Development of XM Statistical Tool for Air Quality Forecasting

Andrew Teakles[1], Sean Perry[2]
1 - Vancouver, British Columbia, Environment Canada
2 - Fredericton, New Brunswick, Environment Canada

Abstract

The XM tool is an envisioned extension of the current Updateable Model Output Statistics (UMOS) framework used to improve post-processing of air quality forecasts. During the course of the project, various statistical kernels will be compared against the current multivariate linear regression technique of UMOS. Their ability to forecast hourly concentrations of NO2, O3, and PM2.5 will be evaluated. The XM tool project will also examine the optimal predictor selection methodology. By the summer of 2011, the most successful method(s) will be transferred to an operational setting and will provide guidance to air quality forecasters via the XSCRIBE matrices.

Introduction

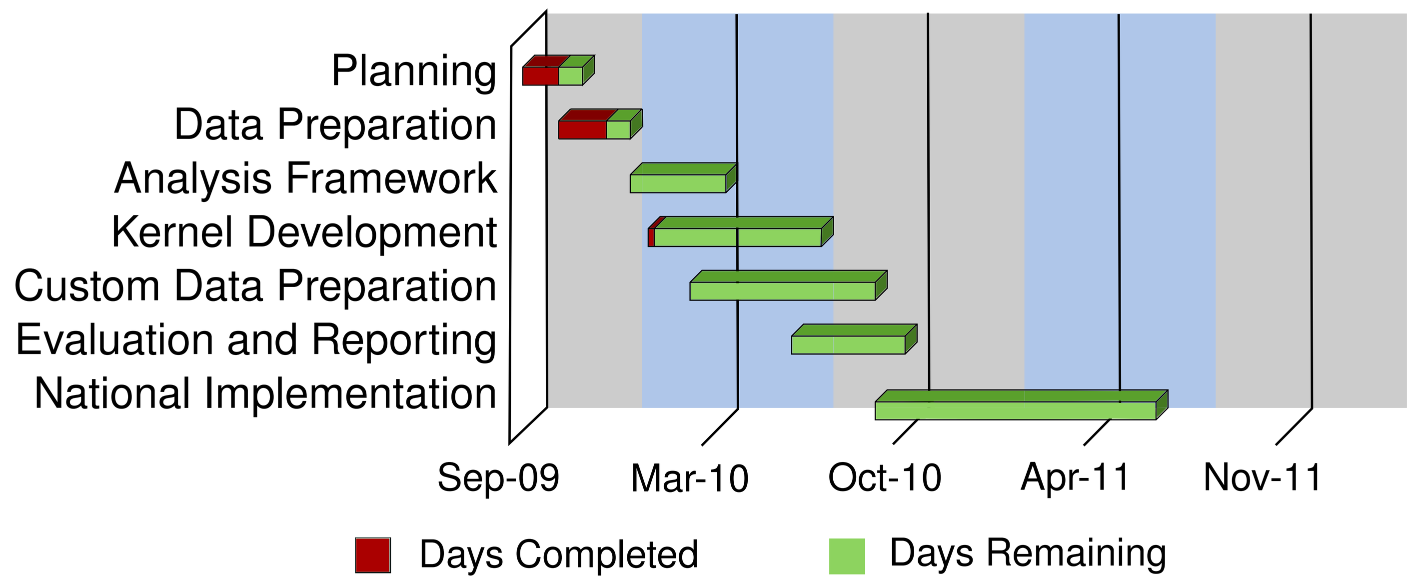
The Air Quality Health Index (AQHI) is the world's first health risk based air quality index. To produce the AQHI, a Canadian epidemiologically study was completed where mortality is related to statistical significant pollutants. The AQHI is calculated based on the relative risk of the combination of NO2, O3, and PM2.5. The XM tool project was initiated to support the Canada's National AQHI Forecast Program. The project aims to optimize the hourly forecast of individual AQHI pollutants for both routine cases and air quality "events".

Objective

The main objective of the XM tool project is to evaluate and compare different statistic methodologies within the Canadian Updateable Model Output Statistics (UMOS) framework. As well as general model performance, evaluation will be based on the following criteria:

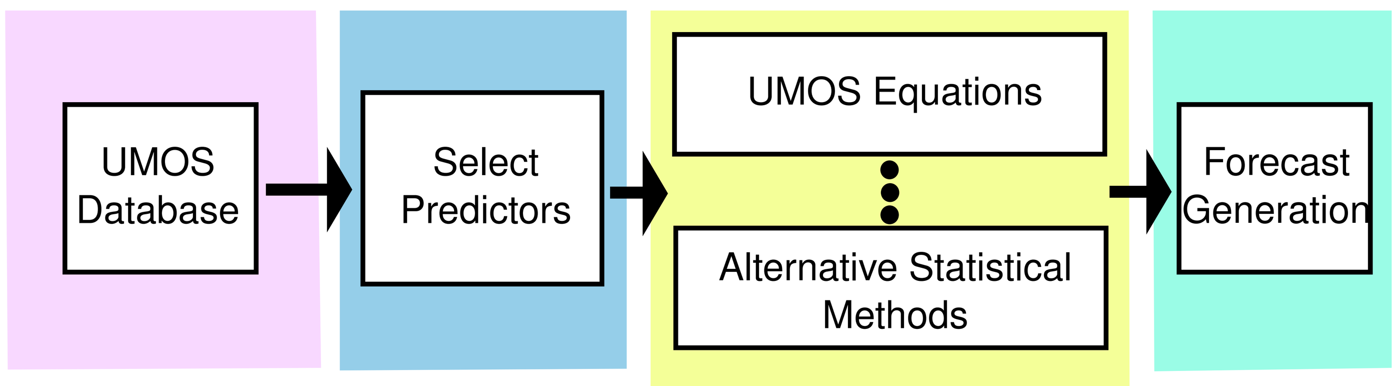
1) Improve guidance for extreme events
2) Improve the guidance for the start and end of air quality episodes
3) Improve the model's first guess (i.e.: by removing bias)

