Assignment 3

Elliot Spears

3/28/2022

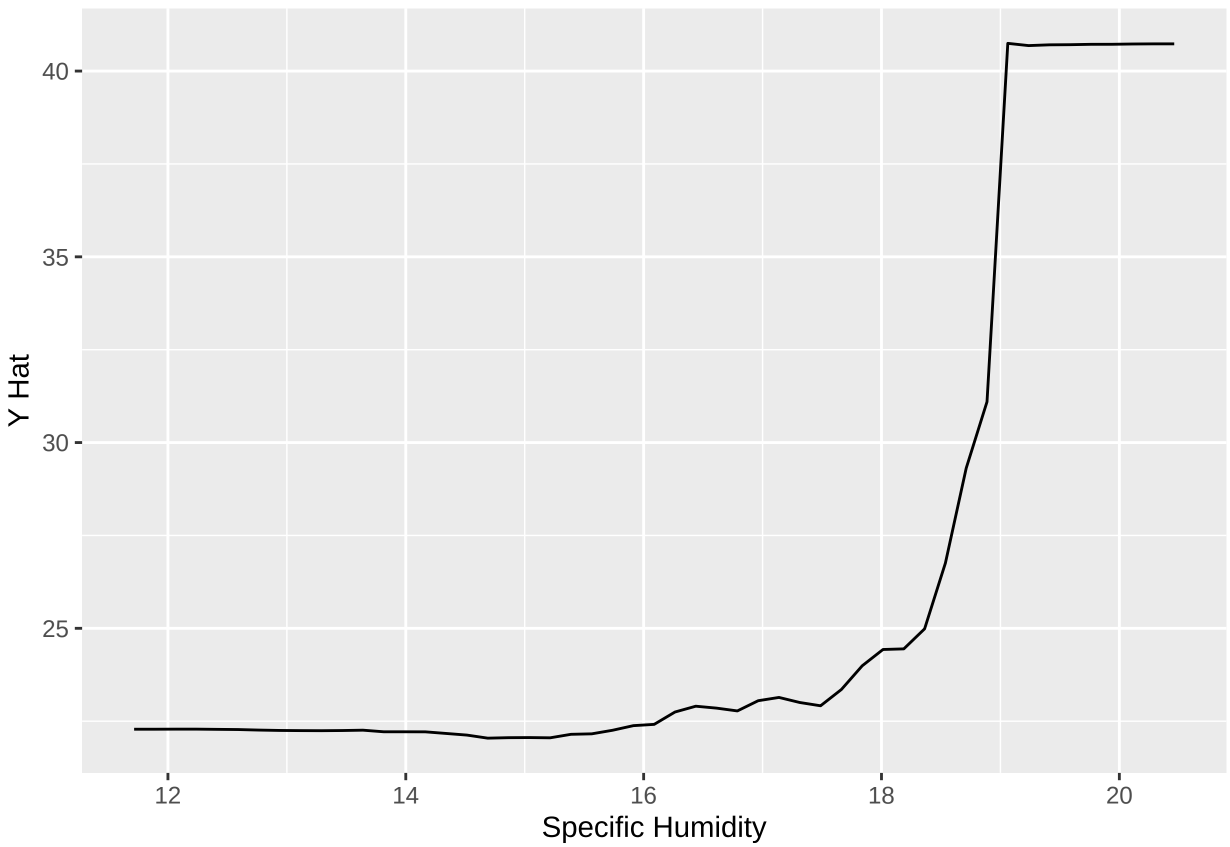
Question 1

We can't simply run a regression using crime as the LHS variable and the number of cops as the RHS variable because this confounds the question as to "what causes what." Cities that have a lot of police officers tend to have a lot of crime, but the reason there is a lot of crime is not because there are a lot of police officers, it's the other way around: there are a lot of police officers because there is a lot of crime. Hence, running a regression of crime on police officers would indicate that having more police officers has a positive effect on crime rates, which is not true. So we need to control for this factor when investigating this question. The researchers investigated what happens to crime rates in Washington D.C. when there are "high-alert days." Since D.C. is susceptible to terrorist attacks, the city's leadership issues many high-alert warnings every year in order to help hedge against a terrorist attack. In practice, this means that the police presence is increased across D.C. This provides us with an opportunity to see what happens to crime when there is a sudden and large increase in police presence all over the city.They found that on high-alert days, there was a drop in crimes when the high-alert took place, so it appears that more police means less crime, which is what we would expect. The Metro ridership controls for the fact that on high alert days we might expect tourism to decrease and that people might stay home instead of going out on the town, including the criminals. The metro ridership didn't change the sign of the high-alert variable, nor did it drastically effect the magnitude of the coefficient, so the overall result still holds. The model in the first column allow for heteroskedastic errors. It also shows that district 1 had a higher drop in crime on high-alert days than other districts did on average, however the other districts variable was not significant at the 5% level. It also shows that metro ridership actually increased on the high-alert days, this was significant at the 1% level.

