19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

An approach to model selection when predicting water quality in NSW using geospatial predictors

C. Badcock [a], D. Ryan [b], I. Growns [b], T. Mount [b]

[a] *Shimsco Consulting,* [b] *Department of Trade and Investment, Regional Infrastructure and Services, NSW Office of Water*
Email: caro_badcock@bigpond.com

Abstract: The NSW Office of Water, within the Department of Trade and Investment, Regional Infrastructure and Services is developing water quality guidelines for regions within New South Wales as part of the state's implementation of the National Water Quality Management Strategy (NWQMS). The new guidelines will be tailored more closely to specific catchments and regions than the current default guidelines in NSW, the *Australian Water quality Guidelines for Fresh and Marine Waters, National Water Quality Management Strategy* (ANZECC & ARMCANZ, 2000). They will also be used to inform various national and state natural resource management targets.

Water quality guidelines typically include reference values that indicate what the "best case" water quality values are for a region: these reference values are usually derived from reference sites, which have been purposefully chosen because they provide the best example of undisturbed conditions within a catchment. Reference sites were not available for this project, hence the Office of Water has used a predictive modeling approach to underpin the development of water quality guidelines by developing statistical estimates of reference condition. The predictive models have drawn on the last 10 years of water quality and in-stream flow records, and associated geospatial information from catchments across NSW. The project has concentrated on building predictive models for five common water quality variables: turbidity, electrical conductivity, water temperature, total nitrogen and total phosphorus.

Water quality is affected by both natural and anthropogenic factors. That is, rainfall and any landscape features not influenced by human behavior versus all land use types, distance to upstream dams and all variables related to vegetation type or cover. Flow variables were categorised separately because they are influenced by both natural and anthropogenic activities. It is possible that current targets for some sites may never be able to be met due to the impact of natural factors. One of the aims of this research is to identify if that is ever possible and, if so, under what conditions. This research will also assist in determining which regions will most benefit by targeted activities to reduce the impact of human behavior on waterways.

An earlier pilot study established that natural and discrete groupings could be formed based on different water quality characteristics alone and that sets of geospatial factors associated with a water quality monitoring station's drainage can be used to explain the water quality characteristics of that station. The current research has built on the pilot study, refining quality control procedures and increasing the scope and number of monitoring stations, whilst the water quality variables of interest remain as total nitrogen, turbidity, total phosphorus, electrical conductivity and water temperature. The range of geospatial variables, although fine-tuned, has still remained a significant number, viz. 72.

This paper will discuss the approach taken to reduce the possible number of geospatial predictors to a number acceptable for a robust prediction of a data series, how the time series of each water quality variable at each water quality monitoring station was summarised to allow investigation of the impact of the geospatial predictor variables, how these predictors were then used to build separate models to predict each of the water quality variables at each water quality monitoring station based on the subset of important geospatial predictors for the corresponding water quality variable and how the models were validated. The resulting predictive models and the estimation of undisturbed water quality values are not discussed in this paper, but will be addressed in future publications.

Keywords: *monitoring, geospatial predictors, model validation*

1. INTRODUCTION

The NSW Office of Water, within the Department of Trade and Investment, Regional Infrastructure and Services (DTIRIS), is developing new water quality guidelines for New South Wales (NSW) as part of the state's implementation of the National Water Quality Management Strategy (NWQMS). The new guidelines will be tailored more closely to specific catchments and regions than the current default guidelines in use within NSW, the *Australian Water quality Guidelines for Fresh and Marine Waters, National Water Quality Management Strategy* (ANZECC & ARMCANZ, 2000). They will also be used to inform various national and state natural resource management targets. .

Water quality guidelines usually incorporate guideline values that management agencies may aspire towards or try to maintain. These will often include values that represent undisturbed or "best case" conditions and many water management agencies have used data from reference sites for this purpose. Reference sites are typically located in areas with minimal disturbance such as catchment reserves and national parks, and were chosen *a priori* to serve as references. Provided that a good deal of care has been given to the association of reference sites with appropriate disturbed sites (based on similarity, e.g. the same catchment, altitude etc.), they provide a real world benchmark for guideline development. On the Australian east coast, both the Queensland and Victorian governments have used reference site data in the development of regional water quality guidelines (Qld DERM, 2009; EPA Victoria, 2003).

