covariances of the random error terms of (1), then the SUR form of this model is

$$y = X\beta + e \tag{3}$$

When Δ is known, then SUR formulation of the regression models produces more efficient regression parameter estimates using generalized least squares estimates given by

$$\hat{\beta} = (X^{'}\Sigma^{-1}X)^{-1}X^{'}\Sigma^{-1}y, \tag{4}$$

than the ordinary least squares estimates in the unrelated form of the model given by

$$\hat{\beta}_j = (X_j^{'}X_j)^{-1}X_j^{'}y_j \tag{5}$$

for each $j = 1, 2, ..., p$. However the efficiency of the estimators in applications is not that simple.

2.1. Efficiency of SUR model for estimating regression coefficients

Note that if either $\sigma_{ij} = 0$ for all $j \neq i$ or $X_i = X_j$ for all $j \neq i$, then the two model formulations produce identical estimates, that is, equation (4) will be identical to (5). The efficiency in the SUR formulation increases, the more the correlations between error vector differ from zero and the closer the explanatory variables for each response are to being uncorrelated (e.g., see Mehta and Swamy [10]). Sparks [16] discussed how to select variables and parameter estimators for the SUR model. The standard errors for the set of the regression parameter estimates in the SUR formulation is given by the diagonal elements of $\left(X^{'}\Sigma^{-1}X\right)^{-1}$ while for the unrelated formulation it is the appropriate diagonal elements of $(X_j^{'}X_j)^{-1}$ for $j = 1, 2, \ldots, p$. These can be used to gain an idea of the relative merits of the SUR model formulation for estimating the regression parameters of the model. Generally when $\sigma_{ij} \neq 0$ and the covariances are known, it can be shown that the diagonal elements of $(X_j^{'}X_j)^{-1}$are larger than the corresponding diagonal elements of $\left(X^{'}\Sigma^{-1}X\right)^{-1}$for each j, and the mean square error of prediction using the generalized least squares estimate is smaller. This is not generally true when the covariances are unknown. The relative benefit of SUR model when covariances are unknown depends on the sample size n (Zellner, [20], Kmenta and Gilbert [6], Revankar [13] & [14], Mehta and Swamy [10], Dwivedi and Srivastava [3], Maeshiro [9]). When fitting regression models with small sample sizes, its unlikely that the seemingly unrelated regression formulation and related generalized least squares estimates are going to add much value. However, the larger the sample size, the more reliable is the estimate of Δ, and hence the more likely an advantage is gained from the seemingly unrelated regression formulation of the model.

3. The Application

In this paper, the response variables are the natural logarithm of one plus the faecal coliform measurement taken at various sites in Port Jackson (Sydney). Faecal coliform measurements are a good indicator of the level of bacteria in the waterways, and therefore they flag risks associated with recreational activities, like swimming and boating. The collection of faecal coliform measurements is expensive, however moderate to heavy local rainfall induce many sewerage overflows into the waterways. Therefore an important explanatory variable of faecal coliform levels in the waterways is the aggregation of previous rainfall values. Aggregation was taken over non-overlapping twelve hour periods (running from 6am to 6pm, and 6pm to 6am). Typically measuring rainfall is considerably cheaper than measuring faecal coliforms. Using a modelling approach we can establish predictions of faecal coliform levels from daily rainfall values, and therefore averting the need to measure faecal coliforms on a daily basis.

Each sample site in Port Jackson differed in distance from the rainfall gauge site. Therefore, forward selection with a backward look (stepwise regression) was used to select the set of rainfall explanatory variables that were best for predicting each site response.

These subset models were included in the seeming unrelated regression formulation. The rainfall data takes away some of the spatial relationship between site response values, however the error term in the models are still going to have some spatial information relating to tidal influences, local rainfall missed by the network of gauges, and flow-down effects in Port Jackson. These spatial relationships determine the value of using the SUR formulation of the model.

The faecal coliform data from sample sites have two incomplete data problems:

- **Missing data:** Not every site was sampled on each day the data were recorded. There were days when a subset of the sites was sampled.
- **Censored data:** The measurement methodology could not measure faecal coliform values below 4 CFU (colony forming units).

Therefore, in this paper methodology is developed to deal with these issues.

3.1. Application of the EM algorithm to deal with missing data and left-censored observations

3.1.1. Dealing with missing data

Missing data occurs when no feacal coliform measurement is collected for least for one of the sites on a sample day.

The EM-algorithm of Dempster, Rubin and Laird [2] is applied to establish maximum likelihood estimates for the regression coefficients and dispersion parameters of the SUR model. The E-step of this algorithm is defined for the various cases below.

Denote y_{ij} as the ith day's *logarithm of faecal coliform measurement plus one* collected at the jth location in the waterway (x_{ij} is defined similarly for rainfall readings). Note that $y_{ij} \sim N(x_{ij}^{'}\beta_j, \sigma_j^2)$ is assumed.

