Censored Regression Trend Analyses for Ambient Water Quality Data

Gordon K. Smyth[1], Melanie Cox[2], Andrew Moss[2]

[1] Centre for Statistics, University of Queensland, Brisbane, Qld 4072, Australia
[2] Waterways Scientific Services, Environmental Protection Agency, Queensland, Australia

Abstract: A method a trend analysis is described for environmental data which accommodates outliers, detection limits, seasonal effects, covariates and long term nonlinear trends.

Keywords: environmental modelling; censored regression; logistic distribution; regression splines; seasonal trends.

1 Introduction

Water is a very important asset. It supports natural environments, including diverse flora and fauna, and is essential to agriculture, industry and economic growth. It also has an important role in recreational activities and in contributing to overall quality of life. In Australia the coastal zone and associated river systems are subject to increasing levels of development and support approximately 75% of the population.
Management of water environments requires an understanding of the impacts on water quality and an understanding of the effectiveness of management actions. Monitoring programs to assess water quality typically aim to assess condition (whether or not water quality meets specified criteria) and trend (whether water quality is getting better or worse). In the state of Queensland in Australia, the Environment Protection Agency (EPA) collects information which is used in government reporting and setting license conditions for industry as well as in determining the effectiveness of environmental policy and management.
The Queensland EPA began a state-wide water quality monitoring program in 1992 and there is now, after nearly 10 years of the program, a strong need to assess what trends are apparent from the accumulated data. Trend analyses should incorporate seasonal patterns and other factors such as changes in rainfall which are likely to impact on water quality. More general questions include (i) assessment of regional trends, (ii) assessment of relationships between indicators, (iv) setting of water quality guidelines for impacted streams and (v) design of future sampling schemes.

Although water quality monitoring is an international concern (Barnet and O'Hagan, 1997; Ford et al, 1993; Skalski, 1990; Urquhart et al, 1998; Wetering and Groot, 1986), methods for assessing trends and for setting reference guidelines are surprisingly undeveloped. This article describes methods for statistical analysis of water quality indicators. A censored regression strategy is used to accommodate arbitrary detection limits for the indicator variables. A heavy-tailed response distribution is assumed to give a high degree of insensitivity to outliers. Harmonic terms are used to model period seasonal trends. Regression splines are used to model nonlinear long-term trends. The use of regression methodology allows covariates such as flow rate, temperature or tidal height to be incorporated into the model.
Independently of our study, Morton (1997) and Nathan et al (1999) have recently developed trend analyses for water quality variables which are similar in spirit to those given here. Our methods differ from theirs in terms of our explicit handling of the detection limits and in the strategy which we use to achieve robustness. Our methods are amongst other things more suited to estimating distributional quantiles which are used to set and assess water quality guidelines because we accommodate heavy-tailed responses rather than rejecting observations which are outside the normal range.

2 The EPA Monitoring Program

The Queensland Environmental Protection Agency began a large scale water quality monitoring program about 10 years ago, collecting monthly data on a range of water quality indicators at over 500 sites throughout the state of Queensland in Australia. The indicators fall into three broad groups. The concentration of chlorophyll-a (CHLA) is an indication of biomass in the water. The concentrations of organic nitrogen (ORGN), ammonia (AMMN), oxidised nitrogen (OXIDN), filterable reactive phosphorus (FRP) and total phosphorus (TP) indicate the available nutrients in the water. Measurements of the physical characteristics of the water include dissolved oxygen (DO), suspended solids (SS), turbidity (TURB), conductivity (COND), temperature (TEMP), acidity (PH) and Secchi depth (SECCHI). The Secchi depth is the depth to which a white disc can be lowered into the water before disappearing from sight, a measure of water clarity. Covariate information such as stream flow rate (FLOW) and tidal cycle information is also available from other sources. In this article we concentrate on data from the Logan River catchment area in South East Queensland and from the Herbert River catchment area in far north Queensland.

3 Initial Data Analysis Considerations

Transformations. CHLA, the nutrient concentrations, SS, TURB and SECCHI were analysed on the logarithmic scale in order to produce roughly symmetric distributions. TURB was offset by 1 unit and SECCHI depth was offset by 0.2m to avoid taking the logarithm of zero. The other variables were analysed on their original scales.

Detection Limits. CHLA and the nutrient indicators are subject to detection limits below which the exact concentration cannot be determined. Such observations are left censored at the detection limit. For several indicators the detection limits changed during the course of the monitoring program as more sophisticated testing procedures were put in place. SECCHI depth is both right and left censored, right censoring occurring when the SECCHI depth exceeds the depth of the stream.

