Forest Stand Density and the Effect of Climate, Soil, Topography and Species Mixing

by Ryan Heiderman

Stat 517 – Final Project, Nov 15, 2018

**1. Problem Statement, Motivation, Research Goals:**

Density management is the control of growing stock (trees) through initial spacing, or subsequent thinning, to achieve specific management objectives, primarily to maintain and enhance forest health and productivity. In order to manipulate growing space effectively, you need to know how it is currently utilized and how a growing space manipulating will affect this utilization. (Kimsey et al.) Understanding the effect of species mixing, climate and topographic variables, as well as soil characteristics on stand density is critical to making sound forest management decisions.

Goals:

1. Prepare stand level data set from over 110,000 plots installed in timber stands across the inland NW.
2. Determine important climatic, edaphic, topographic variables and how they relate to one another.
3. Find underlying patterns related to species mixing.
4. Determine if certain variables have control on tree density (trees per acre). *Regression*
5. Determine if certain variables have control on tree size (basal area). *Regression*
6. Determine if certain variables have control on Stand Density Index. *Regression*
7. Explore possible groupings of stands based on variables. *Clustering*
8. Explore associations. Do certain variables lead to greater proportions of specific species? *Association*

**2. Data Source and Description:**

Stand data for this project was obtained through a collaborative network of Inland Northwest, USA public and private forest land management organizations, the Intermountain Forestry Cooperative, and the United States Forest Service Forest Inventory and Analysis program (www.fia.fs.fed.us).

**3. Literature Review and References:**

Burkhart, H.E., Tomé, M., 2012. Modeling forest trees and stands, Modeling Forest Trees and Stands.

Kimsey, M.J., Shaw, T., and Coleman, M. 2018. Site Sensitive Maxmimum Stand Density Index Models for Mixed Confier Stands Scross the Inland Northwest, USA. *submitted*

Reineke, L.H. 1933. Perfecting a stand-density index for even-aged forests. Journal of Agricultural Research 46:627-638.

Stage, A.R., 1976. An Expression for the Effect of Aspect, Slope, and Habitat Type on Tree Growth Note by A. R. Stage. For. Sci. 22, 457–460.

VanderSchaaf, C.L. and Burkhart, H.E., 2007. Comparison of Methods to Estimate Reineke’s Maximum Size-Density Relationship Species Boundary Line Slope. Forest Science 53(3)

Ducey, M.J., Woodall, C.W., Bravo-Oviedo, A., 2017. Climate and species functional traits influence maximum live tree stocking in the Lake States, USA. For. Ecol. Manage. 386, 51–61.

