Predicting Arsenic Concentration of Groundwater in Bangladesh

Chris Tasich

Final Project

Data Analytics for Engineers

December 11, 2016

I. Abstract

Arsenic in groundwater in Bangladesh presents one of the largest and most immediate humanitarian problems. Groundwater is the most reliable source of freshwater as surface water is heavily polluted with pathogens and waste. Estimates suggest 6-11 million wells servicing ~25 million people are above the Bangladeshi standard for arsenic (50 μg l^{-1}). The Bangladesh Arsenic Mitigation Water Supply Project dataset contains arsenic concentration and well depth data for ~5 million wells across Bangladesh. Using this dataset, regression methods can be used to develop predictive models for arsenic concentration for new well installations. With accurate prediction, resources can be better allocated to ensure clean and safe water for the Bangladeshi population. Our findings suggest support vector machines and regression tree models perform the best ($RMSE \approx 130\, ppb$) using only longitude, latitude, and well depth as predictors. Though the results are poor, these models provide the framework for future research that will utilize more predictors as stratigraphic data becomes available. Furthermore, future work will use classification methods which may provide better predictive accuracy.

II. Introduction

The country of Bangladesh lies on the Ganges-Brahmaputra (GB) deltaplain. The country is one of the most densely populated countries in the world with 156.6 million people living within an area of 147,500 km^2 (approximately the size of Iowa). In the 1970s, surface water was widely abandoned for groundwater due to surface water contamination by pathogens and industrial and organic pollutants. The consequences of this switch were not discovered until the mid-1990s. Nearly one-third of shallow wells (< 300 ft deep) in Bangladesh are contaminated with naturally occurring arsenic (BGS, 2001). These wells exceed the Bangladesh drinking water standard of 50 μg l^{-1} and almost half exceed the World Health Organization standard of 10 μg l^{-1} (BGS 2001; Kinniburgh and Kosmus, 2002; WHO, 2011). Arsenic is a known carcinogen and can lead to a variety of cancers (WHO, 2011). However, in many locations, there are no sustainable alternatives to groundwater. Surface water is easily contaminated by pathogens and accounts for more deaths than arsenic. Furthermore, arsenic is not the only groundwater contaminate. Many arsenic-free wells are often contaminated by other heavy metals or salinity (WHO, 2011). Despite these problems, groundwater offers the best hope for a sustainable freshwater supply. Therefore, there is need to be able to predict arsenic contamination more effectively and completely. Here, we attempt to predict arsenic concentration as a function of well depth, longitude, and latitude. This analysis is only a first attempt at modeling and high mean standard errors are expected. However, this will provide the framework for future analyses using a more robust dataset with more predictors.

Literature Review

There has been much research into the geochemical modeling of arsenic mobility within the delta (Harvey et al., 2002; Ravenscroft et al., 2005; Fendorf et al., 2010) which has elucidated the larger scale spatial patterns of arsenic contamination. Most of the research focuses on determining the

mechanisms through which arsenic becomes mobile without seeking to predict occurrence of arsenic. Other research has discerned the spatial distribution of arsenic through field surveys (BGS, 2001; BAMWSP 2001). However, there has been very little research into prediction of arsenic concentrations using these datasets. This is largely due to the stratigraphic heterogeneity that controls the distribution and connectivity of the shallow aquifer system. Anecdotally, wells separated by as little as 10 m at a similar depth can have vastly different arsenic concentrations. Some research has attempted to predict arsenic contamination with depth at very local levels (Gelman et al., 2004), while others have attempted to determine the lateral extent of arsenic contamination at a regional level (Winkel et al., 2008). However, very little research has attempted to quantify the spatial (both lateral and vertical) using statistical learning techniques. We can leverage these large datasets (BGS, 2001; BAMWSP 2001) and other spatial data to develop predictive models of arsenic within the delta.

III. Modeling Arsenic Concentration

Here, we use the Bangladesh Arsenic Mitigation and Water Supply Program (BAMWSP) dataset ($n=4.69x10^6$) of arsenic concentrations in groundwater. We attempt to build various types of multiple regression models to predict arsenic content based on three predictors – longitude, latitude, and depth of the well.

Bangladesh Arsenic Mitigation and Water Supply Program Dataset

The BAMWSP dataset contains arsenic concentration for ~5 million wells throughout the country of Bangladesh. The data only includes the well depth (ft), arsenic concentration (ppb), and geocode. The geocodes are a derived unit to describe nested administrative units in Bangladesh. These geocodes are associated with the mouza or township unit. In order to assign a more definitive x and y location for wells, we assign the longitude and latitude of the mouza centroid to the wells within each mouza. This results in many wells having the same longitude and latitude. We then filter errant data points (i.e. unrealistic values for predictors) and NAs by removing the entire observation. We also added binary response variables for thresholds of As > 10 ppb, As > 50 ppb, and an additional response variable for log(As). The final dataset contains arsenic concentration both categorical and continuous, well depth (ft), longitude, and latitude (Table 1).

Response Variables				Predictor Variables		
As (ppb)	log(As)	As > 10 ppb	As > 50 ppb	Depth (ft)	Longitude	Latitude
25	3.2188	1	0	180	90.44368	23.12708
51	3.9318	1	1	180	90.44368	23.12708
0	-6.907755	0	0	800	90.44368	23.12708

Table 1 | Sample of arsenic dataset. The dataset was derived from the BAMWSP data.

