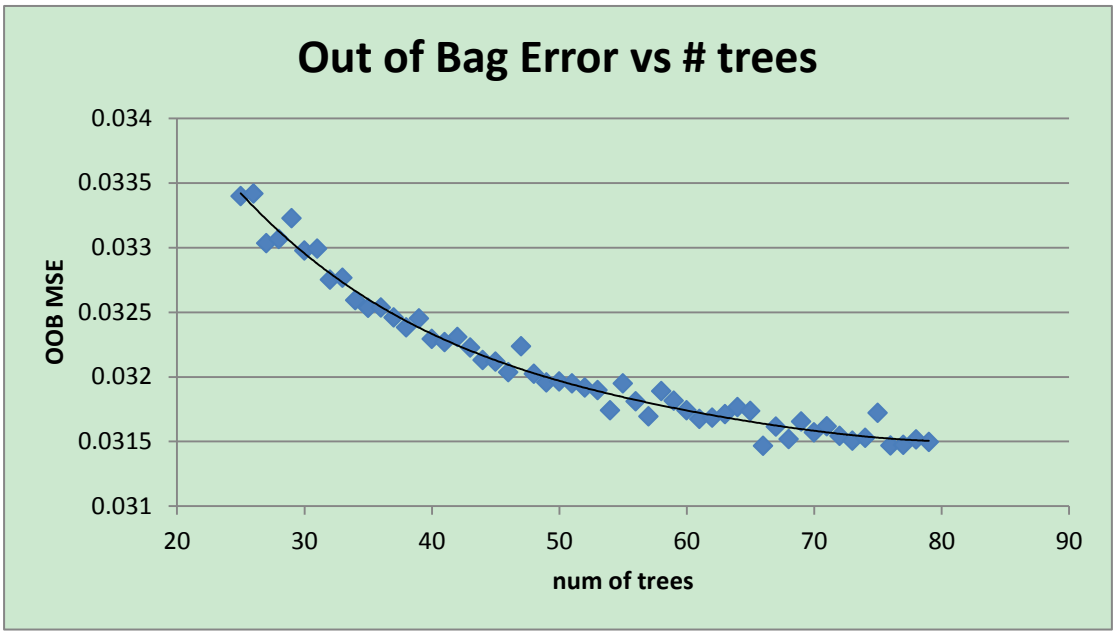
STAT665

Lab2

Bo Song

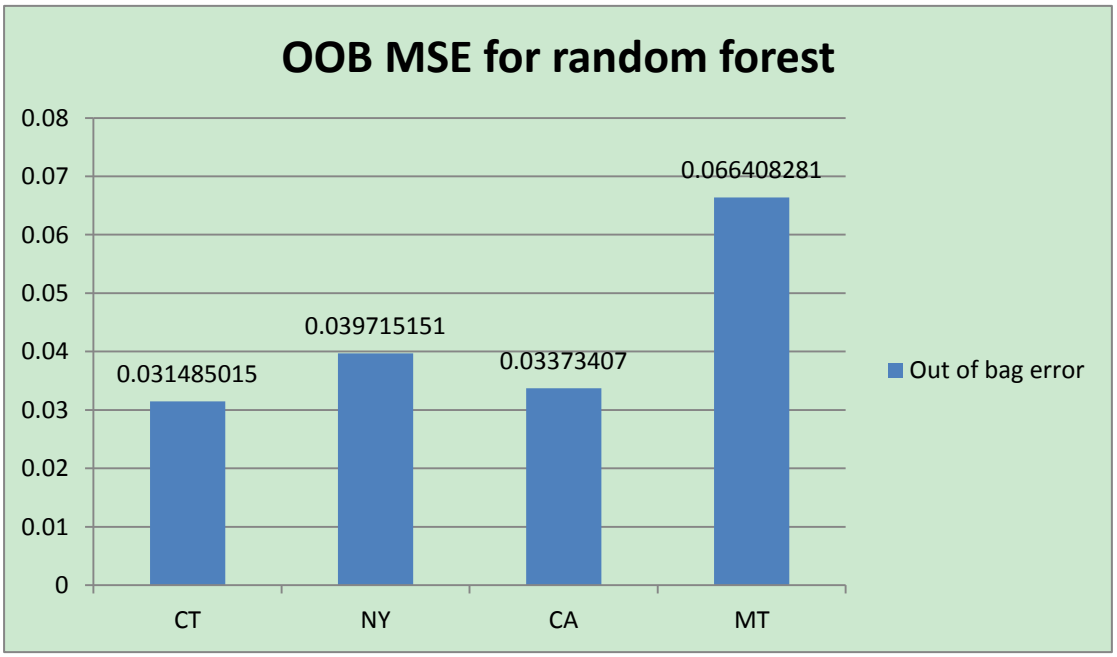
1. **Describe how you decided the ultimate number of trees to use.**