It is been argued in the literature (Haas and Scheff, 1990) that left censored observations can be set equal to half the detection limit for the purposes of computing summary statistics. This strategy is however likely to be satisfactory only for variables which are approximately symmetric on their original scales and when the detection limit is not too far into the tail of the distribution. Neither of these conditions is met for the water quality indicators.

Outliers. Many of the water quality records exhibit a small percentage of outlier observations which are much larger or smaller than the general body of values. Even in the absence of distinct outliers, most of the water quality indicators exhibit variation which is more heavy-tailed than the normal distribution. A primary consideration for any data analysis method is that it be insensitive to outliers.

4 Approaches to Trend Detection

Rank-based methods. We consider three major approaches to detecting trends in the water quality indicators. The most popular method in the environmental literature for testing for trend is Kendall's seasonal trend test, which is a rank-based test for a monotonic trend (Hirsch et al, 1982; Helsel and Hirsch, 1992). This test has many advantages. It is (i) unaffected by whether the data is transformed or not, (ii) insensitive to outliers, (iii) insensitive to what value is substituted for observations below the detection limit and (iv) nearly as powerful as tests based on normality when the data actually is normal. From our point of view the major drawback of Kendall's seasonal trend test is that it does not allow for the possibility of non-monotonic trends and cannot be extended to allow for that possibility. Kendall's test is also technically invalid if the detection limit changes during the period of the study, as it does for the Queensland water monitoring program, as it is not then possible to resolve comparisons between censored

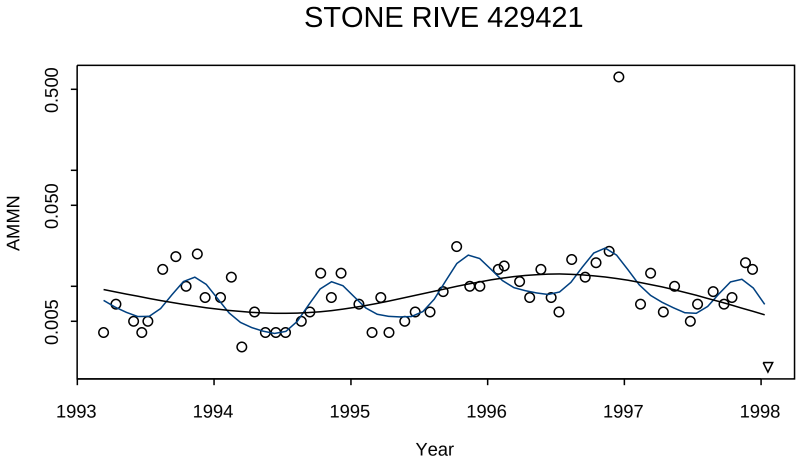

FIGURE 1. Log-ammonia concentration for the Stone River with maximum likelihood seasonal and long term trends.

observations at different detection limits. Other disadvantages are that the test does not naturally allow for the incorporation of covariate information, such as flow data, and that separate methods must be used to estimate the rate of trend, to extract seasonal components, and to estimate quantiles. If a trend is detected using Kendall's seasonal trend test, it is usual to estimate the trend slope and intercept using Theil-Sen's robust line (Helsel and Hirsch, 1992). The slope estimate is simply the median of the slopes through all possible pairs of points.

Figure 1 shows the log-ammonia concentrations in the Stone River just before it flows into the Herbert River. Notice that the last observation is censored and that there is a single very large outlier in December 1996, probably caused by a misreading. Kendall's season trend test gives a P-value of 0.040, providing evidence that ammonia levels have risen in the Stone River. The slope is estimated to be 0.108 on the logarithmic level, corresponding to an increase of 71% over five years. If the large outlier is removed the P-value increases to 0.051 and the slope decreases to 0.097.

Robust Regression. Robust regression is a modification of ordinary least squares regression which down-weights or eliminates observations with large residuals. A promising method is the high efficiency/high breakdown MM-estimation procedure developed by Yohai (1987). We used MM-regression to fit a general trend, given by a regression spline with 3 df, and a periodic seasonal component to the Stone River log-ammonia concentrations. The trend line successfully ignores the large outlier, and also the censored observation which is a less noticeable low outlier. The fitted curve does not indicate an overall increasing trend but is strongly suggestive that there has been non-monotonic variation in ammonia levels during the years of the study. One drawback of robust estimation is that it is relatively difficult to test hypotheses under the robust framework. In this case, tests based on the standard errors of the estimators suggest that neither the linear trend nor the nonlinear trend are significant. On the other hand, the robust F-test suggests that both the linear and the nonlinear trends are significant. The performance of hypothesis tests in the robust regression framework is not yet well understood.