Although the NSW Office of Water has maintained water quality monitoring programs at hundreds of sites around NSW, it has not collected water quality data from designated reference sites. In the absence of reference site data, the authors have used predictive models as a foundation for estimating water quality values under undisturbed conditions, thereby providing statistical – rather than reference site - calculations of reference condition.

The modeling approach developed in this report was based around the assumption that water quality at a given site is largely determined by that site's catchment characteristics, referred to henceforth as geospatial predictors. Water quality at any given site will be due to a mix of natural and anthropogenic geospatial factors within its catchment. If this assumption holds, then predictive models can be used to estimate undisturbed water quality by 'resetting' anthropogenic factors to reflect undisturbed values. For example, modeled unregulated flow and modeled native vegetation coverage can be input in place of actual flow and native vegetation cover. Whilst this approach is clearly ambitious, it nevertheless provides an option for estimating reference condition from a large dataset, in the absence of designated reference sites.

The models drew on water quality and in-stream flow records taken from monitoring sites across NSW and the Murray Darling Basin, and associated geospatial information. The project focused on predictive models for five common water quality indicators: turbidity, electrical conductivity, water temperature, total nitrogen and total phosphorus and has used a bootstrapping approach extensively.

Bootstrapping was used to assist in estimating the number of observations required at each station in order to obtain a representative estimate of that station's water quality, in reducing the number of geospatial predictors to be included in the final model and to validate that model for each water quality indicator.

This report discusses the methods used to develop the predictive models, with particular emphasis on the issue of validating regression models with a high predictor : sample ratio. The outcomes of using the models to predict undisturbed values, and the use of that data in developing regional water quality guidelines, will be discussed in future papers.

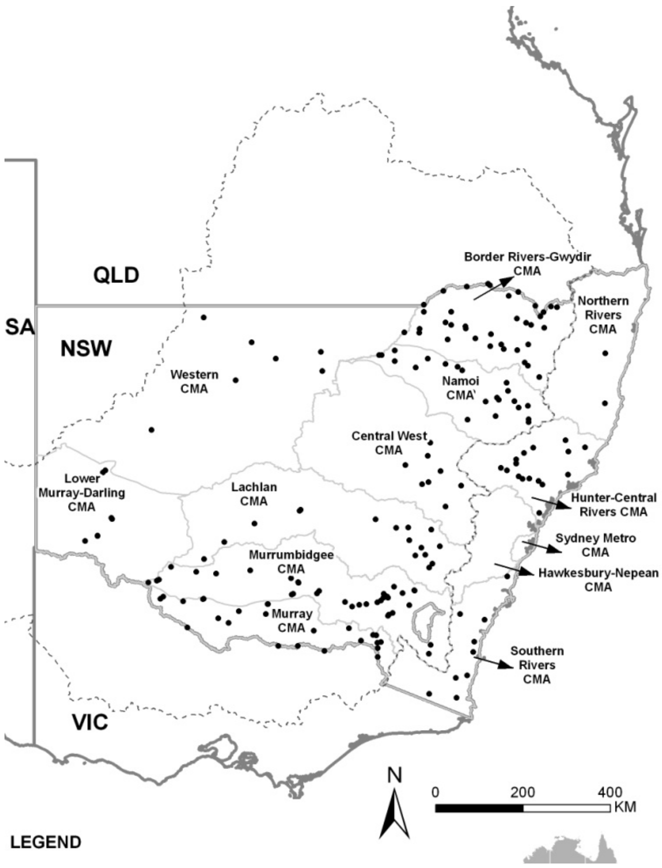

Figure 1: Water monitoring stations included in the analyses

2. DATA

2.1. Water Quality Indicator Variables

The NSW Office of Water has maintained water quality records for approximately 1,000 monitoring stations across NSW. For the purposes of this study, monitoring data were restricted to the period January 2000 to August 2010 as this was, firstly, a period during which a large number of stations were operational and secondly because it minimized the chance that significant changes in geospatial predictors (e.g. land use patterns) could have occurred between the earliest and most recent records.