Case 1: Only one site has missing data. Assume that this jth observed response is missing for day i. The expected value of this missing value, given all other measured values is

$$E(y_{ij}|y_{ik} \text{ for all } k \neq j) = x_{ij}^{'}\beta_j + \sigma_{(-j)}^{'}\Sigma_{(-j)}^{-1}(y_{i(-j)} - Z_{i(-j)}^{'}\beta_{(-j)}) = \widehat{\mu}_{ij.(-j)} \tag{6}$$

where

- $\sigma_{(-j)}^{'}$ is the j th row of Σ with the jth element removed,
- $\Sigma_{(-j)}$ is the matrix Σ with the jth row and jth column removed,
- $y_{i(-j)}$ is the vector of all faecal coliform observations collected on day i with the jth response removed, and
- $\beta_{(-j)}$ is the vector β with the regression coefficients relating to selected explanatory variables of the jth response/site removed, i.e., β_j removed.
- $x_{ij}^{'}$ be the ith row of X_j and
- $Z_i = \begin{bmatrix} x_{i1}^{'} & 0_2^{'} & \cdot\ \cdot\ \cdot & 0_p^{'} \\ 0_1^{'} & x_{i2}^{'} & \cdot\ \cdot\ \cdot & 0_p \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0_1^{'} & 0_2^{'} & \cdot\ \cdot\ \cdot & x_{ip}^{'} \end{bmatrix}$ in which 0_j is a vector of zeros of the same length as vector x_{ij}, then $Z_{i(-j)}^{'}$ is Z_i with the jth row removed.

Case 2: More than one site has missing data. Let C be the set of number sites that have faecal coliform measurements taken on day i. Then the expected value of the missing jth response value, given all observed response values during data collection day i is

$$E(y_{ij}|y_{ik} \text{ for all } k \in C) = x_{ij}^{'}\beta_j + \sigma_{jC}^{'}\Sigma_{CC}^{-1}(y_{iC} - Z_{iC}^{'}\beta_C) = \widehat{\mu}_{ij.C} \qquad (7)$$

where

- Σ_{CC} is the sub-matrix of Σ that corresponds to all responses in C,
- $\sigma_{jC}^{'}$ are the covariances between jth response and all responses in C as a row vector,
- y_{iC} is the vector of all faecal coliform observations collected on day i with response number in C,
- β_C is the vector β with the regression coefficients relating to selected explanatory variables of all response/site variables with numbers not in C removed, and
- $Z_{iC}^{'}$ is a matrix with rows consisting of all rows Z_i that correspond to variable numbers in C.

The expected value of the product of two missing responses on the same day. Denote $\sigma_{jk.C} = \sigma_{jk} - \sigma_{jC}^{'}\Sigma_{CC}^{-1}\sigma_{kC}$. If any pair of response variables are missing on day i, then the expected value of their product is

$$E(y_{ij}y_{ik}|y_{im} \text{ for all } m \in C) = \sigma_{jk.C} + \widehat{\mu}_{ij.C}\widehat{\mu}_{ik.C} \qquad (8)$$

Therefore, when applying the SUR model we need to consider whether to apply the SUR model to:

- all sites with only completed data (discarding observations with any site having missing data), or
- all the data using the EM algorithm (Dempster *et al* [2]) to deal with missing data.

We found using simulation techniques that the latter approach was significantly better in terms of minimizing the mean square error of prediction.

3.1.2. Dealing with left-censored data values.

The measurement process of the faecal coliform values can not report values of 4 colony forming units (CFU) or below, e.g., any “true” zero counts for CFU are reported as below 4 CFU. We therefore do not have complete information for all measurements. Also, the variation in measured values such as:

- sampling variation;
- variation caused by differences in handling (outside the laboratory); and
- variation caused by errors within measurement laboratories,

can be as much as 40% on average at some sites under certain conditions. Therefore, we can not be certain that a near zero measurement is a true reflection of the faecal coliform level at a site.

Four methods of dealing with the censored values and missing values are considered. These are as follows:

Method 1: Simple approach to handling censored data. Here we patch censored data with the quantiles of the estimated distribution below the censored value. To achieve this, the distribution is assumed to be normal (log-normal) and the parameters are estimated from the data as outlined below.

Let $\bar{y}_j$ be the mean of all uncensored y values from the jth site that are not missing, and m_j be the number of missing values in T days sampled. The estimated population mean (Persson and Rootzen [12]) is:

$$\hat{\mu}_j = \bar{y}_j - \alpha \hat{\sigma}_{RML}(j) \tag{9}$$

and

$$\hat{\sigma}^2_{y_j} = \Sigma_{i=1}^{T-m_j} (y^*_{ij} - \bar{y}_j)^2/(T - m_j) - (\alpha \lambda_{(T-m_j)/T} - \alpha^2) \hat{\sigma}^2_{RML}(j) \tag{10}$$

where:

- y^*_{ij} is the ith observed value of the jth response,
- $\alpha = \frac{T}{\sqrt{2\Pi}} e^{-\lambda^2_{(T-mj)/T}}/(T - m_j)$,
- $\lambda_{a/b}$ is the upper $\frac{a}{b}$th quantile of the standard normal distribution, and

- $$\hat{\sigma}^2_{RML}(j) = \frac{1}{2}\{\lambda_{(T-m_j)/T}\bar{\nu}_j + \sqrt{[(\lambda_{(T-m_j)/T}\bar{\nu}_j)^2 + 4\tau_j^2}\} \qquad (11)$$

  in which $\bar{\nu}_j = \Sigma_{i=1}^{T-m_j}(y^*_{ij} - \log(5))/(T - m_j)$ and $\tau_j^2 = \Sigma_{i=1}^{T-m_j}(y^*_{ij} - \log(5))^2/(T - m_j)$.