Question 2

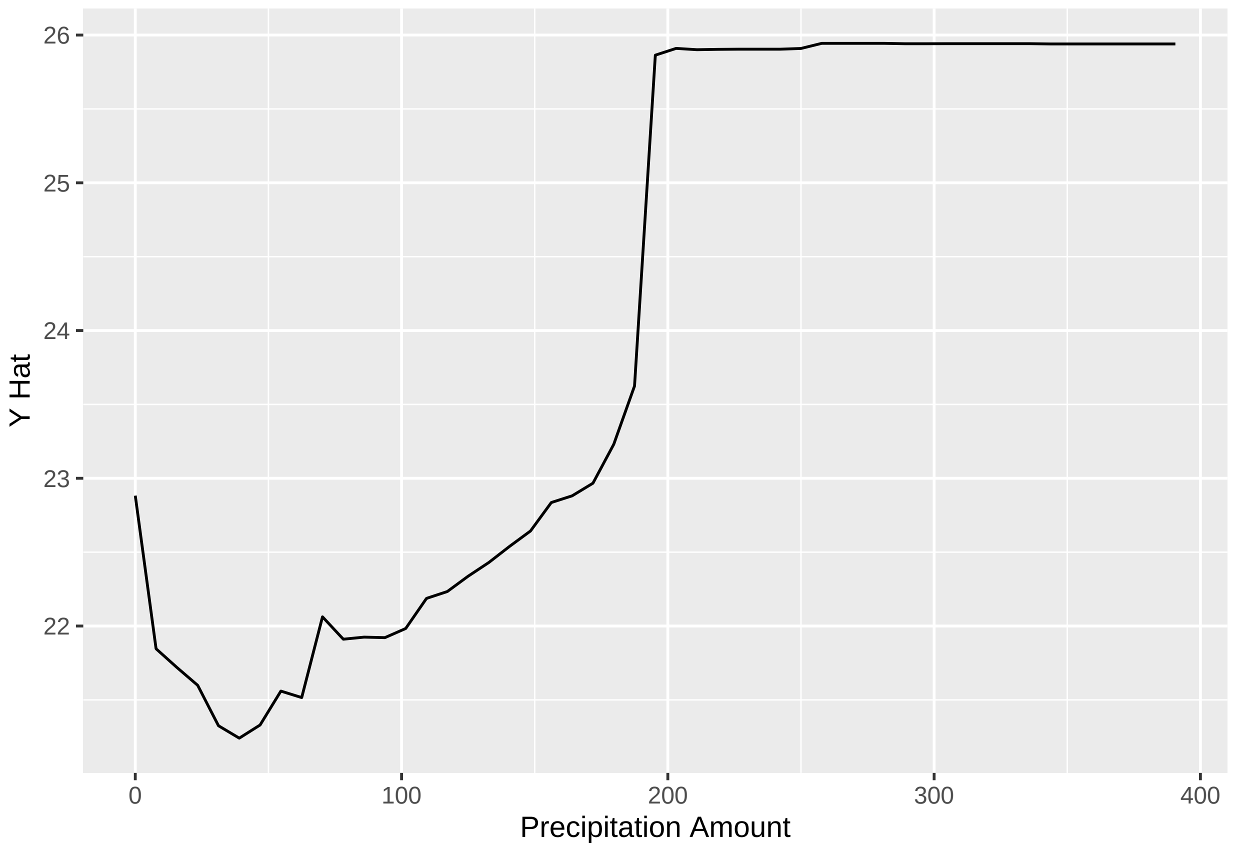
In this question we aim to use CART, random forests, and gradient-boosted trees in order to predict dengue cases based on the features available in the data set. We'll run the various models using the variables that offer the best predictive power and see which one has the lowest RMSE when cross validated against the testing data. We will then create partial dependence plots to illustrate the findings.

