STAT 536 - MIDTERM

Lodgepole Pine Basal Area

Sam Lee

Lodgepole pines are critical to the Uinta National Forest ecosystem, yet their growth is increasingly threatened by environmental stressors such as pine beetle infestations. This study develops a statistical framework to quantify the environmental determinants of Lodgepole pine basal areas, using data from the Forest Inventory Analysis (FIA) in Northeastern Utah. I employ a combination of LASSO regression and spatial autoregressive models to examine the effects of *Elevation*, *Aspect*, *Slope*, and their interactions on basal area. The LASSO model efficiently performed variable selection, isolating key predictors and simplifying the analysis and was ultimately selected for its performance and interpretability. The analysis revealed that there may be an *ideal* level of *Elevation* optimal for Lodgepole pine growth. Finally, I use the evaluated LASSO model to make inferences about observations not yet collected by the FIA. This approach offers a robust, data-driven methodology for understanding and managing forest ecosystems, with implications for targeted conservation strategies.

Introduction

Lodgepole pines (*Pinus contorta*) are a crucial component of the Uinta National Forest ecosystem, providing essential habitat, stabilizing soil, and contributing to the overall health of the forest. However, their growth and sustainability face significant challenges due to environmental changes and increasing threats from pine beetle infestations. Understanding the factors influencing the basal area of Lodgepole pines is vital for effective forest management and conservation efforts. This study examines the relationship between environmental variables and Lodgepole pine basal areas, using data from the Forest Inventory Analysis (FIA) conducted in Northeastern Utah. The dataset captures various attributes, including geographical coordinates (*Longitude* and *Latitude*), average *Slope* of the terrain, orientation of plots relative to north (*Aspect*), and *Elevation*.

Figure 1: Skewness of *Lodgepole* data

Preliminary exploration of the data reveals several complexities. The distribution of basal areas is notably skewed (see Figure 1), suggesting that standard linear regression models might not adequately capture the

underlying patterns. Furthermore, there is evidence of spatial dependencies, where conditions in one plot influence those in neighboring plots, violating the assumption of independent observations. Failing to account for the appropriate dependencies in the data will lead to improper inference on the coefficients. Non-linear interactions between environmental factors, such as *Elevation* and *Aspect*, further complicate the analysis (see Figure 2), indicating that more sophisticated modeling approaches are required. If the model fails to account for non-linearities in the data, then the model runs the risk of being mis-specified, leading to biased coefficients. In short, if the model fails to represent the true relationship between the environmental factors on Lodepole pine growth, the predictions will be inaccurate.

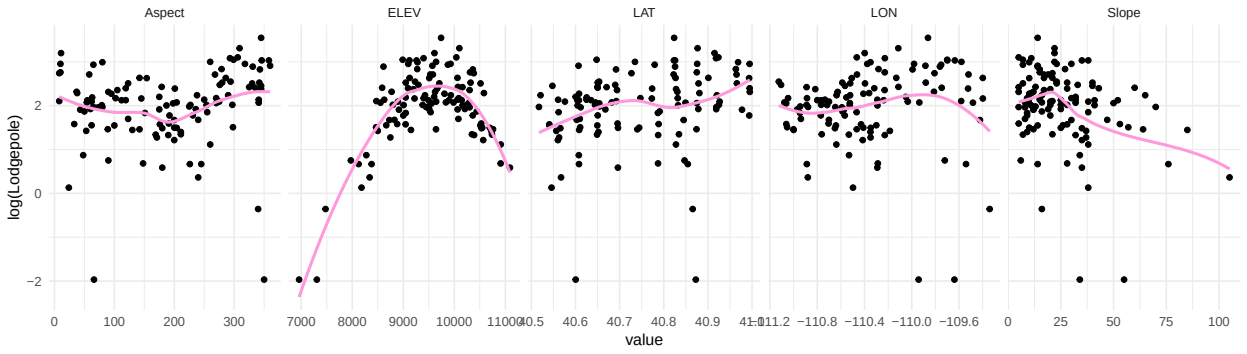

Figure 2: Non-linearities in the data—Scatterplot of each covariate fit with LOESS smoother.

