Data

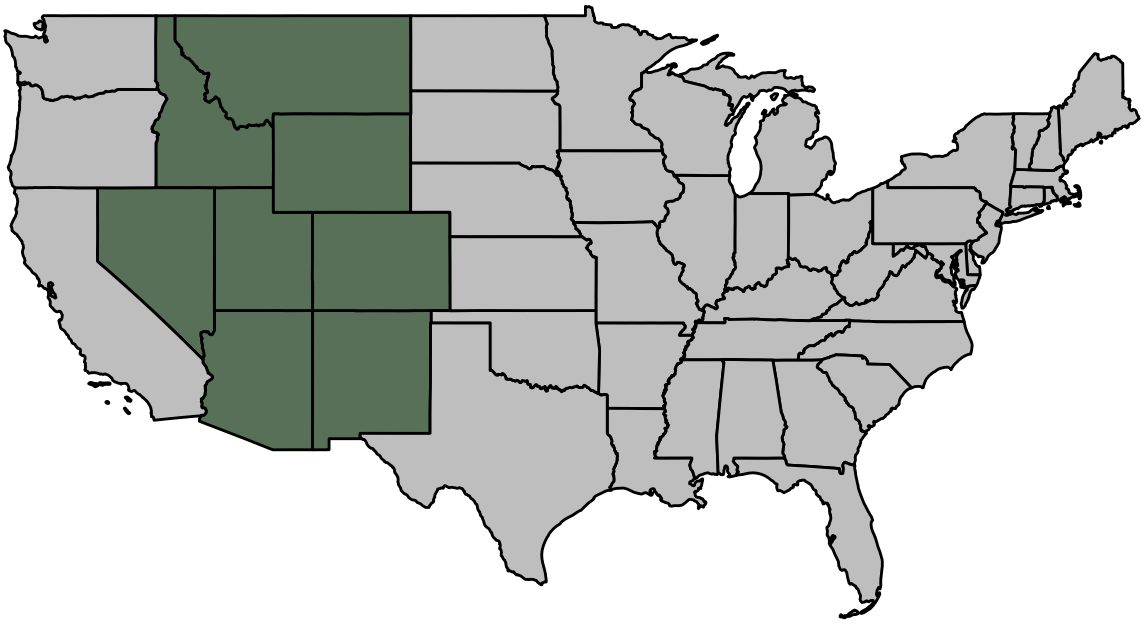
The data used in this thesis was collected by the Forest Inventory and Analysis Program (FIA) in the span of 10 years from 2007 to 2017. While this data was collected over this 10 year period, the analyses done throughout this thesis are under the assumption that this is a "snapshot" of the Interior West at some moment in time. Thus we do not consider any temporal features of this dataset. The data we have is plot-level data for the Interior West region of the United States, where the data for each plot is collected by FIA and its crew members. The units measured by the FIA and their ground crews are approximately 30 m by 30 m hexagonal units. Since the Interior West covers over 140 million acres of forestland, it is extremely impractical for FIA to measure every unit (Source: https://www.fs.fed.us/rm/ogden/lib/interiorwest2.pdf). Instead, they sample from the population of 30 m by 30 m hexagonal units by using a geographically-based systematic sampling design (Source: McConville et al, 2020). The FIA chooses these samples by first overlaying a hexagonal grid over the United States where each hexagon contains 6000 acres of land. Then, they fill these hexagons with much smaller hexagons and randomly sample from the population of small hexagons. Then, ground crews go to these sampled small hexagons and collect variables such as basal area, trees per acre, etc. This plot level data is what we are working with throughout the duration of the thesis.

The dataframe used in this thesis is a joined dataframe derived from two FIA datasets of the Interior West, `spatial` and `response`. The `spatial` dataframe contains 89444 observations and 70 variables, most notably two predictor variables (`forprob` and `forbio`), location information, and ecosubsection. The `response` dataframe contains 86085 observations and 67 variables, most notably four predictor variables collected by FIA crew members (`BALIVE_TPA`, `CNTLIVE_TPA`, `BIOLIVE_TPA`, and `VOLNLIVE_TPA`), location information, and ecosubsection. We join these dataframes by their unique plot number, and subset the number of variables significantly to 19 variables which contain plot information, longitude & lattitude, elevation, predictor variables, response variables, ecosubsection, ecosection, and province. The resulting joined dataframe has 86085 rows as these are the rows which share the same plots between the `response` and `spatial` dataframes. We can see the first few rows of the dataframe with relevant columns selected and digits rounded:

