

Using logistic regression to predict the probability of occurrence of volatile organic compounds in ground water

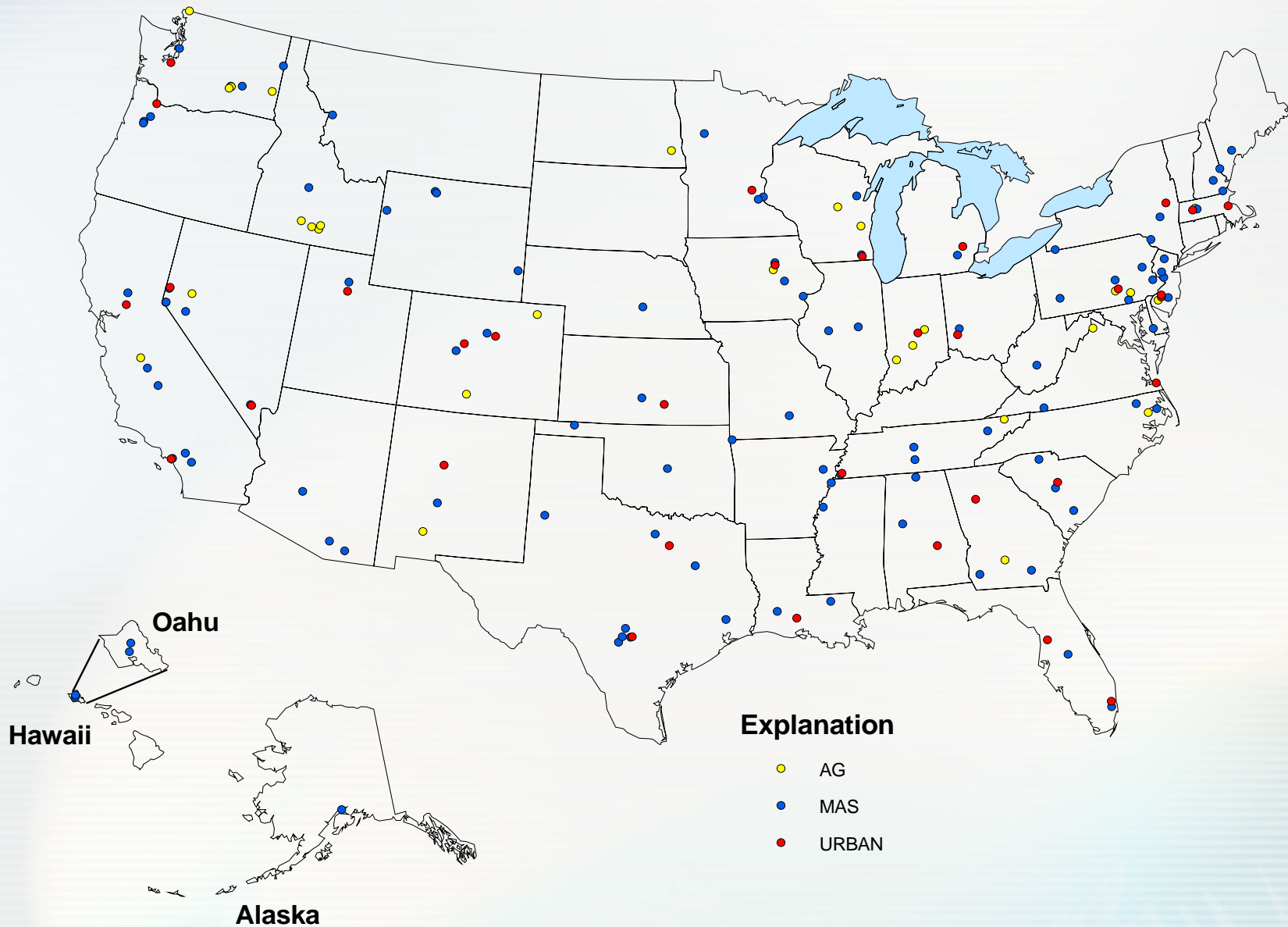
by
Michael Moran
NAWQA VOC National Synthesis