The primary objectives of this study are to identify the key environmental factors that significantly affect Lodgepole pine basal area and to develop models that can predict basal areas in regions where direct measurements are unavailable (see Figure 3). To address the challenges posed by spatial dependencies, non-linearities, and skewed data, the analysis employs a combination of a *least absolute shrinkage and selection operator* (LASSO) regression for variable selection and a spatial autoregressive model to account for spatial correlations. Each method's strengths and limitations are carefully considered to ensure that the models are both interpretable and predictive. The final models are evaluated based on cross-validation accuracy, and the results offer insights into the environmental conditions conducive to Lodgepole pine growth. Ultimately, this research aims to provide reliable conclusions that can inform forest management strategies, enabling targeted actions to mitigate the impact of environmental stressors and infestations.

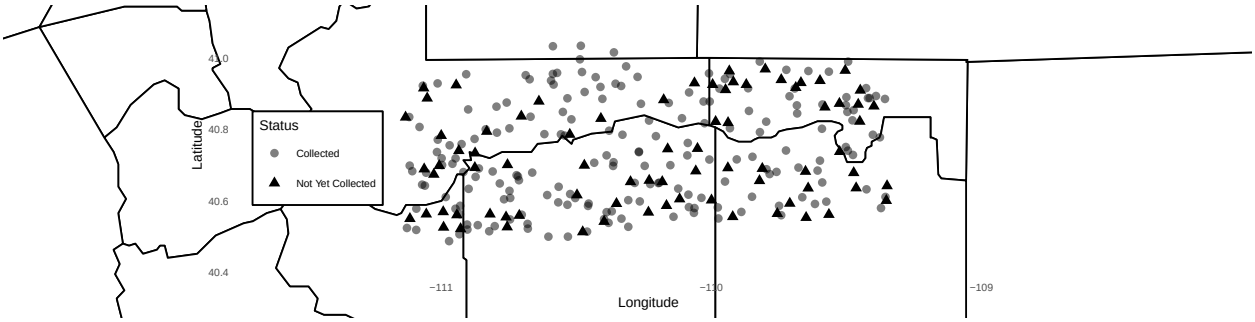

Figure 3: Scatter plot of observed lodgepool pine basal areas collected by the FIA in Northeastern Utah

Methodology & Method Evaluation

I first propose a LASSO regression model due to its ability to perform variable selection and regularization simultaneously, effectively managing high-dimensional data: Given the complex nature of Lodgepole pine basal area data, the initial LASSO model was fit using an expanded set of 50 engineered covariates. This set included not only the main environmental effects—*Longitude, Latitude, Slope, Aspect*, and *Elevation*—but

also all possible pairwise interactions between these variables and their second-degree polynomial terms. The motivation for this expanded covariate set was to capture potential non-linear relationships and interactions that might be significant in explaining variations in basal area. For instance, interactions such as *Elevation* combined with *Aspect* or *Slope* with *Latitude* could reveal more intricate patterns in *Lodgepole* pine growth that a simple linear model would fail to detect. By including polynomial terms, the model also allowed for the possibility of quadratic effects, such as diminishing or increasing returns to factors like *Elevation* and *Slope*, which may arise in this ecological context. LASSO regression was thus used to narrow down the set of 50 covariates to a more parsimonious fit by zero'ing some of the covariates out. By initially including a comprehensive set of covariates, the LASSO was able to determine which interactions and polynomial terms significantly contributed to explaining the variability in Lodgepole pine basal area. This approach ensures that key environmental drivers are not overlooked due to overly restrictive model assumptions. I define the LASSO model using the transformed distribution of *Lodgepole* values in Equation 1.

$$\arg\min_{\beta}\left\{\sum_{i=1}^{n}\left[\log(\text{Lodgepole}_i) - x_i'\beta\right]^2 + \lambda\sum_{p=1}^{P}|\beta_p|\right\} \tag{1}$$