INVYR	PLOT	LON_PUBLIC	LAT_PUBLIC	ELEV_PUBLIC	forgrp	forprob	demLF
2014	83657	-111.3261	35.02106	6680	180	1	2080
2014	87963	-109.9398	36.59399	5550	0	0	1700
2014	84186	-109.9925	36.27860	7510	180	1	2305
2014	87499	-109.9058	35.32838	5630	0	0	1717
2014	88091	-109.9024	34.83752	5510	0	0	1672
2014	80842	-109.9774	33.52990	5920	180	1	1828

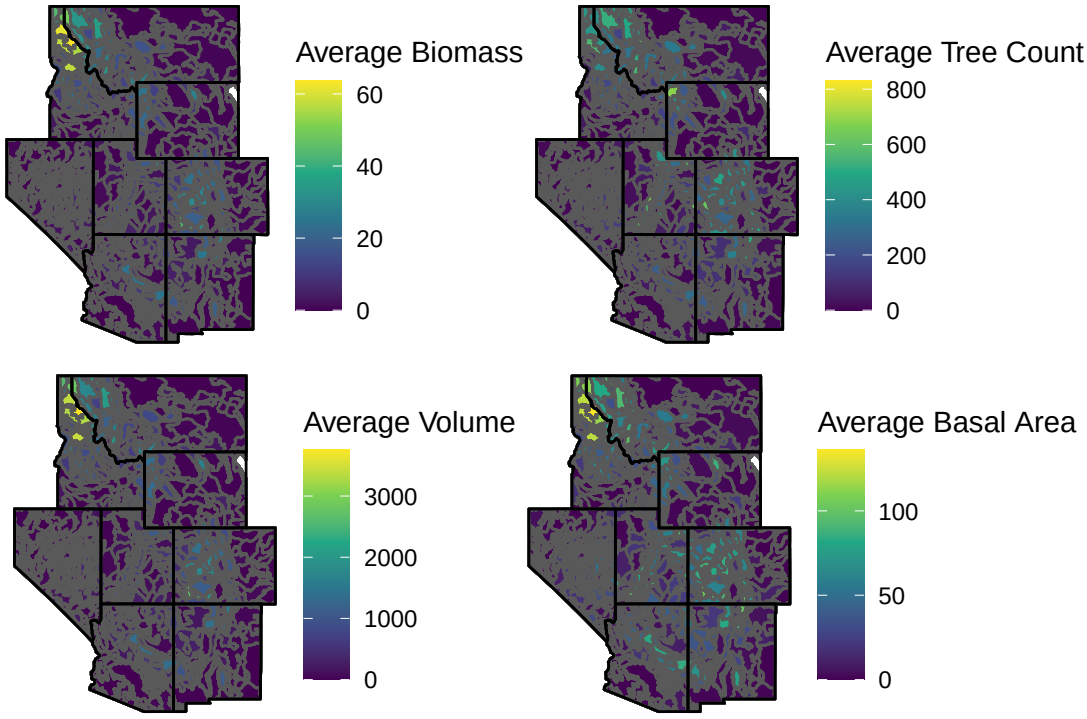
forbio	BALIVE_TPA	CNTLIVE_TPA	BIOLIVE_TPA	VOLNLIVE_TPA	subsection	section	province
12.67684	236.1169	555.3678	43.08024	2066.3352	M313Ak	M313A	M313
0.00000	0.0000	0.0000	0.00000	0.0000	313Ac	313A	313
10.57715	105.3212	108.3248	18.02096	988.9084	313Ap	313A	313
0.00000	0.0000	0.0000	0.00000	0.0000	313Db	313D	313
0.00000	0.0000	0.0000	0.00000	0.0000	313Db	313D	313
14.46534	149.7189	273.5608	21.76656	1381.8511	313Cd	313C	313

Again, the data we have is from the Interior West, and the FIA defines the Interior West as Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming. For reference we have provided the Interior West colored green on a map of the continental United States:



