Artificial Intelligence
Emerging Approaches for Environmental Data Analysis

The dynamic and complex nature of the environment presents environmental professionals with significant challenges when modeling an environmental system. It is imperative that decision-makers are provided with a careful synthesis of the monitored/colllected environmental datasets that includes proper quantification of the relationships between the associated independent and dependent variables. Traditional statistical modeling techniques such as regression analysis have limited application when modeling a multivariate environmental system due to the discrete nature of the relationships between environmental variables. This limitation can be overcome by adopting the use of artificial intelligence.
The need to develop innovative methods to solve complex multivariate environmental problems, coupled with the rapid advancements made in the use of computer technology during the early 1990s, guided the environmental research and scientific communities to experiment with the use of artificial intelligence (AI) methods in examining a wide range of environmental problems. AI is a branch of computer science that provides research and professional communities with a variety of robust computational approaches that exhibit the characteristics associated with intelligence in human behavior (i.e., understanding, learning, and reasoning to solve problems and make decisions). The number of environmental studies using AI methods has significantly increased in the past decade compared to those in the 1990s.

This article provides an overview of the current status in the use of AI methods when analyzing multivariate environmental data. Furthermore, this article emphasizes the need to use AI methods in future studies when analyzing multivariate environmental data.

AI Methods for Environmental Data Analysis
The following summarizes the role of AI methods in the three fundamental components of an environmental data analysis: database development, identification of influencing variables and patterns, and modeling.

Environmental Database Development
The first stage in any multivariate environmental data analysis is to monitor and collect the associated independent and dependent variables. Ideally, one would prefer to have a complete database with no missing values. However, in a real-world scenario there are always situations with missing variables due to instrumentation problems during field tests. Missing values may create an uncertainty with regard to statistical inferences that may or may not have an impact on the management decision-making process.

Environmental professionals adopted the use of decision trees (classification and regression trees) and artificial neural networks (ANNs) in cases where the application of conventional interpolation techniques is limited for multivariate environmental data analysis.

A decision tree is a powerful non-parametric modeling approach that provides a flow-chart representation of the recursive partition(s) of the primary environmental dataset (parent node) into smaller datasets (child/end node) to improve the fit as best as possible.

ANNs are computational models that perform based on the operation of biological neural networks and gain knowledge through the “learning process.” ANNs comprise of highly interconnected processing elements (neurons) organized into three layers, namely, the input layer, the hidden layer, and the output layer. Each layer consists of a number of neurons that receive one or more inputs and produce an output with the knowledge being stored in the interneuron connected synaptic weights.

The use of decision trees and ANNs in imputing the values of missing variables is well documented in the literature. ANNs are computationally more time consuming and follow a black box approach, while the decision trees are less time consuming and provide a flow-chart visualization of the rules for future prediction.

Key Factors and Patterns Affecting the Environmental Process
The second stage in multivariate environmental data analysis is to identify the key factors and patterns affecting the environmental process. Determination of the optimal set of factors affecting an environmental process can help reduce the computational burden and redesign the monitoring studies that reduce monitoring costs.

Several studies adopted the use of regression trees and ANNs in determining the optimal set of factors affecting an environmental process. One study demonstrated the methodology to identify the influential environmental variables using regression tree analysis, which was further extended to rank the influential variables using the analysis of variance as a complimentary sensitivity analysis to the developed regression trees. This approach proved to be valuable, considering that statistically significant relationships were established.

Regression trees also proved to be useful in detecting the outliers. The selection of algorithm for ANN model inputs to prevent model under-specification/over-specification must be based on the criteria that it is model-free (statistical non-linear measures of dependence are used to determine the strength of relationships between inputs and outputs), accounts for redundancy (optimize input...
The environmental database is first divided into training, validation, and test subsets. ANN model development is dependent on the training and validation processes. The accuracy of ANN performance is obtained by comparing the model predictions with actual observations from the test subset. Determining model accuracy based on a single test subset is commonly referred as the holdout validation and is widely used with environmental data. However, it is essential to adopt the use of multi-fold (k-fold) cross validation (environmental database is divided into “k” subsets and model accuracy is averaged over each subset being used as a test case) to ensure that the ANNs performance is representative of the algorithms adopted and not due to chance.

Figure 1 illustrates a flowchart of the general methodology used in developing ANNs. Incorporation of the key environmental variables identified from regression tree analysis as inputs in the development of evolutionary neural networks (BPN+GA) provided superior results in comparison with results obtained from individual ANNs.7,8
Summary

AI methods have provided environmental professionals with ample capabilities in solving complex nonlinear multivariate environmental data systems that could not be analyzed using traditional methods. The use of decision trees and ANNs have proved to be beneficial for identifying key environmental variables and patterns. Based on the observations made in reference to the literature cited in this article and the level of user-friendliness, it is suggested that beginners adopt the use of decision trees when modeling complex multivariate environmental systems, while advanced users could focus on the simultaneous use of decision trees and ANNs.

While computer professionals may be inhibited in their research due to the lack of comprehensive diversified databases, environmental professionals and graduate students could thoroughly examine the performance of different AI methods with the diverse sets of available databases. More research is needed to refine the current AI applications. It is hoped that this article will encourage environmental professionals to aggressively pursue the application of AI methods when analyzing their respective multivariate environmental data.

Acknowledgment:
Some of the conclusions in this article were made by the authors with experience of working on an alternate fuel grant funded by the U.S. Department of Transportation to the Intermodal Transportation Institute of The University of Toledo. The views expressed in this article are those of the authors only.

References
4. Kadiyala, A.; Kumar, A. An examination of the sensitivity of sulfur dioxide, nitric oxide, and nitrogen dioxide concentrations to the important factors affecting air quality inside a public transportation bus; Atmosphere 2012, 3, 266-287.

Contact Alison Lizzi for more information on Sponsorship or Exhibit Opportunities, please call 412-904-6003 or email alizzi@awma.org.