Electric Car Modeling

Drew Cornmesser, Levi Johnson, Colin Short

Dec. 12, 2023

Statement of Project

A common issue with electric vehicles (EV) and EV owners is the limited availability of chargers. Gas stations are far more accessible and easy to find than EV charging stations. This results in owners opting to take their gas cars on trips instead of the EV's that they own, or is a potential limiting factor in people opting to purchase gas cars instead of EV's. To resolve this, it is necessary to understand hotspots of where EV's are located across the country so that we can in turn make educated decisions on where to place EV chargers. By employing strategic planning for EV charging station deployment we are allowing for chargers to be more accessible and available to the general population so that they might opt to take their EV's on trips instead of gas vehicles. The data utilized has approximately 160,000 samples and is sourced from Kaggle. It only includes EV data from the State of Washington so this technology is primarily aimed at employing strategic planning for EV charging station deployment in the State of Washington.

This project also aims to identify spatial patterns in relation to urban versus rural areas so that we can further employ strategic planning for EV charging station deployment in the State of Washington.

This ultimately makes the primary question:

- How are EV's distributed across the State of Washington, and what spatial patterns emerge, particulary in relation to urban versus rural areas?

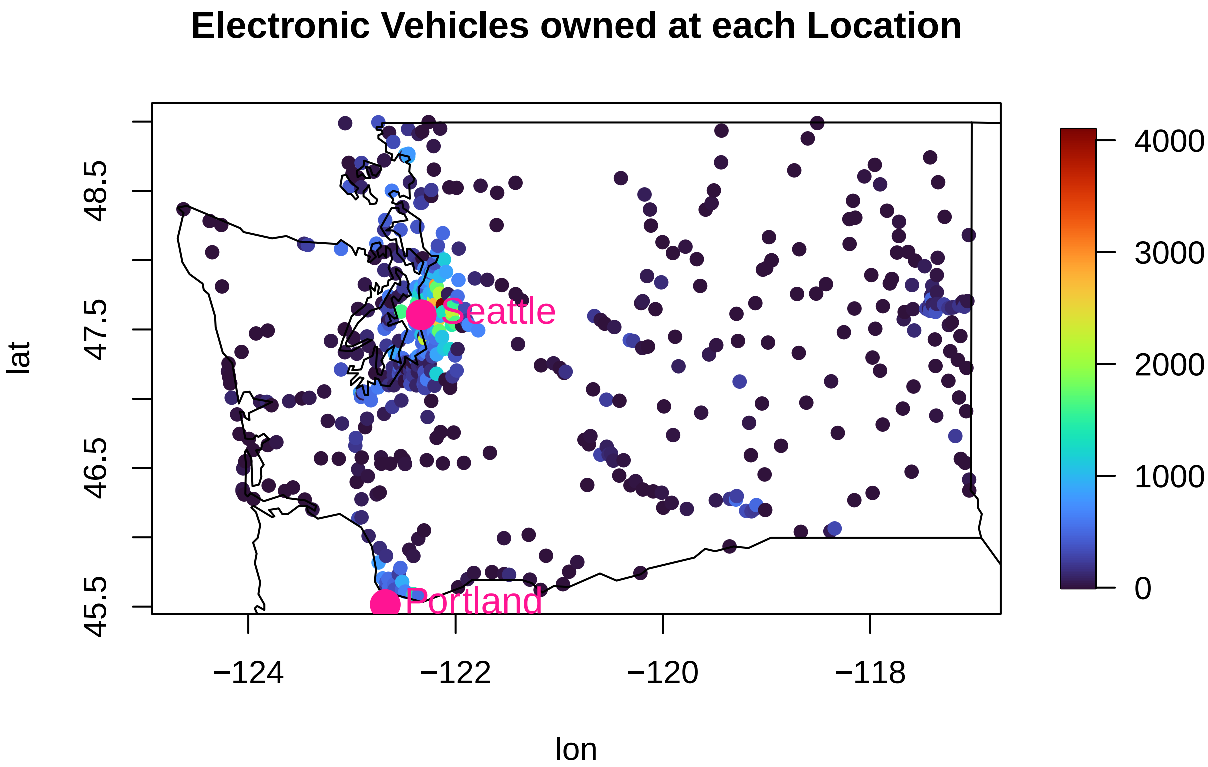
Data Description

The data is a rather large dataset as it has roughly 159,458 rows of data (or samples). This large of a dataset allows for us to have confidence in our spatial model and to be able to accurately predict and show the spatial trend of electric vehicles. It should also be noted that all of the samples in the data set are in the State of Washington. This is due to the raw nature