USGS
science for a changing world

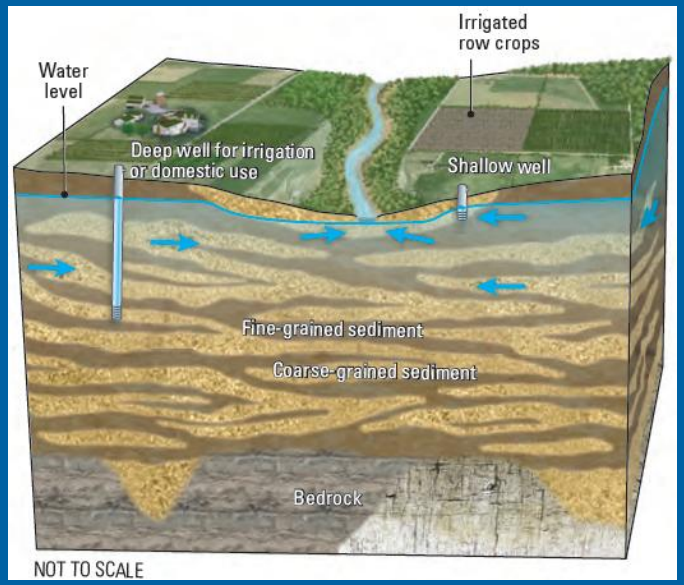
Advances in Statistical Modeling and Mapping of Groundwater Contaminants

Mindy L. Erickson, PhD, PE
Hydrologist

U.S. Department of the Interior
U.S. Geological Survey

Current USGS work

- Statistical probability mapping of several constituents across Glacial Aquifer System
  - Arsenic
  - Manganese
  - pH
  - Redox/DO
- Complex statistical modeling now possible
  - Machine learning (ML) approaches: Random Forest (RF) and Boosted Regression Tree (BRT)



USGS

Statistical predictive modeling

- Advantages of new techniques
  - Does not assume a relationship
  - Can fit nonlinear relations
  - Accommodates predictor variables of any type (e.g. categorical, numeric, binary)
  - ‘Boosting’ focuses on the unexplained deviance of previous tree

USGS

Motivation

- **2015 USGS circular about water quality in the US Glacial Aquifer System:**

  "Contaminants from geologic sources – in particular arsenic and manganese – in groundwater used for drinking are a potential concern for human health"

- **MN arsenic and manganese**
  - **Arsenic ≥ 10 µg/L in 10.7% of new potable wells**
  - **Manganese ≥ 100 µg/L in 50% of ambient wells; ≥ 300 µg/L in 22%**
  - **Unequally distributed across state**
  - **Redox- and pH-sensitive**