NAWQA VOC Monitoring

Cycle 1 Ground Water Sampling

- 💧 3,883 NAWQA-sampled wells analyzed for 55 target VOCs
- 💧 1,185 wells sampled by other agencies (Retro) and analyzed for up to 55 VOCs
- 💧 161 total networks
 - 💧 98 MAS networks
 - 💧 33 Urban networks
 - 💧 30 Agricultural networks

NAWQA VOC Monitoring



Statistical Analyses

Hypothesis Testing

- 💧 Contingency tables
- 💧 Mann-Whitney test
- 💧 Kolmogorov-Smirnov test

Correlation

- 💧 Spearman correlation

Associations

- 💧 Multivariate logistic regression

Logistic Regression

Advantages

- 💧 Response is ideal for censored data
- 💧 Can be univariate or multivariate
- 💧 Can identify variables associated with response
- 💧 Can predict probability of binary or ordinal response relative to independent variables

Disadvantages

- 💧 Independent variables often generalized
- 💧 Can be difficult to extrapolate probability beyond sample point

Logistic Regression

Logistic Regression Equation

$$P = \frac{e^{(\beta_0 + \beta_1 X_1 + \dots \beta_i X_i)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \dots \beta_i X_i)}}$$

where,

P = probability of detecting VOC,

β_0 = y-intercept,

β_i = slope coefficient of X_1 to X_i explanatory variables,

X_i = 1 to i explanatory variables

VOC Analyses

Prediction

- 💧 Any VOC for all networks – National Scale
[ES&T, vol. 33, no. 23, p. 4176-4187]

Process Understanding

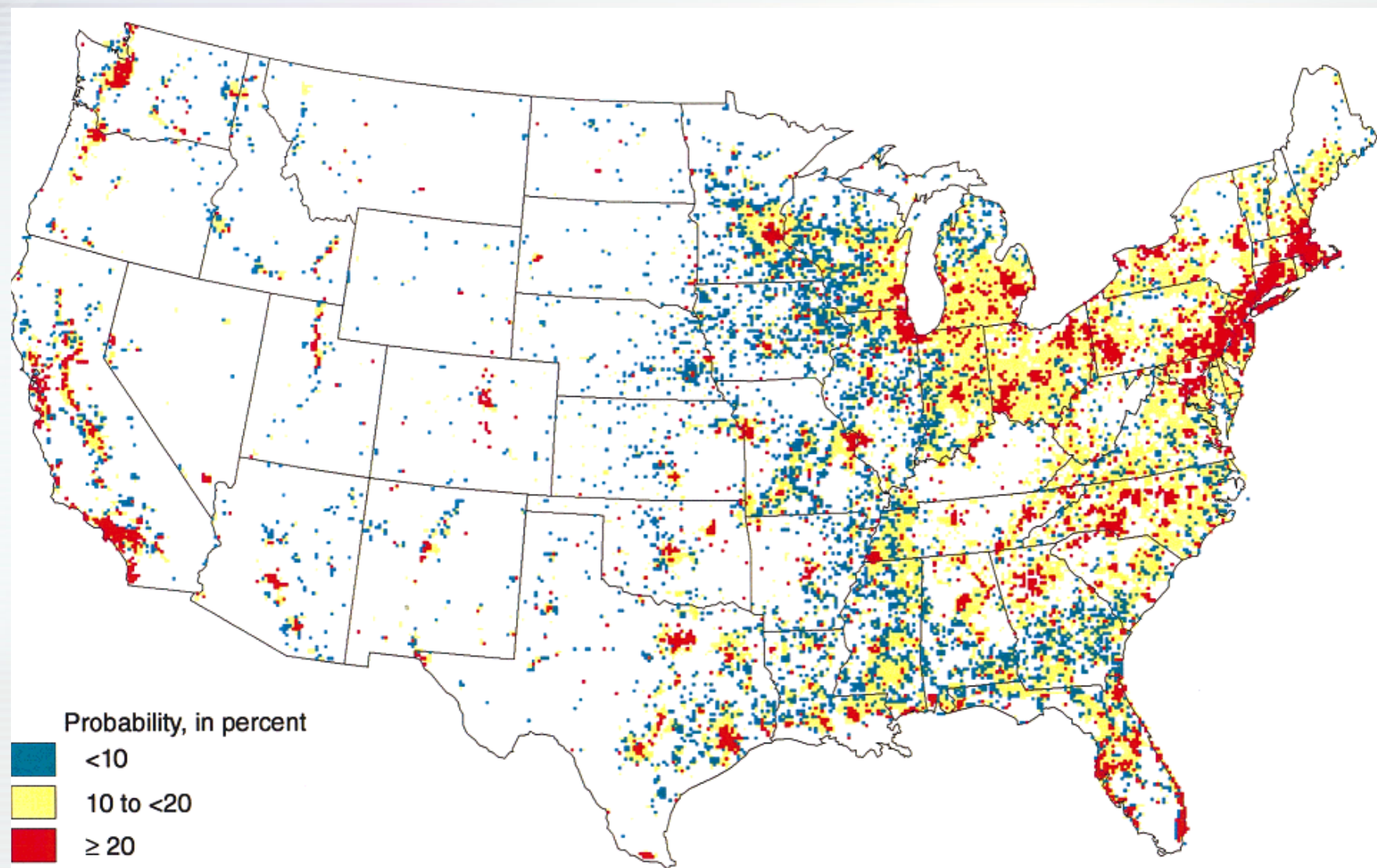
- 💧 Individual VOCs in urban land use studies
[ES&T, vol. 38, no. 20, p. 5327-5338]
- 💧 Most frequently detected VOCs analyzed
- 💧 MTBE and 4 solvents; DCM, PCE, TCA, TCE
- 💧 Regressions were done to establish associations
- 💧 No attempt to predict probability of occurrence