For our example we know that faecal coliforms are positively associated with local rainfall. Therefore for the jth site, the censored days are ranked according to total rainfall in the region and they are patched with the respective quantile value of a normal distribution with mean $\hat{\mu}_j$ and variance $\hat{\sigma}^2_j$. This is repeated for each response j. These patched censored values are treated as if they were known, and the parameters of the model (excluding missing observation) are estimated for each site individually using (5).

Method 2: Simple approach to handling censored data and missing data. We carry out all steps in Method 1. That is, we patch censored data with the quantiles of the estimated distribution below the censored value. Then the patched censored values and ordinary least squares are used to estimate regression parameter for each site, independently. These estimated models are used to estimate missing data for each site, and then the SUR model is used to improve the estimate of the regression model (i.e., using (4)).

Method 3: EM algorithm used to handle both censored data and missing data. It seems convenient to use the same methodology for dealing with both sets of incomplete information. Therefore, the EM algorithm is used to “patch” left-censored value as well as the missing data via an expectation step, and then these “patched” values are used to establish maximum likelihood estimates of the parameters. These two steps are applied iteratively until estimates converge.

We build in the information recorded for left-censored values by finding the conditional expectation of responses given censored values are less than the low detect values. However, there is information available from neighboring sites and local rainfall records, and this can also help with patching “censored” values.

Case 1: *All other response values for the ith day are either uncensored or not missing.* Put $\sigma_{jj.(-j)} = \sigma_{jj} - \sigma'_{(-j)}\Sigma^{-1}_{(-j)}\sigma_{(-j)}, Z_{(-j)}$, and let y denote the log faecal coliform measurement plus 1. Censored value y_{ij} is replaced by (see Appendix A)

$$E(y_{ij}|y_{ik} \text{ for all } k \neq j \ \& \ y_{ij} < \log(5)) = \widehat{\mu}_{ij.(-j)} - \sigma_{jj.(-j)}\phi(z_1)/\Phi(z_1) \qquad (12)$$

where:

$$z_1 = [\log(5) - \widehat{\mu}_{ij.(-j)}]/\sqrt{\sigma_{jj.(-j)}} \tag{13}$$

Case 2: *All other response values for i th day are uncensored but some sites on the day have missing values.* Censored value y_{ij} is replaced by (see Appendix A)

$$E(y_{ij}|y_{ik} \text{ for all } k \neq j \ \& \ y_{ij} < \log(5)) = \widehat{\mu}_{ij.C} - \sigma_{jj.C}\phi(z_2)/\Phi(z_2) \tag{14}$$

where:

$$z_2 = [\log(5) - \widehat{\mu}_{ij.C}]/\sqrt{\sigma_{jj.C}} \tag{15}$$

Case 3: *Some other response values for ith day are censored and other sites on the day have missing values (see Appendix B).* Let S be all sites with censored values on day i. Censored value y_{ij} is replaced by (see Appendix B)

$$E(y_{ij}|y_{ij} < \log(5) \ \& \ y_{i\ell} < \log(5) \text{ for all } \ell \in S) \tag{16}$$

which is solved using simulation methods.

Case 4: *Some other response values for ith day are censored and sites on the day have observed uncensored values (see Appendix B).* Censored value y_{ij} is replaced by (see Appendix B)

$$E(y_{ij}|y_{ik} \text{ for all } k \neq j, y_{ij} < \log(5) \ \& \ y_{i\ell} < \log(5) \text{ for all } \ell \in S) \tag{17}$$

which is solved using simulation methods.

Method 4: E-Step of the *EM algorithm is used to handle both censored data and missing data, but this E-step is simplified to condition on all variables not included in the expectation.* It would be convenient if we could ignore the difficult conditioning on other censored values as in Case 3, and condition on them as if they were observed, by using the previous estimates of censored values in iterations as the"observed values". This method obviously ignores some of the information the censored values provide, but we want to know what is lost.

4. Comparison of Methods by Simulation

4.1. Example 1

The following SUR model was simulated

$$w_{i1} = \exp(2.5 + 2x_{1j} + x_{2i} + e_{1i} + 0.4e_{2i}) - 1$$

and

$$w_{i2} = \exp(3 + 2x_{3i} + 2x_{4i} + e_{3i} + 0.4e_{2i}) - 1$$

where $x_{1i}, x_{2i}, x_{3i}, x_{4i}, e_{1i}, e_{2i}$ and e_{3i} are generated as independent standard normal random variates. Ten percent of the random variables w_{i1} and w_{i2} were hidden at random, i.e., treated as missing. Any of w_{ij} values below 4 were censored. The resulting data were fairly typical of some sites in Port Jackson. The number of observations simulated on each occasions is 40. The response variables in the fitting process was taken as $y_{ij} = \log_e(w_{ij} + 1)$ for $j = 1, 2$.