   Draw a chart to see the relationship between # trees and OOB MSE



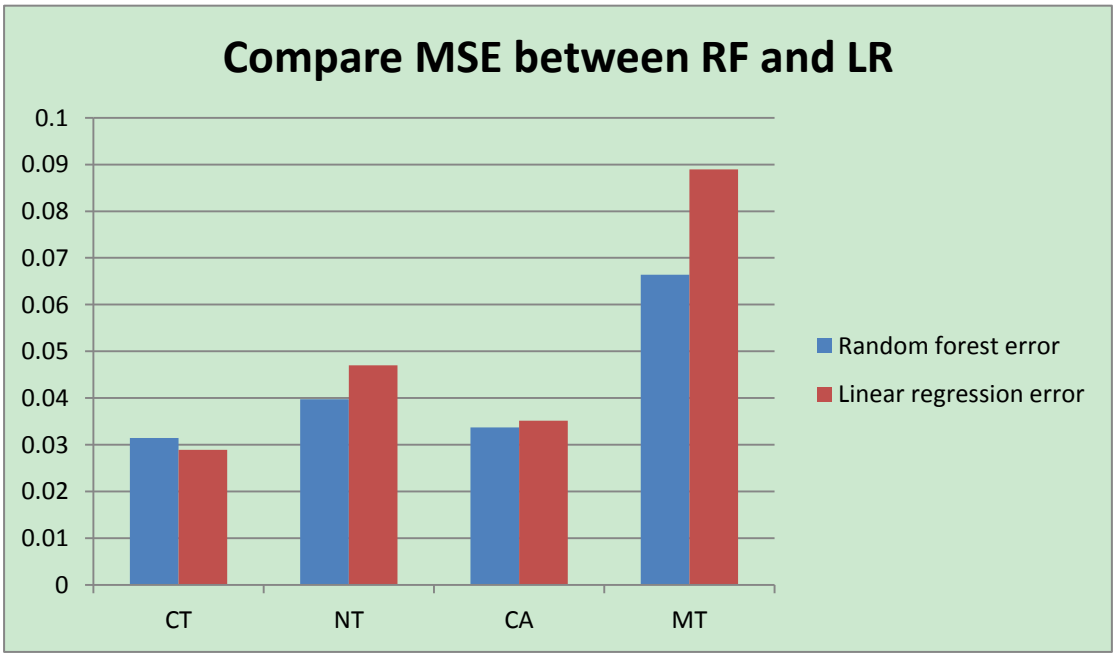
   It seems as the number of trees increases, the OOB MSE decreases gradually and becomes stable at some value. Since Taylor said we don't need to be crazy to tune this parameter, I chose 79 as the number of trees in my experiment. Though a larger number may performs better.

2. **Compare the 'out-of-bag' mean squared error in CT to the errors of NY, MT, and CA. Describe what patterns or surprising features arise.**



The model trained by the data of CT predicts well on the data CA, predicts a little worse on NY, predicts worst on MT.

3. **Now fit a linear regression using the same variables on the CT data and predict the results from the other three states. How do these mean squared errors compare to the random forest model? Note: The raw data has perfect multicolinearity because each set of variables add up to one; either drop the last count from each set, or, in R, just ignore the warnings as this is what will be done automatically for you.**



Linear regression performs better on its training data than CT, but performs worse on other states' data (NT, CA, MT). Generally speaking, for this question, random forest is better.

4. **The field h_geocode in the datasets gives a unique identifier for the census block from**

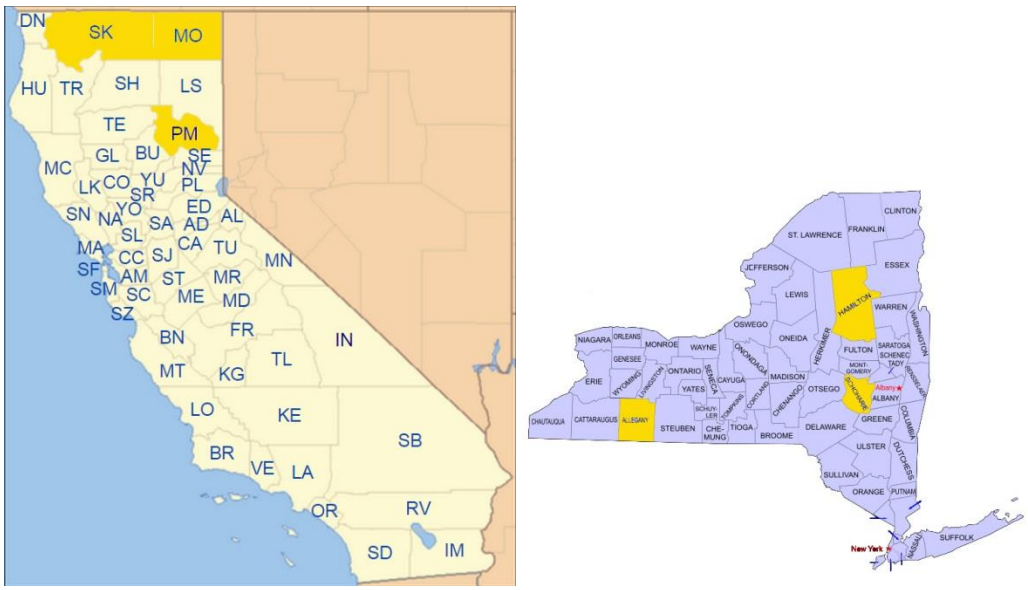
which the data are taken. The first two digits give the state FIPS code and the next 3 digits give the county FIPS code. You can easily find sources that give a mapping from county FIPS codes to county names. What are the three worst counties in terms of squared error under the random forest model in New York and California? Explain any patterns you see and (if applicable) suggest a possible solution.