Monitoring stations were available from thirteen Catchment Management Authorities, incorporating the Murray Darling Basin through to coastal catchments in northern and southern NSW. In order to be included in the study, each monitoring station was required to:

- be of fixed location;
- be visited repeatedly for water quality sampling from 1 January 2000;
- have complete data in terms of all candidate geospatial predictors;
- have data covering at least 5 years between January 2000 and August 2010;
- be located alongside or very close to a flow gauge with a record of daily flows that were time-matched to the water quality sampling record;
- meet the minimum number of required observations as estimated in Section 3.1. (Selected stations are shown in Figure 1.)

For regression analysis purposes the total nitrogen, total phosphorus, turbidity and electrical conductivity series from each water quality monitoring station were log transformed to the base 10 as is the usual approach for these variables in order to stabilize the variance (Snedecor and Cochran (1993). Water temperature was not log transformed. Medians were then calculated for each water quality variable for each station. The use of medians is common to water quality guidelines and hence, was the chosen summary measure. Additionally, the median does not differ with scale, meaning that a log-scale median value derived from the predictive models will back transform without distortion into a median on the native scale.

2.2. Water Quality Predictors (Geospatial Variables)

Water quality predictors were defined as any measurable characteristic of a monitoring station's drainage or location that was considered to have a direct conceptual link to one or more of the water quality variables. A total of 73 water quality predictors were investigated and consisted of variables related to:

- Station easting; northing; altitude; distance to upstream dam; stream order; distance to source; stream length;
- Drainage area; slope; altitude;
- Stream flow; rainfall (yearly, seasonal); groundwater flow (local, intermediate, regional);
- Land use (grazing, intensive animal production, cropping/horticulture, mining and quarrying, urban, wetlands, tree and shrub cover, native vegetation cover);
- Land use proximity effects (distance to centroids for intensive animal production, mining and quarrying, urban, effluent and sewage);
- Geological outcrop productivity index (scores of 1 to 5, negligible productivity to high productivity, % cover for each class). This index, developed by the NSW Office of Water, rates the weathering capacity of rock types according to mineralogy (T. Mount, NSW Office of Water, unpublished data).;
- Geological outcrop productivity proximity effects (distance to centroids);
- Electrical conductivity and turbidity

Predictors were only included if measurements were available for every station in the analysis: because some stations had catchments extending into adjacent states (Queensland and Victoria), some desirable GIS layers were incomplete or unavailable outside of the NSW state border.

2.3. Number of observations within the data series at each station

The number of observations varied considerably between monitoring stations and there was concern regarding the minimum number of observations required to obtain a representative median estimate. A bootstrapping approach was developed by the authors to estimate the minimum number of observations required.

The bootstrapping process utilized data from a subset of 68 stations that had been used in an earlier pilot study (Ryan et al. 2011). Stations with a minimum of 65 observations were included. Rather than just take the median of the original series as an estimate of the 'true median', the bootstrapped 99% confidence limits were generated to reflect the range for the 'true median' and acknowledge the possible imprecision of the estimate given the available series.

Table 1: **The number of samples required to obtain percentile estimates within the 99% confidence limits of the original series**

Water Quality Variable	10th and 90th Percentiles	15th and 85th Percentiles	20th and 80th Percentiles
Electrical Conductivity (EC)	60	40	30
Water Temperature	60	40	30
Turbidity	60	40	30
Total Nitrogen	50	40	30
Total Phosphorus	40	40	20

One hundred bootstrapped samples with replacement were generated for each monitoring station for sample sizes of 10, 20, 30 and so on up to 160, and the median of each sample was estimated. Various percentiles were calculated from each distribution of bootstrapped medians for each sample size for each water quality monitoring station and were compared with the bootstrapped 99% confidence limits for the median of the station's original series (Table 1). Looking at EC for example, for bootstrapped samples with 60 observations, 80% of the median estimates were within the 99% confidence limits of the original series (i.e. the 10th and 90th percentiles of the bootstrapped distribution of the medians were between the 99% confidence limits of the median for the original series).

The number of stations with at least 60 observations whilst meeting the selection criteria was too few to provide the desirable spatial coverage so the percentile range was relaxed to 70% (i.e. 15th and 85th percentiles) which lowered the minimum number of samples to 40 for all five water quality variables and increased the number of suitable stations to 178.