of the data which was sourced off of Kaggle. The dataset also has 13 columns. These columns include the *County*, and *City* in the state of Washington. The *State* column is also present but every value has *WA* since every row corresponds to the State of Washington. Additional location information are a *Latitude* and *Longitude* column. The remaining columns are descriptors for the particular electric vehicle. These include the *Model*, *Make*, and *Year* of the EV. A *Vehicle Type* column is also included. The values in this column are either *Plug-in Hybrid Electric Vehicle (PHEV)* for Hybrid-Electric cars or *Battery Electric Vehicle (BEV)* for fully electric cars. The last two columns are *County EV Count* and *City EV Count*. These columns contain the amount of EV's that are present in a respective county/city.

To reproduce the analysis access to the cleaned_electric_car_data.csv file is needed. This file will be included in the submission. Additionally, it is necessary to use the *fields*, *dplyr*, and *scales* libraries in *R* in order to reproduce the analysis.

Statistical and Graphical Methods

Refer to the below *bubblePlot* of the count of electric vehicles in the State of Washington. Notice The first thing that the group noticed in this preliminary spatial analysis is the high pocket of EV's clusters in that one area. After plotting the city of Seattle on the map we can see that the pocket of electric vehicles are primarily clustered and dense in the urban area of Seattle. Notice Portland at the bottom of the map as well shows higher amounts of EV's in its suburbs. Portland is technically in Oregon but it is on the border and its suburbs do reside in Washington so becasue of that it does add to the trend we see of more urban centers have higher amounts of EV's. The rest of the plot shows low EV counts in more rural areas as well. The below plot leads us to believe that urban areas will see higher amounts of EV's than rural areas. N