Table 1 Worst 3 performance countries in NY and CA

State	Country Name	FIPS	MSE
NY	Hamilton	041	0.13588796464149869
	Schoharie	095	0.11349735494302644
	Allegany	003	0.111050974957451
CA	Modoc	049	0.12742915219576989
	Siskiyou	093	0.11158637217716787
	Plumas	063	0.10153327126806813

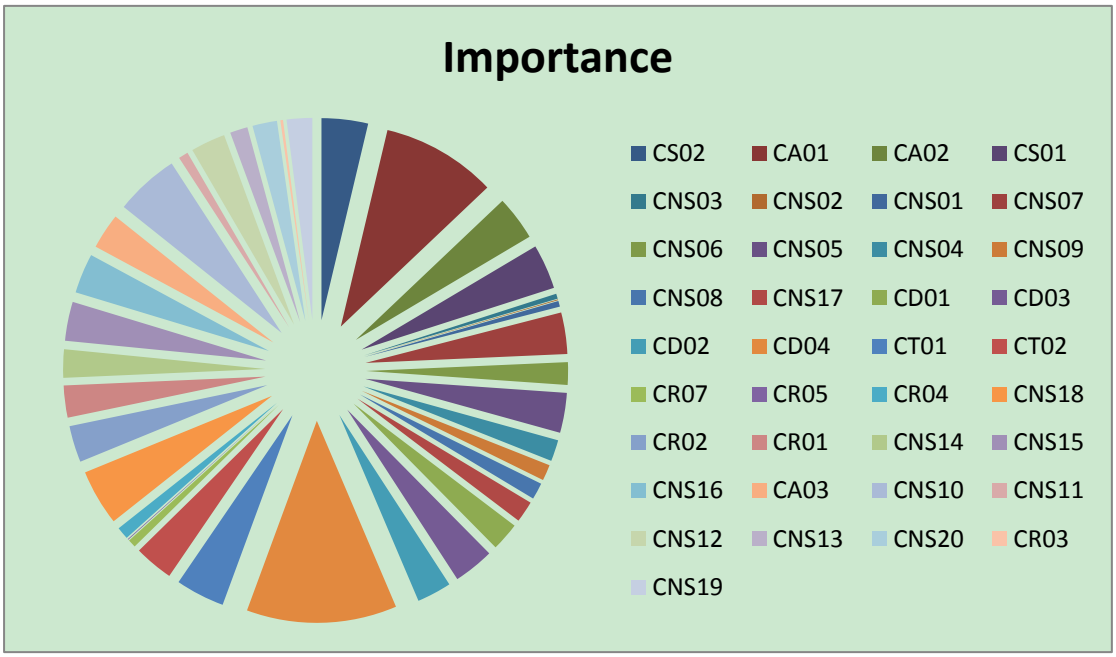
Geographically speaking, the worst 3 countries in CA all locate at the north of CA, the worst 3 in NY just is scattered in NY.

Picture 1 Country map for CA(left) and NY(right) (yellow area indicates worst countries)



The properties of these 6 countries may be quite different from that of CT. Say the mainstay industry, component of population and so on.

5. **Calculate the variable importance scores for the random forest model. Describe the most important variables; does it make sense that these would help identify areas with a high proportion of high income earners?**



It's hard to distinguish the most important 3 features from the pie chart above. Actually there is not a dominated feature among these scores of features. However, if we quantify them, we get that CD04, CA01, CNS10 are 3 most important features. Their means are as follows,

Table 2 Most important 3 features

CD04	Number of jobs for workers with Educational Attainment: Bachelor's degree or advanced degree	0.12024425
CA01	Number of jobs for workers age 29 or younger	0.0920325
CNS10	Number of jobs in NAICS sector 52 (Finance and Insurance)	0.05120127

It's not surprising since the workers with higher education degree(CD04), young age(CA01) to be energetic and enthusiastic and working in finance and insurance usually can get a salary for more than $3333 per/month.