The Use of Machine Learning Models to Predict Groundwater Quality in the Confined Claiborne Aquifers of the Mississippi Embayment

Katherine Knierim
James Kingsbury, Connor Haugh

Lower Mississippi-Gulf Water Science Center
National Water Quality Assessment Project – Intensive Principal Aquifer Analysis
Mississippi River Valley Alluvial & Mississippi Embayment Aquifers

Preliminary Information—Subject to Revision. Not for Citation or Distribution
MERAS Groundwater Quality – salinity
Airborne Electromagnetic Survey

Chloride

- 0 - 10
- 11 - 50
- 51 - 250
- 251 - 3000

- MRVA
- MERAS
Machine-Learning Methods: Boosted Regression Trees

*ensemble method* - build lots of simple trees (weak learners)

*boosting* - use residual of previous tree

**PROS:** non-linear, non-monotonic, no hypothesis testing assumptions, handles missing data, can use many predictors

**CONS:** susceptible to over-fitting, can be harder to interpret (black box?)

**Model Hyperparameters**
- interaction depth (how deep to split trees)
- minimum observations per node
- learning rate (how much of previous tree to use)
- number of trees

**Solution:** find simpler models with similar predictive performance
Modeling 3D WQ in Mississippi Embayment

Machine learning to model and map groundwater quality
Enables groundwater quality to be placed into a groundwater-flow system context

**EXPLANATORY VARIABLES (400+)**

- WELL GEOMETRY
  - [Map Image]
  - [Map Image]
  - [Map Image]

- SURFACE VARIABLES
  - [Map Image]
  - [Map Image]
  - [Map Image]

- GW MODEL VARIABLES
  - [Map Image]
  - [Map Image]
  - [Map Image]

**Machine Learning:**
- Boosted Regression Tree

**RESPONSE VARIABLES**
- SC, Cl, pH, redox, Mn, As

Map GW WQ across MRVA, MCAQ, and LCAQ

Preliminary Information - Subject to Revision. Not for Citation or Distribution
Final Prediction Framework

MRVA (one layer)

CLBG (6 layers)
Modeling 3D WQ in Mississippi Embayment

Machine learning to model and map groundwater quality
Enables groundwater quality to be placed into a groundwater-flow system context

**EXPLANATORY VARIABLES (400+)**

- Aquifer
- Depth to Screened Interval
- Total Screen Length
- Position in study area
- Confinement

**WELL GEOMETRY**

**SURFACE VARIABLES**

**GW MODEL VARIABLES**

**RESPONSE VARIABLES**

SC, Cl, pH, redox, Mn, As

Machine Learning: Boosted Regression Tree

Map GW WQ across MRVA, MCAQ, and LCAQ

Preliminary Information: Subject to Revision. Not for Citation or Distribution
Modeling 3D WQ in Mississippi Embayment

Machine learning to model and map groundwater quality
Enables groundwater quality to be placed into a groundwater-flow system context

EXPLANATORY VARIABLES (400+)
- WELL GEOMETRY
- SURFACE VARIABLES
- GW MODEL VARIABLES
- Land Use
- Soils (phys. params)
- Soil (chemistry)
- Climate
- Recharge

Machine Learning: Boosted Regression Tree

RESPONSE VARIABLES
- SC, CI, pH, redox, Mn, As

Map GW WQ across MRVA, MCAQ, and LCAQ

Preliminary Information - Subject to Revision. Not for Citation or Distribution
Surface Variable Example: Land Use/Land Cover

https://github.com/brclark-usgs/zonepy

<table>
<thead>
<tr>
<th>Well</th>
<th>Soybeans</th>
<th>Corn</th>
<th>Cotton</th>
<th>WetInd</th>
<th>Crop</th>
<th>Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>72</td>
<td>9</td>
<td>-</td>
<td>13</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>27</td>
<td>-</td>
<td>7</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>61</td>
<td>-</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>
Modeling 3D WQ in Mississippi Embayment

Machine learning to model and map groundwater quality
Enables groundwater quality to be placed into a groundwater-flow system context

EXPLANATORY VARIABLES (400+)

WELL GEOMETRY

SURFACE VARIABLES

GW MODEL VARIABLES

- Travel Time
- Heads
- Flux
- Water Use
- Groundwater Age

Machine Learning: Boosted Regression Tree

RESPONSE VARIABLES

SC, Cl, pH, redox, Mn, As

Map GW WQ across MRVA, MCAQ, and LCAQ
GW Model Variables: GW Levels and Age

Particles added to the GW flow model $\rightarrow$ time of travel approximates (relative) groundwater age.
Workflow – automation via Python & R