For the purpose of this study, we chose a subset ($n=4.66x10^5$) of the original data (Fig. 1) for two reasons: 1) it was computationally less expensive to implement and test our models and 2) the subset represents a portion of the country where we have collected stratigraphic data which we will use as predictors in future model iterations.

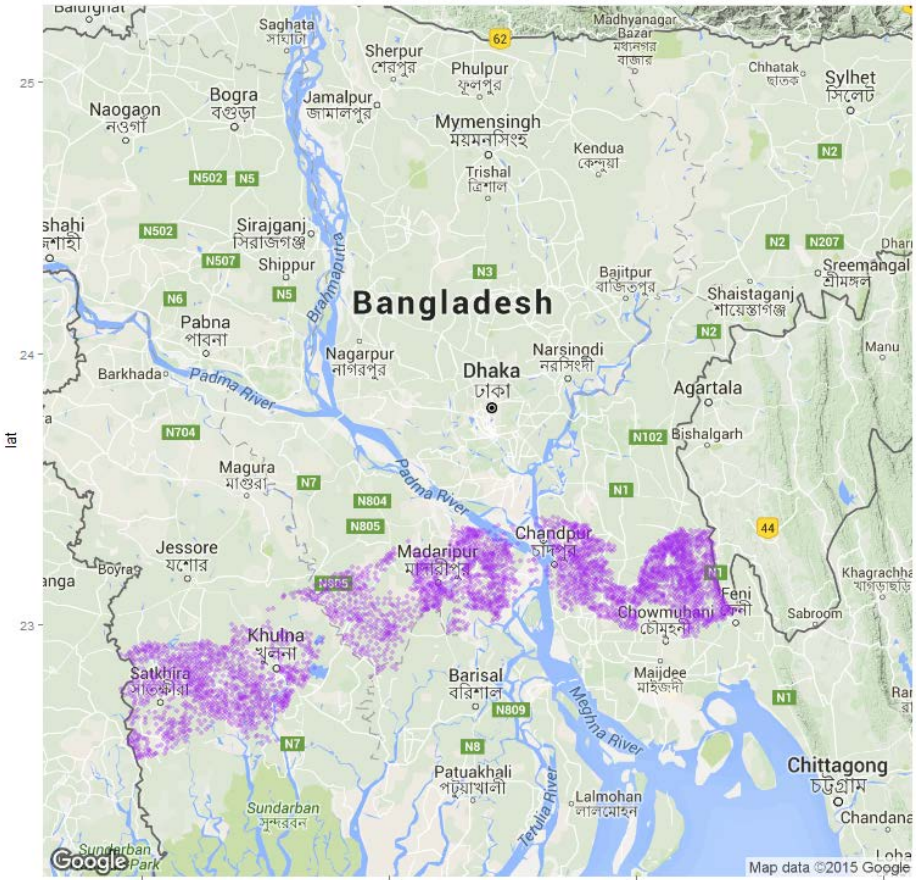

Figure 1 | Map view of well distribution in southern Bangladesh. The transect cuts along the tidally dominated network in the southwest and the fluvial network in the central and southeast.

A pairs plot of the model variables was used to explore relationships in the data (Fig. 2). The arsenic concentrations appear discretized due to the sampling method used for ~50% of the observation. The samples were taken by trained NGO workers who used a two-step Hach EZ arsenic kit. The kit uses an arsine gas (AsH_3) reaction with a mercuric bromide test strip. The reacted test strip is then compared visually to a reference scale (0, 10, 25, 50, 100, 250, and 500 ppb) and hence, the discretized distribution of the data (van Geen et al., 2005). Despite the discretization of arsenic concentrations, the dataset has been effectively leveraged using statistical learning techniques (van Geen et al., 2006).