The lowest RMSE is derived from the random forest. Hence, we will use this model to create our partial de-

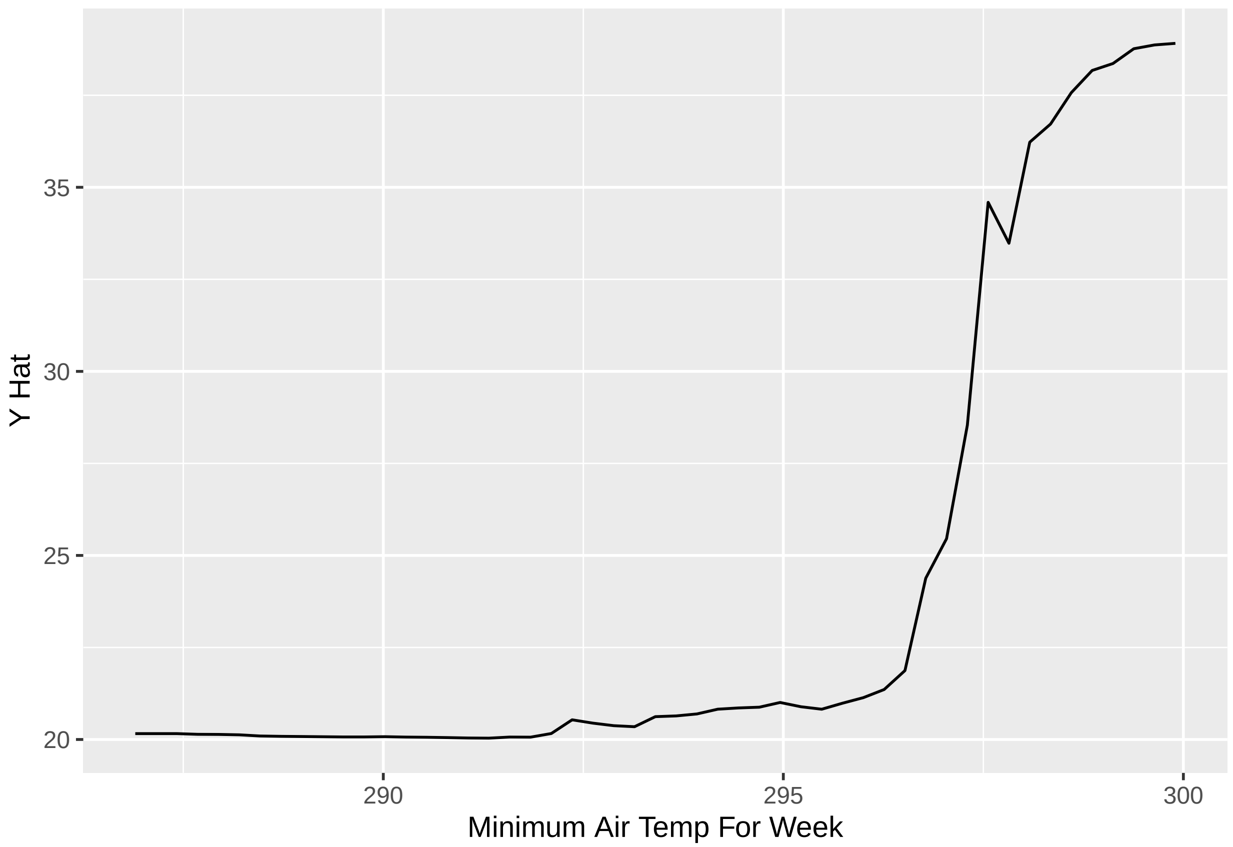


pendence plots.

The above plot demonstrates that once specific humidity reaches about 15 units of grams of water per kilogram of air for the week, we have an increasing number of cases up until about 19 units of grams of water per kilogram of air for the week, at which point the cases begin to level off.



Above we can see that cases tend to go down when rainfall increases from zero up to about 50 millimeters, at which point the cases begin to increase all the way up to about 200 millimeters, where the cases level out for all additional precipitation amounts.