VOC Analyses

Prediction – National Scale

- 💧 Prediction was applied to analysis of any VOC for all networks
[ES&T, vol. 33, no. 23, p. 4176-4187]
- 💧 Logistic regression equation was univariate
- 💧 Probability of occurrence of any VOC was only related to population density
- 💧 Although little was learned about STF, the probability of occurrence could be extrapolated

VOC Analyses



Taken from: Squillace, P.J., Moran, M.J., Lapham, W.W., Price, C.V., Clawges, R.M., and Zogorski, J.S., 1999, Volatile organic compounds in untreated ambient groundwater of the United States, 1985-1995: Environmental Science & Technology, v. 33, no. 23, p. 4176-4187.

Process Understanding – Urban Areas

- 💧 Prediction could not be applied to analysis of individual VOCs in urban land-use studies
[ES&T, vol. 38, no. 20, p. 5327-5338]
- 💧 Logistic regression only yielded associations
- 💧 Many equations included variables that could not be extrapolated beyond the sample point
 - 💧 Example: dissolved oxygen was related to many VOCs but the concentration of dissolved oxygen was unknown beyond the sampled well

Process Understanding – Major Aquifers

- 💧 Source and transport variables often surrogates
 - 💧 Example: sources often were population density and urban land use
- 💧 For some variables, extrapolation beyond the sample point may not be possible
 - 💧 Example: transport variable depth to top of open interval

VOC Analyses

Process Understanding – Various VOCs in Ground Water

- estimation is possible
 - estimation may be difficult

VOC	Associated variables
Methyl <i>tert</i> -butyl ether (MTBE)	population density use of MTBE in gasoline recharge aquifer consolidation soil permeability leaking underground storage tanks
Perchloroethene (PCE)	urban land use sand content of soil dissolved oxygen soil erodibility RCRA sites within 1km depth to top of screened interval septic systems within 1km
Trichloroethene (TCE)	population density dissolved oxygen RCRA sites within 1km CERCLA sites within 1km casing diameter
1,1,1-Trichloroethane (TCA)	dissolved oxygen urban land use population density depth to top of screened interval recharge CERCLA sites within 1km septic systems within 1km
Methylene chloride (DCM)	population density bulk density of soil sand content of soil urban land use median year of home construction

Observations

Logistic Regression **IS** useful for:

💧 Identifying variables associated with occurrence of VOCs in ground water

⇒ This may be useful for determining the factors that cause or influence occurrence