Denote $\hat{\beta}_p$ as the estimator gained by using the E-step of EM-algorithm to patch missing and censored values as in equations (5), (6), (7), (10), (11) and (12). $\hat{\beta}_{np}$ is the estimator gained by first setting censored values of y_{ij} to $\log_e(4+1)$ and then the EM algorithm to establish parameter estimates. The simulation was run 100 independent times with each simulation calculating $|\hat{\beta}_p - \beta_{true}| - |\hat{\beta}_{np} - \beta_{true}|$ where $\beta'_{true} = \begin{bmatrix} 2.5 & 2 & 1 & 3 & 2 & 2 \end{bmatrix}$. The results are presented in Fig. 1, where the parameter estimates are number 1, 2, 3, 4, 5 & 6 referring to estimates of $2.5, 2, 1, 3, 2$, and 2, respectively. It is clear from Fig. 1 that we gained by using the EM algorithm to adjust

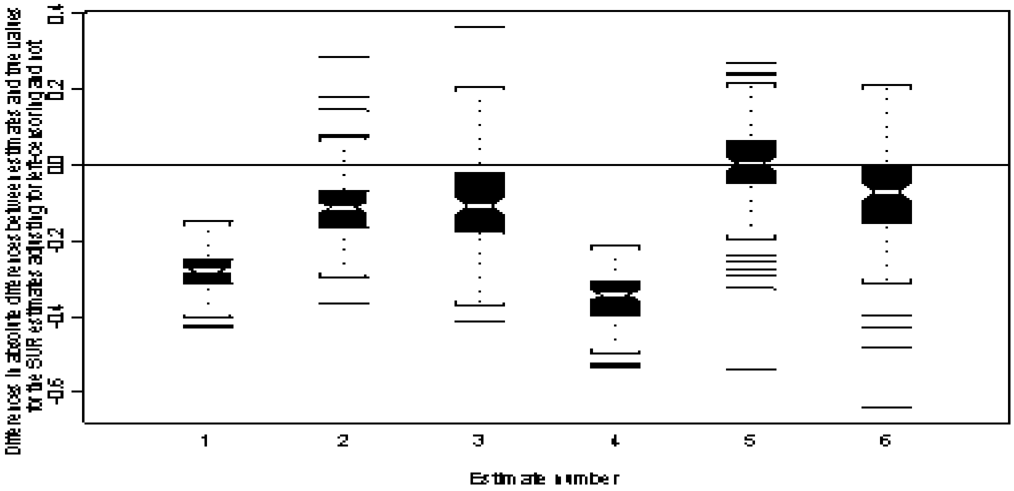


Figure 1. Comparison of regression estimates when missing are patched, and censored data are patched and not patched.

4.2. Example 2

The same simulation example as Example 1 was used, except this simulation compared the various methods defined earlier for dealing with missing and censored values.

Fig. 2 illustrates that there is very little difference between Method 3 and Method 4, and both these methods are superior to Method 1 and Method 2. Method 2 has only a slight advantage over Method 1. Method 4 requires much less computational effort than Method 3, of the order of $\frac{1}{400}$th of the effort, and this is likely to increase as the amount of censoring increases. In terms of balancing effort and accuracy, Method 4 seems to be preferable. Very little is lost by avoiding the complicated evaluation of conditional expectations (16) and (17).

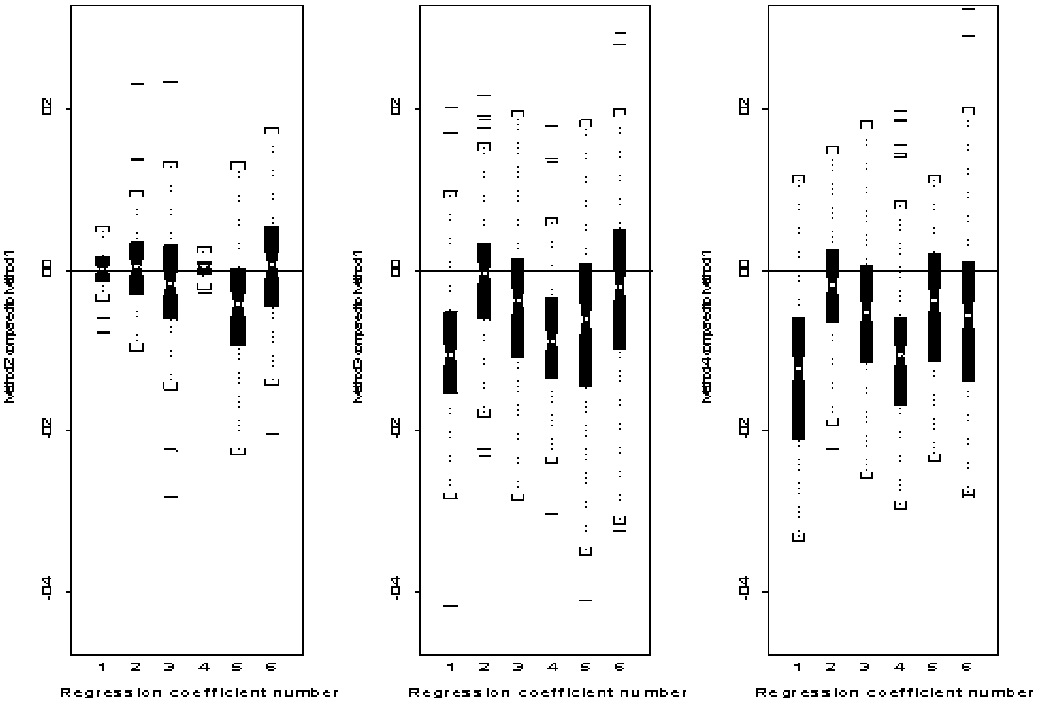


Figure 2. Comparing methods of estimating model parameters using a simulation study.

5. Examples of Application

Recall the application section earlier in the paper. This outlined the intention of producing efficient predictions of faecal coliform (fc) values at various recreational sites in Port Jackson (Sydney). Three different geographical locations in Port Jackson are used to illustrate the value of these models (see Table 1):

Table 1. Sample site information and model fit.