Hence, we will make predictions with this model using the following general form by back-transforming Equation 2:

$$\widehat{\text{Lodgepole}} = \exp\left(X\hat{\beta} + \lambda^*\sum_{p=1}^{P}|\hat{\beta}_p|\right) \tag{2}$$

Where X is the the set of 50 factors (consisting of the main effects of *Longitude, Latitude, Aspect, Elevation,* and *Slope* and the engineered combinations between them in addition to the intercept—the covariates that the model didn't zero-out (upon evaluation) are summarized in Table 2). I use λ^* as the computed optimal penalty parameter as determined by k-fold cross validation[1]. The ability for LASSO to select (on average) the "best" set of parameters wins favor over other alternatives. Additionally, the LASSO model is relatively computationally simple and conceptually tractable, making the results easier to understand.

Secondly, I propose a *hedonic spatial autogressive model* to account for the violation of independent data. This allows for every data point to be influenced by every other data point in the data set *weighted by how close* the data are[2]. This model is outlined below in Equation 3. The spatial autoregressive model is advantageous for handling spatially correlated data, providing insight into how regional trends and local interactions impact *Lodgepole* pine growth. However, it relies on assumptions about the spatial structure, which could lead to limitations if the true spatial process is more complex than captured by Ω. However, while this model is more parsimonious than the proposed LASSO model as it includes fewer covariates, this model imposes rather strict parametric assumptions for correct model specification. This model imposes a strict assumption on ε necessary for (MLE) estimation (although Normality may be reasonable given that we perform our analysis on the log-transformed *Lodgepole* values). Additionally, computations required to simultaneously estimate both ρ and β are more intensive than the LASSO.

$$\begin{aligned} Y &= \rho\Omega Y + X\beta + \varepsilon \\ Y &= (I_n - \rho\Omega)^{-1}(X\beta + \varepsilon) \\ \varepsilon &\sim \mathcal{N}(0, I_n\sigma^2) \\ \implies Y|X &\sim \mathcal{MVN}\left((I_n - \rho\Omega)^{-1}X\beta, \sigma^2(I_n - \rho\Omega)^{-1}(I_n - \rho\Omega')^{-1}\right) \end{aligned} \tag{3}$$

[1] During my particular cross-validation procedure, I used $k = 10$, and estimated an optimal λ (selecting the optimal λ within one standard error of the minimum cross-validated error for greater parsimony) of 0.0252.

[2] In the spatial regression model, the measure of how close each observation i to another observation j in the data set is accounted for in the $n \times n$ weighting matrix, Ω. Thus, the entry, $\Omega_{[i,j]}$ denotes the spatial weight between data points i and j. I first calculate the Haversine distance matrix D, using the longitude and latitude distance between all pairs in the data. Ω is then calculated as $1/D$, where we assign zeros to the diagonal (that is, $\Omega_{[i,i]} = 0$) since the distance between any observation and itself is zero.

Given that the data are skewed, we use the log-transformed values of *Lodgepole* for Y. To address non-linearity in each of the covariate structures, polynomial factors of degree 2 are included. Therefore, using Maximum Liklihood Estimation (MLE) to derive estimates for ρ and the seven coefficients of interest, β, we can use Equation 4 to predict new values[3] for *Lodgepole*.

$$\begin{aligned} log(\widehat{\text{Lodgepole}}_i) = \hat{\rho}\Omega_{[i,-i]}log(\text{Lodgepole}_{-i})+ \\ (\beta_0 + \beta_1\text{Slope}_i + \beta_2\text{Aspect}_i + \beta_3\text{ELEV}_i + \beta_4\text{Slope}_i^2 + \beta_5\text{Aspect}_i^2 + \beta_6\text{ELEV}_i^2) \end{aligned} \tag{4}$$

For the LASSO regression, k-fold cross-validation[4] was employed to select the optimal penalty parameter to balance model complexity and prediction accuracy. Maximum likelihood estimation was used to estimate the optimal parameters of the spatial regression model, but no further cross-validation was needed to estimate any other hyperparameters. Both models were assessed using the adjusted-R-squared for the in-sample fit and a leave-one-out cross-validation (LOOCV) procedure to assess the out of sample predictive power[5] (using root mean-squared error as the measure of fit). These results are summarized below in Table 1.