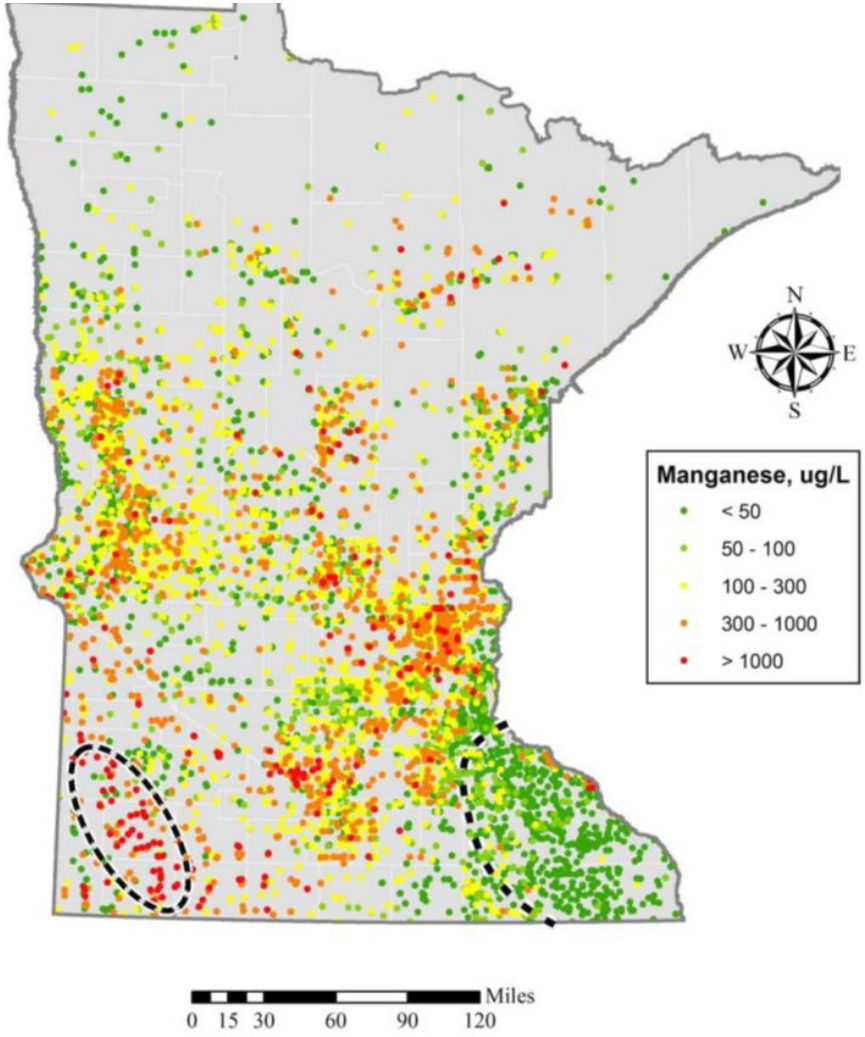
USGS

Manganese in Minnesota's Groundwaters

Emphasizing the Health Risks of Manganese in Drinking Water

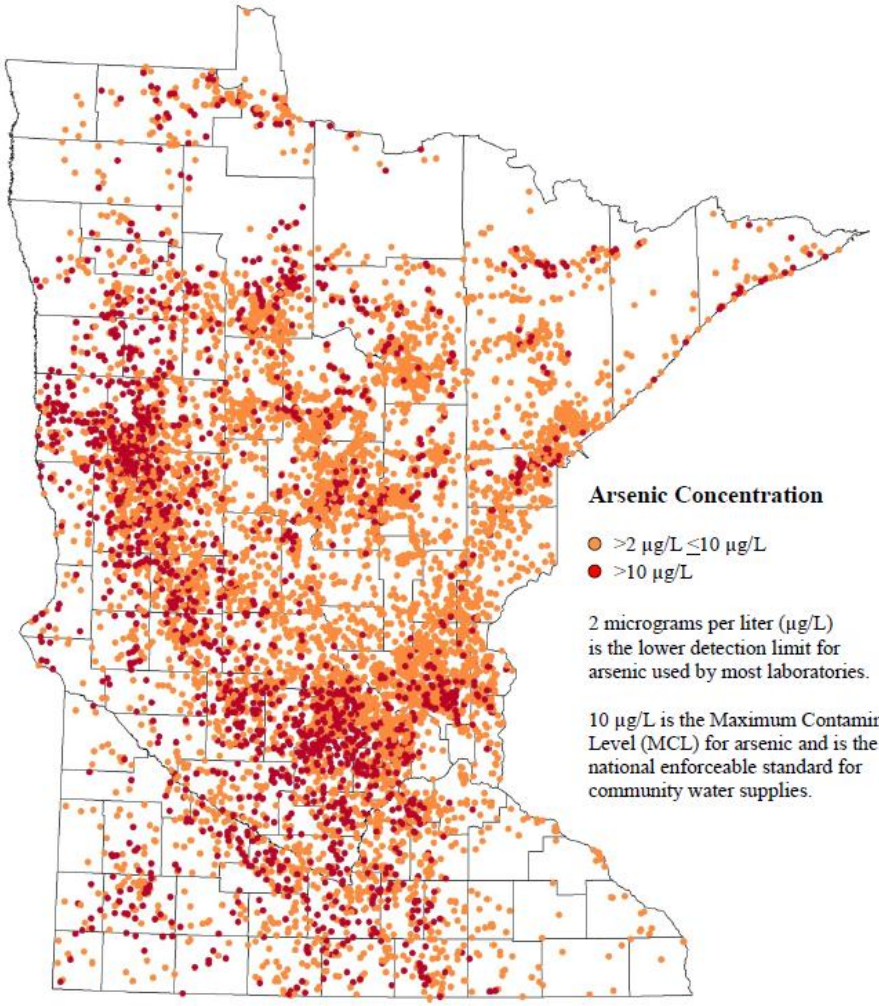
Prepared for the Minnesota Ground Water Association