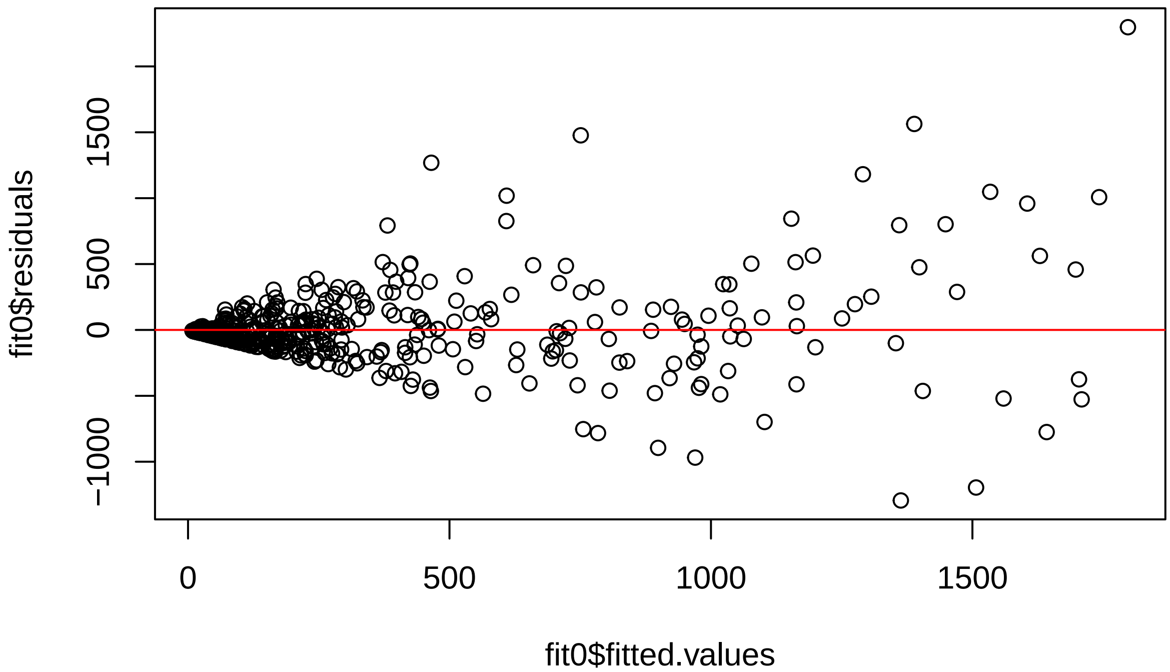


This looks much better and definitely will be more useful. When looking at this you can sort of see that there is really close to zero values in most of the rural areas. However, by citys such as Seattle or Portland the values are much much higher. This is probably due to multiple things. There is larger populations at cities so it work makes sense there is more cars and therefore more EV's. Even if there is a lower percentage of EV's there still will be more due to the increased population. As well with the increased infrastructure there is mostlikely more charging stations and therefore more viable to have an EV. If we want to account for the first part we could potentially model percentage of population that has an EV by taking our ev Values devided by total car ownership at the location.

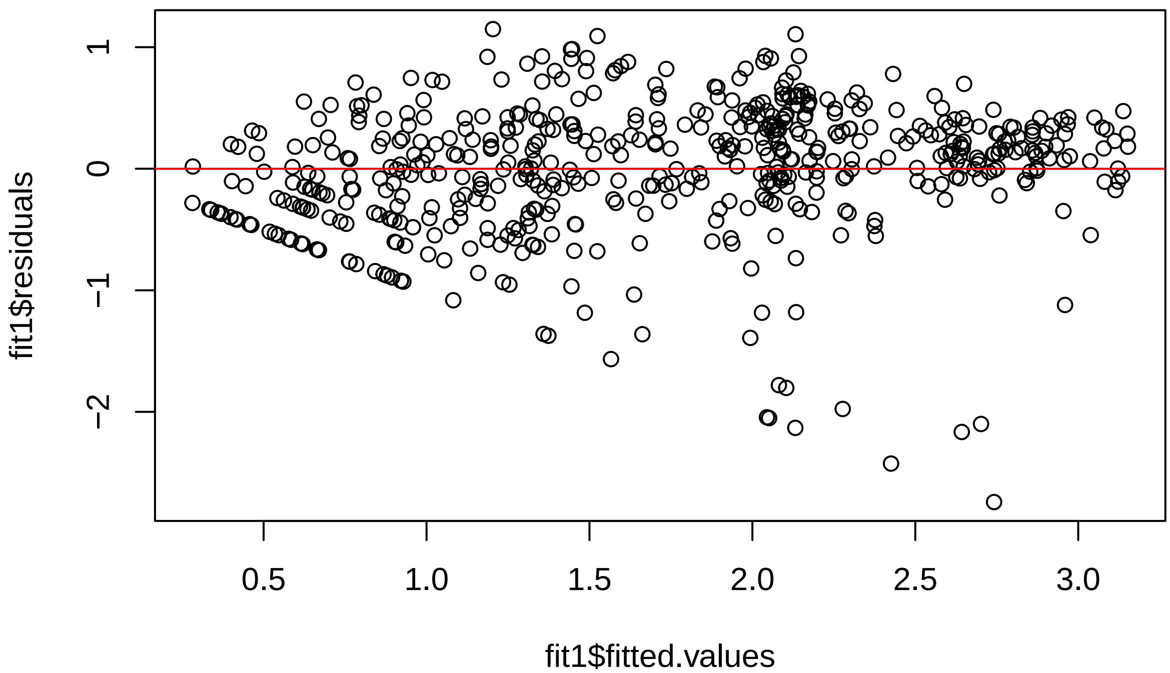
Based on the size of this location as well as the number of datapoints across the state, we are confident that we will be able to fit spatial process right away. Because of that there should not be a need to utilize various spatial analysis techniques such as *fasttps* or *lattusKrig* to cut down on computation time and get a better fit.

