Predicting nitrates in drinking water

Jared Kraczek
Mentored by Richard Latham

SMA Science and Mathematics Academy

Booz | Allen | Hamilton

Introduction

Drinking water is essential to human life and clean drinking water is essential for human health. Even small concentrations of contaminates can have negative health consequences (World Health Organization, 2011). Ensuring water quality requires widespread time consuming and expensive testing. The goal of this project was to create a predictive model that could be used to better allocate human resources to testing high-risk areas, instead of having to test all areas. Data mining, the process of extracting useful information from large quantities of data, could be used to improve the process of monitoring drinking water. The data mining process is defined by three main steps: data collection, data preprocessing, and data analysis (Aggarwal, 2015). The purpose of this project was to predict nitrate levels in drinking water for all counties in Maryland based on data related to natural features of the testing locations, historical environmental data, and data related to current human activities around where the water was tested.

Materials and Methods

Model building and data handling was done using the R programming language in the RStudio integrated development environment (IDE). Data was sourced, collected, and preprocessed for analysis. During this process, the focus was on finding complete data sets that would contain sufficient data for future analysis. As much of the open source data available was incomplete, this proved to be a more difficult and time-consuming process then was originally anticipated. As a result the original scope of the project, looking at nitrate levels in the state of Maryland was pared down to the county scale for data collection and analysis. Data was sourced from a variety of governmental groups and organizations. Table 1 shows the sources of data, type of data collected, and how the data was used in the predictive model.

The Maryland agricultural census data was already organized by county, requiring minimal preprocessing. In the Environmental Protection Agency (EPA) water contamination data, each contaminate was linked to a Public Water System Identification Number (PWSID). During the data preprocessing phase, the PWSID had to be tied to a location in Maryland for county level analysis. The waste handling facilities data was already sorted by county and had location attached. The preprocessed data from Table 1 was used to create a linear model (Figure 1) to predict nitrate values in drinking water by county.

Materials and Methods

Source Name	Data Provided	Model Parameter(s)
Six-Year Review of Drinking Water Standards	Water contamination	Concentration (mg/L) for nitrate in water (N)
US Census	Population	Median income of county residents (x_1), number of commuting workers (x_2), population density (x_3)
Maryland Agricultural Census	Fertilizer and chemical use	Density of applied fertilizer (x_4), density of applied herbicide (x_5)
Homeland Infrastructure-Foundation Level Data	Waste disposal site locations	Number of waste disposal sites per county (x_6)
Superfund Sites	Hazardous waste sites	Number of Superfund sites per county (x_7)
EPA Sites in Maryland	Land and water area of counties	Area of water (x_8), land area used to calculate herbicide, fertilizer, and population density

Table 1: Summary of data sources used to create the predictive model.

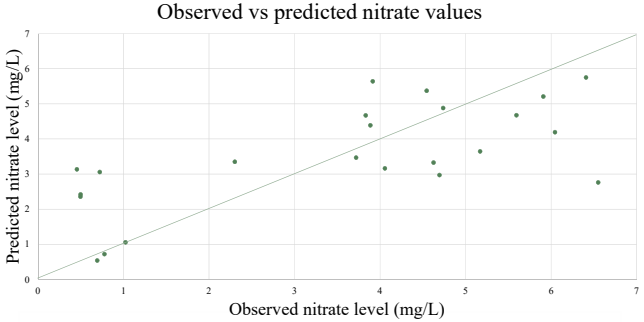
$$N = -1.671 \cdot 10^{-5}x_1 - 1.765 \cdot 10^{-5}x_2 + 5.656 \cdot 10^{-3}x_3 + 0.0714x_4 - 0.0407x_5 + 0.351x_6 + 0.0997x_7 + 0.0410x_8 - 0.730$$

Figure 1: Linear model created to predict nitrate levels (N) using a variety of sourced data.

Results

The observed nitrate levels (mg/L) in drinking water and predicted nitrate levels (mg/L) in drinking water were moderately, positively correlated, $r(23) = 0.69$ (Graph 1). The observed values had a larger and more even spread then the predicted values. Of the parameters used to predict nitrate levels it was found that water area was the most significant predictor followed by fertilizer density and population density.

Results



Graph 1: Scatterplot of observed and predicted nitrate levels for each county. The model has a smaller spread and tends to predict more moderate values. The line represents ideal model behavior.

Conclusions

The purpose of this project was to create a mathematical model to predict nitrate levels in Maryland drinking water on a county scale. The final linear model provided limited utility due to limitations in the number and frequency of features available to predict nitrate levels accurately. The difference in distributions between the observed and predicted values suggests nonlinear response of nitrate to inputs and/or missing factors in the model. Furthermore, the scale of predictions, county level, is too large to be useful in most cases. A model on a smaller scale would require data on a municipality level, which was not accessible. This model did however confirm the established relationship between the presence of nitrate in drinking water and fertilizer use (Tesoriero, 2017). Possible follow-up studies could include repeating this process for different contaminants and including additional data across a larger geographic area.

References

Aggarwal, C. C. (2015). *Data Mining: The textbook*. Cham, Switzerland: Springer.

Tesoriero, A. J., Gronberg, J. A., Juckem, P. F., Miller, M. P., & Austin, B. P. (2017). Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. *Water Resources Research*, *53*(8), 7316–733.

World Health Organization. (2011). *Guidelines for drinking-water quality*. Geneva.