Table 1: Summary of Model Comparisons

Model	Adjusted-R-squared (In-sample)	LOOCV RMSE
LASSO	0.5943	0.2597
Autoregressive Spatial Model	0.8945	0.2707

The autoregressive spatial regression model performs notably better when the adjusted-R-squared is used as the in-sample measure of fit. On this front, we acknowledge the possibly large number of (engineered) covariates in the LASSO model adjusting the R-squared value down. Both models perform similarly when the out-of-sample RMSE is considered, with the LASSO model having the slight edge. Although the spatial regression model addresses the spatial correlation the best, we will prefer the LASSO model for this analysis due to its robustness in its predictive power and interpretability. We also note that since the the spatial model is autoregressive, it requires other observations in the data to make a prediction for a given ith observation (refer back to Equation 4). The LASSO model does not require this.

Results

With the selected model, we estimated the standard errors through bootstrapping[6] to assess the the 95% confidence intervals on $\hat{\beta}$. The most significant coefficients are summarized in Table 2 in the Appendix for full coefficient results—*Note that these results use the scaled factors for each covariate, and by implication, the*

[3] This is made under the assumption that the weighting matrix, Ω, is sufficiently dense. That is, we have enough observations to begin with in Y to make new predictions. This assumption may be violated if the new observations themselves are related to each other in a way that's not explained through the weighting matrix Ω (which tells us how related the existing data are to each other and to a new ith observation). Note that the notation, $\Omega_{[i,-i]}$ denotes the partition of Ω consisting of the i row of Ω and all the cells on that row excluding the cell that belongs to the ith column of Ω. Similarly, $Lodgepole_{-i}$ denotes all of the observations in the vector *Lodgepole*, excluding the ith observation in that vector, that is, the one we're wanting to predict.

[4] During my particular cross-validation procedure, I used $k = 10$, and estimated an optimal λ (selecting the optimal λ within one standard error of the minimum cross-validated error for greater parsimony) of 0.0252.

[5] Note that RMSE for both models is on the log-scale of *Lodgepole*.

[6] To estimate the standard errors of $\hat{\beta}$, we first computed $B = 10,000$ bootstrap samples from Y (the water flow metric) and our covariate matrix X with replacement of size $K = N = 114$ where N was the total number of observations in the data set. Using the optimal penalty parameter, λ^*, as computed through our cross-validation step previously, we estimated B # of LASSO Regression models and computed the standard error of each $\hat{\beta}_j$ for $j = 1, ..., P$ given our $\hat{\beta}$ vector of dimension P through the following computational sequence: (1) For each $\hat{\beta}_j$, compute $\bar{\hat{\beta}}_j = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_j$. (2) $SE(\hat{\beta}_j) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_j - \bar{\hat{\beta}}_j)^2}$. (3) We compute the 95% C.I. as $\left(2\hat{\beta}_j - \hat{\beta}_{j_{\text{boot}}}^{(0.975)}, 2\hat{\beta}_j - \hat{\beta}_{j_{\text{boot}}}^{(0.025)}\right)$, where $\hat{\beta}_{j_{\text{boot}}}^{(0.975)}$ and $\hat{\beta}_{j_{\text{boot}}}^{(0.025)}$ are the 97.5th and 2.5th quantiles of the bootstrapped distributions of $\hat{\beta}_j$, respectively.

magnitudes of each covariate reflects relative importance to Lodgepole base area. To assess feature importance, Table 2 also records the *inclusion frequency*[7].

From these results, we infer that *Elevation* is the most influential factor for healthy Lodgepole pines. Given that *Elevation* consists of a two-degree polynomial, the second degree being negative being significant, suggests that there exists a range such that *Lodgepole* pine growth is ideal. The analysis found this elevation to be at approximately[8] 9581 (9502, 9664) feet.