In the *spatialProcess* the s variable are the *Longitude* and *Latitude* columns in the dataset. The y response variable is the total count of electric vehicles at that respective location in the dateset.

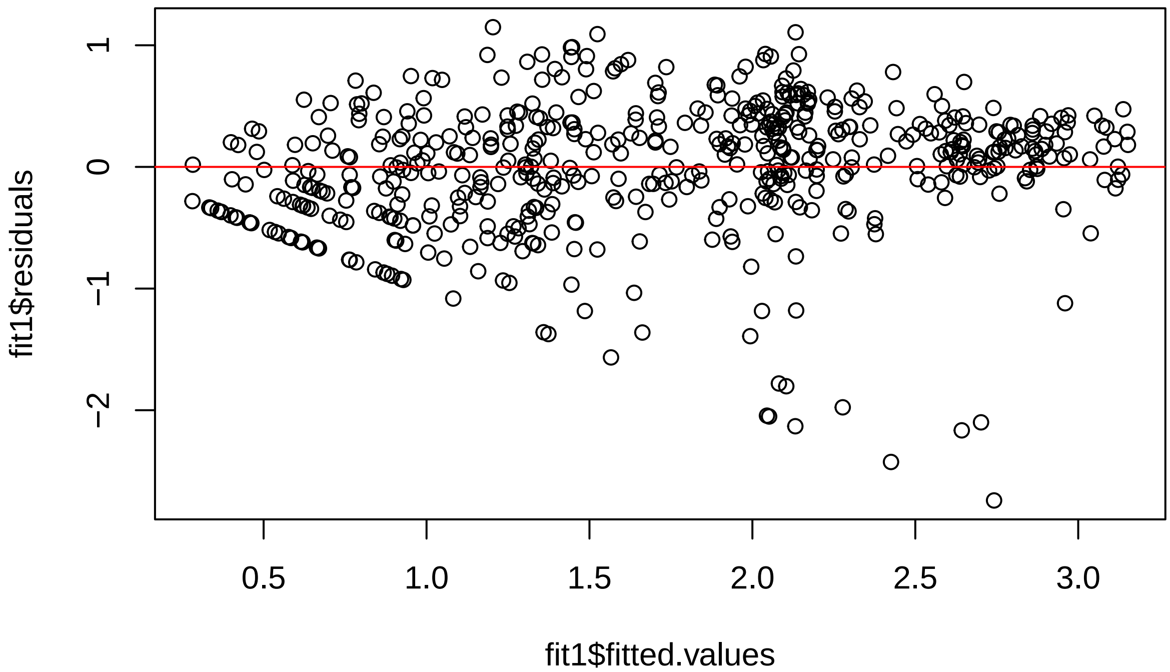
The below is the residuals versus fitted plot from the spatial process. The presence of a cornucopia in the residuals shows that there is a close relationship between small residuals and fitted values being pretty close while larger values are much more varied. However, this will most likely be a hard problem to overcome as the rural locations in Washington State are around 0 or 1 making these locations easy to predict for the spatial model. There is more variance in the cities, which ultimately makes sense as there is more data at those locations leading to a larger variance in the bigger city centers.