Censored Regression. Censored regression is a maximum likelihood regression technique which explicitly allows for the fact that an exact value is not known for the censored observations. Censored regression is not ordinarily a robust technique. However it is possible to make the method relatively insensitive to outliers by choosing a response distribution which is heavier-tailed than the normal. Censored regression is implemented in S-Plus through the function censorReg (Meeker and Duke, 1981). We fit censored regressions using the logistic (or log-logistic) response distribution. The logistic distribution has a very similar shape to the normal distribution except in the tails of the distribution where it has much heavier tails.

Censored regression is the most powerful and the most flexible of the three methods we consider. It (i) will detect smaller trends with greater precision than the other methods, (ii) allows covariate information and nonlinear trends to be incorporated easily, (iii) allows hypotheses to be reliably tested and (iv) produces estimates for any desired quantile of the water quality indicators. Censored regression, as does robust regression, assumes that the water quality indicators have approximately symmetric variation on a suitably be transformed scale. The transformations suggested in Section 3 have been found satisfactory for this purpose. Neither robust nor censored regression require the observations to be equally spaced at monthly or any other intervals.

We fitted a censored regression with seasonal and nonlinear trends to the Stone River ammonia concentrations. We treat the concentrations as having a log-logistic distribution. Figure 1 shows the fitted trend with both seasonal and nonlinear components. It is apparent that the maximum likelihood fit is resistant to the outlier. The estimated trend is virtually identical to that from the robust regression.

Using the maximum likelihood approach it is possible to reliably test for the significance of the trends. Both the seasonal component ($P = 0.0006$) and the nonlinear trend ($P = 0.0004$) are highly significant. However the trend does not have a strong linear component ($P = 0.07$). We conclude that the ammonia concentration has varied over the five years of the study but there is no overall increasing trend. We feel that Kendall's trend test is somewhat mislead by the non-monotonic nature of the trend in this data sequence. The above P-values change slightly when the large outlier is removed. The changes do not affect the conclusions and are no larger than the corresponding change for Kendall's test.

5 Trend Analyses

Time Trends. Many of the indicator variables have substantial annual cycles. We estimate these periodic cycles using a cosine term with a period of one year plus a harmonic term with a 6 month period. The seasonal component consumes 4 df in the model fits. The overall trend was estimated by a linear trend in time. Long term variation from year to year was estimated by a regression spline on year with 3 df. This was found sufficient for the 10 year period considered in this study. The linear component is included in the regression spline, so the 3 df for long term trend can be decomposed into 1 df for linear trend and 2 df for nonlinear trend.
Covariate Information. Much of the seasonal variation and, sometimes, the long term trend as well can be explained by variation in stream flow rate, temperature and tide height. Flow rate is highly variable as it includes annual monsoon rains and occasional major storms. Some streams experience zero flow at certain times and at other times are in flood conditions. The effect of flow rate on indicator variables can be expected to be nonlinear. We have used a regression spline with 3 df on log-flow rate offset by 1 unit to incorporate possible nonlinear effects for flow rate. Temperature and tide height are adequately explained using linear trend terms.
Autocorrelation. We have found that the correlation between successive observations was low after suitable modelling of the trend. We have therefore treated the monthly measurements as independent. If observations were made at smaller intervals in time, then the methodology would need to be revised to incorporate a time series error structure.

6 Results

We give brief results for one site, the Logan River at a site 77km upstream. Table 1 gives the percent of each indicator which is explained by seasonal variation and by linear and nonlinear trends. The covariates temperature and flow rate were also included in the regression. The percents are computed using percent changes in the log-likelihood as the predictors were added sequentially into the regression. Figure 2 gives time series plots for four of the indicators, together with seasonal, linear and nonlinear trend lines where these are significant at 5%.
We find that all of the physical indicators are strongly affected by flow rate, as also are TP, ORGN and OXIDN. Variables such as SS and TURB increase with flow rate which PH decreases. CHLA, AMMN and FRP were little affected by flow rate. CHLA and several of the nutrient levels, ORGN, FRP and TP were found to increase with temperature. FRP and COND also had strong seasonal patterns unexplained by temperature and flow. Strong linear trends are apparent over the period for CHLA, FRP and COND and strong nonlinear trends are apparent for CHLA and COND.