We use Equation 2 to make predictions for the remaining 78 observations in the FIA data set. These are summarized visually in Figure 4. The predictions are given numerically in Table 3, with 95% bootstrapped[9] confidence intervals given as well.

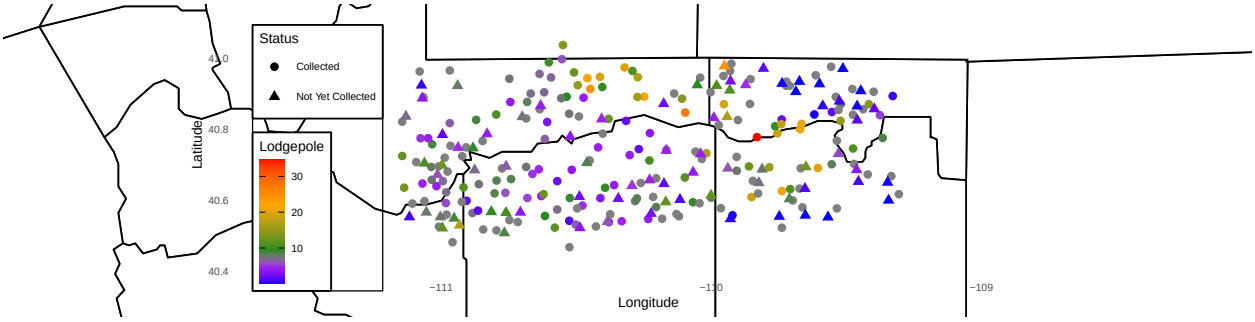

Figure 4: Predicted lodgepool pine basal areas imposed on the scatter plot of observed lodgepool pine basal areas collected by the FIA in Northeastern Utah

Discussion & Conclusion

This study examined the environmental factors affecting Lodgepole pine basal areas in the Uinta National Forest, using LASSO regression and spatial autoregressive models. The analysis revealed that *Elevation*, *Aspect*, and *Slope* are key determinants of Lodgepole pine growth, along with significant interactions between these variables. Specifically, basal areas tended to decrease at higher elevations, perhaps due to harsher conditions such as lower temperatures and reduced soil moisture. *Slope* effects were complex; while steeper slopes generally reduced basal areas due to potential erosion and drainage issues, interactions with *latitude* showed that these effects varied across different parts of the study area, reflecting localized environmental conditions. While the LASSO model used in the analysis is believed to provide a robust and parsimonious we address its limitations here. The LASSO model, in contrast from the proposed autoregressive spatial model, did not account for spatial autocorrelation, which could lead to biased inferences if local dependencies are significant. Notably, the models relied on log-transformation to approximate normality, though this assumption may not hold perfectly across all observations. Polynomial terms were included to account for non-linearities, but there may still be higher-order effects or interactions not captured by the current models. Future research could benefit from exploring more flexible spatial models or integrating additional environmental variables, guiding targeted forest management strategies that promote the sustainability of Lodgepole pine populations.

[7] This is a metric of robustness to variation in random sampling. Hence, a larger inclusion frequency indicates a stronger relationship with *Lodgepool*. We use inclusion frequency in part to assess how well these factors explain the overall cumulative basal area of lodgepole pines. For a given jth factor, using the bootstrap distributions, we calculate the inclusion frequency as $\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(\hat{\beta}_j^{\text{lasso, } b} \neq 0)$.

[8] This was calculated using Equation 2. We first hold all other factors besides *Elevation* constant, then, using the fitted λ^*, we compute B number number of new LASSO regression models using a bootstrap sample, each time estimating the predicted *Lodgepole* with respect to changes in *Elevation*. The resulting interval is the 95% quantile interval on the bootstrapped prediction distribution.

[9] A Monte carlo standard error is applied to obtain 95% prediction intervals. That is, I use the approximation of $SE(\hat{y}_i) \approx \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{y}_i^b - \bar{\hat{y}}_i\right)^2}$ for each observation, y_i.