For the plot I chose the minimum air temperature for the week to see what happens to cases. We would expect really low minimum air temperature weeks to see lower cases as they likely take place in cold months on average. Conversely, we would expect high minimum air temperature weeks to take place in the summer, where mosquitoes are more prevalent. That's exactly what we observe in the graph above.

Question 3

term	estimate	std.error	statistic	p.value
(Intercept)	-739.00	242.52	-3.05	0.00
CS_PropertyID	0.00	0.00	-1.88	0.06
size	0.00	0.00	7.80	0.00
empl_gr	1.36	2.30	0.59	0.55
stories	8.52	3.60	2.37	0.02
age	0.84	2.29	0.37	0.71
renovated	-84.07	67.41	-1.25	0.21
class_a	345.71	58.87	5.87	0.00
class_b	200.89	44.71	4.49	0.00
LEED	129.36	443.25	0.29	0.77
Energystar	172.80	477.07	0.36	0.72
green_rating	-26.46	479.75	-0.06	0.96
amenities	145.65	32.79	4.44	0.00
cd_total_07	-0.01	0.02	-0.37	0.71
hd_total07	0.07	0.01	6.23	0.00
total_dd_07	NA	NA	NA	NA

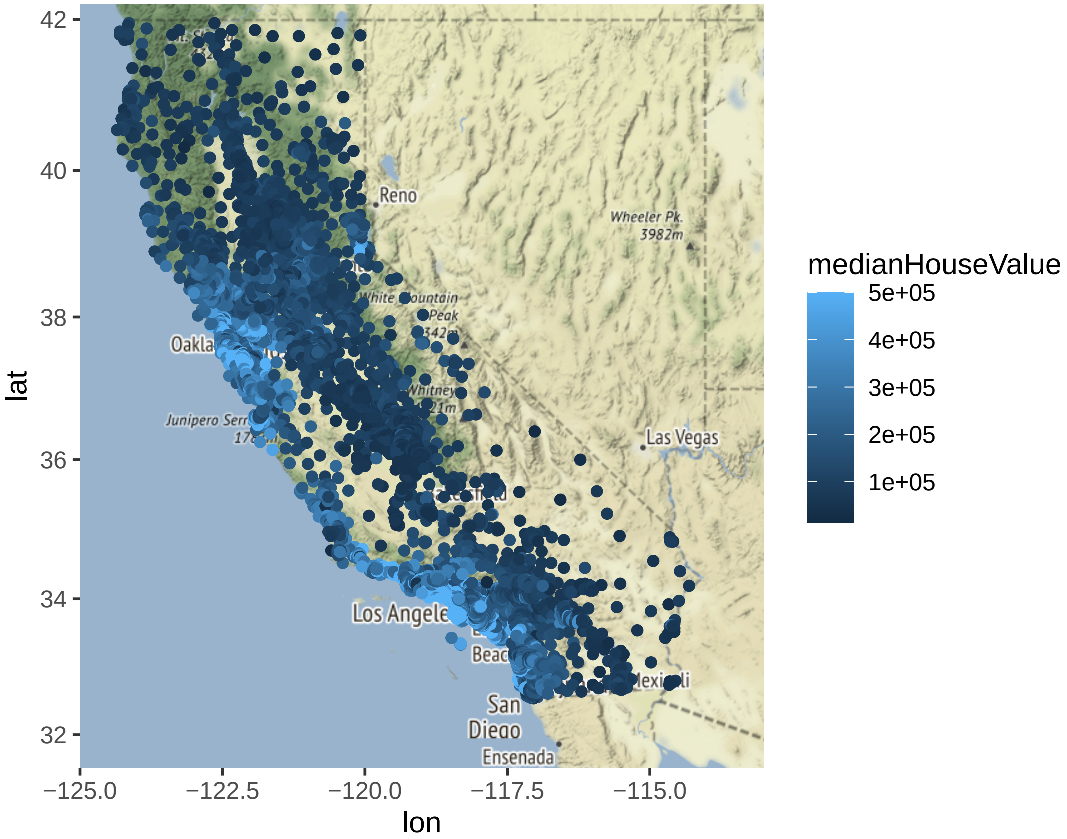
term	estimate	std.error	statistic	p.value
Precipitation	-47.21	12.10	-3.90	0.00
Gas_Costs	-27070.85	10011.67	-2.70	0.01
Electricity_Costs	20244.20	3199.15	6.33	0.00
City_Market_Rent	97.95	1.83	53.65	0.00
I(ageˆ2)	-0.02	0.02	-1.19	0.23
I(Precipitationˆ2)	0.75	0.18	4.12	0.00
I(storiesˆ2)	-0.24	0.07	-3.50	0.00
size:age	0.00	0.00	-1.69	0.09
age:renovated	1.59	1.07	1.48	0.14