2.4. Predictive Models

Given that there were 178 stations meeting the selection criteria and 73 water quality predictors under consideration, any regression model containing the full set of predictors would be overfitted and have poor predictive power. Model reduction procedures were therefore necessary and the authors have largely adhered to the approaches advocated by Harrell (2010), which advocate a maximum ratio of predictors to observations of 1:10, based on the findings of Harrell et al. (1984; 1985).

Pre-model reduction procedures included: literature/knowledge reviews to eliminate weak conceptual links; removal of predictors with narrow distributions; identification of redundant and uninformative predictors. Statistical procedures were minimized in keeping with the general principle that such investigation require special adjustments must be made to P-values, standard errors, test statistics and confidence intervals in order for these statistics to have the correct interpretation (Harrell 2010). For the same reason, decision trees (Breiman et al.1984) were not used to explore the data as these result in overfitting in three directions: searching for the best predictors, for best splits and searching multiple times.

Predictor reduction

Once conceptual links had been reviewed, principal components analysis was used to further compress the predictor set along thematic associations. The five flow variables were combined into a Flow PCA group, 25 rainfall variables combined into a Rainfall PCA group and 12 catchment-scale variables combined into a Catchment PCA group.

PC axes were retained as predictors if the eigen value was greater than 1. Two flow, three rainfall and three catchment PCs were retained for the model reduction process, with well over 90% of the total variance preserved in each case. This process reduced 42 of the original predictors down to eight principle components.

The candidate predictor list still exceeded the 10:1 rule (31 for electrical conductivity, 35 for turbidity, 19 for water temperature, 33 for total nitrogen and total phosphorus), hence model reduction was also required.

Model reduction

The reliability of commonly used model reduction techniques such as forward, backward and stepwise selection has been questioned because there can be more than one 'best model', R^2 values tend to be biased high, significance tests have less than the assumed degrees of freedom, error in regression coefficients is biased low and collinear predictions can be arbitrarily interchanged (Harrell, 2010). Sauerbrei and Schumacher (1992) suggested a bootstrapped approach as an aide in selecting predictors for the final reduced model. The principle is that the frequency of predictor selection across all of the bootstrapped models is an indication of importance: important predictors will appear very frequently and unimportant predictors infrequently. In this way, the final predictor list is selected quite literally by weight of numbers. It is important, however, to note that the choice of a cut-off frequency is arbitrary (Harrell 2010).

In this study a similar bootstrapping process to Sauerbrei and Schumacher (1992) was applied, along with a novel selection procedure for choosing the final model predictors. The selection procedure was designed to address the arbitrary nature of cut-off points by qualifying it against the bootstrapping results and finally against the number of observations. The procedure can be summarised in the following steps:

1. Bootstrap the medians for each station 1,000 times and obtain the median for each bootstrapped sample for each station
2. Estimate the reduced model for each the water quality indicator for each bootstrapped sample set using backwards regression. The optimum model was determined when the Schwarz Bayesian Information Criterion (SBC) achieved its lowest value (Schwarz, 1978).
3. Calculate summary statistics for all candidate predictors. That is, calculate the frequency of every predictor and the number of predictors in the 1,000 bootstrapped reduced models and then estimate the mean and standard deviation of predictors for the bootstrapped samples.

Once the set of bootstrapped reduced models was created, the list of predictors for a final model was selected by sequentially applying the following three rules:

Rule 1: Identify predictors that occurred in greater than 50% of the bootstrapped samples (based on the logic that important predictors should be present in the majority of bootstrapped samples).

Rule 2: Adjust the number of predictors so that it is representative of the mean number of predictors in the bootstrapped models. Given that the number of predictors in each bootstrapped model was selected according to its optimal SBC value, it was assumed that the average number of predictors from the 1,000 bootstrapped models would indicate the optimum number of predictors that should be included in a final model.