Appendix

Table 2: Lasso Coefficient Estimates

Covariate	Description	Estimate	95% CI	Inclusion Frequency
Intercept		1.481*	(1.016, 1.539)	1.000
LON	Longitude coordinate of the plot	2.469*	(1.584, 4.939)	0.972
Slope	Average slope of the plot in degrees (0=flat, 90=vertical)	-1.127*	(-2.255, -0.347)	0.961
Aspect	Counterclockwise degrees from north facing (90=west, 180=south, 270=east)	1.876*	(1.283, 2.909)	0.998
Aspect (squared)	Counterclockwise degrees from north facing (90=west, 180=south, 270=east) (squared)	2.676*	(1.945, 4.254)	0.995
ELEV	Elevation of plot centroid in feet	16.348*	(14.947, 25.607)	0.999
ELEV (squared)	Elevation of plot centroid in feet (squared)	-14.727*	(-23.474, -13.71)	0.998
LON (squared):LAT	Interaction effect between LON (squared) and LAT	-31.729*	(-44.551, -19.606)	1.000
LON:Slope (squared)	Interaction effect between LON and Slope (squared)	2.893	(-4.199, 5.786)	0.487
LON (squared):Slope (squared)	Interaction effect between LON (squared) and Slope (squared)	-10.22	(-20.439, 5.204)	0.946
LON:ELEV	Interaction effect between LON and ELEV	9.443	(-39.022, 18.886)	0.913
LON:ELEV (squared)	Interaction effect between LON and ELEV (squared)	-7.237	(-14.474, 32.008)	0.761
LON (squared):ELEV (squared)	Interaction effect between LON (squared) and ELEV (squared)	3.863	(-14.802, 7.726)	0.765
LAT:Aspect (squared)	Interaction effect between LAT and Aspect (squared)	4.321	(-6.308, 8.641)	0.668
LAT (squared):Aspect (squared)	Interaction effect between LAT (squared) and Aspect (squared)	-1.484	(-2.967, 9.008)	0.482
LAT (squared):ELEV	Interaction effect between LAT (squared) and ELEV	3.075	(-18.461, 6.15)	0.461
LAT (squared):ELEV (squared)	Interaction effect between LAT (squared) and ELEV (squared)	0.33	(-19.118, 0.66)	0.521
Aspect (squared):ELEV	Interaction effect between Aspect (squared) and ELEV	8.895	(-25.859, 17.791)	0.853

Table 3: Predictions for the remaining FIA Lodgepole pine base areas

Longitude	Latitude	Slope	Aspect	Elevation	Lodgepole Prediction	Lower Bound	Upper Bound
-109.520	40.872	6	306	7430	0.628	0.389	1.012
-110.952	40.560	6	90	9874	8.748	7.896	9.690
-110.995	40.563	6	156	9278	7.955	7.071	8.950
-110.636	40.872	6	70	10880	4.142	3.347	5.126
-110.761	40.696	7	327	10824	6.212	5.092	7.577
-109.582	40.869	8	43	7366	0.431	0.260	0.714
-110.406	40.833	8	3	11151	4.693	3.495	6.303
-110.482	40.518	9	111	8474	3.396	2.851	4.047
-109.570	40.556	10	250	6128	0.005	0.002	0.018
-109.932	40.829	10	352	10601	16.466	14.167	19.137
-110.998	40.518	11	260	9821	11.672	10.525	12.943
-111.056	40.567	12	170	9125	7.927	7.014	8.959
-111.123	40.834	12	303	9309	7.067	6.287	7.944
-110.226	40.652	13	151	10627	4.512	3.864	5.270
-110.765	40.517	14	254	10105	8.596	7.763	9.518
-109.465	40.686	15	212	8184	3.849	3.056	4.848
-109.706	40.597	17	205	8798	8.478	7.378	9.742