Areas of Port Jackson	Site locations	No. of obs.†	**Percentage**		
			values censored	values missing	variation explained by model
Parramatta River	PJ02, PJ03, PJ04	146	20%	4%	72%
Lane Cove River	PJ05, PJ06, PJ07	150	11%	3%	73%
Middle Harbor	PJ29, PJ30, PJ31, PJ32	153	31%	1%	80%

† days fc values were measured

Recall the explanatory variables are the lagged 12-hourly rainfall totals. Up to five lags at each of eleven rainfall stations are considered. Using forward selection with a backward look (stepwise regression) we selected the significant rainfall explanatory variables for each site. The selected subset models are then written in SUR model form. The parameters are estimated using the SUR model. The results are recorded in Tables 2, 3, and 4.

The following are worth noting in these three regions of Port Jackson.

- The variable selection process selects 12-hour rainfall totals lagged 1, 2, 3, 4 and 5 from time fc was measured as significant explanatory variables approximately 9%, 9%, 19%, 23%, and 40% of the time, respectively. This was a surprise because prior to this analysis, experts at Sydney Water suggested that rainfall now would not influence faecal coliform levels 48 hours later. For this reason, we only considered explanatory rainfall variables up to lag 5 (48 to 60 hours later). The modelling process did not seem to adhere to conventional wisdom, and future models should consider lags beyond 60 hours.

- Note that the explanatory variables relating to lag 5 values generally corresponded to, on average, higher regression parameter estimates in

Table 2. Estimated model for sites in the Parramatta River.

	Method 3	**Method 2**	**Method 1**
PJ02			
Constant	3.0132	2.9653	2.9331
Rainfall site 22 lag 5	0.2382	0.2639	0.3123
Rainfall site 25 lag 1	0.1869	0.1787	0.1662
Rainfall site 37 lag 2	0.0703	0.0719	0.0723
Rainfall site 64 lag 4	0.1903	0.1961	0.1985
Rainfall site 70 lag 3	0.6155	0.6134	0.6624
PJ03			
Constant	2.2788	2.1949	2.1602
Rainfall site 25 lag 1	0.0664^{ns}	0.0674	0.0680
Rainfall site 29 lag 5	0.1026	0.1054	0.0952
Rainfall site 37 lag 2	0.0484	0.0505	0.0555
Rainfall site 37 lag 3	0.1691	0.1854	0.2079
Rainfall site 37 lag 5	0.1398	0.1395	0.1337
Rainfall site 70 lag 3	0.1466	0.1657	0.2190
Rainfall site 70 lag 5	0.3135	0.3287	0.3585
Rainfall site 87 lag 4	0.0593^{ns}	0.0616	0.0608
PJ04			
Constant	1.8708	1.7810	1.7914
Rainfall site 25 lag 5	-0.0629	-0.0679	-0.1053
Rainfall site 64 lag 4	0.0769	0.0821	0.0860
Rainfall site 70 lag 5	0.1300	0.1345	0.1553
Rainfall site 87 lag 4	0.1769	0.1897	0.2110

the Lane Cove River. Explanatory variables relating to lag 3 or 5 rainfall values generally corresponded to, on average, higher regression parameter estimates in Middle Harbor. Explanatory variables relating to lag 3, 4 or 5 rainfall values generally corresponded to higher regression parameter estimates in Port Jackson.

- Variables that are significant when fitting independent multiple regression models can become insignificant when fitting the seeming unrelated regression model (e.g., those parameters with a superscript *ns* on the left-hand side of the estimate are nonsignificant).
- Some estimates change very little when all the information is used (SUR formulation), while others change significantly (e.g., Rainfall site 25 lag 5 of PJ30 changes dramatically).
- Estimates of the explanatory variable regression coefficients are generally closer to zero using the SUR model (Method 3), than the respective

coefficients estimated from separate models – Method 1 (in more than 80% of the cases).

Table 3. Estimated model for sites in the Lane Cove River.

	Method 3	**Method 2**	**Method 1**
PJ05			
Constant	2.0164	1.8858	1.8801
Rainfall site 25 lag 5	-0.0860	-0.0805	-0.0886
Rainfall site 29 lag 1	0.1948	0.2052	0.1967
Rainfall site 37 lag 4	0.0973	0.1027	0.1070
Rainfall site 64 lag 5	0.2061	0.2102	0.2107
Rainfall site 70 lag 5	0.2308	0.2273	0.2618
PJ06			
Constant	2.8925	2.7821	2.7773
Rainfall site 22 lag 2	0.1976	0.2029	0.2391
Rainfall site 22 lag 5	0.4554	0.4765	0.4931
Rainfall site 25 lag 5	-0.1985	-0.2060	-0.2197
Rainfall site 66 lag 1	0.1792	0.1882	0.1629
Rainfall site 87 lag 3	0.1785	0.1670	0.1608
Rainfall site 87 lag 4	0.0499	0.0641	0.0698
PJ07			
Constant	3.5448	3.4947	3.4823
Rainfall site 17 lag 4	0.1887	0.1924	0.1956
Rainfall site 29 lag 3	0.3449	0.3479	0.3617
Rainfall site 66 lag 5	0.1167	0.1217	0.1223

6. Conclusion

The SUR model can be usefully applied to many point estimation problems in environmental settings. This model proves useful in spatio-temporal settings where temporal information is built into models via the appropriate explanatory variables and lagged response variables, while the spatial correlations are used via the model errors to improve parameter estimates and predictions.