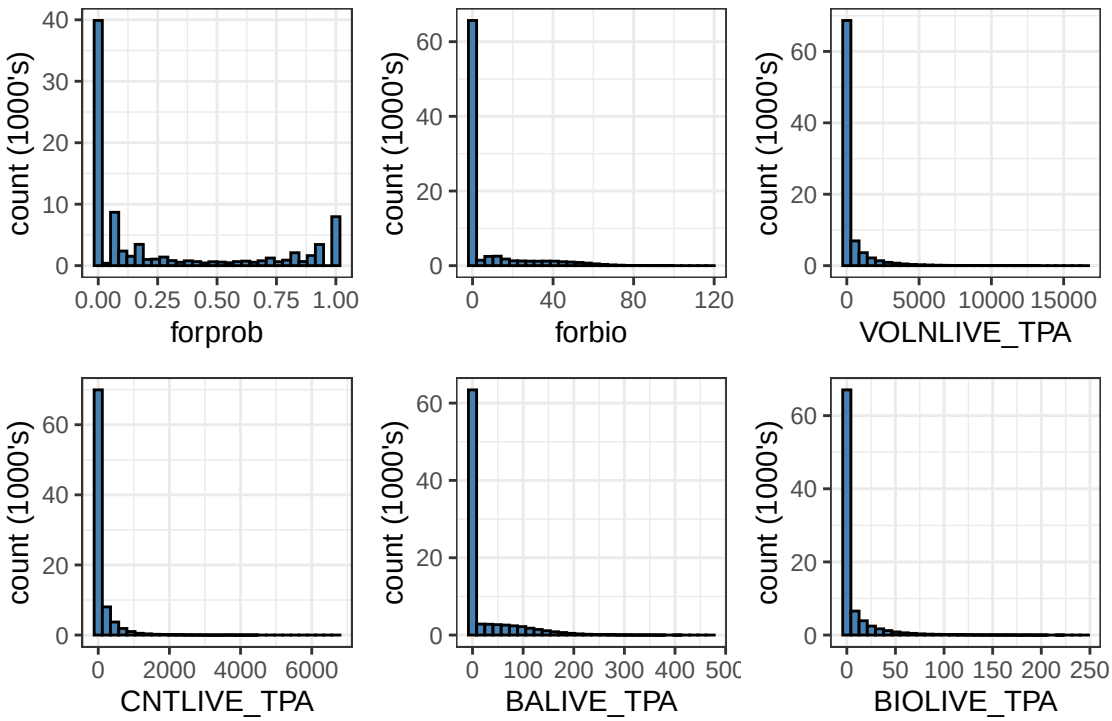
While the data covers the Interior West as a whole, we have very granular information, as each row represents a plot sampled by FIA. The data also includes variables that subset the Interior West into provinces which contain ecosections, and these ecosections contain ecosubsections. In our data, on average, each ecosection contains approximately 7.06 ecosubsections, and each province contains an average of 4.86 ecosections. So, an average province then contains just over 34 ecosubsections. The data we have covers a total of 14 provinces, 68 ecosections, and 480 ecosubsections. The hierarchical struture of the data and nestedness of the ecosubsections within ecosections within provinces lends itself to be able to create hierarchical models which borrow strength from surrounding areas.

While this data contains a multitude of variables, the analyses done in this thesis focus on four key response variables and two explanatory variables. The response variables used are basal area (square-foot), trees per acre, above-ground biomass (lbs), and net volume (ft^3). These variables are coded as `BALIVE_TPA`, `CNTLIVE_TPA`, `BIOLIVE_TPA`, and `VOLNLIVE_TPA`, respectively. We can look at the average of these variables across the Interior West region by ecosubsection in the plots below.



While we have four variables which we will model as response variables throughout the analyses, we also have two predictor variables which will be of much use to us. In particular, forest probability and forest biomass (coded as `forprob` and `forbio`.) These variables which we will treat as predictors are remotely sensed variables, meaning that they were not collected by FIA crew members, but rather with aerial photography and/or satellite imagery. However, we will be using these variables to attempt to predict our response variables in order to understand how good of estimates we can make with this remote data that does not require as much effort to collect. While it may seem unnatural to attempt to predict forest biomass with forest biomass, the differences in the data collection process between the ground level data and remotely sensed data are quite different.

These variables are almost all right-skewed, and all take value zero quite often. This is because there is lots of land in the Interior West which is not forest, hence our forestry variables should take the value zero. To see these phenomena, we can look at histograms of our six key variables:



It is notable that the `forprob` variable is bimodal and modes zero and one, while all other variables are extremely right-skewed. We can also summarize the data to see some summary statistics of our six key variables:

variable	mean	quantile_25	median	quantile_75	min	max	na_count
forbio	6.66	0	0.00	0.00	0	118.00	0
forprob	0.27	0	0.07	0.56	0	1.00	1
BIOLIVE_TPA	6.23	0	0.00	1.98	0	244.35	0
BALIVE_TPA	22.75	0	0.00	14.75	0	469.39	0
CNTLIVE_TPA	98.60	0	0.00	30.09	0	6677.93	0
VOLNLIVE_TPA	342.32	0	0.00	74.69	0	16435.55	0