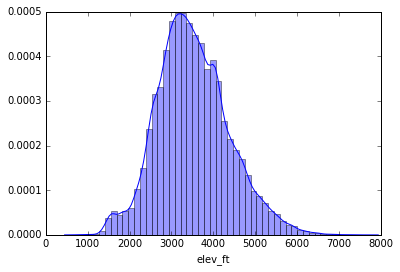
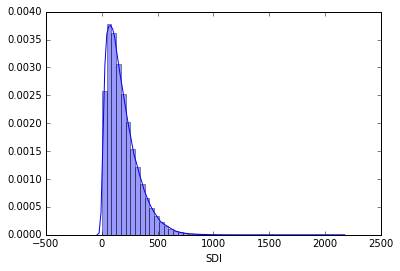
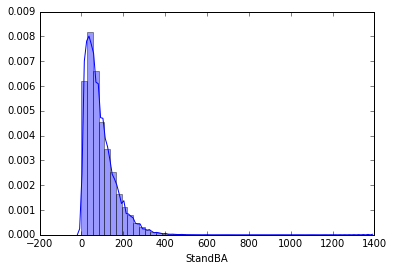
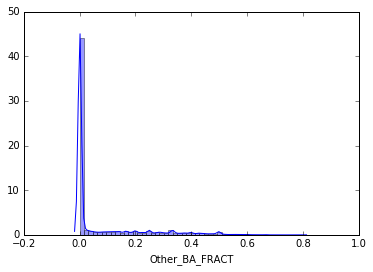
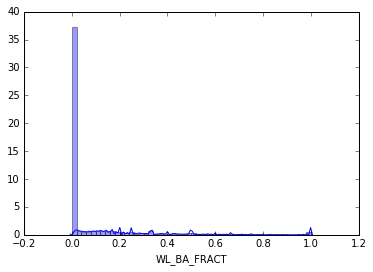
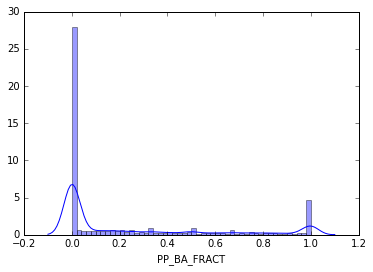
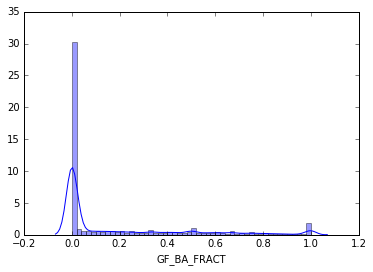
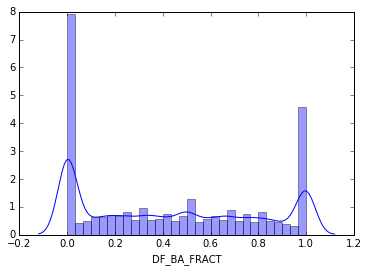
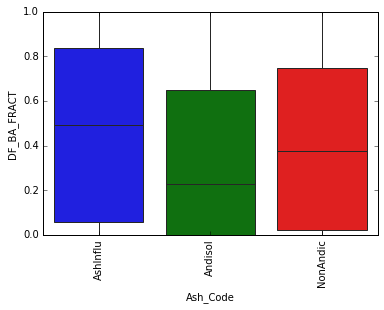
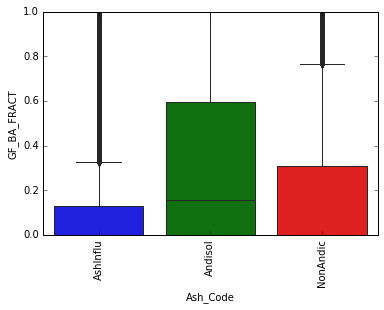
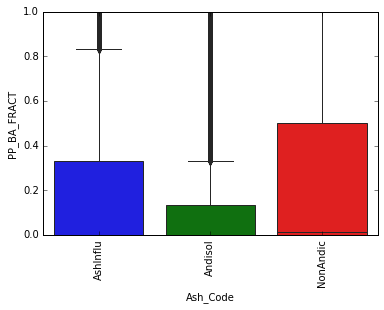
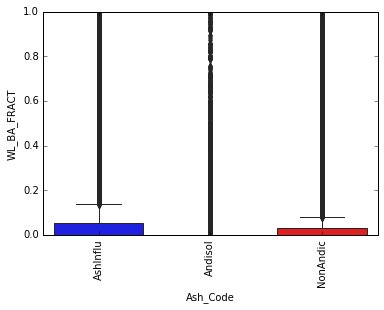
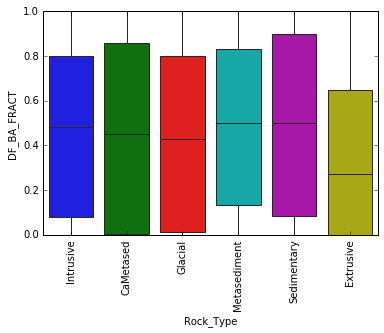
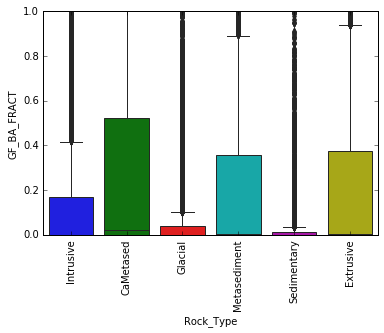
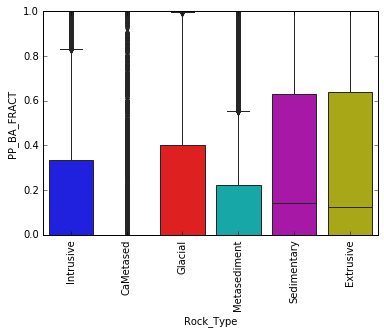
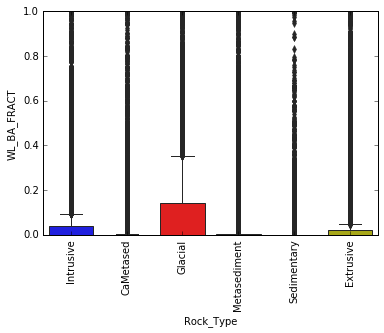
**4. Preliminary EDA:**

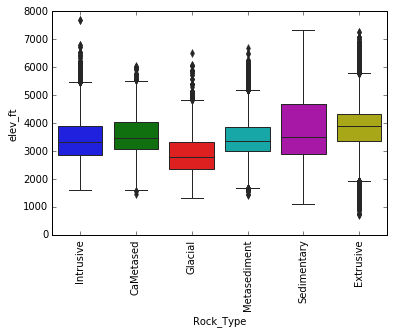
1. What is the shape of your data set - how many rows (observations), and how many columns (variables/features)? *Original data set was 110,500 rows of observations with 34 variables, but after cleaning and additional variables the final set was 92,406 rows with 41 variables*
2. What are the names of these variables, and its full descriptions?