Often these environmental settings have missing data and left censoring observations. This paper provides an approach to fitting SUR models when faced with these difficulties. Several methods of handling these are explored in the paper, and the simple approach of applying the E-step of the EM algorithm by conditioning on all observations (whether observed, censored or missing), and iterating until estimates converge (Method 4)

Table 4. Estimated model for sites in Middle Harbor.

	Method 3	**Method 2**	**Method 1**
PJ29			
Constant	1.9208	1.8250	1.8517
Rainfall site 25 lag 5	-0.0657^{ns}	-0.588	-0.0976
Rainfall site 64 lag 5	0.1525	0.1563	0.1750
Rainfall site 70 lag 3	-0.3198	-0.2565	-0.2852
Rainfall site 70 lag 5	0.2300	0.2282	0.2465
Rainfall site 87 lag 3	0.2857	0.2882	0.2924
Rainfall site 87 lag 5	0.1339	0.1418	0.1428
PJ30			
Constant	1.6241	1.5312	1.5508
Rainfall site 25 lag 5	-0.0181^{ns}	-0.0246	-0.0760
Rainfall site 37 lag 5	0.0709	0.0739	0.0923
Rainfall site 70 lag 4	0.1626	0.1831	0.1898
Rainfall site 70 lag 5	0.3772	0.3849	0.4195
Rainfall site 87 lag 3	0.6791	0.6967	0.7063
PJ31			
Constant	1.4817	1.4038	1.4035
Rainfall site 22 lag 2	0.2505	0.2727	0.3084
Rainfall site 29 lag 4	0.0693	0.0764	0.0618
Rainfall site 37 lag 3	0.3255	0.3259	0.3205
Rainfall site 66 lag 5	0.1226	0.1299	0.1312
Rainfall site 70 lag 3	0.3399	0.4350	0.4118
Rainfall site 70 lag 4	0.0835^{ns}	0.0769	0.0889
Rainfall site 70 lag 5	0.3046	0.3058	0.3025
PJ32			
Constant	1.7820	1.7777	1.7609
Rainfall site 17 lag 2	0.1888	0.1956	0.1967
Rainfall site 17 lag 3	0.6784	0.6714	0.6733
Rainfall site 17 lag 4	0.1532	0.1579	0.1584
Rainfall site 29 lag 5	0.3904	0.3727	0.4346
Rainfall site 64 lag 5	0.1477	0.1470	0.1486
Rainfall site 70 lag 4	0.1518	0.1556	0.1405

is computationally efficient and reasonably accurate. In those few cases where accuracy is paramount use Method 3.

References

1. Blattberg, R., and George, E.I. (1991), Shrinkage estimation of price and optional elasticities: Seemingly Unrelated Equations, *J. Amer. Statist. Assoc.*,86, 304-315.

2. Dempster, A.P., Laird, N. M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Stat. Soc. B* 39, 1-38.
3. Dwivedi, T.D. and Srivastava, V.K. (1978). Optimality of least squares in the seemingly unrelated regression equation model. *Journal of Econometrics*, 7, 391-395.
4. Greene, W. (1993). *Econometric Analysis (2nd Ed.)* New York: Macmillan.
5. Hasenauer, H., Monserud, R.A. and Gregoire, T.G. (1998). Using simultaneous regression techniques with individual-tree growth models. *Forest Science*, 44(1), 87-95.
6. Kmenta, J. and Gilbert, R.F. (1968). Small sample properties of alternative estimators of seemingly unrelated regressions. *J. Amer. Statist. Assoc.*, 63, 1180-1200.
7. Kurata, H., and Kariya, T. (1996). Least upper bound for covariance matrix of a generalized least squares estimator in regression with applications to a seemingly unrelated regression model and a heteroscedastic model. *The Annals of Statistics*, 24, 1547-1559.
8. Liu, J.S. and Wang, S.G. (1999). Two-stage estimate of the parameters in seemingly unrelated regression model. *Progress in Natural Science*, 9.
9. Maeshiro, A. (1980). New evidence on small sample properties of estimators of SUR models with autocorrelated disturbances. *Journal of Econometrics*, 12, 177-187.
10. Mehta, J.S. and Swamy, P.A.V.B. (1976). Further evidence on the relative efficiencies of Zellner's seemingly unrelated regression estimator. *J. Amer. Statist. Assoc.*, 71, 634-639.
11. O'Donnell, C.J., Shumway, C.R. and Ball, V.E. (1999). Input demands and inefficiency in US agriculture. *American Journal of Agricultural Economics*, 81, 865-880.
12. Persson, T. and Rootzen, H. (1977). Simple and highly efficient estimators for a type I censored normal sample. *Biometrika*, 64, 123-128.
13. Revankar, N.S. (1974), Some finite sample results in the context of two seemingly unrelated regression equations. *J. Amer. Statist. Assoc.*,69, 187-190.
14. Revankar, N.S. (1976). Use of restricted residuals in SUR systems: Some finite sample results. *J. Amer. Statist. Assoc.*,71, 183-192.
15. Samonte-Tan, G.P.B. and Davis, G.C. (1998). Economic analysis of stake and rack-hanging methods of farming oysters (Crassostrea iredalei) in the Philippines. *Aquaculture*, 160,239-249.
16. Sparks, R.S. (1987). Selecting estimators and variables in the seemingly unrelated regression model. *Commun. Statist.-Simula.*, 16, 99-127.
17. Srivastava, V.K. and Giles, D.E.A. (1987). *Seemingly Unrelated Regression Equations Model: Estimation and Inference*, New York:Marcel Dekker.
18. Wilde, P.E., McNamara, P.E., and Ranney, C.K. (1999). The effect of income and food programs on dietary quality: A seemingly unrelated regression analysis with error components. *American Journal of Agricultural Economics*, 81(4), 959-971.
19. Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Amer. Statist. Assoc.*, 57, 348-368.
20. Zellner, A. (1963). Estimators for seemingly unrelated regression equations: some exact finite sample results. *J. Amer. Statist. Assoc.*, 58, 977-992.