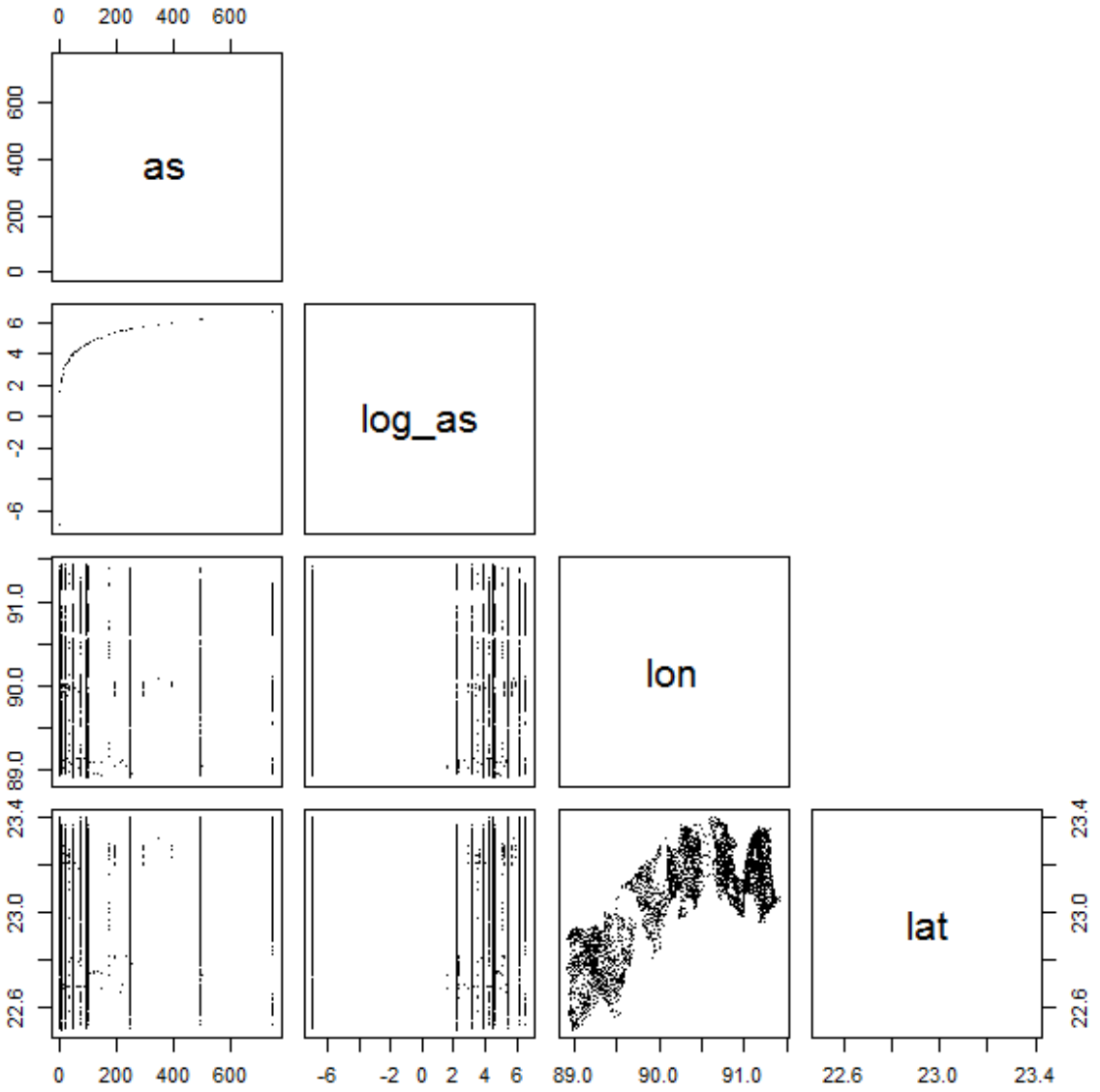

Figure 2 | Pairs plot of processed dataset. Variables include arsenic concentration (as), the log of arsenic concentration (log(as)), longitude (lon), and latitude (lat). The true distribution of arsenic concentration is continuous. However, here we can see the effect of discretized sampling methods (test strips).

Predictor Selection

Predictors were selected using best subset selection. We chose this method as there were only three predictors and this was computationally simple to achieve. In the future, we will have a larger set of predictors at which point a stepwise selection method may be more appropriate (James et al., 2013). All predictors were selected regardless of the metric used to evaluated the best model (Fig. 3). However, C_p and BIC sometimes favored excluding longitude. After extensive model testing, models including

longitude always outperformed ones excluding longitude. Therefore, we chose to include longitude in our final analysis.