It looks like the generic forest model takes the cake. In the case of the linear model, we removed the net variable as well as cluster variable, but kept all the other variables plus interactions between age and size, as well as age and renovation status. I also included three quadratics: age, precipitation, and stories. I figured that the effects of those variables die off eventually. This model seems to have the best predictive power. This was the best linear model I could make and it still was bested by the generic forest model by a wide margin. It also appears that electricity costs have the largest effect upon the revenue per square foot per calendar year. The next largest positive effect came from whether or not the building was a “Class A” building. the green rating does not have as large of an effect as we would have expected, nor is it even statistically significant in the linear model, which is problematic since that is the main variable we are trying to estimate. It seems however, based upon the findings from the random forest, that we can likely sign the green rating variable with a positive sign, which means that the green rating variable likely has a positive effect on revenue per square foot per calendar year.The linear model has such a huge standard error compared to the value of the coefficient that this isn’t certain, though. We can conclude that while green rating has some mild predictive power for the revenue per square foot per calendar year. If I were a commercial real estate developer building apartments, I would be more concerned with getting my apartment complex the highest class rating possible in order to maximize revenue.

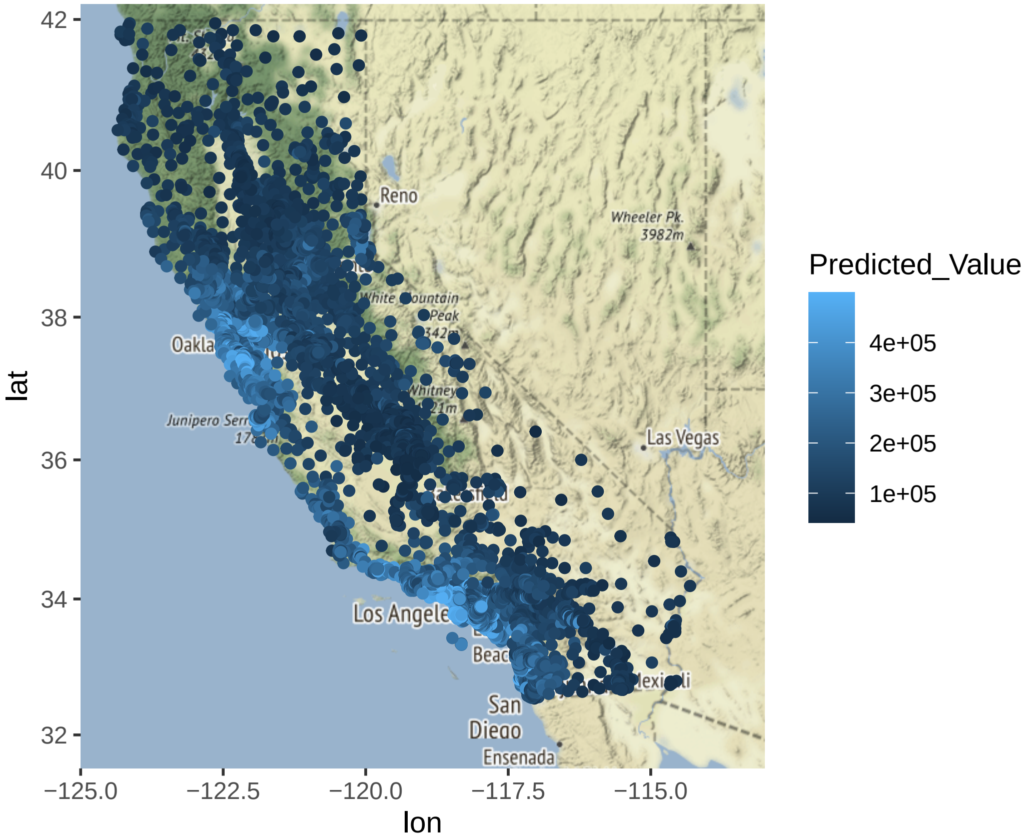
Question 4

term	estimate	std.error	statistic	p.value
(Intercept)	-3553196.81	69368.15	-51.22	0
longitude	-42219.29	791.03	-53.37	0
latitude	-41757.99	746.45	-55.94	0
housingMedianAge	1153.32	48.02	24.02	0
totalRooms	-9.81	0.89	-11.01	0
totalBedrooms	123.50	7.68	16.08	0
population	-36.49	1.17	-31.10	0
households	39.68	8.28	4.79	0
medianIncome	40906.41	375.39	108.97	0