September 2015



MINNESOTA DEPARTMENT OF HEALTH
MDH
Minnesota Department of Health
Well Management Section
P.O. Box 64975
St. Paul, Minnesota 55164-0975
651-201-4600 or 800-383-9808
www.health.state.mn.us/divs/eh/wells

Arsenic Occurrence in New Wells
August 2008 – July 2013



Each dot represents a single well. Wells which were sampled and had less than 2 µg/L arsenic are not shown on this map.

Current USGS statistical modeling and mapping work

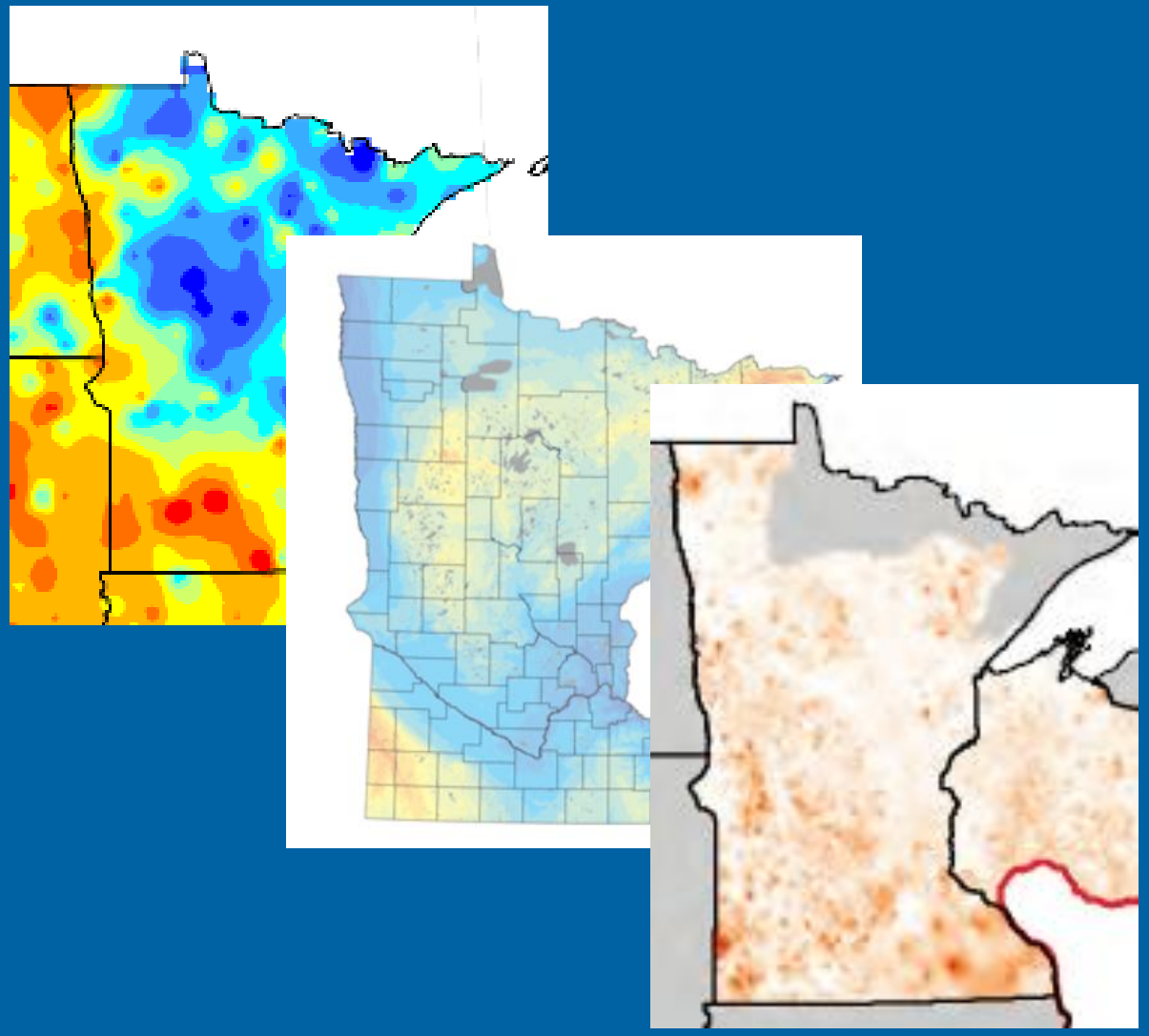
Types of variables considered for Machine Learning