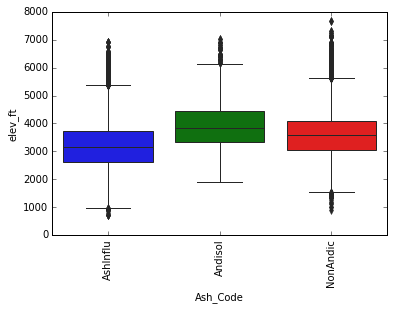
|  |  |
| --- | --- |
| **VARIABLE** | **DESCRIPTION** |
| Stand\_ID | Unique stand identifier |
| Latitude | Lat coordinates |
| Longitude | Long coordinates |
| Rock\_Type | Underlying rock type |
| Ash\_Code | Ash influence on soil |
| StandBA | Stand Basal Area |
| StandQMD | Stand Quadratic Mean Diameter (in) |
| SDI | Stand Density Index |
| SDI\_m | Stand Density Index\_metric |
| lnQMD | Natural log of QMD (in) |
| QMDcm | Stand Quadratic Mean Diameter (cm) |
| lnQMDcm | Natural log of QMD (cm) |
| StandTPA | Stand Trees per Acre |
| lnTPA | Natural log of Trees per Acre |
| StandTPH | Stand Trees per Hectare |
| lnTPH | Natural log of Trees per Hectare |
| Aspect | Aspect of slope |
| Slope | Slope percentage |
| elev\_ft | Elevation of stand in feet |
| mat | Mean annual Temp |
| map | Mean annual Precip |
| gsp | Growing season Precip |
| mtcm | mean temp coldest month |
| mtwm | mean temp warmest month |
| ffp | frost free period |
| dd5 | degree days >5C |
| gsdd5 | growing season days >5C |
| d100 | Julian date the sum of degree days>5 reaches 100 |
| dd0 | Degree days <0 |
| smrsprpb | Summer/Spring precip balance (jul+aug)/(apr+may) |
| sdi | Summer dryness index |
| adi | Annual dryness index |
| tan\_slope | Topographic variable |
| cos\_aspect | Topographic variable |
| tan\_slope\_cos\_aspect | Topographic variable |
| sin\_aspect | Topographic variable |
| tan\_slope\_sin\_aspect | Topographic variable |
| DF\_BA\_FRACT | Douglas Fir BA fraction |
| GF\_BA\_FRACT | Grand Fir BA fraction |
| PP\_BA\_FRACT | Ponderosa Pine BA fraction |
| WL\_BA\_FRACT | Western Larch BA fraction |
| Other\_BA\_FRACT | other species BA fraction |

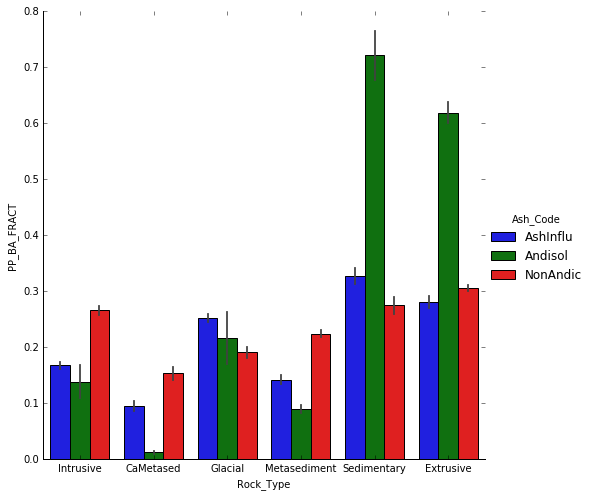
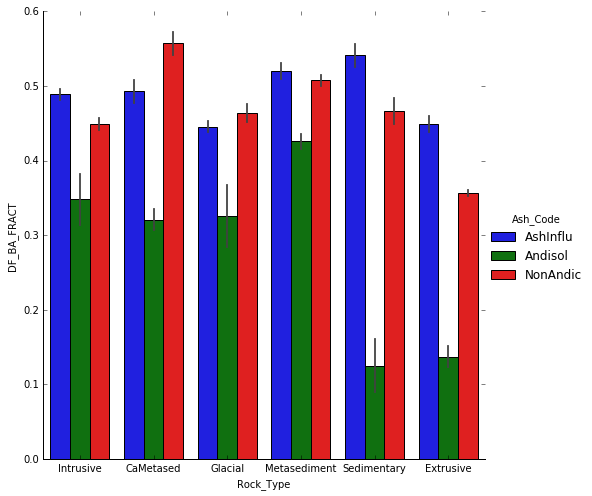
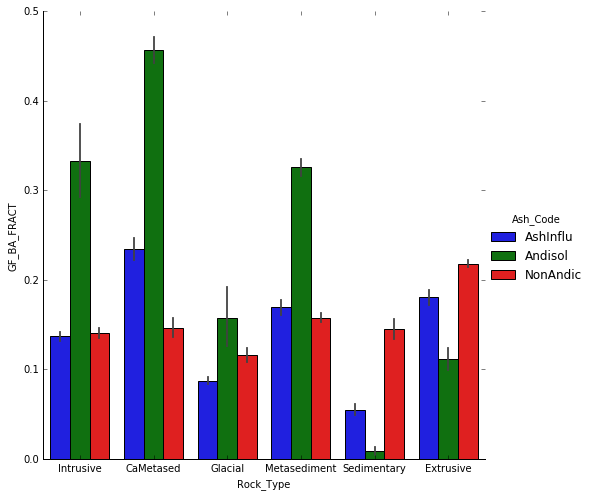
1. How many numerical variables? *All the climatic, topographic and stand/tree variables*
2. How many categorical variables? *Rock Type and Ash Code*
3. How many text variables? *Stand ID*
4. What variables other than the above are involved? *Lat/Long are location coordinates*
5. What methods have you used to preprocess/clean/explore the data and why?
   1. I have first deleted records with suspect data or missing data. I have also deleted any records where the QMD is less than 1 inch. This is in line with methods proposed in many of stand density index papers.
   2. I have also added some columns.
      1. Created a column called SDI which is Stand Density Index, which is an index based on the relationship between the number of trees per unit area and the quadratic mean diameter. Reineke (1933) noted that plotting the logarithm of TPA against the logarithm of QMD generally resulted in a straight-line relationship with a negative slope that could define the limit of maximum stocking
      2. Converted QMD and TPA to metric
      3. Log transformed QMD and TPA (both English and metric units)
   3. This has led to the data shape of 92,406 records with 41 variables.