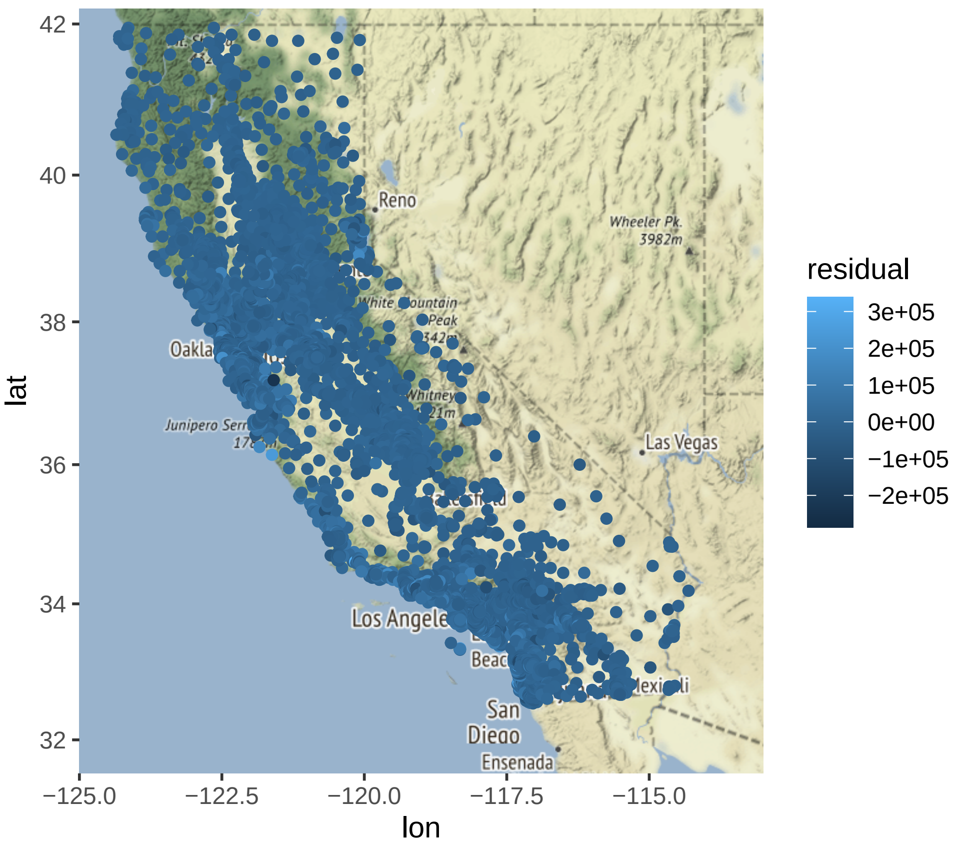
I tried to tinker with the model a little bit and create my own customized forest with an interaction between rooms and bedrooms, as well and a quadratic terms for median house age, assuming that the age of a house has an increasing, but diminishing effect upon the house value. My custom model turned out to have a slightly lower RMSE when compared to the original “stock” forest model. It appears that median income and the median age of a home have the largest positive effect upon the expected value of a home’s value. All of the variables are statistically significant at the 5% level. It is surprising that as median home age increases there is an increase in home value. I don’t have an explanation for this as I’m not familiar with the California real estate market, but it certainly is noteworthy. Not surprisingly, the median income of the people that make up the tract has a huge effect upon the expected home value. As you increase income, you increase the size and quality of home that one can afford and, hence, home price/value. Below, we include figures to help the reader get a sense of how home values change with longitude and latitude.



The above map plots the original data, and provides us with a visual of how median house value changes by longitude and latitude.



The above map shows us how my custom forest model helps us predict median home values by longitude and



latitude.

This last plot shows the model's errors versus longitude and latitude.