Appendix A

Let log of the faecal coliform values plus one be denoted by the random variable y and assume that this is normally distributed. Let $\phi(z) = e^{-\frac{1}{2}z^2}/\sqrt{2\pi}$

and $\Phi(z) = \int_{-\infty}^{z} \phi(x)dx$, then

$$\begin{aligned} E(y|y<c) &= \textstyle\int_{-\infty}^{c} ye^{-\frac{1}{2\sigma^2}(y-\mu)^2}/(\sqrt{2\pi}\sigma)dy/\Phi(\frac{c-\mu}{\sigma}) \\ &= \textstyle\int_{-\infty}^{c} \frac{(y-\mu)}{\sigma}e^{-\frac{1}{2\sigma^2}(y-\mu)^2}/\sqrt{2\pi}dy/\Phi(\frac{c-\mu}{\sigma}) + \mu \\ &= -\sigma e^{-\frac{1}{2\sigma^2}(y-\mu)^2}/\sqrt{2\pi}|_{y=-\infty}^{y=c}/\Phi(\tfrac{c-\mu}{\sigma}) + \mu \\ &= -\sigma\phi(\tfrac{(y-\mu)}{\sigma})/\Phi(\tfrac{c-\mu}{\sigma}) + \mu \end{aligned}$$

Appendix B

Log of faecal coliform values plus one at site j is denoted by the random variable y_j. Assume the normal distribution. Let $S = (\ell_1 ... \ell_S)$ be all response variable numbers corresponding to censored values and $C = (C_1, ..., C_{p-S-1})$ be all the response variable numbers that correspond to uncensored values,
$y_r^{'} = [\ y_j \ \ y_{\ell_1} \ \ . \ \ . \ \ . \ \ y_{\ell_S}\ ]$,
$y_C^{'} = [\ y_{C_1} \ \ y_{C_2} \ \ . \ \ . \ \ . \ \ y_{C_{p-S-1}}\ ]$,
$cov([\ y_r^{'} \ \ y_C^{'}\ ]) = \begin{bmatrix} \Sigma_{rr} & \Sigma_{rC} \\ \Sigma_{Cr} & \Sigma_{CC} \end{bmatrix}$, $E([\ y_r^{'} \ \ y_C^{'}\ ]) = [\ \mu_r^{'} \ \ \mu_C^{'}\ ]$ $\Sigma_{rr.C} = \Sigma_{rr} - \Sigma_{rC}\Sigma_{CC}^{-1}\Sigma_{Cr}$, $\mu_{r.C} = \mu_r - \Sigma_{rC}\Sigma_{CC}^{-1}(y_C - \mu_C)$,

$$\Phi(\mathbf{1}c) = \int_{-\infty}^{c} ... \int_{-\infty}^{c} e^{-\frac{1}{2}(y_r-\mu_r)^{'}\Sigma_{rr}^{-1}(y_r-\mu_r)}/\{(2\pi)^{p/2}\,|\Sigma_{rr}||^{1/2}\}dy_j dy_S,$$

where $dy_S = dy_{\ell_1} ... dy_{\ell_S}$, and

$$\Phi_{r.C}(\mathbf{1}c) = \int_{-\infty}^{c} ... \int_{-\infty}^{c} e^{-\frac{1}{2}(y_r-\mu_{r.C})^{'}\Sigma_{rr.C}^{-1}(y_r-\mu_{r.C})}/\{(2\pi)^{p/2}\,|\Sigma_{rr.C}|^{1/2}\}dy_S$$

1. The case when the only data available within a day is censored data.

$$\begin{aligned} &E(y_j|y_j < c \,\&\, y_\ell < c \text{ for all} \ell \in S) \\ &= \textstyle\int_{-\infty}^{c} ... \int_{-\infty}^{c} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} y_j e^{-\frac{1}{2}(y-\mu)^{'}\Sigma^{-1}(y-\mu)}/\{(2\pi)^{p/2}\,|\Sigma|^{1/2}\} \\ &\quad dy_j dy_S/\Phi(\mathbf{1}c) \\ &= \textstyle\int_{-\infty}^{c} ... \int_{-\infty}^{c} y_j e^{-\frac{1}{2}(y_r-\mu_r)^{'}\Sigma_{rr}^{-1}(y_r-\mu_r)}/\{(2\pi)^{p/2}\,|\Sigma_{rr}|^{1/2}\} \\ &\quad dy_j dy_S/\Phi(\mathbf{1}c). \end{aligned}$$