💧 Ranking variables associated with occurrence by importance

⇒ This may be useful for determining the relative importance of factors that cause or influence occurrence

Observations

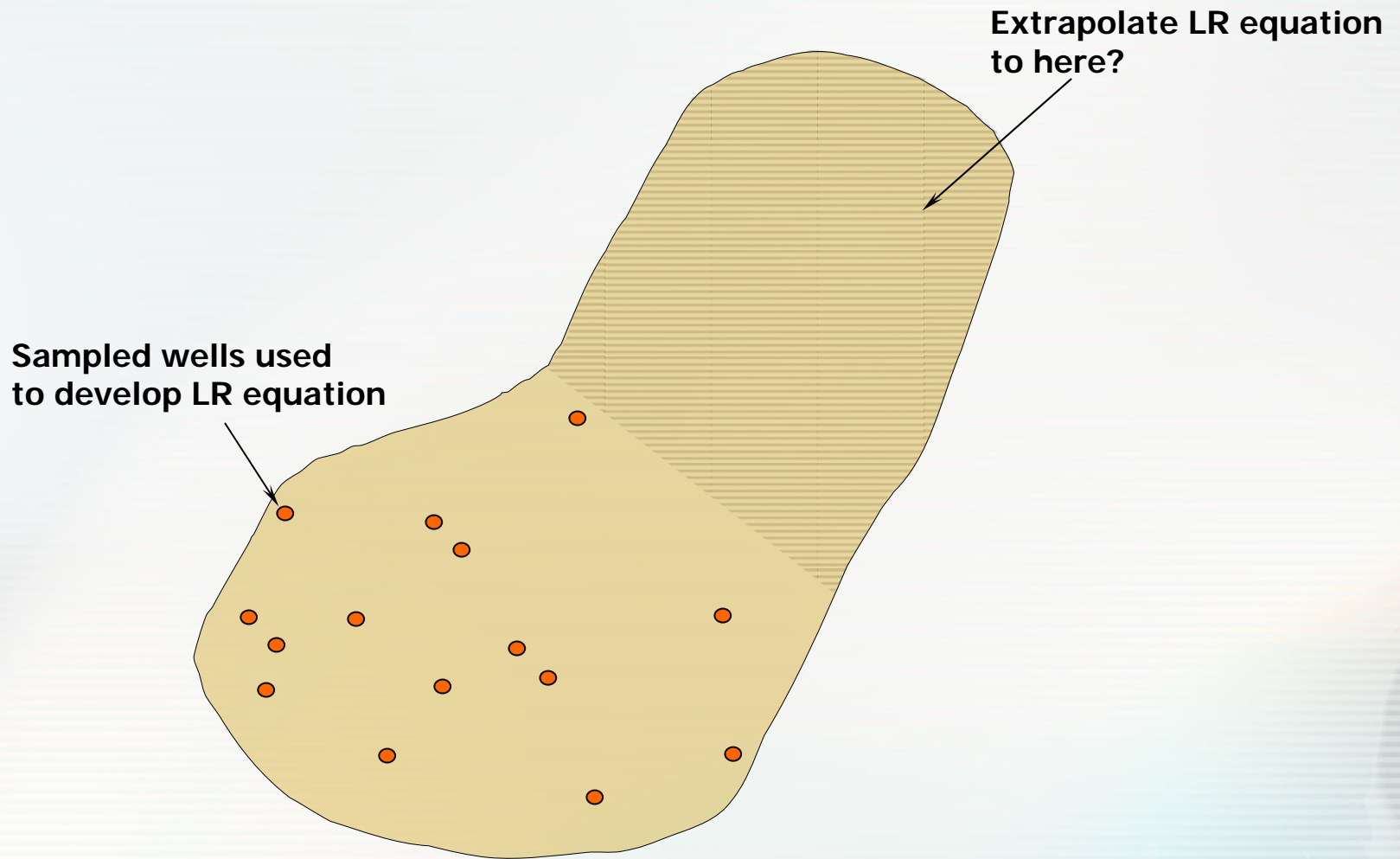
Logistic Regression **MAY BE** useful for:

