A Hybrid Boosted Regression Tree Model to Predict and Visualize Nitrate Concentration Throughout the Central Valley Aquifer

Katherine M. Ransom, Bernard T. Nolan, Jon Traum, Claudia C. Faunt, Andrew M. Bell, Jo Ann M. Gronberg, David C. Wheeler, Celia Rosecrans, Bryant Jurgens, Gregory E. Schwarz, Kenneth Belitz, Sandra Eberts, George Kourakos, and Thomas Harter
Study Goals and Overview

- To map groundwater nitrate concentration “wall to wall and top to bottom”
- Gain understanding of the system
- Groundwater age, field scale nitrogen input, oxidation/reduction potential
- Boosted Regression Trees
Nitrate in Groundwater - Sources

Domestic wastewater is a potential source in rural and urban areas from septic tanks or leaky sewer lines (Bremer and Harter, 2012, and Viers et al., 2012).

Natural sources (organic matter decay) contributes a minimal amount.

*Nitrogen Cycle image: Modified from University of Wisconsin Integrated Pest and Crop Management, shown on http://fyi.uwex.edu/discoveryfarms/page/6/*
Nitrate in Groundwater - US

## Nitrate in Groundwater – Models

<table>
<thead>
<tr>
<th>Authors</th>
<th>Scale</th>
<th>Method(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nolan, Hitt, and Ruddy, 2002</td>
<td>National</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Nolan and Hitt, 2006</td>
<td>National</td>
<td>Non-linear Regression</td>
</tr>
<tr>
<td>Nolan et al., 2014</td>
<td>Central Valley</td>
<td>Logistic Regression, Random Forest</td>
</tr>
<tr>
<td>Nolan, Fienen, and Lorenz, 2015</td>
<td>Central Valley</td>
<td>Boosted Regression Trees, Bayesian Networks, Artificial Neural Networks</td>
</tr>
<tr>
<td>Ransom et al., 2017</td>
<td>Central Valley</td>
<td>Boosted Regression Trees</td>
</tr>
</tbody>
</table>


Ransom et al., 2017. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA, Science of the Total Environment, 601-602, 1160-1172.
Building on Previous Work

Hybrid Approach

- Oxidation/reduction potential
- Groundwater age
- Nitrogen loading – field scale

3D map

- Predictions mapped at depth
- Interpolation between predictions
Machine Learning for Nitrate

Pros
- Relations need not be linear or follow a particular data distribution
- Screens large numbers of variables
- Handles missing data
- Results not affected by collinearity
- Automatically incorporates interactions and thresholds
- Useful for inference

Cons
- Overfitting the data
- Model is harder to interpret
- Perceived as “black box”

Modified from: B.T. Nolan, 2017
Statistical Methods - Workflow

- Predictor variables attributed to wells, 145 total
- Boosted regression tree modeling
- Predictors ranked based on importance (variable reduction routine)
- Top 25 variables kept for final
- Predictions made at 17 depths, 3D map created

Measured concentrations

- 15.24 m deep
- 30.48 m deep
- 45.72 m deep
- 60.96 m deep
Well Data and Predictor Variables
EXPLANATION
Nitrate concentration in groundwater, in milligrams per liter, as N
- 0 to 2
- >2 to 4
- >4 to 6
- >6 to 8
- >8 to 10
- >10

A) Shallow
- 1400 wells
- Domestic wells
- 180 ft/54.9 m
- 27% exceedance

B) Deep
- 2108 wells
- Public wells
- 400 ft/121.9 m
- 6% exceedance

1662 “Hold-out” wells (not shown)
Probability of Anoxic Condition

EXPLANATION
Probability of DO < 0.5 ppm

- < 0.15
- 0.15 - 0.3
- 0.3 - 0.45
- 0.45 - 0.6
- 0.6 - 0.75
- > 0.75
MODFLOW/MODPATH Estimates of Groundwater Age with Depth

- Key component not included in previous models.
- “Proxies” such as well depth or depth to water.

Field-Scale Nitrogen Leaching Flux - 1975

Based on nearly 200 land use types, including 60 crop types.

County-Scale Nitrogen Input

Total landscape nitrogen input, 1992 (kg)

- <=2000
- >2000 - 4000
- >4000 - 6000
- >6000 - 8000
- >8000 - 10000
- >10000
Statistical Methods - Software

Variable Processing

Modeling and Prediction

3D Visualization

Packages

- caret
- gbm
- raster
- sensitivity
- boot
Statistical Methods - Boosted Regression Trees

- aka Gradient Boosting Machine
- An ensemble method: collection of many small models (boosting)
- Based on classification trees
- Each new tree built on the residuals of the previous tree (gradient)
- Randomness added by subsampling data
- Trees controlled by tuning aka metaparameters

Example Apartments Dataset

<table>
<thead>
<tr>
<th>m2.price</th>
<th>construction.year</th>
<th>surface</th>
<th>floor</th>
<th>no.rooms</th>
</tr>
</thead>
<tbody>
<tr>
<td>5897</td>
<td>1953</td>
<td>25</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1818</td>
<td>1992</td>
<td>143</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>3643</td>
<td>1937</td>
<td>56</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3517</td>
<td>1995</td>
<td>93</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>3013</td>
<td>1992</td>
<td>144</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5795</td>
<td>1926</td>
<td>61</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>2983</td>
<td>1970</td>
<td>127</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>2346</td>
<td>1985</td>
<td>105</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>4745</td>
<td>1928</td>
<td>145</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>4284</td>
<td>1949</td>
<td>112</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

Simple Regression Tree
Results – Model Performance

Training RMSE: 0.705
Training R²: 0.825
Hold-out RMSE: 1.132
Hold-out R²: 0.443

Residual Comparison
• To 1600 ft below ground surface
• 17 predicted layers
• Linear interpolation
• 1 m vertical resolution
Results – Predictions at Specified Depths

EXPLANATION
Nitrate - N (mg/L)
- < 2
- 2 - 4
- 4 - 6
- 6 - 8
- 8 - 10
- > 10

West Fans
East Fans
Basin

CALIFORNIA

0 180 360 Miles
0 160 320 Kilometers
Secondary Results – Partial Dependency Plots

Probability of Anoxic Conditions - DO

Probability of dissolved oxygen < 0.5 ppm

Probability of Anoxic Conditions - Mn

Probability of manganese > 50 ppb
Secondary Results – Partial Dependency Plots

Distance to River

Natural and Water Land Use, 1990s

Distance to river with stream order > 3, m

Area surrounding well as natural land use, m²
Secondary Results – Partial Dependency Plots

Natural and Water Land Use
Prob of DO < 0.5 ppm

Natural and Water Land Use
Prob of Mn > 50 ppb
Summary and Conclusions

- Mapped nitrate tended to decrease with depth
- Alluvial fans region had higher nitrate concentrations than basin subregion
- Anoxic conditions highly related to nitrate concentration
- Patterns on partial plots make intuitive sense
- Coming soon: updated national nitrate and arsenic maps

Locating High Risk Domestic Wells

- Cookie cutter national models (updated or current) for full coverage
- Use estimates from current national arsenic model (Ayotte et al., 2017)
- Develop new California specific model
- Consider multiple constituents together (multinominal BRT)?
- Nitrate, arsenic, uranium, others?
- Overlay with well locations

Questions?

Article available at:

Data raster grids available at:
https://www.sciencebase.gov/catalog/item/58c1d920e4b014cc3a3d3b63
Statistical Methods – Cross Validation

Credit: Hastie et al., 2009. The Elements of Statistical Learning.
Increase in Prediction Errors to Hold-out Data

RMSE
1.15
1.20
1.25
1.30
1.35
1.40
1.45

Variable Index

Percent Difference RMSE
0
5
10
15
20
25

Variable Index

1% difference
Results – Prediction Intervals

199 models made with bootstrapped sets of the training data

199 predictions made to hold-out data
Results – Prediction Interval Width

EXPLANATION
Relative prediction interval width
- < 4
- 4 - 8
- 8 -12
- > 12

CALIFORNIA
- East Fans
- West Fans
- Basin

Private well depth
Public well depth

West Fans
East Fans
Basin
Results – Sobol Sensitivity Indices