2. When both censored and uncensored values are available

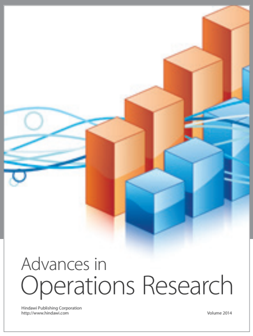
$$\begin{aligned} &E(y_j|y_j < c\,,\, y_\ell < c \text{ for all } \ell \in S \text{ and } y_C) = \\ &\textstyle\int_{-\infty}^{c} ... \int_{-\infty}^{c} y_j e^{-\frac{1}{2}(y_r-\mu_{r.C})^{'}\Sigma_{rr>c}^{-1}(y_r-\mu_{r.C})}/\{(2\pi)^{p/2}\,|\Sigma_{rr.C}|^{1/2}\} \\ &\quad dy_j dy_S/\Phi_{r.C}(\mathbf{1}c) \end{aligned}$$

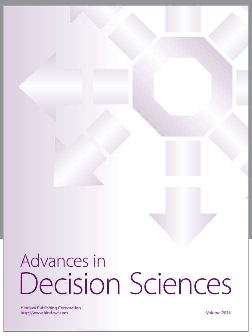
The above integrals can be solved by simulation. That is,

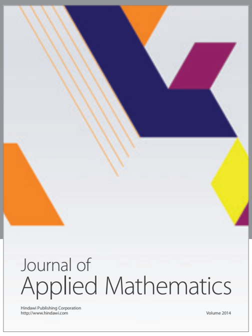
1. Simulated the response vector from a multivariate normal random variate with mean vector μ and covariance matrix Σ. Retain all values that are censored. Repeat this many times. Then find all simulations that have $y_j < c$ & $y_\ell < c$ for all $\ell \in S$, then take the average of all the y_j values in this group to give the required values.

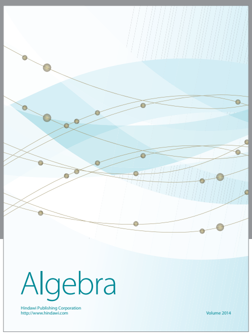
2. Simulated y_r conditional on y_C given they are jointly multivariate normal. Retain only observations with all values in y_r censored (less than c). Then calculate the average of these values these retained values to give the expected values of for each variable in y_r.

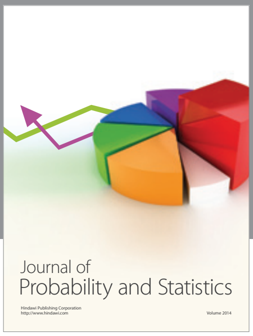
These approaches are computationally fairly expensive and the parameter estimation process is slow.


Advances in
Operations Research

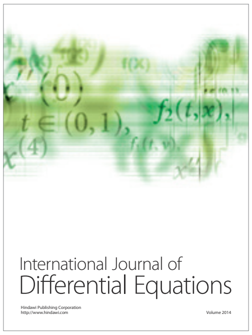

Advances in
Decision Sciences

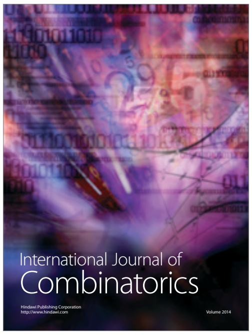

Journal of
Applied Mathematics

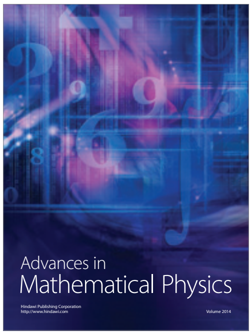

Algebra

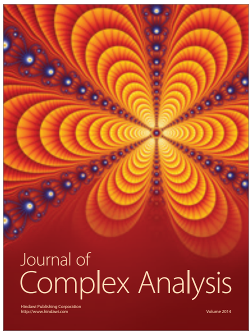

Journal of
Probability and Statistics

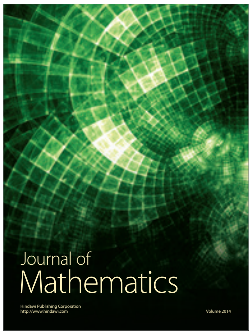
The Scientific
World Journal

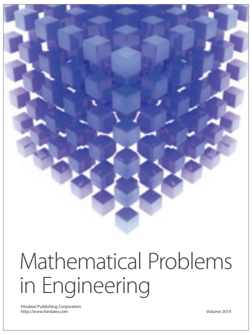

International Journal of
Differential Equations

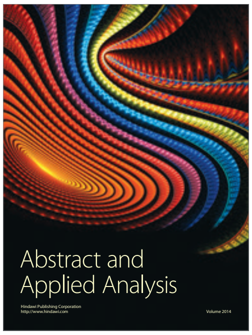

International Journal of
Combinatorics

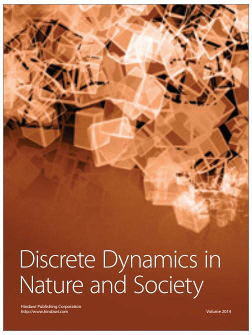

Advances in
Mathematical Physics

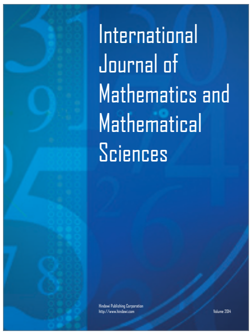

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

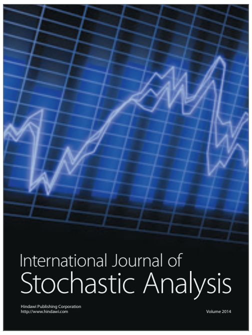

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization