Warren Keil & Will Cranford

Dr. Cook

Project Write-up

November 22, 2017

The Effects of Injection Drilling on Seismic Activity

Our project consists of analyzing the data about injection well drilling and its effect on seismic activity in the state of Oklahoma. We have a couple of reasons for choosing this topic to research. For one, the activity of fracking -another term associated with injection well drilling- has received a lot of media attention in recent years, and many speculate that it is directly responsible for the increase in the number earthquakes in the Oklahoma area. The other reason we are interested in this problem is that predicting earthquakes is an inherently difficult and complicated problem. Researchers have been trying for decades to effectively model and predict earthquakes but without much success. We feel that even though this problem is acknowledgedly hard, it is still a very important one. Earthquakes sometimes cause large number of casualties. Successfully modeling earthquakes can potentially save a lot of lives.

We have a few specific questions that we are aiming to provide answers to. Is the location and magnitude of earthquakes dependent on injection well activity? Is the number of earthquakes in a certain area dependent upon injection well activity? And is the magnitude of earthquakes in our region influenced by injection well activity? These three questions are the main focus of our project. The main predictors we will be studying are the number of wells drilled, the psi of the fluid pumped into the ground at each well, and the number of barrels per day each well pumps out of the ground.

The dataset used in our study consists of two comma-separated-value files obtained from kaggle.com. One file contained injection well data while the other contain earthquake data. Each file contained around 16,000 observations and 20 predictors.

Methodology

Since the three questions that we are trying to answer are very different in nature, we will provide the methodology and results independently for each of the three approaches we took to our problem. The three different approached will be labeled as time series, magnitude prediction, or locational approach. We will provide the whole methodology and results for each approach before

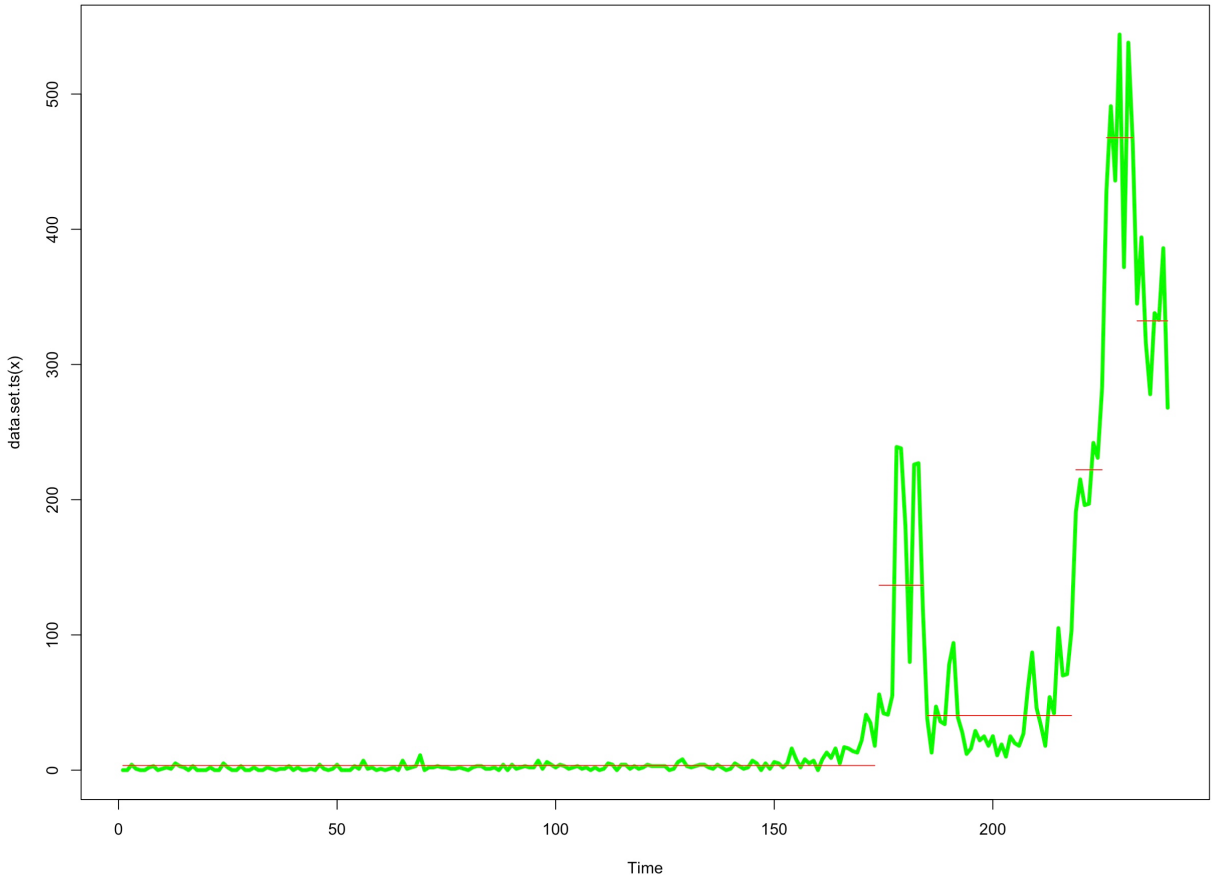
Time Series

Data Processing - (Time Series). When observing the data, we first noticed that the two files contained different date formats and at least one of them was not in the standard format. To remedy this problem we quickly parsed the content of these columns and reformatted into the standard date format. We then chose to create a new predictor for each dataset to be the number of months since the 1936. We chose this because 1936 was the date of our earlier observation and having a column that has an integer value for months since that date allows us to easily manipulate time periods. We then had to throw out just a few observations that had no coordinate information. Any observation of data without a location would not be useful for our analysis. The next issue we had to deal with was three or four variables that definitely should have been classified as numeric were shown as factor variables in R. After re-categorizing these predictors many times, we were finally able to be stored as numeric variables. We also noticed that some of both

our well data and earthquake data looked incomplete. For example, we saw that many wells has zero or missing psi data. To fix this, we made every observation that had a zero or missing psi element equal to the mean of every other observation that had complete information. This is also known as imputation. We are aware that this may artificially influence our results but we feel it will be less damaging than leaving missing or zero values. We applied this method of filling in incomplete data to the magnitude predictor, and to the barrels per day predictor.

We then decided to make a new data frame that had all of the important information categorized with respect to time in months. To do this we created new variables based off existing variables to portray information per month. These variables included total earthquakes per month, total medium and big earthquakes per month of magnitude 2.5 or higher, total big earthquakes per month of magnitude 3.0 or higher, total wells drilled per month, total wells in the past six months, the cumulative number of wells since 1936, total psi of fluids pumped in ground, total psi of fluids pumped in ground in last six months, cumulative psi of fluids pumped in ground since 1936, total barrels of fluid pumped in ground, total barrels of fluid pumped in ground over past six months, and the cumulative number of barrels pumped in ground since 1936.

Techniques - (Time Series). The first technique used to analyze the time series data was changepoint analysis. Changepoint analysis is a technique used to detect significant changes in the mean, variance, or both. We felt this technique was appropriate because is widely publicized information that the number of earthquakes have dramatically increased the Oklahoma area. With our changepoint analysis, we can give precise times of when the mean number of earthquakes did increase. Our changepoint analysis results in R found there was significant changes in



the mean number of earthquakes in Oklahoma during March and October of 2014. We used this information to help us construct of next technique, the ARIMA model.

The ARIMA model is model that can attempt to forecast data in a time-series format. ARIMA itself stands for Auto Regressive Integrated Moving Average. The ARIMA model makes various assumptions about data such as it is stationary and non-seasonal. So we had to make adjustments to our data before running the model. The first thing we did was adjust our data for seasonality. We then tested for stationarity with the Augmented Dickey-Fuller Test. We then ran the auto.arima function in R and it suggested to use a second degree auto-regressive function, and first degree differencing and a second degree moving average. However, we found that we

still failed the Dickey-Fuller test for stationarity when using these parameters. On the ACF plot, we found the 8th lag number was causing our model to fail the test. We then chose eight as the order of our auto-regressive function since this was the order of the lag that was giving us problems. We also elected to use 2 for our degree of differencing and 0 for the degree of moving averaging. We were justified in doing this because the ACF plot showed no signs of non-stationarity after making these changes and the Dickey-Fuller test gave favorable results also.

Comparison Between Methods - (Time Series). The main differences between our time series methods of changepoint analysis and ARIMA and the other methods used in this project such as random forests and support vector machines is that the ARIMA model makes some strong assumptions based off time. That is, the model assumes that each data point is correlated and to some extent dependent upon previous data points. The ARIMA model predicts future values in sequential order. Thus the further a prediction value is away from the time of prediction, the greater the standard deviation for that prediction will be for that prediction.

R packages -(Time Series). The packages used for our time series analysis include:

changepoint - Allows us to run the changepoint function in R.

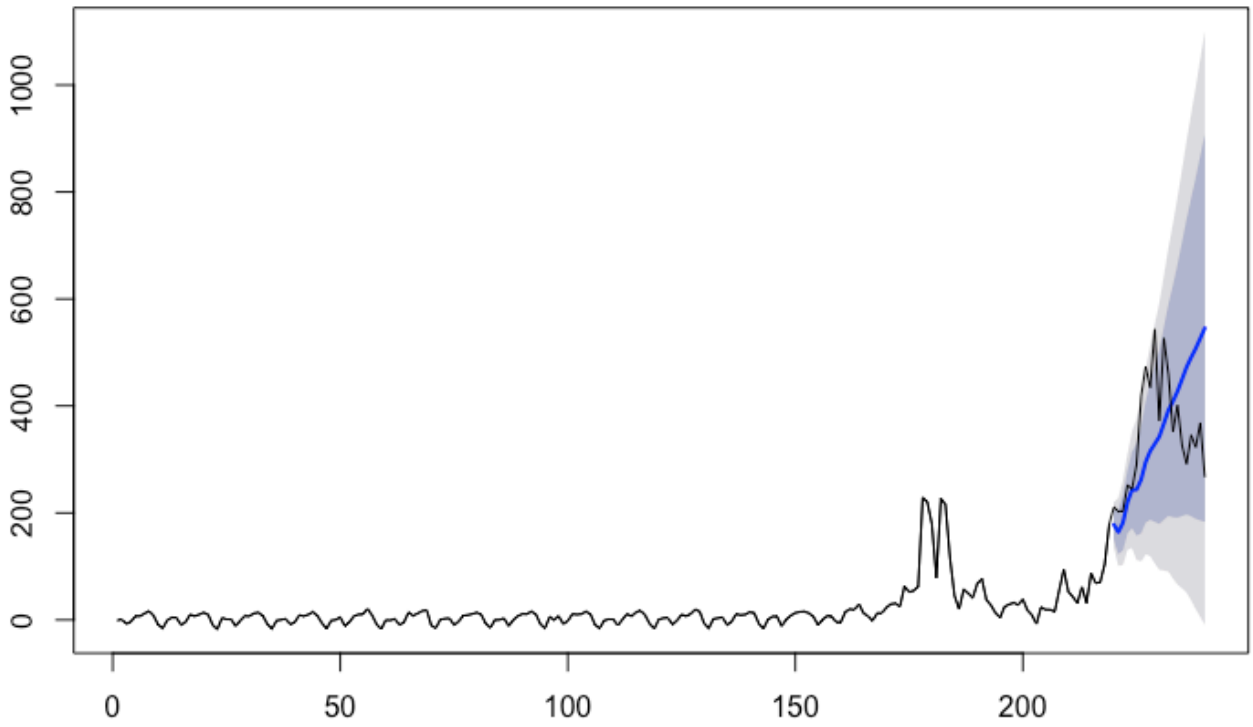
forecast - Contains the arima function and its associated forecast function.

tseries - Contains multiple functions that test for seasonality and stationarity.

Results -(time series Approach). We got significant results from both the changepoint analysis and the ARIMA model. The changepoint analysis showed us some information that was obvious -that the number of earthquakes dramatically increased after 2009- but it also gave us useful in-

formation. We were able to use the times of change in variance in 2014 to pick a good point to try to forecast using our ARIMA model.

We also got favorable results from the ARIMA model. We correctly predicted the actual number of quakes within one standard deviation of our model's forecast. We also ended with root mean squared error of 21.12696.



Magnitude Prediction

Data Processing - (Magnitude Prediction) The following method of data organization was used to predict whether or not an earthquake is significant (magnitude > 2.5) given the amount of fluid injected and pressure that it is injected within a certain radius around the earthquake. First, a function was created that calculated the distance from a point to another point using the longi-

tude and latitude values of each point. Then a nested for loop with a conditional if statement was used to segregate the injection well dataset. The loop took each individual earthquake and picked out each injection well that was within a certain distance from the earthquake and was in operation prior to the earthquake occurring. It separated these wells into datasets so that each earthquake had its own dataset of injection wells. The datasets were then saved into a list. This list contained 7966 datasets since this is the number of earthquakes there were. We ran the loop for distances of 5,10, and 15 miles in order to see what the maximum distance is that a well can effect an earthquake. The number of total iterations for each loop was the number of earthquakes, 7966, times the number of injection wells, 9636. This amounted to almost 77 million iterations which took around 2 and a half hours for the computer to run through. Since this was done for three different distances, the total iterations was around 254 million and a total runtime of 7 and a half hours. We then used the “lapply” function to apply the “sum” and “mean” functions to each dataset that pertained to each earthquake in order to obtain the “bbls_sum”,”bbls_mean”,”psi_sum”, and “psi_mean” variables. The “bbls_sum” variable gave the total amount of average fluid injected within a certain distance from the earthquake. The “bbls_ave” variable gave the average amount of fluid injected per well within a certain distance from each earthquake. The “psi_sum” variable gave the total amount of pressure that the fluids were injected, and the “psi_ave” variable gave the average rate per well that the fluids were injected within a certain distance from the earthquake. We then used the “unlist” function on each variable and the “as.vector” function on each variable. After that, we used “cbind” to attach the values for each variable to the earthquake data set. This gave us a dataset where each earthquake had values that showed the total amount of fluid injected and rate injected within a certain dis-

tance, and the average amount of fluid injected and rate injected per well within a certain distance. We then created a dummy variable in order to separate significant earthquakes from insignificant earthquakes. The dummy variable gave the value of "0" if the magnitude for a given earthquake was less than 2.5, and it gave the value of "1" if the magnitude of the earthquake was greater than or equal to 2.5.

Techniques - (Magnitude Prediction) The techniques used to analyze this newly organized dataset were plotting, simple linear regression, random forests, and general additive modeling. The whole purpose of organizing the data in this particular fashion is to accurately predict whether an earthquake will be significant (magnitude >2.5) based on our drilling related predictors. In order to do this, we had to understand what variables significantly affected the magnitude of the earthquakes. First, we made simple box plots showing the values of depth and magnitude of each earthquake in relation to total and average fluid injected and total and average psi amounts. Next, we used simple linear regression which gave us an idea of what predictor variables are significant for magnitude and depth of each earthquake. We then used general additive modeling to show the relationship of each individual variable to whether the earthquake was significant or not. Lastly, we used random forests to predict whether the earthquake was significant given the values of our predictor variables for each earthquake.

Comparison of Methods - (Magnitude Prediction) The box plots were used to give a visualization of the contrast between the magnitudes of earthquakes where there was minimal injection well activity versus magnitudes of earthquakes with increasing amounts of injection well activity and the magnitudes of the earthquakes given the injection well activity. The simple linear model

approach was used because based off of the box plots we could infer that there was a linear relationship with injection well activity versus the magnitude of the earthquakes. Based on the p-values of the predictor variables we could see what variables had a significant relationship with the magnitude and depth of the earthquake. The general additive modeling technique was used to show a visualization of the effect each individual predictor variable had independent of the others in determining if each earthquake was significant or not. This technique affirmed our opinions we had formed from observing the box plots. The random forest technique was used to get an idea of the accuracy of our data, and the significance of each variable, and also if there were other variables that our dataset was missing that would determine the significance of an earthquake. Because of the way the random forest processes the data from the predictor variables with the tree method, if the accuracy of the predictions of the random forest model is very good, then we could confidently say that we understand what variables are effecting the magnitude of the earthquakes.

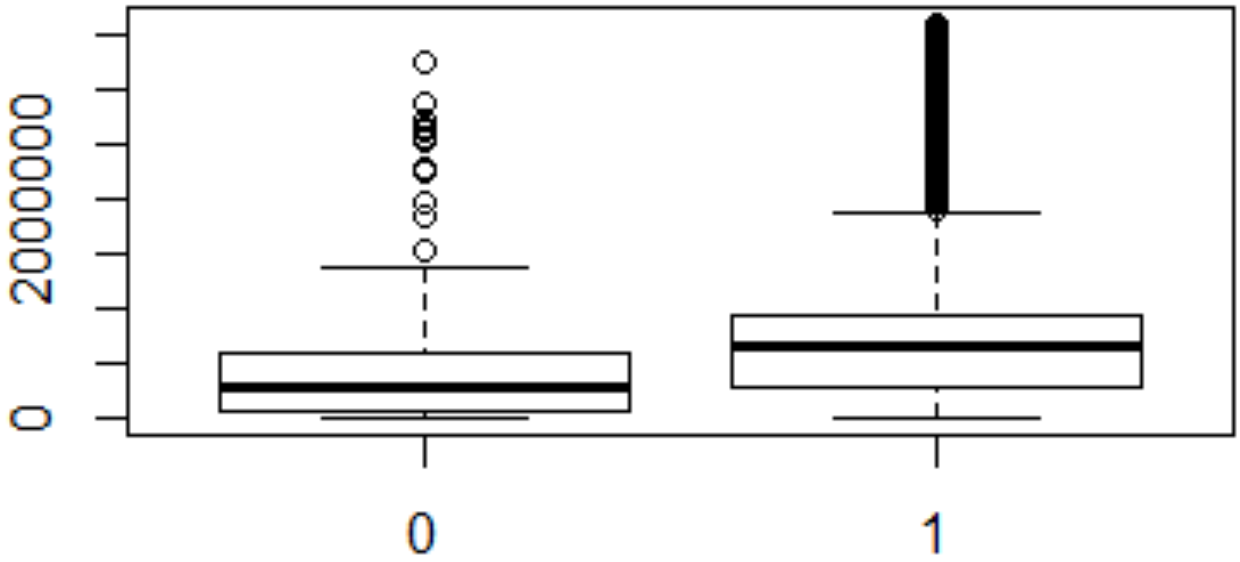
R Packages - (Magnitude Prediction) The R packages used in the magnitude prediction were:

plyr - This lets us manipulate and clean data with ease.

lubridate - This package contains functions to change date formats.

randomForest - This contains the random forest model.

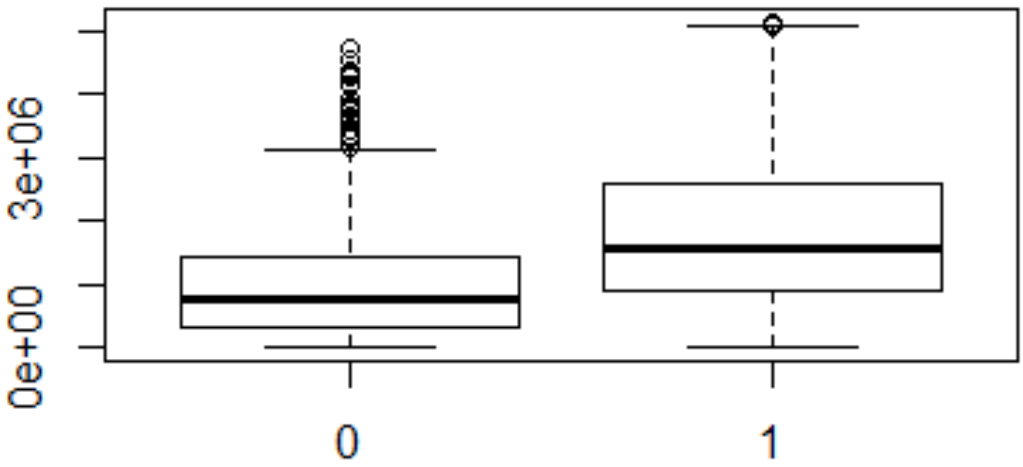
gam - The contains the generalized additive model.



Results- (Magnitude Prediction) The box plots showed significant evidence that the amount of fluid injected within a 5, 10, and 15-mile radius significantly effects whether an earthquakes magnitude will be above or below 2.5.

The above plot shows the amount of fluid injected within a 10-mile radius of each earthquake and whether or not the magnitude is equal to or above 2.5. The "1" is the earthquakes that have a magnitude at or above 2.5. The amount of fluid injected for the average amount of significant earthquakes is higher than the amount of fluid injected for the average amount of insignificant earthquakes. We could also see that a large percentage of earthquakes that had more than 2 million barrels of fluid injected in a 10-mile radius were significant. In fact, 98.5 percent of earthquakes that occurred under this condition had magnitudes higher than 2.5. The plot below shows a 15 mile radiuses around each earthquake. We could see that this plot also showed that

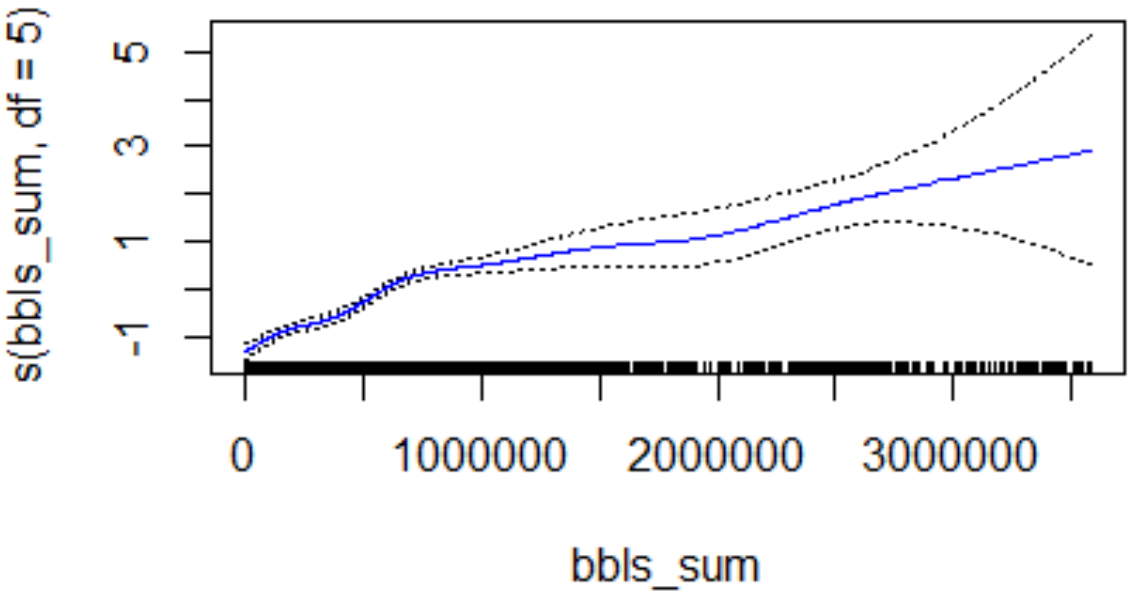
the percentage of significant earthquakes that occurred in areas with increased well activity was higher than the percentage of significant earthquakes in areas with lesser well activity.



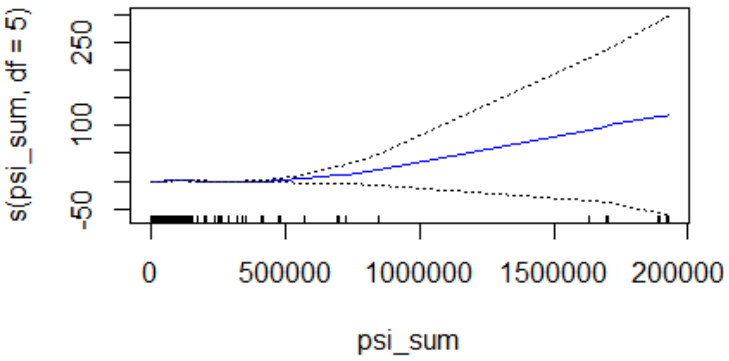
By using the linear regression model, we wanted to get an understanding of which predictor variables had a significant value in determining the magnitude of the earthquake based on the p-value. It was found that the total amount of fluid injected was a variable that had an extremely small p-value of 1.85e-09. The “bbls_ave” and “psi_sum” variables were somewhat significant as well with p-values of .201 and .381 respectively. Interestingly, the p-value for the “psi_ave” was the highest at .678. When we ran the linear model again with the depth of the earthquake included as a predictor variable it had an extremely small p-value as well at 1.34e-05. When we ran a linear regression model with depth as the response variable the predictor variable with the lowest p-value was “psi_ave” at .00618. The variable “bbls_sum” had a significant p-value as well. This shows that the predictor variable “psi_ave” along with “bbls_sum” could be useful in predicting a depth variable that could then be used in predicting the magnitude of an earthquake. The depth variable itself is not useful since it is not available until after the earthquake has occurred.

General Additive Modeling

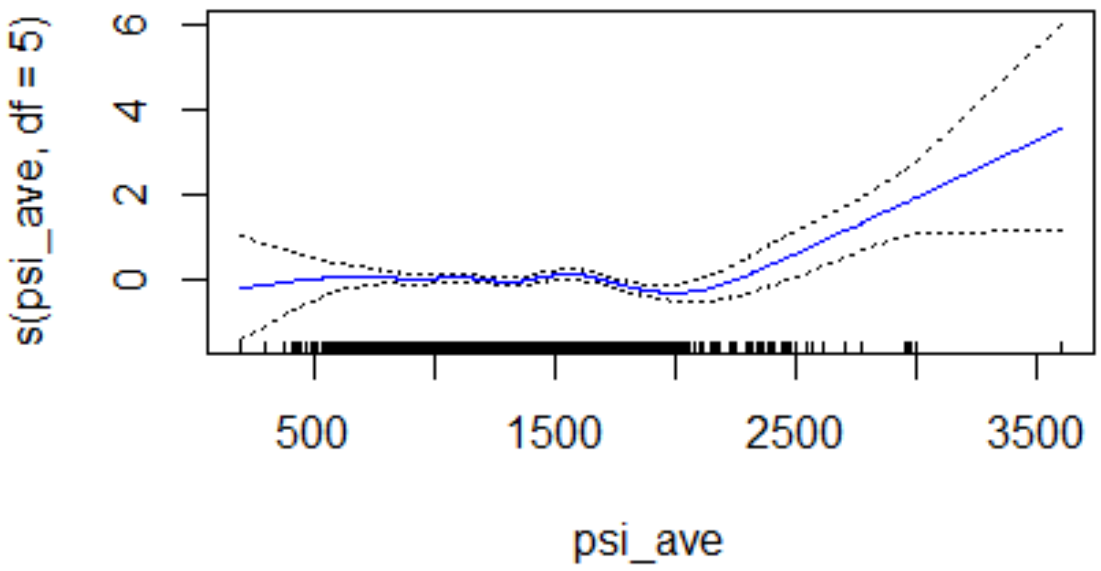
The general additive models showed us a good picture of how the predictor variables were affecting the magnitude. Using these visualizations, we can get a picture of how the magnitude increases or decreases given the value of the predictor variable.



Given the small confidence interval for the "bbls_sum" variable we can see that as this variable increase the magnitude of the earthquake increases. The "psi_sum" appears to increase as the magnitude increases however the confidence interval is so wide this relationship can't be assumed. The "bbls_ave" appears to generally increase as the magnitude increases except for the extremely large values of "bbls_ave". The confidence interval is narrow for most of the data except for these large values of "bbls_ave", so it is hard to infer if this drop in magnitude due to large



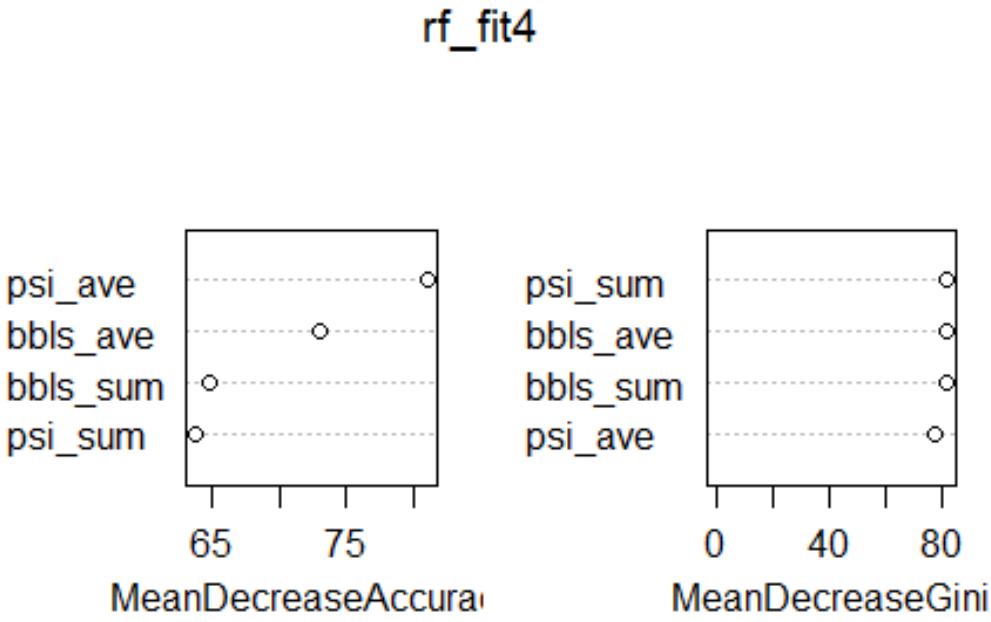
amounts of "bbls_ave" is genuine. The "psi_ave" variable is interesting because it shows relatively little change in the magnitude until the extremely large values of "psi_ave" where the magnitude seems to increase a lot. The confidence interval is somewhat large, but even taking the minimum value in the confidence interval for the larger values of "psi_ave" shows that the magnitude does increase somewhat for these values.



Random Forest

Using the random forest method we were able to see how accurately we could predict whether or not an earthquake was significant based on the predictor variables we had. We trained our random forest model on the first 6966 earthquakes that have occurred in Oklahoma and then tested our model on the last thousand of earthquakes that have occurred. We used 500 trees in our model, mtry=2, and class weight values of .04,1. This gave us a classification accuracy of 93.5 percent. It predicted 73.7 percent of the earthquakes below magnitude 2.5 correctly and the

total amount of earthquakes below 2.5 magnitude was 21.3 percent of our test data. More importantly, only 5 percent of the time the model predicted an earthquake would be below 2.5 magnitude when the actual earthquake was above 2.5 magnitude. Here is a plot of the variable importance of the random forest model.



This variable importance chart is interesting because it shows that "bbls_ave" and "psi_ave" are important variables for the random forest model. This goes against the resulting p-values from the linear regression models that showed "bbls_sum" as the most important variable.

Locational Approach

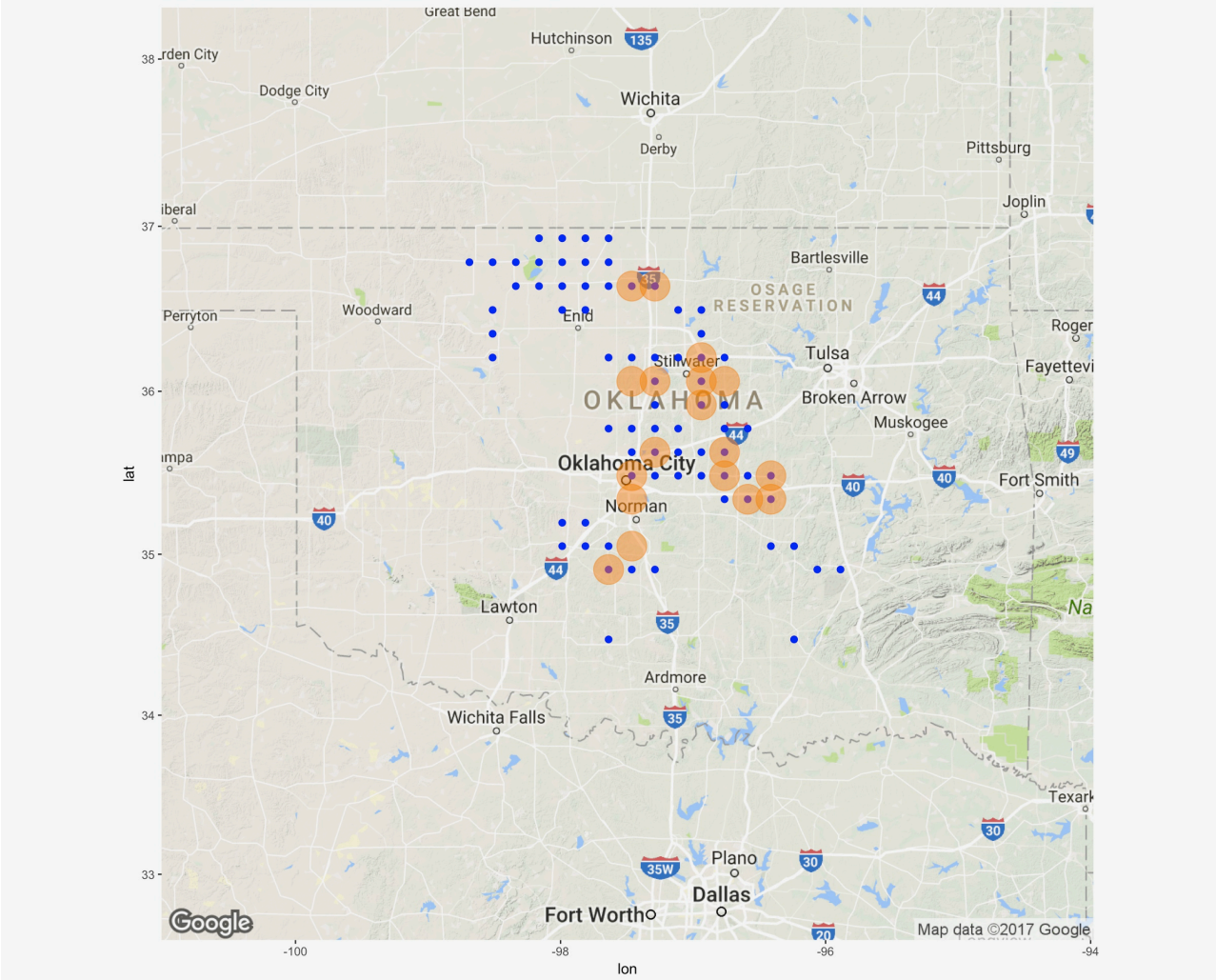
Data Processing - (Locational Approach) The locational approach to this problem involved dividing up the main part of Oklahoma into 540 evenly spaced circles. The reason we used circles to divide our area is because we used a spherical distance function in R to calculate the location of various things and we used this function to determine whether an earthquake or drilling

well was within the radius of a given point on the map. We used the rows of our data frame as the different areas of Oklahoma. Since we did not have room for the year information, we made each data frame one year and made twenty data frames. We chose to conduct this study for the times starting in January 1996 to December 2015.

To get the data in the right format to conduct our location approach, we had to do a lot of work to get the data in a suitable place. The first thing we did was to convert each variable to a numerical type. All of the variables were categorized as factors when first loaded into R. We then ran numerous loops to create new variables that stored information for each location in Oklahoma for each year. We also ran a random forest model and got interesting results. So we decided to use a stacked model and use the random forest prediction Our ending variables for each of the twenty data frames were: number of earthquaked in each location in Oklahoma, average magnitude of earthquakes for each location in Oklahoma, sum of the magnitude of earthquakes for each location in Oklahoma, number of wells drilled for each location in Oklahoma, sum of the barrels pumped in the ground for each location in Oklahoma, sum of the psi of the fluids pumped in the ground for each location in Oklahoma, the random forest prediction for the average magnitude, and the categorical variable of whether a given location would have at least one 2.5 magnitude earthquake in a given year.

Techniques - (Locational Approach) Our technique was a stacked model consisting of first running a random forest model to predict the number of earthquakes and then using the predictions of this random forest model along with the other predictors to use a support vector machine model to try to predict whether each location in Oklahoma would have at least one magnitude

2.5 earthquake. The other variables included besides the random forest prediction were the number of wells, and the sum of the barrels pumped into the ground. We ran this stacked model for each of the twenty years. We saved the prediction results of the SVM prediction as another column of each data frame in order to make it easier to access the results.



Comparison of Methods - (Locational Approach) When comparing this stacked model to our other models we found that it outperformed the random forest model. It is unfair to say the SVM outright did better than random forest since we did not try a different stacked model with random forest as our final output. It is safe to assume that we may have had similar success if we did use

random forest as the final model for our stacked model. We cannot really compare this model to the time series model since they are predicting different things and are inherently different.

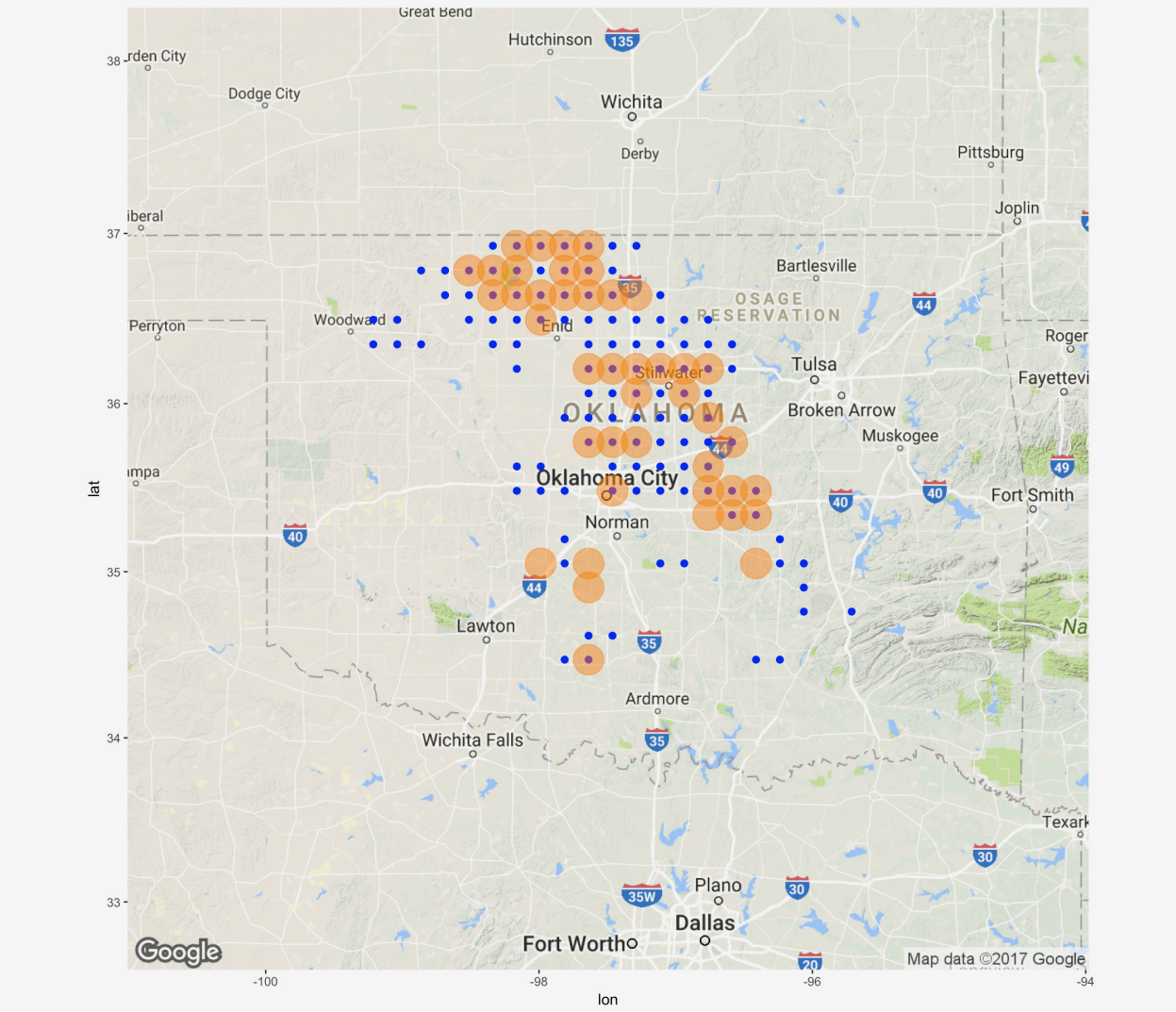
R Packages- (Locational Approach) The packages used in our locational approach are:

lubridate - contains various date formatting functions.

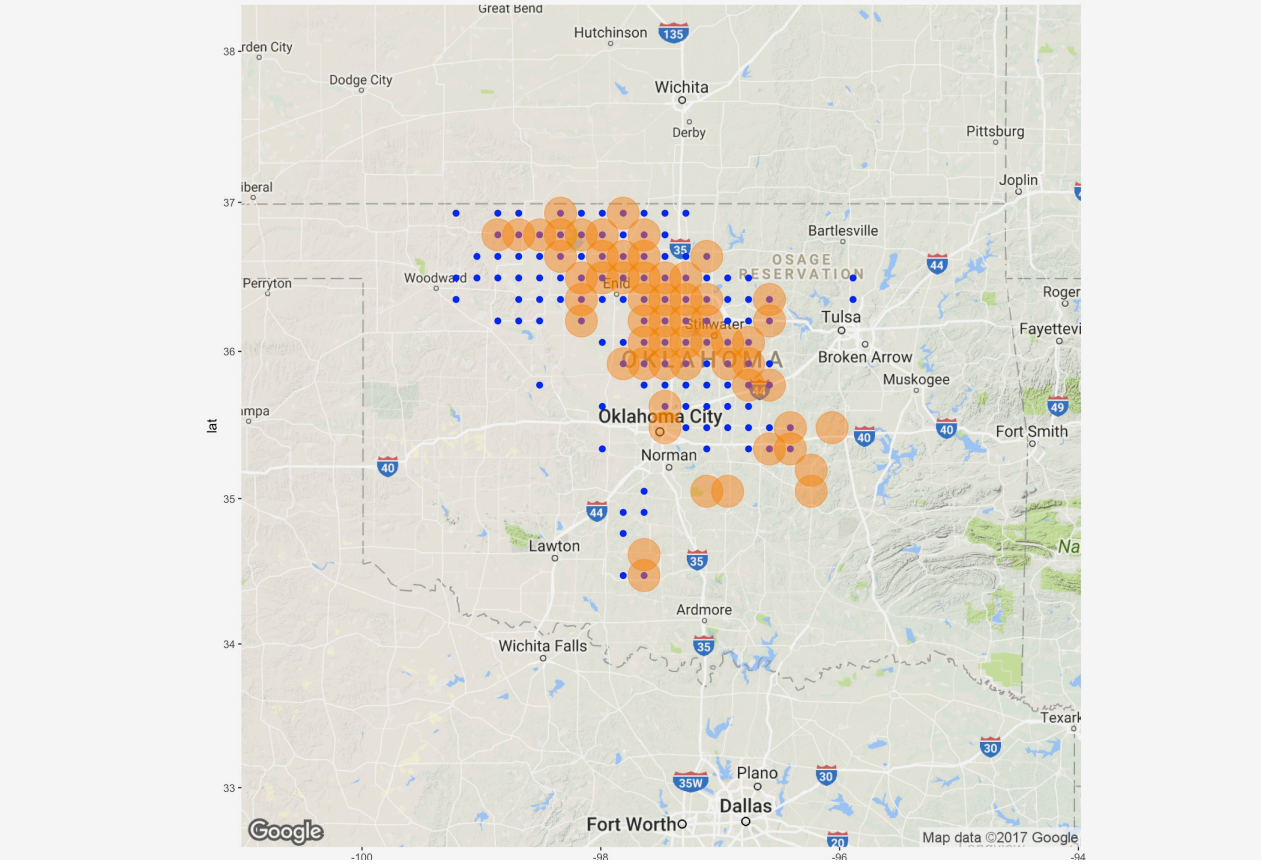
randomForest - contains the random forest fit model.

e1071 - contains the SVM fit model along with other functions.

ggplot2 - gives the user more plotting utilities and capabilities.



Results - (Locational Approach) The results obtained from our stacked model were very impressive during the years with increased activity. We found that in 2013 our model only had an error rate of 11.48%. In 2014 the error rate was 15.19%. And in 2015 the error rate was 14.07%. We also noticed the since most locations in Oklahoma do not directly experience a big earthquake every year, then then model had a smaller error rate when ever it under predicted earthquakes. Thus, we were more impressed by the result in 2014 and 2015 even though the model had a smaller error rate in 2013.



Discussion

After completing our time series analysis we felt that our model confirmed what the many people already suspected, that there has been significant increases in the size and magni-

tude of earthquakes in Oklahoma. While it is already common knowledge that seismic activity has increased in this region, the changepoint analysis given by our model gave us the exact months where the mean and variance of earthquakes has changed significantly. Changepoint analysis also gave us good suggestions on which months to use as testing data for our other time series model, the ARIMA model.

Our ARIMA model proved to be very good at forecasting the future number of earthquakes per month but only when there was a large amount of drilling and number of barrels of fluid pumped into the ground. This shows that our ARIMA model is viable for forecasting seismic activity when there a lot of drilling activity in the same region. It should be noted that most places do not have the same amount of drilling activity as Oklahoma, and our ARIMA model should not be expected to have similar results in other areas.

One thing the our ARIMA model provided besides a forecast of earthquakes was another confirmation of increase of the number of earthquakes. The steps we took in building our ARIMA model ensure that the any seasonality and non-stationarity was removed from the data. After doing this, we were able to still show a huge spike in earthquakes from 2009 to 2016. This gives further evidence that this increase in earthquakes was not part of some seasonal trend or caused by something else inherent in the data.

When evaluating how well our model predicted the size of a given earthquake, we were very please with the results. The linear model was interesting because of the relationship with “psi_ave” and depth of the earthquake. We assumed beforehand that the total amount of fluid injected would have a significant p-value as a predictor variable for magnitude, but “psi_ave” hav-

ing a significant relationship with depth was unexpected. The GAM method showed an unexpected result as well with regard to the relationship of large values of "psi_ave" with an increase in magnitude. The random forest model reinforced that the variables we were using were significantly related to the magnitude of the earthquake based on the accurate prediction rates. The box plots from the data also showed a clear relationship of "bbls_sum" to magnitude of the earthquakes for different distances around the earthquakes.

We suspect that there is evidence of some inaccuracy in the datasets we used however due to the earthquakes that had magnitudes greater than 5 not having a significant amount of well activity in their vicinity. Also, we believe that adding the predictor variable of drilling depth could increase the accuracy of our models. We believe combining drilling depth, "psi_ave", and "bbls_sum" could adequately predict the depth of the earthquake and then we could use the predicted values as another predictor variable as a substitute for actual depth values and then use the predicted depth values as a predictor variable for magnitude.

When evaluating our locational model, we feel that the results were very good. Accurately predicting the location of big earthquakes in a given year both gives evidence that earthquakes are caused by drilling activity and it suggests the thresholds for the number of wells and barrels of fluid it takes to cause big earthquakes. While our model gave impressive results, we were limited in making more progress by two factors. For one, our model only was able to predict location with good accuracy in years that had enormous amounts of drilling activity. Thus this model was not very useful in most years. The other thing that limited our ability to push this model to its full capability is time. We used a mixed stack model in this approach but we only used one

intermediate model in the ensemble. Ideally, we would use at least six or more models as predictors in our model. We would also test each one as the final model in the ensemble. We would also take more time to try to optimally tune the parameters for each model.

If we were to start this project over from the start, it would be beneficial to start making the meta ensemble predictors earlier so that we have time to properly tune each one. It would also help to introduce a wider range of create predictor variables. For example, we created a variable that counted if a location had at least one earthquake of magnitude 2.5 or higher. It may have allowed us to predict earthquakes in times of lower drilling activity if we made variables that counted earthquakes of smaller magnitude. This is just one example of something we could have done to potentially improve our model. There is a large number of possible changes we could have made to try to improve the model. These are ideas that we will implement in the future as we continue research on the effects of seismic drilling in Oklahoma.