- GW travel time: well depth, depth below the water table, recharge, hydrologic position
- Soil drainage: hydrologic groups, soil drainage class, soil texture
- Geology: Surficial geology, thickness of fine grained material, age of bedrock deposits
- Reduction potential: organic matter content
- Other: land use, soil chemistry

USGS

Online geospatial data sources

- SSURGO
- USGS
- States
- Etc.



USGS

Current USGS statistical modeling and mapping work

Variable Selection for Machine Learning

- Include all variables (60 – 80 or more) in initial runs
- Variables ranked by importance
- Remove least important variables
- Final models include the 12 – 20 most important variables

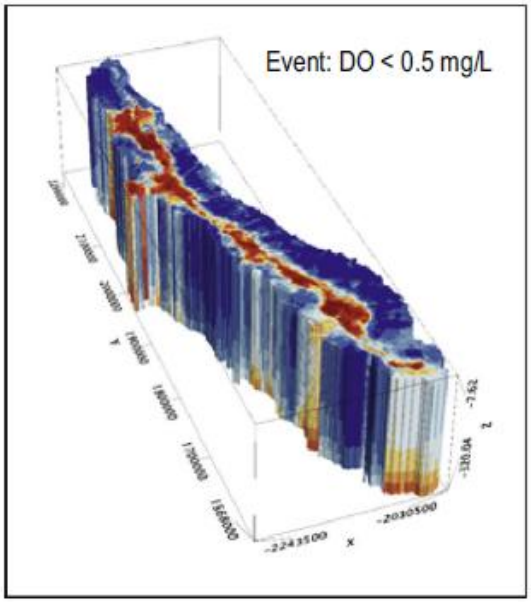
USGS

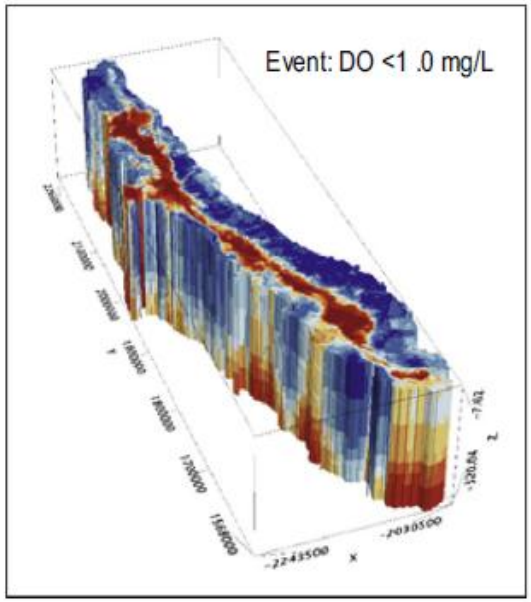
Current USGS manganese work

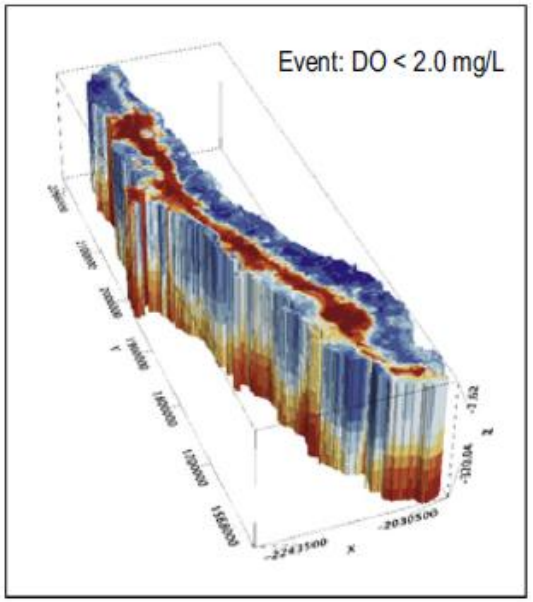
- Reasonable performance for Manganese, DO, Arsenic prediction in Central Valley, CA
  - overall accuracy 90%
  - Top predictor variables: related to older groundwater and anoxic conditions
    - Lateral position
    - Depth to water table
    - Sum of poorly-drained soils
    - Average porosity
    - Average organic matter