Table 2: Frequency of predictor occurrences in bootstrapped backwards selection regression models for Total Nitrogen (truncated list not showing all 33 predictors). The highlighted predictors were retained in the final model

Predictor	No. Samples	% samples
Distance to Centroid, OPI Class 5	871	87.1
Distance to Upstream Dam	812	81.2
Distance to Centroid, Intensive Animal Prod.	799	79.9
Catchment PC 2	780	78
% Cover Intermediate Groundwater Systems	720	72
% Cover Trees and Shrubs	700	70
% Cover Regional Groundwater Systems	682	68.2
Distance to Centroid, urban	659	65.9
% Cover Local Groundwater Systems	616	61.6
% Cover OPI Class 2	567	56.7
% Cover OPI Class 3	555	55.5
Catchment PC 3	551	55.1
% Cover, Wetlands	534	53.4
% Cover OPI Class 5	532	53.2
Catchment PC 1	494	49.4
% Cover OPI Class 4	459	45.9

The number of predictors selected by Rule 1 was compared to the mean number of predictors for the bootstrapped set: if the total selected by Rule 1 was within one standard deviation of the mean, then it was retained. If it differed by more than one SD, then predictors were added or removed *irrespective of whether they occurred in more than 50% of samples*. For example, if 13 predictors occurred in more than 50% of the models, then Rule 1 would return a value of 13. If, however, the average number of samples across the bootstrapped models was 10 with a standard deviation of 2, then 13 predictors would exceed the Rule 2 limit. One predictor would need to be removed.

Rule 3: The ratio of predictors to samples could not exceed 1:10. This rule ensured that the final arbiter in predictor selection was the strength of the data itself. If the ratio exceeded 1:10 after Rules 1 and 2 were applied, then predictors were removed in order of frequency until the ratio was equal to or less than 1:10. For example, after applying Rules 1 and 2, a reduced model may have 12 predictors and 150 samples. The ratio in this case is 1:12.5, which is lower than the 1:10 rule. Therefore, a final model with 12 predictors would not violate Rule 3.

Table 2 gives an example for the total nitrogen model. 33 different predictors were selected in at least one bootstrapped model, the average number of predictors was 16 and SD = 4. A total of 14 predictors were selected in over 50% of bootstrapped models, which was within one standard deviation of the mean. The total nitrogen model comprised 112 stations, producing a predictors/samples ratio of 1:8. This violated Rule 3 of the selection procedure so the last three predictors were omitted and the final model retained the first 11 predictors only. In effect, the final total nitrogen model has been penalized by its sample size.

All bootstrapping and regression procedures were performed with the statistical software SAS Version 9.2 (2008).

Model Validation

The model was validated using a bootstrapped technique based on Efron (1983, 1986), Efron and Gong (1983), and Efron and Tibshirani (1986) that used bootstrapped sample sets to test the predictive accuracy of each regression model. Each bootstrapped data set was fitted to the final model (as determined earlier) to generate a R^2 estimate, and repeated 1,000 times to produce an average bootstrap R^2. The difference between the average bootstrapped R^2 and the R^2 from the original model was used to estimate the model's predictive bias, or 'optimism' (Harrell 2010). If the mean bootstrapped R^2 was less than the original model's R^2, then this implied overfitting and an overly optimistic estimate of predictive accuracy. The difference between the two R^2 estimates was subtracted from the original model R^2 to provide a final model that has been 'bias-corrected' or 'overfitting-corrected' (Harrell 2010). The same procedure was used to estimate the optimism of each of the regression coefficients. (Results not presented here.)

Table 3: Summary of model R^2 and mean bootstrapped R^2 from the reduced models

Variable	Model R^2	Bootstrapped R^2
Electrical Conductivity	0.70	0.67
Turbidity	0.81	0.79
Water Temperature	0.54	0.52
Total Nitrogen	0.57	0.51
Total Phosphorus	0.72	0.68

The bootstrapped validation procedure was as follows:

Step 1: Bootstrap the original set of medians (with replacement) 1,000 times, using a different random number seed value each time.

Step 2: Fit each bootstrapped sample to the final regression model and calculate mean statistics. Calculate average bootstrapped coefficients and standard errors, so that bias estimates can be calculated for each predictor. The average bootstrapped model R^2 was also calculated with the same intention of calculating a bias estimate after comparison with the original model.

Step 3: Use bias estimates to 'bias-correct' the original model parameters.

Results for the model R^2 are summarised in Table 3 and show that in all cases, the difference between the model R^2 and the mean bootstrapped R^2 was very small.

3. CONCLUSIONS AND RECOMMENDATIONS

The bootstrapping approach was used on three occasions as part of this project.