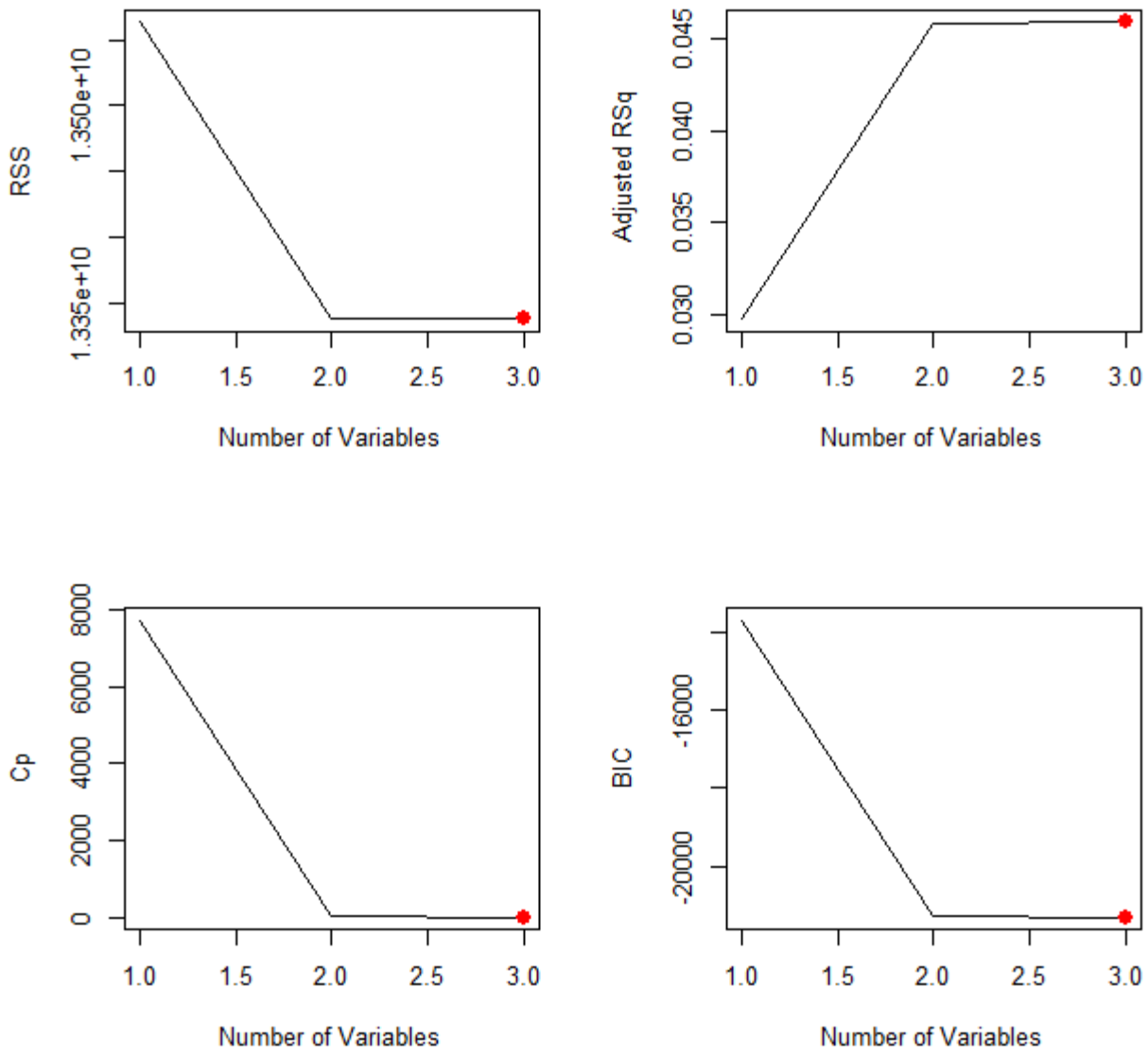

Figure 3 | Plots of predictor selection methods. All metrics suggest the inclusion of all three variable – depth, lat, lon.

Regression Modeling of Arsenic Concentrations

We modeled arsenic concentrations using linear regression, regression trees, non-linear polynomial regressions, and support vector machines (SVM). We also applied log transformations to the response variable in some of our models to simplifying any complex relationships with depth or latitude. Furthermore, we used k-fold cross-validation (K=10; K=100) to validate our models. We then

compared the model fits based on RMSE which is in terms of arsenic concentration (ppb) (Fig. 4; Table 2).