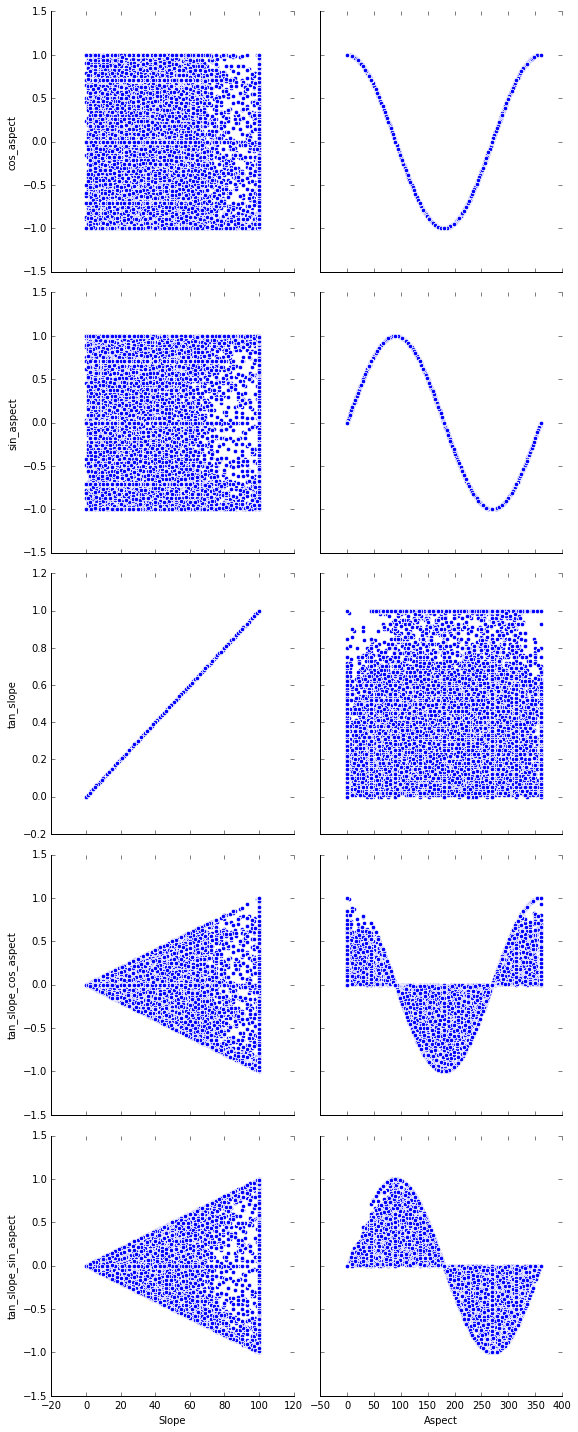
Methods used to explore the data:

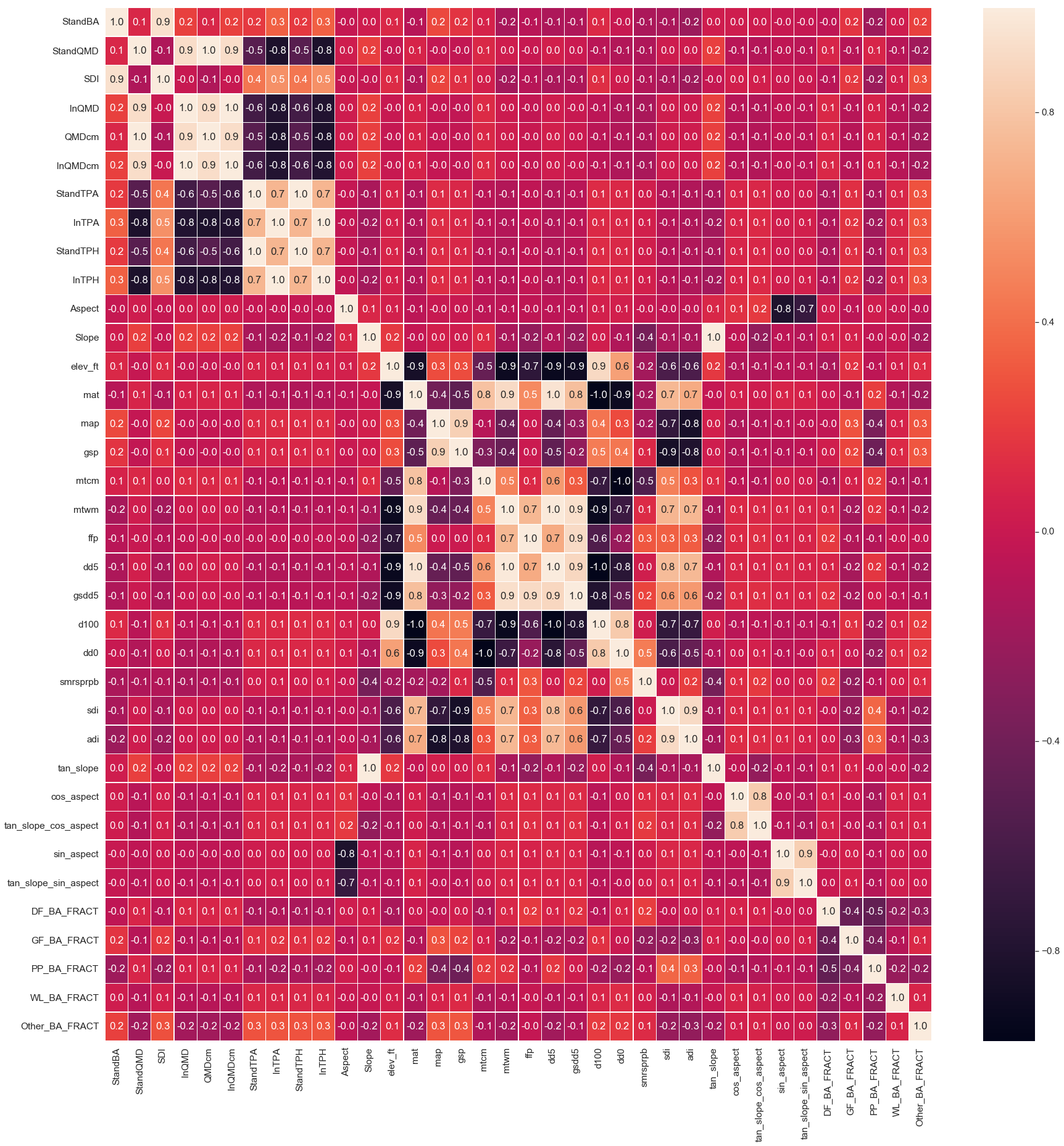
1. I first made histograms and ran data summaries of some of the important variables to see the distribution and range of values.
   1. Elev
   2. SDI
   3. Basal Area
   4. Species BA Fractions
2. I produced box plots to see the distribution of the species BA fraction variables against the categorical variables of Ash Code. These were interesting results as it showed me which species of trees seem to dominate which soil types influenced or not by volcanic ash. Volcanic ash influences soil water holding capacity. A soil with an ash cap, or a soil formed from volcanic ash will have a higher water holding capacity. Ponderosa Pine tends to dominate in hotter, drier areas and it is no surprise that the non-andic soils have the greatest proportion of Pine on them. These boxplots also show that Douglas Fir seems to be the dominant tree found in the inventories that make up this data set, as was seen in the histograms as well. Larch seems to have a outliers occurring on andisols but the majority of the BA is not found on these soil types. The other species BA tend to be less
   1. Ash Code v DF\_BA\_FRACT – AshInfluenced soil carries slightly more Douglas Fir BA
   2. Ash Code v GF\_BA\_FRACT – Andisols carry the most Grand Fir BA 
   3. Ash Code v PP\_BA\_FRACT – Non-Andic Soils carry the most Ponderosa Pine BA 
   4. Ash Code v WL\_BA\_FRACT – variable but lowest mean on Andisols 
3. I produced similar boxplots as above, this time relating the subsurface soil parent materials with the species BA fractions to see what sort of relationships exist. Douglas Fir had a fairly uniform distribution across parent material types, this isn’t a surprise considering Doug Fir is an intermediate species that can tolerate a wide range of site conditions. Grand Fir had the widest BA range on calcium baring metasedimentary parent material. Ponderosa Pine had greatest basal area rage on sedimentary and extrusive parent material types. Western Larch had greatest basal area range on soils with glacial parent material.
   1. Soil Parent Material v DF\_BA\_FRACT 
   2. Soil Parent Material v GF\_BA\_Fract 
   3. Soil Parent Material v PP\_BA\_FRACT 
   4. Soil Parent Material v WL\_BA\_FRACT 
4. Some other interesting findings exploring the categorical variables:
   1. Rock Type v Elevation



* 1. Ash Code v Elevation 

1. Using categorical plots to show the interaction of rock type and ash code on species basal area fractions. Certain species basal area fractions seemed to dominate particular combinations of rock type and ash code. For example, the highest basal area percentage of all combinations was Ponderosa Pine found on sites with sedimentary parent material and andisol soil properties at more than 70% basal area fraction. Followed closely by Ponderosa Pine on extrusive andisols at greater than 60%. These sites are clearly dominated by Pine. As shown in the previous boxplots, Doulas Fir is the most common of all the species, but where it is the least common is the aforementioned sites dominated by Ponderosa Pine.
   1. Interaction of Rock Type and Ash Code on Ponerosa Pine basal area fraction 
   2. Interaction of Rock Type and Ash Code on Douglas Fir basal area fraction 
   3. Interaction of Rock Type and Ash on Grand Fir basal area fraction 
2. Stage (1976) suggested a transformation of aspect (azimuth direction) using sine and cosine to show a measure of East-West (-1 to 0 West, and 0 to 1 East) and North-South (-1 to 0 South, and 0 to 1 North) slope effect respectively, as well as a tangent transformation of slope (tan slope = % slope) multiplied by the cosine and sine transformations of aspect, to show the additive effect of steeper slopes. Using pairwise plots, the relationship between these variables and their transformations can be seen.



1. Variable correlation was explored using pairwise plots between all variables (not shown here) as well as a heatmap. Correlation between the climatic variable is apparent. In some instances this is due to particular variables being used to calculate others to get at indexes or other ratios. For example, summer and annual dryness indexes use degree days and precipitation to calculate. In other cases it is the nature of the variables and the natural correlation. For mean annual temperature (and in return all the degree days calculations) are strongly negatively correlated with elevation. Elevation not surprisingly is a strongly correlated variable with many climatic features. 

**5. Modeling Process:**

Using my dataset, I will cover the three learning areas discussed in STAT517, including supervised, cluster and association analyses. Supervised learning will involve regression on some key variables of interest, including stand density index (SDI), stand basal area and stand trees per acre. I will utilized linear regression, lasso and ridge regression, decision tree and random forest regressor and also k nearest neighbor and neural network regressors. In the stand density index literature a few different techniques have been proposed to get at a maximum stand density, including quantile regression. Using python’s statsmodels module, I also performed quantile regression on the log transformation of TPA and QMD. For clustering analysis, I utilized a dimensionality reduction technique and then attempted Kmeans to find clusters. For association analysis, I first transformed the data into binary. For categorical variables this was simple dummy coding. For continuous variables, the majority of the variables in this dataset, I used a binning process to separate the data in low, medium and high and then dummy coded those binned numbers. Other binning was done for particular variables such as elevation, which I broke into 1000 ft sections and the aspect transformations which where broke into North and South, or East and West. Slope was broken into a few bins with a flat, slight slope, steep and very steep depending on slope percentage. The basal area fractions were broken up into 25% sections of low, medium, high and full.

**Regression**

The first regression was run to predict lnTPA (the log transformation of TPA). I dropped all directly correlated tree measurement variables, keeping only lnQMD an independent tree variable. Split my data into training and testing sets, with lnTPA as the y. Then ran linear, ridge and lasso regressions. Printing the training and testing score, as well as number of variables used. Once models were run, I found the coefficients of each variable to see the influence.

treer=pd.get\_dummies(trees, columns=["Rock\_Type", "Ash\_Code"])

treer=treer.drop(['Stand\_ID','StandBA','StandQMD','QMDcm','lnQMDcm','StandTPA','StandTPH','lnTPH','SDI','SDI\_m'], axis=1)

from sklearn.model\_selection import train\_test\_split

X=treer.drop('lnTPA',axis=1)

y=treer.lnTPA

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=1)

from sklearn.linear\_model import LinearRegression

lr = LinearRegression().fit(X\_train, y\_train)

print("lr.intercept\_: {}".format(lr.intercept\_))

print("Training set score: {:.2f}".format(lr.score(X\_train, y\_train)))

print("Test set score: {:.2f}".format(lr.score(X\_test, y\_test)))

print("Number of features used: {}".format(np.sum(lr.coef\_ != 0)))

lr\_coef = pd.Series(data=lr.coef\_,index=X\_train.columns)

lr\_coef.sort\_values(ascending=False,inplace=True)

from sklearn.linear\_model import Ridge, RidgeCV

ridge = Ridge().fit(X\_train, y\_train)

print("ridge.intercept\_: {}".format(ridge.intercept\_))

print("Training set score: {:.2f}".format(ridge.score(X\_train, y\_train)))

print("Test set score: {:.2f}".format(ridge.score(X\_test, y\_test)))

print("Number of features used: {}".format(np.sum(ridge.coef\_ != 0)))

ridge\_coef = pd.Series(data=ridge.coef\_,index=X\_train.columns)

ridge\_coef.sort\_values(ascending=False,inplace=True)

from sklearn.linear\_model import Lasso

lasso = Lasso().fit(X\_train, y\_train)

print("lasso.intercept\_: {}".format(lasso.intercept\_))

print("Training set score: {:.2f}".format(lasso.score(X\_train, y\_train)))

print("Test set score: {:.2f}".format(lasso.score(X\_test, y\_test)))

print("Number of features used: {}".format(np.sum(lasso.coef\_ != 0)))

lasso\_coef = pd.Series(data=lasso.coef\_,index=X\_train.columns)

lasso\_coef.sort\_values(ascending=False,inplace=True)

Linear and ridge produced similar results, with the regularization of ridge scaling the coefficients and intercept. The lasso did not produced strong scores. As expected, lnQMD has a strong influence on lnTPA, as trees get larger they outcompete nearby neighbors causing a decline in the number of trees as the stand ages. Other influential variables included climatic variables of summer/spring precip balance and temperature, both positively influencing trees per acre. The other strong influencing variables were the species basal area fractions.

lr.intercept\_: -32375.5239278

Training set score: 0.67

Test set score: 0.68

Number of features used: 38

ridge.intercept\_: 4.96064161779

Training set score: 0.67

Test set score: 0.68

Number of features used: 38

lasso.intercept\_: 5.79820656244

Training set score: 0.05

Test set score: 0.06

Number of features used: 7

The next models run were Decision Tree and Random Forest Regressors. These both performed well on the dataset. These make more sense than straight linear regressions because of the range of data points. Again, no surprise that lnQMD was the most important feature with 0.709771 (71% influence) and the BA fractions of other, grand fir and douglas fir the next most important with scores around 0.03 (3% influence).

from sklearn.tree import DecisionTreeRegressor

dtr = DecisionTreeRegressor().fit(X\_train, y\_train)

Decision Tree regressor score on training set: 1.00

Decision Tree regressor on test set: 0.62

from sklearn.ensemble import RandomForestRegressor

rfr=RandomForestRegressor(n\_estimators=200).fit(X\_train,y\_train)

Accuracy of Random Forest regressor on training set: 0.97

Accuracy of Random Forest regressor on test set: 0.81

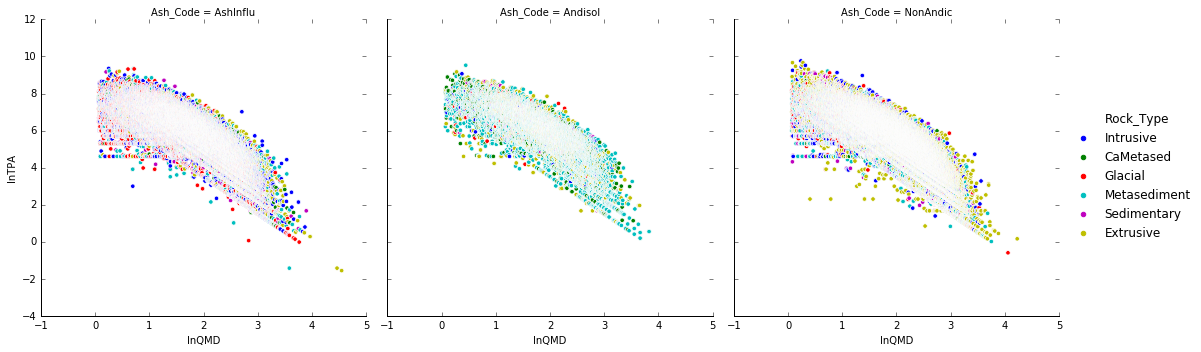
Nearest neighbor and neural networks performed poorly, with R^2 training scores of 0.40 and -0.13 respectively, leading to even worse testing set scores.

In the end the random forest regressor performed the best. I will continue to tune the parameters to see if the score can gain any more.

When I swapped lnTPA for lnQMD in my regression models, they all performed similarly. Again the random forest regressor was the highest scoring, with 0.97 and 0.78 respectively.

**Quantile Regression**

Reineke (1933) showed that when log transformations of QMD and TPA were plotted, an outer boundary shows a nice straight line relationship. This outer boundary is deemed the maximum stand density index, in particular the maximum trees per acre with QMD = 10 inches. While Reineke hand fit a line and estimated the slope, the advent of computers has led to more statistically sound fitting of this line. Quantile regression involves placing a regression line with some level x% of the data falling below and above the line (Ducey et al., 2017). First the 0.5 quantile is found which predicts the conditional median of the data. Then the level can be modified depending on application of the analysis. In the case of maximum stand density, the interest is in the furthest possible line creating the boundary of the data. I ran the median regression line through lnTPA on lnQMD, then found the 95%-99% quantiles lines. The slope of the 99% quantile is right in line with Reineke’s findings, as well as numerous studies published on maximum stand density index.

First plots of lnQMD v lnTPA, with interaction of rock type and ash code. These plots show the outer boundary well, that is the maximum stand density line. 

Code for running quantile regression:

import statsmodels.api as sm

import statsmodels.formula.api as smf

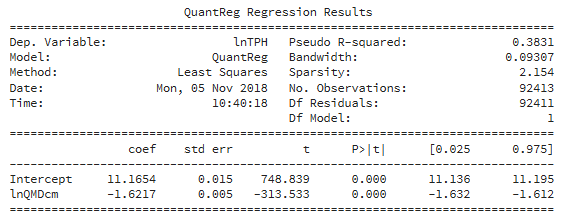
from statsmodels.regression.quantile\_regression import QuantReg

First the median quantile line:

mod = smf.quantreg('lnTPH ~ lnQMDcm', trees, groups=trees["Ash\_Code"]) #lnQMDcm

res = mod.fit(q=.5) #, vcov='robust', kernal='epa', bandwidth='hsheather'

print(res.summary())



Then using this median line, the .95 - .99 quantiles are fit with the following code:

quantiles = np.arange(.95, .99, .01)

def fit\_model(q):

res = mod.fit(q=q)

return [q, res.params['Intercept'], res.params['lnQMDcm']] + \

res.conf\_int().loc['lnQMDcm'].tolist()

models = [fit\_model(x) for x in quantiles]

models = pd.DataFrame(models, columns=['q', 'a', 'b','lb','ub'])

ols = smf.ols('lnTPH ~ lnQMDcm', trees).fit()

ols\_ci = ols.conf\_int().loc['lnQMDcm'].tolist()

ols = dict(a = ols.params['Intercept'],

b = ols.params['lnQMDcm'],

lb = ols\_ci[0],

ub = ols\_ci[1])

print(models)

print(ols)

Results:

q a b lb ub

0 0.95 12.162622 -1.584774 -1.591716 -1.577832

1 0.96 12.212545 -1.583830 -1.590955 -1.576706

2 0.97 12.276207 -1.583985 -1.591567 -1.576403

3 0.98 12.383495 -1.592602 -1.600820 -1.584383

4 0.99 12.513622 -1.598165 -1.607397 -1.588932

{'a': 11.065727555167737, 'b': -1.6277913148131153, 'lb': -1.6359901871280331, 'ub': -1.6195924424981976}

Graphing the median and 95% - 99% quantiles:

x = np.arange(trees.lnQMDcm.min(), trees.lnQMDcm.max(), .5)

get\_y = lambda a, b: a + b \* x

fig, ax = plt.subplots(figsize=(8, ))

for i in range(models.shape[0]):

y = get\_y(models.a[i], models.b[i])

ax.plot(x, y, linestyle='dotted', linewidth='3', color='magenta')

y = get\_y(ols['a'], ols['b'])

ax.plot(x, y, color='red', label='OLS')

ax.scatter(trees.lnQMDcm, trees.lnTPH, alpha=.2)

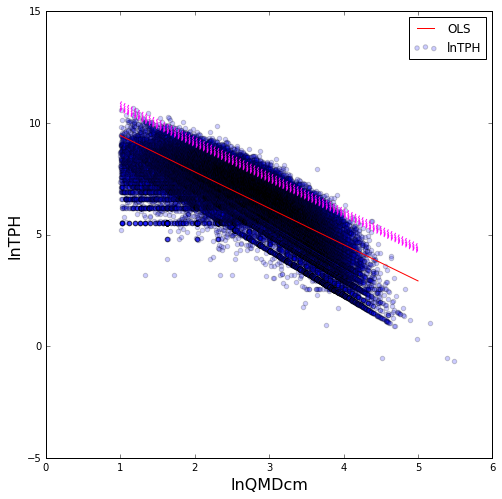
ax.set\_xlim((0, 6))

ax.set\_ylim((-5, 15))

legend = ax.legend()

ax.set\_xlabel('lnQMDcm', fontsize=16)

ax.set\_ylabel('lnTPH', fontsize=16);



This graph illustrates the outer boundary well with the 95% - 99% quantile lines tit to the outer edge of the data points. Using the output of the 99% quantile, the maximum stand density index (number of trees at 25.4 cm) is:

**1546.8** - 25.4 cm trees per hectare = **626.2** - 10” trees per acre

**Association Analysis**

Association analysis involved first transforming the data into binary. This was done as previously discussed in the introduction to the modelling section. Some example code is as follows:

treed=pd.get\_dummies(trees, columns=["Rock\_Type", "Ash\_Code"])

treed['elevation']=pd.cut(treed.elev\_ft,bins=[0,1000,2000,3000,4000,5000,6000,7000,8000],labels=["less\_1000","btw\_1\_2thou","btw\_2\_3thou","btw\_3\_4thou","btw\_4\_5thou","btw\_5\_6thou","btw\_6\_7thou","great\_7000"])

treed=pd.get\_dummies(treed, columns=["elevation"])

treed['DF\_BA']=pd.cut(treed.DF\_BA\_FRACT, bins=[0,.25,.5,.75,1],labels=["low","medium","high","full"])

treed['GF\_BA']=pd.cut(treed.GF\_BA\_FRACT, bins=[0,.25,.5,.75,1],labels=["low","medium","high","full"])

treed['PP\_BA']=pd.cut(treed.PP\_BA\_FRACT, bins=[0,.25,.5,.75,1],labels=["low","medium","high","full"])

treed['WL\_BA']=pd.cut(treed.WL\_BA\_FRACT, bins=[0,.25,.5,.75,1],labels=["low","medium","high","full"])

treed['Other\_BA']=pd.cut(treed.Other\_BA\_FRACT, bins=[0,.25,.5,.75,1],labels=["low","medium","high","full"])

treed=pd.get\_dummies(treed, columns=["DF\_BA","GF\_BA","PP\_BA","WL\_BA","Other\_BA"])

treed['temp']=pd.qcut(treed.mat,3,labels=["low","med","high"])

treed=pd.get\_dummies(treed,columns=["temp"])

treed['NS\_Aspect']=pd.cut(treed.cos\_aspect,bins=[-1,0,1],labels=["South","North"])

treed=pd.get\_dummies(treed,columns=["NS\_Aspect"])

treed['EW\_Aspect']=pd.cut(treed.sin\_aspect,bins=[-1,0,1],labels=["West","East"])

treed=pd.get\_dummies(treed,columns=["EW\_Aspect"])

treed['Slope']=pd.cut(treed.Slope,bins=[0,10,30,50,101],labels=["flat","SlightSlope","Steep","VerySteep"])

treed=pd.get\_dummies(treed,columns=["Slope"])

Then using frequent patterns apriori to find frequent item sets, and association rules to find antecedents and consequents, association analysis can look for rules of interest. I am still running through the association analysis but some of the common rules are climatic. For example, different levels of precipitation are associated with particular temperatures and degree day accumulations. This is no surprise given the nature of these climatic variables.

from mlxtend.frequent\_patterns import apriori

freq\_treed=apriori(treed, min\_support=0.1, use\_colnames=True)

freq\_treed['length'] = freq\_treed['itemsets'].apply(lambda x: len(x))

from mlxtend.frequent\_patterns import association\_rules

rules=association\_rules(freq\_treed)

rules.sort\_values(['lift'],ascending=False).head(500)

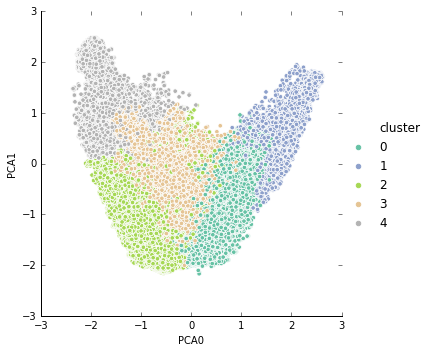
Exploring the rules allows for particular variables of interest to be searched. For example, what leads to a low annual dryness index?

lowADIrules=rules[ rules['consequents'] == {'a\_dry\_index\_low'} ]

The resulting rules are interesting but not unexpected. For example a westerly aspect with high precipitation will lead to a low dryness index with a lift score of 2.48

**Clustering**

First dimensionality reduction was performed using PCA. Roughly 80 components made up 100% variation. Then Kmeans was used to cluster data. 5 clusters were found. I am still in the process of tuning these cluster numbers and exploring the significance (if any) of these clusters.



**6. Project Progress, Timeline, and Achievement**:

My presentation date is scheduled for November 27, allowing under two weeks from turning in this preliminary report. I will turn my focus, for now, on my in class presentation. Then I can continue to focus on regression, but also see if I can work out the clustering and association analysis.

**7. Conclusions and Possible Future Work:**

Supervised regression is the most appropriate analysis when looking into forest stand variables of tree diameter and number of trees per acre. I will continue to tune my models and produce results appropriate for the dataset.