Timeline



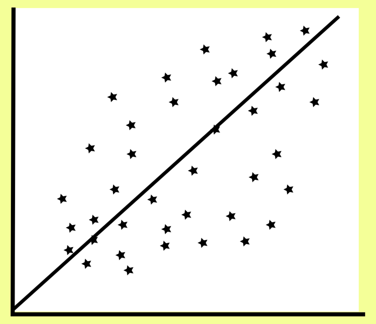
Methodology

The Canadian UMOS currently relies on multivariate linear regression technique to produce statically post-processed forecast guidance. Using the current UMOS framework it is possible to evaluate different statistical methods. These statistical methods, described below, may better suit the objectives of this project. Dependent on the statistical method applied, different predictor selection methodologies will be evaluated.

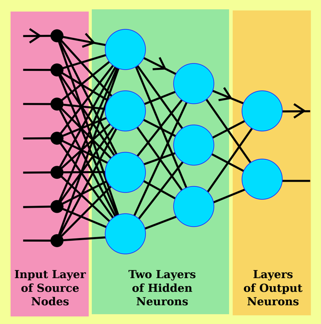


UMOS

An established national methodology for statistical post-processing based on the linear relationships (MLR) between predictors and predictand. UMOS is an effective method for removing model bias, however may develop an under-fit model due to the presumed linear response to predictors. UMOS has the advantage of having an established methodology to transition between model designs and varying predictand response. The effectiveness of the methodology for predictors varies substantially between model designs [O3, NO2, PM2.5] and changing emission sources.



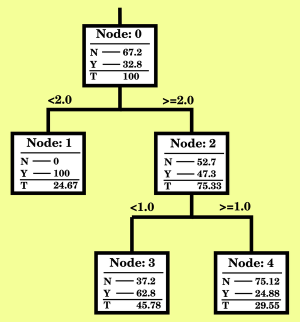
Neural Network



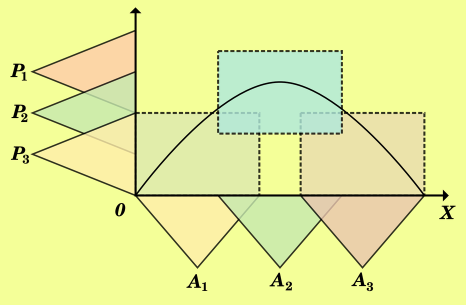
A neural network is a non-linear, black-box model that is very efficient at finding the complex relationships between predictors and predictands. A neural network consists of processing nodes, called neurons, connected in a network architecture. Model training is typically accomplished by iteratively adjusting each neuron's synaptic weights and bias values to minimize overall model error. Neural networks are commonly used for either pattern recognition or function approximation. A ensemble neural network is a collection of different neural networks whose outputs are averaged to generate the final prediction. Ensemble averaging typically improves model performance by stabilizing model predictions due to data randomness and the presence of local minima in the error surface.

Classification and Regression Tree

Robust classification technique that has the potential to improve the forecast for extreme events. However, classification methods are subject to misclassification error due to discreet cutoff. The model may be overfit if the Regression Tree is not trimmed sufficiently. Fuzzy logic or ensemble version of CART may improve the reliability of this methodology. One additional benefit of this technique is that typically each branch of the regression tree develops surrogate rules for splitting.



Fuzzy Logic



Fuzzy logic develops a set of if-then rules linking parameters to a defined set through the concept of membership. Unlike a classical logic scheme, the association of a parameter in a set is not limited to True or False statements (1 or 0 values). The fuzzy logic analogy to the CART methodology would be to allow for the final CART splits to overlap. One of the strengths of fuzzy logic is that it reduces the errors associated with the potential for misclassification in CART.

UMOS Database

UMOS-AQ predictors will be used as a primary source of data. The UMOS-AQ database consists of 84 predictors extracted from the GEM-MACH15 GRIB data. The predictors include both meteorological and air quality forecast values in the lowest layers of the model.
Antecedent values (last 24 hours) are available. These include both antecedent meteorological data and antecedent air quality data. Custom datasets will test the adaptability of the kernels to a model change, the computational efficiency, and how quickly the technique learns with new data.

Select Predictors

A key goal of this project is to assess different predictor selection methodologies. Currently UMOS uses a multiple linear regression (MLR) technique to select predictors. This method linearly compares the statistical significance of each variable in the model. An alternative to MLR is a General Additive Model (GAM). The GAM approach examines the non-linear interactions between predictors. The GAM can be achieved though the linear summation of non-linear terms which inherently resolves some of the complex interaction between predictors. In this way, it is better suited as a predictor selection methodology in non-linear models. A third predictor selection approach is integrated innately in some model types. For example, the CART method determines the statistically significant predictors as part of its sorting criteria.

Forecast Generation

The best statistical methodology may be selected by site, pollutant or seasonality. Once the ideal statistical technique has been chosen and integrated into UMOS framework, the post-processed pollutant values will be distributed via the XSCRIBE Matrices. This allows for seemless integration with Environment Canada's current data systems.

Conclusion

An exciting new project that should establish improved AQHI forecasts across Canada, the XM tool project will also provide findings and a development framework that will assist future endeavours in statistical air quality forecasting.

Acknowledgements

The following Environment Canada staff have already contributed to the future success of the XM Tool Project; Stavros Antonopoulos, Veronique Bouchet, Qian Li, Rachel Mintz, Sean Perry, Jacques Rousseau, Andrew Teakles, and Aiming Wu.