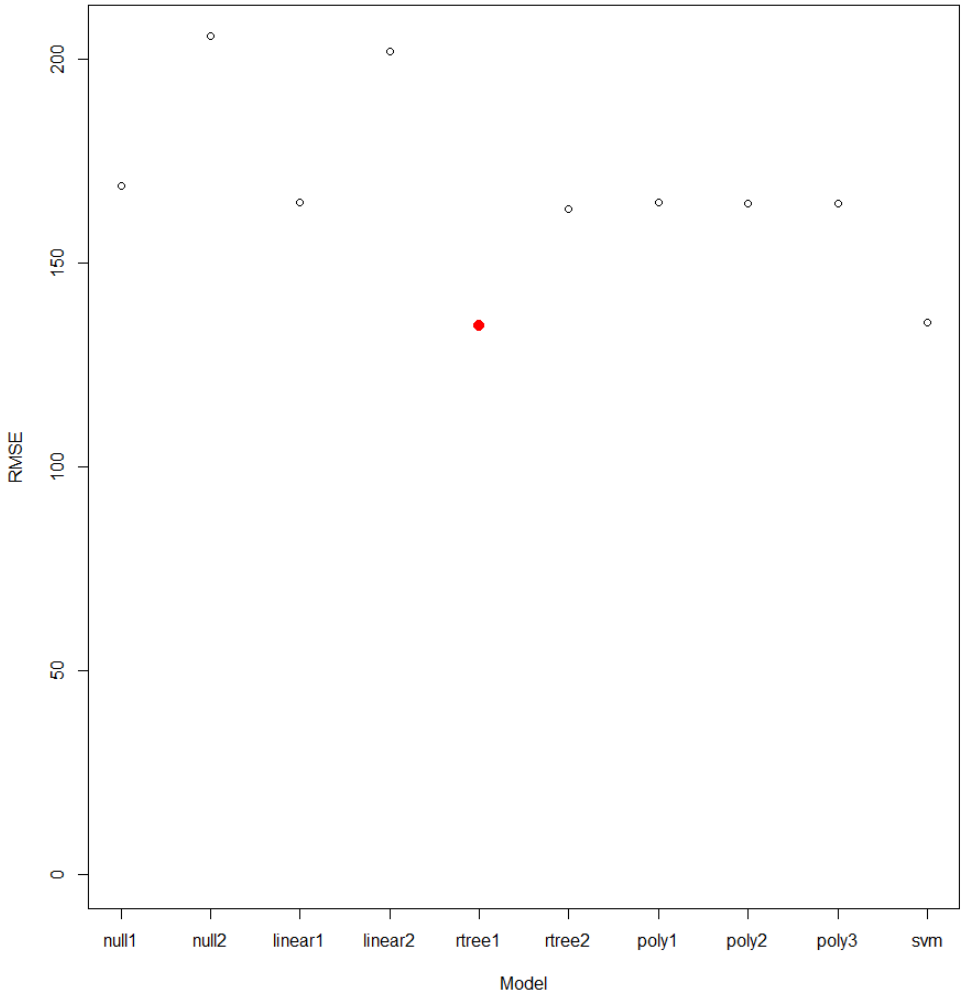

Figure 4 | Plot of all models with their associated root-mean-square error. Most models have RMSEs of 170-200 though both the SVM and regression tree models perform slightly better.

Model Type	Formula	RMSE
Null	mean(As)	169.0042
Null	mean(log(As))	205.7938
Linear	As ~ depth + lon + lat	164.8969
Linear	log(As) ~ depth + lon + lat	202.0204
Tree	As ~ depth + lon + lat	134.6268
Tree	log(As) ~ depth + lon + lat	163.1576
Polynomial	As ~ poly(depth,3(+ lon + lat	164.8131
Polynomial	As ~ poly(depth,4) + lon + lat	164.4951
Polynomial	As ~ depth + lon + poly(lat,2)	164.5765
SVM	As ~ depth + lon + lat (cost=100; gamma=10)	135.4405

Table 2 | Table of model types with formulas and associated RMSE. The models with the most predictive power were regression trees and support vector machines.

The most successful models were the tree regression and the SVM models. Though, root-mean-square errors (RMSE) were very high for all models. The overall best model was a tree model (As ~ depth + lon + lat) (Fig. 5).

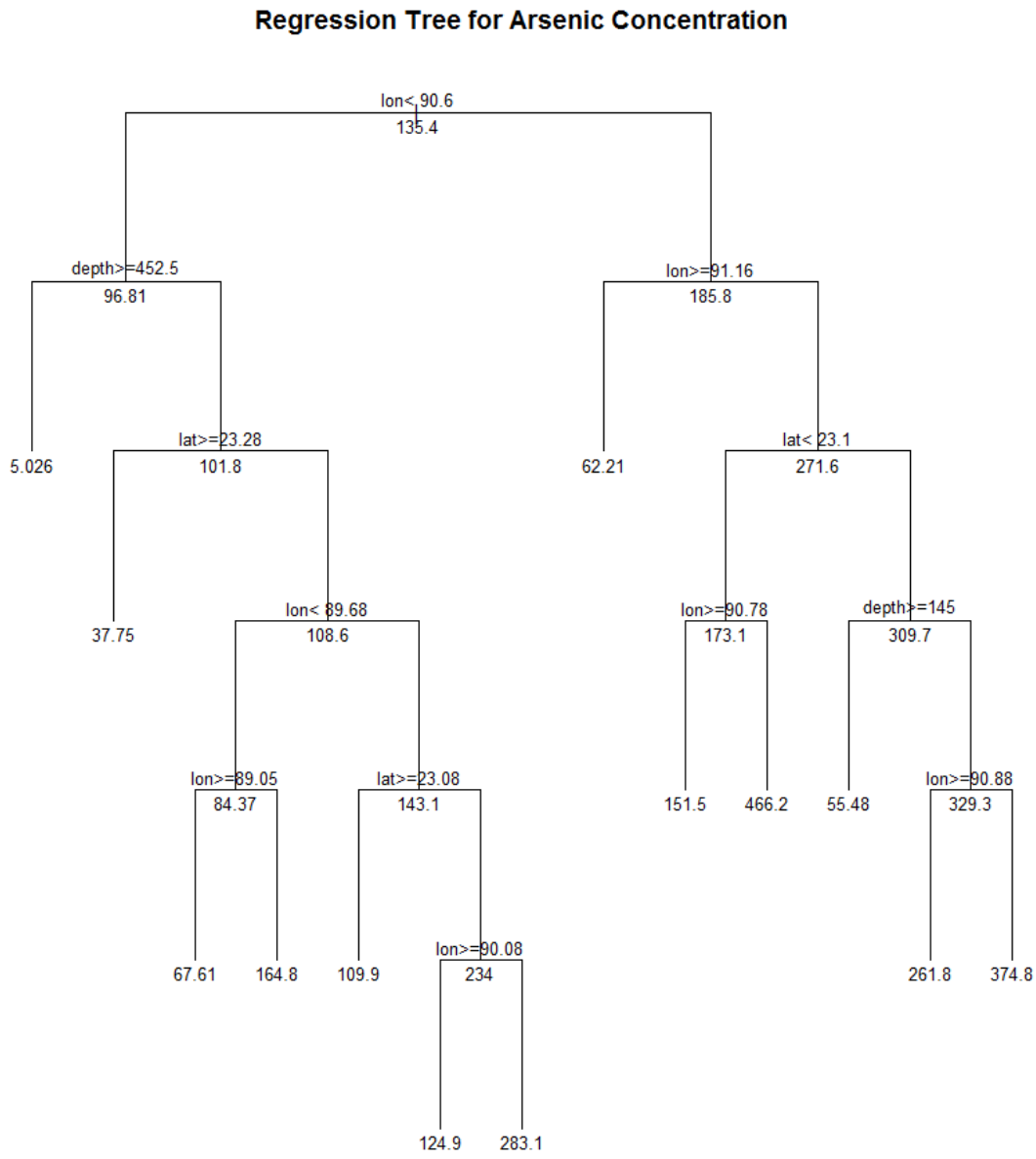

Figure 5 | Regression tree model with lowest RMSE. Some larger scale patterns become apparent (e.g. deep wells and wells found in the NE portion of Bangladesh are low in arsenic concentration).

IV. Discussion

Our modeling effort were quite unsuccessful at predicting arsenic content within a meaningful range. Safe values of arsenic are 0 to 50 ppb, however, our best model had RMSEs of ~130. A successful model will have a resolution of 1-10 ppb. Despite these issues, these models served as a first attempt. There are a number of ways through which we can improve our models.

Predictors

Our model only utilized 3 predictor variables – longitude, latitude, and well depth. This dataset was quite basic and can easily be expanded to use more predictors. Presently, collaborators are processing stratigraphic core samples along our modeled transect. This data will contain geologic data with depth. From this, we can use spatial interpolation methods (e.g. kriging) to develop a three-dimensional stratigraphic model. We can then extract this information to be used as a predictor within our model.

Another consideration is that arsenic content tends to be higher in sediment that has been deposited from the Ganges and much lower if not nonexistent in Brahmaputra sediment (BGS, 2001). We can use proximity to the current and historic flow paths of the Ganges and Brahmaputra as a predictor. Furthermore, we could combine this proximity data with our stratigraphic data to yield distance in three-dimensional space to Ganges sediment packets. Sediment provenance can be derived from strontium signals in sediment (Pickering et al., 2014). Lastly, depth to Pleistocene surface (as measured through core samples) has been shown to be a good indicator of arsenic content (Ravenscroft et al., 2005).

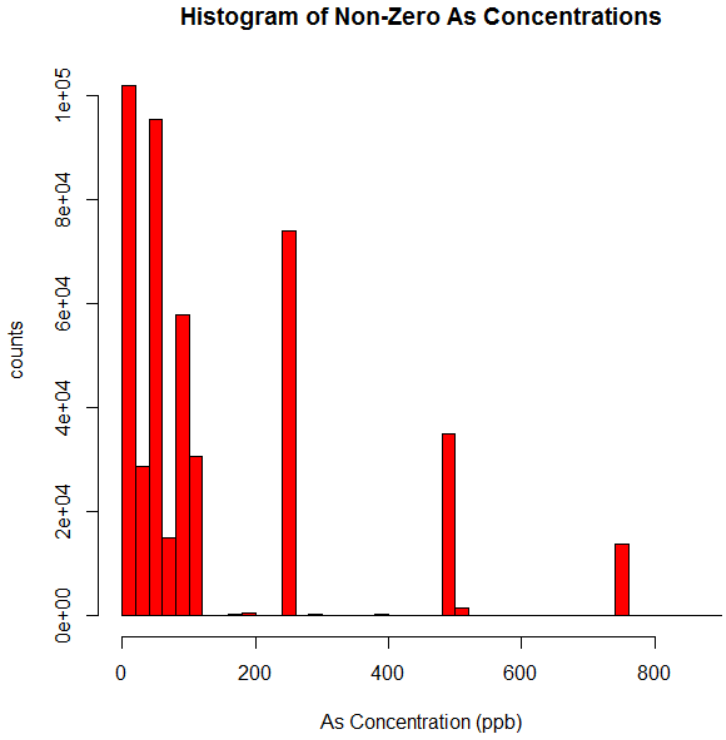

Figure 6 | Histogram of arsenic concentrations. Concentrations appear multi-modal. This is due to sampling using test strips.

Data Integrity

More analysis must be done into understanding the integrity of the data. Presently, the data is discretized (Fig. 6), despite arsenic content realistically being continuous. However, it has been shown that this dataset can be used effectively (van Geen at al., 2006).

Classification

A better approach to modeling this data would be a classification scheme as we only need to know if a well is safe or unsafe. In our dataset, we created binary response variables around threshold values (As > 10 ppb; As > 50 ppb) to indicate well safety. Furthermore, we attempted to classify wells using logistic regression, but were unsuccessful in building a stable model largely due to limited predictor variables. As we build the dataset and add more predictors, we believe this will yield more predictive accuracy than the regression methods described in this study. While not useful in a predictive sense, this analysis has provided insight into areas of improvement that will increase the likelihood of a successful predictive model in subsequent model iterations.

V. References

Bangladesh Arsenic Mitigation Water Supply Project (BAMWSP). (2001). Rapid Assessment of Household Level Arsenic Removal Technologies.