Zonepy ➔ loop thru exp. var. rasters and attribute all wells

Python (pandas) & R (corr) ➔ remove variables with linear correlation $r^2 > 0.8$

Python (Glob) ➔ merge multiple text files into one attribute table (400+ variables)

R (parallel & caret) ➔ Train models on YETI USGS supercomputer in parallel
Specific Conductance Models

Training

\[ n = 1,834 \]
\[ r^2 = 0.95 - 0.99 \]

Holdout

\[ n = 456 \]
\[ r^2 = 0.67 - 0.71 \]
Specific Conductance Models

Pre-processing = remove linear correlated variables

Holdout

- full, best
- pre, best
- full, simpler
- pre, simpler
- full, reduced
- pre, reduced
### Specific Conductance Models

<table>
<thead>
<tr>
<th>Model</th>
<th>$r^2$</th>
<th>No. Vars</th>
<th>Int. Depth</th>
<th>Min. Obs.</th>
<th>Lrn. Rate</th>
<th>No. Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full, best</td>
<td>0.71</td>
<td>431</td>
<td>16</td>
<td>8</td>
<td>0.004</td>
<td>4500</td>
</tr>
<tr>
<td>Pre-processed, best</td>
<td>0.68</td>
<td>225</td>
<td>18</td>
<td>8</td>
<td>0.006</td>
<td>4000</td>
</tr>
<tr>
<td>Full, simpler</td>
<td>0.69</td>
<td>431</td>
<td>6</td>
<td>8</td>
<td>0.014</td>
<td>3000</td>
</tr>
<tr>
<td>Pre-processed, simpler</td>
<td>0.67</td>
<td>225</td>
<td>6</td>
<td>10</td>
<td>0.010</td>
<td>5000</td>
</tr>
<tr>
<td>Full, reduced</td>
<td>0.69</td>
<td>56</td>
<td>6</td>
<td>8</td>
<td>0.014</td>
<td>3000</td>
</tr>
<tr>
<td>Pre-processed, reduced</td>
<td>0.67</td>
<td>50</td>
<td>6</td>
<td>10</td>
<td>0.010</td>
<td>5000</td>
</tr>
</tbody>
</table>

#### Holdout

- Full, best
- Pre, best
- Full, simpler
- Pre, simpler
- Full, reduced
- Pre, reduced

![Scatter plot of predicted vs. observed values with regression line](image)
Relative Influence of Explanatory Variables

- Position in System
- GW Flow Model / Age
- Soil Chemistry / Physical Properties
- Climate
- Hydrogeology
- Precip. Chemistry
- Well Geometry
Well altitude (DEM)

Explanation
Land Surface Altitude Value
-19 - 100
101 - 200
201 - 300
301 - 500
501 - 1,500

Well Altitude (ft)

Ln Spec. Cond.

partial dependence plots
Hydrogeology

GW path length (min)

Distance to Fault

Ln Spec. Cond.
## Comparing Explanatory Variables

<table>
<thead>
<tr>
<th>Specific Cond.</th>
<th>Chloride</th>
<th>pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well Altitude</td>
<td>Longitude</td>
<td>Altitude Top of Screen</td>
</tr>
<tr>
<td>Geomorph (uplands)</td>
<td>GW (min path length)</td>
<td>GW Age (minimum)</td>
</tr>
<tr>
<td>GW Level (1930)</td>
<td>Soil Kaolinite</td>
<td>Longitude</td>
</tr>
<tr>
<td>Longitude</td>
<td>GW Level (1930)</td>
<td>GW Age (30th %)</td>
</tr>
<tr>
<td>GW Age (30th %)</td>
<td>Thickness Conf. Unit</td>
<td>Geomorph (uplands)</td>
</tr>
<tr>
<td>Depth to Water (change)</td>
<td>Well Altitude</td>
<td>Soil Chemistry (Ti)</td>
</tr>
<tr>
<td>Soil pH</td>
<td>Normalized Position</td>
<td>Altitude Bottom of Screen</td>
</tr>
<tr>
<td>Precip (July)</td>
<td>GW Age (90th %)</td>
<td>Distance to Fault</td>
</tr>
<tr>
<td>Soil Clay %</td>
<td>Soil (% fine)</td>
<td>Normalized Position</td>
</tr>
<tr>
<td>Thickness Conf. Unit</td>
<td>GW (max path length)</td>
<td>Latitude</td>
</tr>
</tbody>
</table>
Next steps… 3D prediction

QUESTIONS?

Kathy Knierim
kknierim@usgs.gov