💧 Predicting the probability of occurrence of VOCs in ground water provided:

- 1) Data are not spatially limited
- 2) Data are not from highly specific sampling approaches
- 3) Explanatory variables can be extrapolated

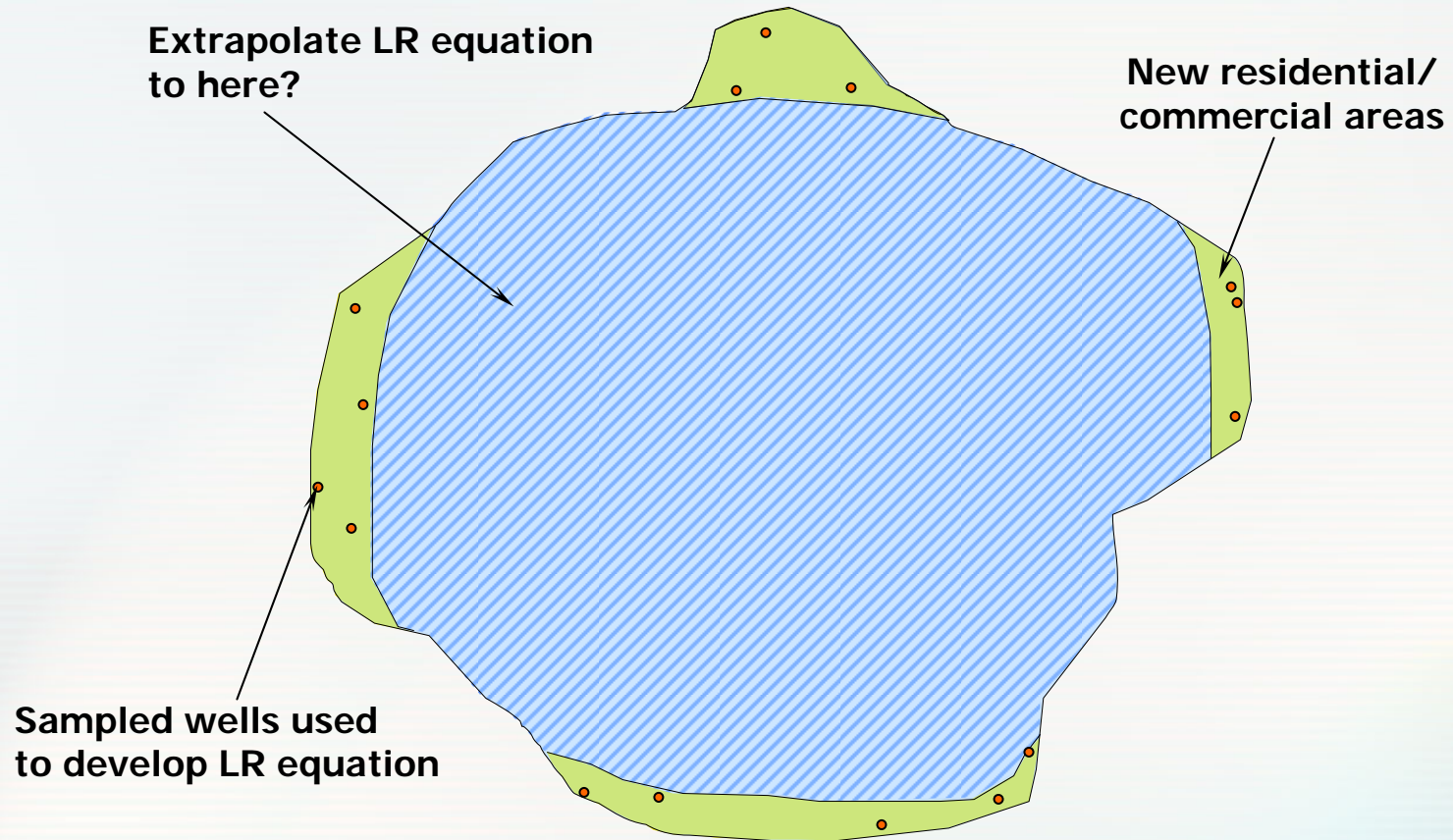
Hypothetical Cases

Hypothetical Aquifer Study Area



Hypothetical Cases

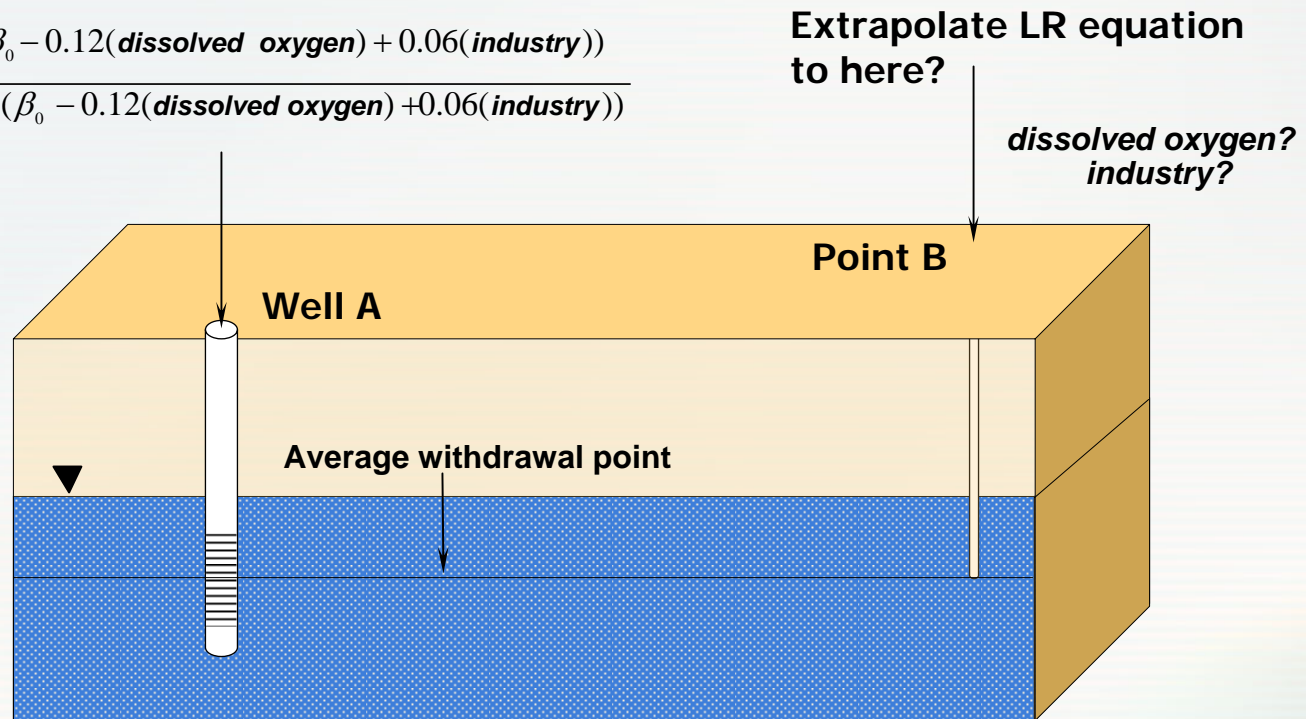
Hypothetical Urban Area



Hypothetical Cases

Benzene Occurrence

$$P = \frac{e^{(\beta_0 - 0.12(\text{dissolved oxygen}) + 0.06(\text{industry}))}}{1 + e^{(\beta_0 - 0.12(\text{dissolved oxygen}) + 0.06(\text{industry}))}}$$



Conclusions

- 💧 Logistic regression is useful for determining associations that may aid in understanding sources, transport, and fate
- 💧 For determining associations, logistic regression should use explanatory variables that are as specific as possible
- 💧 Although associations found using logistic regression may be insightful, they do not imply a direct cause and effect

Conclusions

💧 Logistic regression can be useful in predicting the probability of occurrence of ground water contaminants if:

- 1) Data are not highly localized
- 2) Data are not from highly specific sampling approaches

💧 For prediction purposes, logistic regression should use explanatory variables that:

- 1) Are ubiquitous to the area of interest
- 2) Can be estimated or extrapolated