First it was used to assess the minimum number of observations required at each water quality monitoring station to obtain a representative estimate of the median result for that station. The spread of geographical regions for stations in this analysis was crucial to ensure that the resulting sample size was as applicable as possible to stations other than those used to estimate the sample size. In particular it should be recognized that this process only included stations in the Murray Darling Basin (Ryan et al. 2011) and the subsequent predictive models were built including stations on the eastern side of the Great Dividing Range. The additional stations were not expected to require more observations to obtain a representative median than those used in this sample size estimation. Bootstrapping allowed the development of a distribution of medians for each water quality variable for each sample size of interest; from which various percentiles were able to be extracted and compared with the 99% bootstrapped confidence limits for the median of the original set of medians. Although a compromise was reached based on the 15th and 85th percentiles of the bootstrapped medians being within the 99% confidence limits of the median for the original data, i.e. 40 observations was required, the process of obtaining the sample size was regarded as successful. Since the publication of these results, comments have been received as to estimating sample sizes assuming the median of the original series is the 'true' median and obtaining bootstrapped medians for smaller sample sizes as done here. Then, instead of comparing percentile intervals with the 99% confidence interval of the original series assess the coverage probabilities by comparing the 'true' median with the bootstrapped confidence intervals from the range of sample sizes. This approach has yet to be explored.

In the model reduction process, bootstrapping across the station medians with replacement and repeating the backwards selection procedure potentially allowed models to be selected based on multiple observations with similar geographical predictors. For example, a bootstrapped sample may include a small number of stations many times and as a result the reduced model will be heavily dependent on the geospatial predictors for that small group of stations. If this pattern of geospatial data across the samples is rare this reduced model will not occur very often. Hence, intuitively the model's ability to predict water quality for a new station should be improved. After obtaining the results of the backwards selection models the three step process developed by the authors to determine the final model proved objective and easy to apply. All final reduced models selected using this process contained geospatial predictors deemed to be appropriate for the water quality variable of interest.

Lastly, bootstrapping was used to validate the final reduced model for each water quality variable. The models did not appear to over-fit the data or overly bias estimates according to Harrell's optimism criteria for the estimated R^2 or regression coefficients.

The authors found the process described above provided them with a framework to obtain the final predictive models. However, as with any project care must be taken in understanding the data on hand, the appropriateness and validity of collapsing response data into summary statistics, as done in this project with the medians, and the role of the predictors.

The interpretation of the reduced models and how the results will feed into water quality guideline development will be provided in future publications as this work is ongoing.

REFERENCES

Australian and New Zealand Environment and Conservation Council (ANZECC) and Agriculture and Resource Management Council of Australia and New Zealand (ARMCANZ), (2000). *Australian Water Quality Guidelines for Fresh and Marine Waters, National Water Quality Management Strategy.* Australian and New Zealand Environment and Conservation Council, Canberra.

Breiman, L., Friedman, J.H, Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees, Monterey, CA: Wadsworth & Brooks/Cole

EPA Victoria (2003). *Water Quality Objectives for Rivers and Streams – Ecosystem Protection.* Environmental Protection Agency, Victorian State Government

Harrell, Jr. F. E., Lee, K. L., Califf, R. M., Pryor, D.B. and Rosati, R. A. (1984). Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine,* 3: 143-152

Harrell, Jr. F. E., Lee, K. L., Matchar, D. B. and Reichert, T.A. (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports,* 69: 1071-1077

Harrell, Jr., F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression and survival analysis.* Springer-Verlag, New York, USA

Queensland Department of Environment and Resource Management (2009). *Queensland Water Quality Guidelines, Version 3*. ISBN 978-0-9806986-0-2.Quinn,G.& Keough, M. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press.

Ryan, D.A., Badcock, C.A, Doyle, G, Muschal, M. Mount, T. (2011). *Water Quality Guidelines for Regions within New South Wales: Preliminary Research and Development of Statistical Approaches.* Publication Number EEP 2010-11.068. NSW Office of Water

SAS Institute Inc, (2008). SAS/STAT® 9.2 *User's Guide*. Cary, NC: SAS Institute Inc.

Sauerbrei, W. and M. Schumacher (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11: 2093-2109,

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*. 6:461-464

Snedecor G. W. and Cochran W. G., (1993). Statistical Methods 8th Ed. Iowa State University Press, Ames Iowa