British Geological Survey (BGS). (2001). Arsenic contamination of groundwater in Bangladesh (British Geological Survey Technical Report). (D. G. Kinniburgh & P. L. Smedley, Eds.). Keyworth: British Geological Survey. Retrieved from http://www.bgs.ac.uk/research/groundwater/health/arsenic/Bangladesh/reports.html

Fendorf, S., Michael, H. A., & van Geen, A. (2010). Spatial and Temporal Variations of Groundwater Arsenic in South and Southeast Asia. Science, 328(5982), 1123–1127. https://doi.org/10.1126/science.1172974

Gelman, A., Trevisani, M., Lu, H., & van Geen, A. (2004). Direct data manipulation for local decision analysis as applied to the problem of arsenic in drinking water from tube wells in Bangladesh. Risk Analysis, 24(6), 1597–612. https://doi.org/10.1111/j.0272-4332.2004.00553.x

Harvey, C. F., Swartz, C. H., Badruzzaman, a B. M., Keon-Blute, N., Yu, W., Ali, M. A., … Ahmed, M. F. (2002). Arsenic mobility and groundwater extraction in Bangladesh. Science (New York, N.Y.), 298(5598), 1602–6. https://doi.org/10.1126/science.1076978

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 103). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kinniburgh, D. G., & Kosmus, W. (2002). Arsenic contamination in groundwater: some analytical considerations. Talanta, 58(1), 165–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18968743

Pickering, J. L., Goodbred, S. L., Reitz, M. D., Hartzog, T. R., Mondal, D. R., & Hossain, M. S. (2014). Late Quaternary sedimentary record and Holocene channel avulsions of the Jamuna and Old Brahmaputra River valleys in the upper Bengal delta plain. Geomorphology, 227(October 2013), 123–136. https://doi.org/10.1016/j.geomorph.2013.09.021

Ravenscroft, P., Burgess, W. G., Ahmed, K. M., Burren, M., & Perrin, J. (2005). Arsenic in groundwater of the Bengal Basin, Bangladesh: Distribution, field relations, and hydrogeological setting. Hydrogeology Journal, 13(5–6), 727–751. https://doi.org/10.1007/s10040-003-0314-0

van Geen, A., Cheng, Z., Seddique, A. A., Hoque, M. A., Gelman, A., Graziano, J. H., … Ahmed, K. M. (2005). Reliability of a Commercial Kit To Test Groundwater for Arsenic in Bangladesh. Environmental Science & Technology, 39(1), 299–303. https://doi.org/10.1021/es0491073

van Geen, A., Trevisani, M., Immel, J., Osman, N., Cheng, Z., & Ahmed, K. M. (2006). Targeting Low-arsenic Groundwater with Mobile-phone Technology in Araihazar, Bangladesh. Journal of Health, Population and Nutrition, 24(3), 282–297.

van Geen, A., Ahsan, H., Horneman, A. H., Dhar, R. K., Zheng, Y., Hussain, I., … Graziano, J. H. (2002). Promotion of well-switching to mitigate the current arsenic crisis in Bangladesh. Bulletin of the World Health Organization, 80(9), 732–7. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567605&tool=pmcentrez&rendertype=abstract

Winkel, L., Berg, M., Amini, M., Hug, S. J., & Annette Johnson, C. (2008). Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. Nature Geoscience, 1(8), 536–542. https://doi.org/10.1038/ngeo254

World Health Organization (WHO). (2011). Arsenic in drinking water: Background document for development of WHO Guidelines for Drinking-water Quality. Geneva, Switzerland.