-109.993	40.831	17	280	11036	4.229	3.268	5.473
-110.171	40.739	17	300	11007	4.340	3.383	5.568
-109.756	40.555	20	175	6982	0.119	0.060	0.238
-110.944	40.927	20	335	8909	8.058	7.064	9.193
-110.529	40.789	20	87	10816	3.788	3.112	4.612
-110.876	40.688	20	100	10081	6.936	6.266	7.677
-109.936	40.695	21	75	10605	6.180	5.314	7.188
-109.646	40.551	22	192	6546	0.027	0.010	0.068
-110.766	40.563	22	17	9609	10.476	9.394	11.681
-109.938	40.919	22	174	9225	10.800	9.585	12.168
-111.019	40.682	22	151	10240	5.894	5.295	6.560
-109.402	40.862	23	338	7755	1.707	1.198	2.432
-110.877	40.742	23	8	10304	9.180	8.212	10.263
-109.346	40.645	24	160	6761	0.080	0.036	0.180
-109.927	40.554	25	80	7379	0.338	0.205	0.556
-111.059	40.695	25	290	9749	10.635	9.574	11.814
-111.002	40.701	25	320	9944	12.212	11.035	13.516
-110.491	40.620	25	96	10943	2.538	2.015	3.195
-110.047	40.739	25	220	10739	4.379	3.663	5.235
-110.171	40.877	25	123	11147	2.323	1.732	3.115
-109.344	40.601	28	192	6280	0.013	0.004	0.040
-111.116	40.558	28	187	7530	0.851	0.550	1.317
-110.292	40.651	28	22	10755	5.084	4.237	6.100
-109.683	40.909	30	2	6422	0.014	0.005	0.039
-109.808	40.964	30	16	7831	1.726	1.243	2.397
-110.064	40.685	30	137	10715	3.859	3.246	4.589
-109.743	40.936	31	359	6778	0.093	0.042	0.206
-109.644	40.690	31	246	9222	12.585	11.168	14.180
-110.229	40.563	32	165	8359	2.612	2.150	3.173
-109.943	40.968	33	347	9336	24.219	21.561	27.205
-109.819	40.689	33	138	10217	7.949	7.151	8.837
-109.461	40.645	34	256	7524	0.978	0.631	1.517
-111.062	40.881	35	230	8769	3.674	3.190	4.231
-109.511	40.962	37	241	6402	0.010	0.003	0.027
-110.936	40.522	37	344	9805	16.801	15.144	18.638
-109.455	40.823	38	140	7902	1.474	1.085	2.002
-109.676	40.933	40	142	6246	0.005	0.002	0.016
-111.064	40.924	40	158	7673	0.484	0.330	0.709
-110.111	40.607	40	218	7844	1.094	0.791	1.513
-110.388	40.541	40	55	9153	6.258	5.542	7.066
-109.878	40.917	42	72	8457	4.873	4.079	5.821
-109.525	40.731	42	244	8439	5.490	4.582	6.577
-109.988	40.919	45	210	9538	11.088	9.925	12.388
-110.055	40.921	45	165	10000	9.427	8.521	10.430
-110.823	40.562	49	311	10202	9.589	8.632	10.652
-109.453	40.872	50	50	6980	0.095	0.048	0.189
-110.168	40.596	50	331	10010	11.806	10.671	13.061
-109.921	40.934	51	255	8068	2.600	2.006	3.370
-109.995	40.606	51	355	8928	11.853	10.401	13.508
-110.349	40.601	53	223	8225	2.143	1.718	2.675
-110.463	40.699	55	355	10674	8.371	7.101	9.867
-109.456	40.915	60	320	6069	0.003	0.001	0.009
-109.820	40.651	60	73	8896	7.269	6.367	8.298
-110.823	40.785	60	147	10381	4.099	3.641	4.615
-109.606	40.936	65	160	6738	0.033	0.015	0.075
-110.177	40.646	70	65	8071	1.769	1.366	2.291
-110.709	40.835	74	336	10595	7.392	6.367	8.582
-109.645	40.642	75	118	7809	1.150	0.822	1.609
-110.998	40.788	75	64	8241	1.582	1.272	1.968
-110.702	40.567	75	114	8685	3.562	3.070	4.134
-110.939	40.746	80	220	9544	4.874	4.363	5.444