USGS

Boosted Regression Tree Example



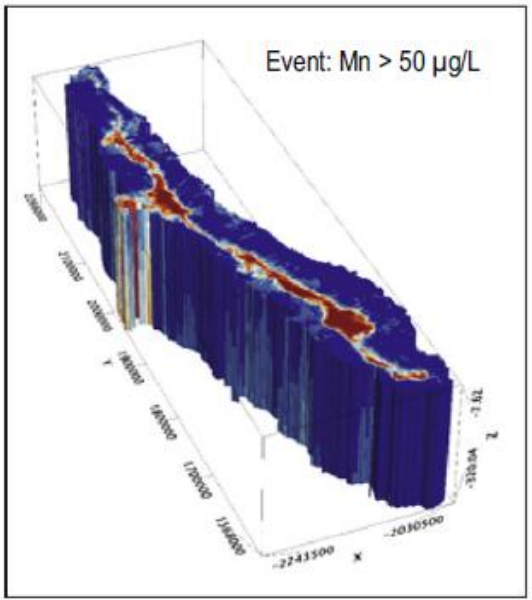


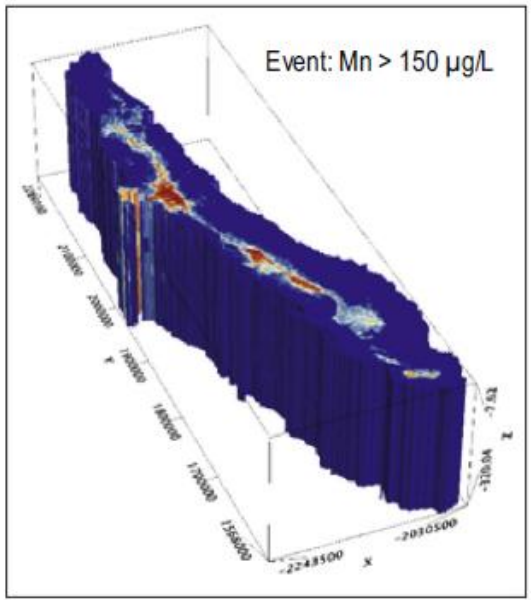


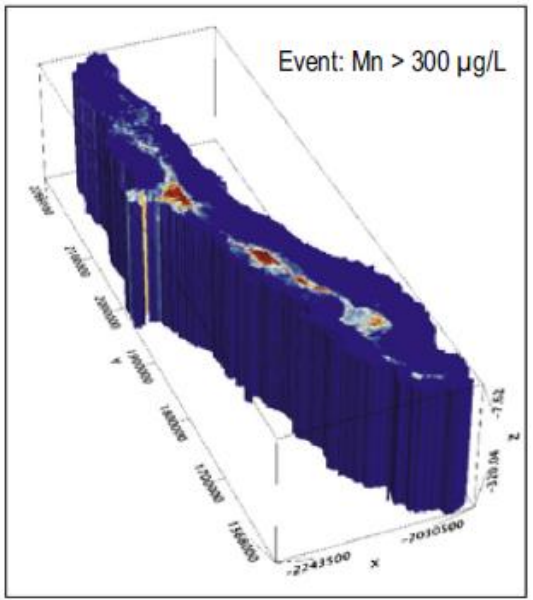
EXPLANATION

Probability of Event

- 0 - 0.1
- >0.1 -0.2
- >0.2 - 0.3
- >0.3 - 0.4
- >0.4 - 0.5
- >0.5 - 0.6
- >0.6 - 0.7
- >0.7 -0.8
- >0.8 - 0.9
- >0.9 - 1.0







Questions?

Mindy L. Erickson, USGS
merickso@usgs.gov

USGS