VI. Appendix

```
### LOAD LIBRARIES ###
library(pacman)
pacman::p_load(ggplot2,dplyr,rpart)
library(e1071)
library(leaps)

### LOAD DATA ###

setwd('C:/Projects/Vanderbilt/bng_arsenic/')

as.df = tbl_df(read.csv('data/bamwsp.csv',header=T))
as.df = rename(as.df,As_ppb = arsenic_ppb)

xy.df = tbl_df(read.csv('data/Mouza_25km.csv'))
xy.df = rename(xy.df,geocode = GEO2,lon = X_COORD,lat = Y_COORD)

df = left_join(as.df,xy.df,by='geocode')
df$geocode = as.factor(as.df$geocode)

d = df %>%
  filter(!is.na(lon)) %>%
  filter(!is.na(lat)) %>%
  filter(depth_ft>0) %>%
  filter(depth_ft<3000) %>%
  filter(As_ppb<1000) %>%
  select(As_ppb,depth=depth_ft,lon,lat,geocode) %>%
  mutate(As_10 = as.factor(ifelse(As_ppb>=10,1,0))) %>%
  mutate(As_50 = as.factor(ifelse(As_ppb>=50,1,0))) %>%
  mutate(as = ifelse(As_ppb==0,0.001,As_ppb),log_as=log(as))

### VISUALIZE DATA ###

# Pairs plot with 10% of data
d.sample = sample_frac(d,0.1,replace=F)
pairs(~as+log_as+lon+lat,data=d.sample,upper.panel=NULL,pch='.')

# Histogram of >0 As Concentrations
d.gt0As = filter(d,as>0)
hist(d.gt0As$as,breaks=60,xlab='As Concentration
(ppb)',ylab='counts',col=2,main='Histogram of Non-Zero As Concentrations')

## Plot As Concentrations vs Other Variables

# Plot As concentrations vs Depth
plot(d$as,-d$depth,axes=F,ann=F,col='darkgrey')
axis(3)
axis(2)
box()
mtext('Arsenic Concentration (ppb)',side=3,line=3,cex = par("cex.lab"))
mtext('Depth (ft)',side=2,line=3,cex = par("cex.lab"))

# Plot As concentrations vs Lon
plot(d$lon,d$as,xlab = 'Longitude',ylab='Arsenic Concentration
(ppb)',col='darkgrey')
```

```r
# Plot As concentrations vs Lat
plot(d$as,d$lat,ylab = 'Latitude',xlab='Arsenic Concentration
(ppb)',col='darkgrey')

## Determine variables using Best Subset Selection

regfit.full = regsubsets(as~depth+lon+lat,d)
reg.summary = summary(regfit.full)

# Plot model statistics

par(mfrow=c(2,2))
plot(reg.summary$rss ,xlab="Number of Variables",ylab="RSS",type="l")
best.rss = which.min(reg.summary$rss)
points(best.rss,reg.summary$rss[best.rss],col="red",cex=2,pch=20)

plot(reg.summary$adjr2 ,xlab =" Number of Variables ",ylab=" Adjusted
RSq",type="l")
best.r2 = which.max(reg.summary$adjr2)
points(best.r2,reg.summary$adjr2[best.r2],col="red",cex=2,pch=20)

plot(reg.summary$cp ,xlab ="Number of Variables",ylab="Cp",type='l')
best.cp=which.min(reg.summary$cp)
points (best.cp,reg.summary$cp[best.cp],col="red",cex=2,pch=20)

plot(reg.summary$bic,xlab ="Number of Variables",ylab="BIC",type='l')
best.bic = which.min(reg.summary$bic)
points(best.bic,reg.summary$bic[best.bic],col="red",cex=2,pch=20)

par(mfrow=c(2,2))
plot(regfit.full,scale ="r2")
plot(regfit.full,scale ="adjr2")
plot(regfit.full,scale ="Cp")
plot(regfit.full,scale ="bic")

### MODELS ###

# num of folds
k <- 10

# create index
folds <- rep_len(1:k, nrow(d))

# shuffle index using sample
folds <- sample(folds, nrow(d))

# number of models
nmodels <- 11

model.error <- matrix(data=NA, nrow=nrow(d), ncol=nmodels)
for (i in 1:k){
  # select rows in fold
  fold <- which(folds==i)

  # subset training data
  train <- d[-fold,]
```

```
  # model 1 = null model of As
  null1 <- mean(train$as)

  # model 2 = null model of log(As)
  null2 <- mean(train$log_as)

  # model 3 = linear model of As
  linear1 <- lm(as ~ depth + lat + lon, data=train)

  # model 4 = linear model of log(As)
  linear2 <- lm(log_as ~ depth + lat + lon, data=train)

  # model 5 = regression tree of As
  rtree1 <- rpart(as ~ depth + lat + lon, data=train)

  # model 6 = regression tree of log(As)
  rtree2 <- rpart(log_as ~ depth + lat + lon, data=train)

  # model 7 = polynomial model
  poly1 <- lm(as ~ poly(depth,3) + lat + lon, data=train)

  # model 8 = polynomial model
  poly2 <- lm(as ~ poly(depth,4) + lat + lon, data=train)

  # model 9 = polynomial model
  poly3 <- lm(as ~ depth + poly(lat,2) + lon, data=train)

  # model 10 = SVM
  svm <- svm(as ~ depth + lat + lon, data=train)

  # test set
  test <- d[fold,]

  # store observed and predictions for folds
  model.error[fold,1] = exp(test$log_as)
  model.error[fold,2] = null1
  model.error[fold,3] = exp(null2)
  model.error[fold,4] = predict(linear1,test)
  model.error[fold,5] = exp(predict(linear2,test))
  model.error[fold,6] = predict(rtree1,test)
  model.error[fold,7] = exp(predict(rtree2,test))
  model.error[fold,8] = predict(poly1,test)
  model.error[fold,9] = predict(poly2,test)
  model.error[fold,10] = predict(poly3,test)
  model.error[fold,11] = predict(svm,test)
}

# store errors in a dataframe
model.results <- data.frame(obs=model.error[,1],
                            null1=model.error[,2],
                            null2=model.error[,3],
                            linear1=model.error[,4],
                            linear2=model.error[,5],
                            rtree1=model.error[,6],
                            rtree2=model.error[,7],
                            poly1=model.error[,8],
                            poly2=model.error[,9],
```

```
                                poly3=model.error[,10],
                                svm=model.error[,11])

# rmse function
rmse <- function(obs,est){sqrt(mean((obs-est)^2,na.rm=T))}

# calculate rmse for models
rmse.all <- apply(model.results[,2:nmodels],2, rmse, obs=model.results[,1])

### PLOTS ###

# Models with RMSE
plot(rmse.all,xaxt='n',xlab='Model',ylab='RMSE',ylim=range(0:max(rmse.all)))
axis(1,at=1:10,labels=names(rmse.all))
points(5,rmse.all[5],col="red",cex=2,pch=20)

# Best Model
plot(rtree1, uniform=TRUE, main="Regression Tree for Arsenic Concentration")
text(rtree1, all=TRUE, cex=.8)
```