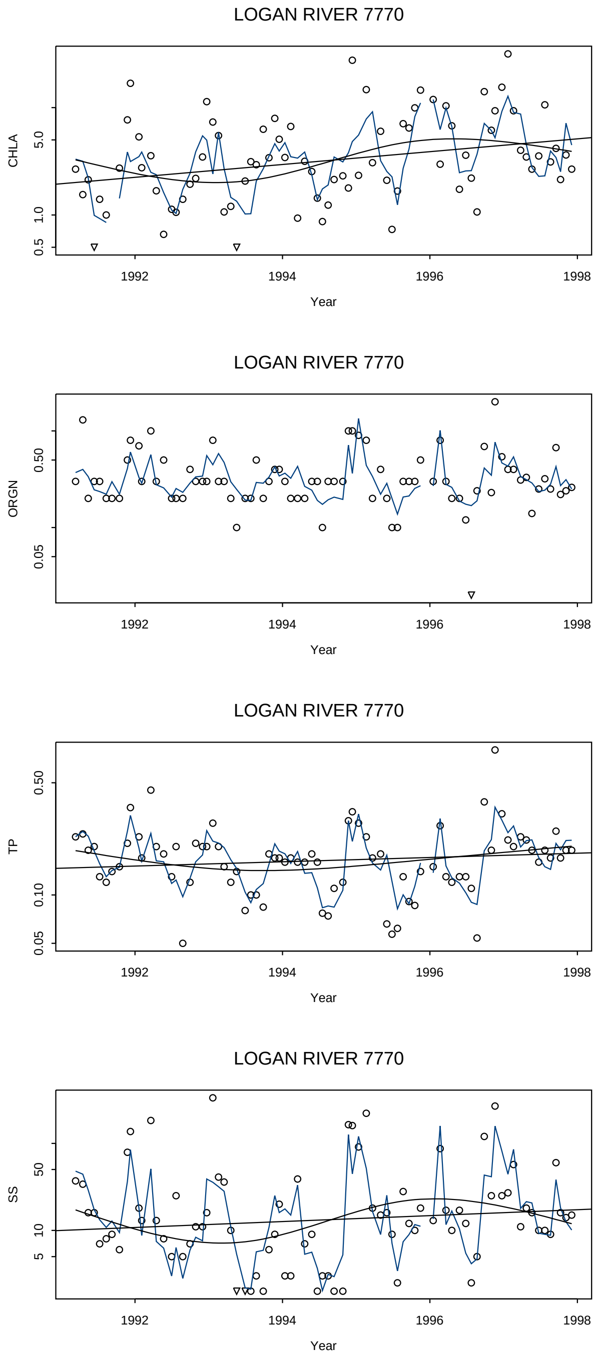

FIGURE 2. Logan River 77km upstream.

TABLE 1. Logan River 77km. Percent of variability of indicator predicted by various predictors.

Indicator	Temp	Flow	Seas	Linear	Nonlin
CHLA	28.6	4.6	6.8	7.8	7.2
ORGN	11.8	23.2	3.5	0.0	2.3
OXIDN	7.3	19.0	1.6	5.1	1.9
AMMN	2.3	1.1	9.8	2.5	3.6
FRP	13.8	3.5	13.7	12.9	2.0
TP	11.4	33.2	7.4	1.8	3.9
SS	8.2	44.1	2.1	2.0	5.2
PH	1.2	31.5	3.0	2.4	2.9
DO	2.8	30.5	4.0	0.2	3.1
COND	3.1	17.4	11.9	10.6	12.9
TURB	7.6	38.2	8.0	2.5	4.5

References

Barnet, V., and O'Hagan, A. (1997). *Setting Environmental Standards.* Pollution. Chapman and Hall, London.

Ford, J., Stoddard, J. L., and Powers, C. F. (1993). Perspectives on environmental monitoring: an introduction to the U.S. EPA Long-Term Monitoring Project. *Water, Air and Soil Pollution* **67**, 247–255.

Haas, C.N., and Scheff, P.A. (1990). Estimation of averages in truncated samples. *Environmental Science and Technology* **24**, 912–919.

Helsel, D.R., and Hirsch, R.M. (1992). *Statistical Methods in Water Resources.* Elsevier, Amsterdam.

Meeker, W. Q. and Duke, S. D. (1981). CENSOR - A user-oriented computer program for life data analysis. *Amer. Statist.* **35**, 112.

Morton, R. (1997). Semi-parametric models for trends in stream salinity. Report Number CMIS 97/71, CSIRO Mathematical and Information Sciences, Canberra, Australia.

Nathan, R.J., Nandakumar, N., and Smith, W.E. (1999). An the application of generalised additive models to the detection of trends in hydrologic time series data. *Water 99 Joint Congress Handbook and Proceedings*, Institute of Engineers, Australia, Barton ACT, pages 437–440.

Skalski, J. R. (1990). A design for long-term status and trends monitoring. *J. Environ. Manag.* **30**, 139–144.

Urquhart, N. S., Paulsen, S. G., and Larsen, D. P. (1998). Monitoring for policy-relevant regional trends over time. *Ecological Applications* **8**, 246–257.

Wetering, B. G. M. v. d., and Groot, S. (1986). Water quality monitoring in the state-managed waters of the Netherlands. *Water Research* **20**, 1045–1050.

Yohai, V.J. (1987). High breakdown-point and high efficiency estimates for regression. *Ann. Statist.* **15**, 642–665.