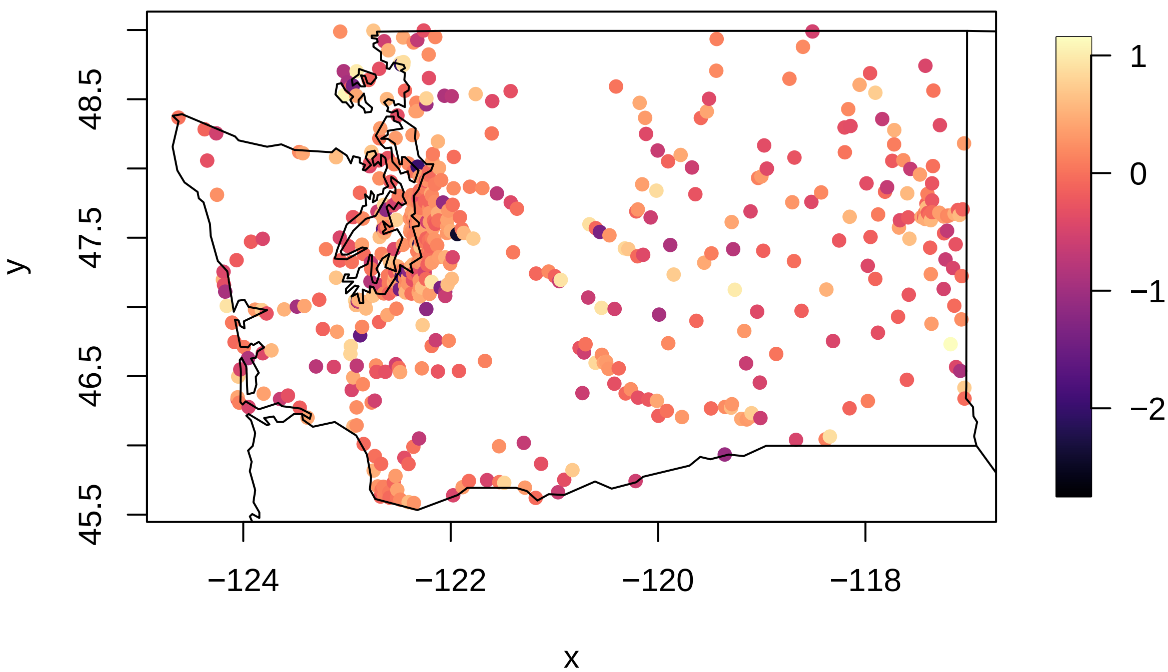
Potential solutions to recitfy the above issue include a log transformation of the response variable (count of EV's). This is done below:



The above residuals versus fitted plot below of the logged response variable looks much improved from the initial plot. As such we can proceed with this model.



Below is a bubble plot of the residuals for the spatial process with the logged response variable. Overall, the residuals appear to be well distributed with no real trend in where the errors are occuring. Logging the response variable for our spatial process definetly created a better spatial model.



The above spatial process model with the logged response variable uses a default $\nu = 1$ (the Matern Covariance). Below we try an Exponential smoothness ($\nu = 0.5$) and a stronger smoothness ($\nu = 1.5$) and assess which model performs better. We are still logging the response variable (count of electric vehicles) as well to stay consistent with the incorporated spatial process models thus far.

Below are the MLEs for the spatial model utilizing the Matern Covariance ($\nu = 1$).

```
## lnProfileLike.FULL
```

```
##          -573.6487
```

```
## lnProfileREML.FULL
##          -579.2239
```

Below are the MLEs for the spatial model utilizing the Exponential Smoothness ($\nu = 0.5$).

```
## lnProfileLike.FULL
##          -575.4982
```

```
## lnProfileREML.FULL
##          -580.8355
```

Below are the MLEs for the spatial model utilizing the larger smoothness ($\nu = 1.5$).

```
## lnProfileLike.FULL
##          -573.4956
```

```
## lnProfileREML.FULL
##          -579.1969
```

The MLE values above show that using a larger smoothness ($\nu = 1.5$) allows for higher MLE's and as such the higher model. This makes sense since there are rural locations with a sparse amount of data found at them. This allows for a smoothness of $\nu = 1.5$ to work well on those sparse locations.

The higher smoothness value leads us to wonder if even higher smoothness ($\nu > 1.5$) values would better represent the data.

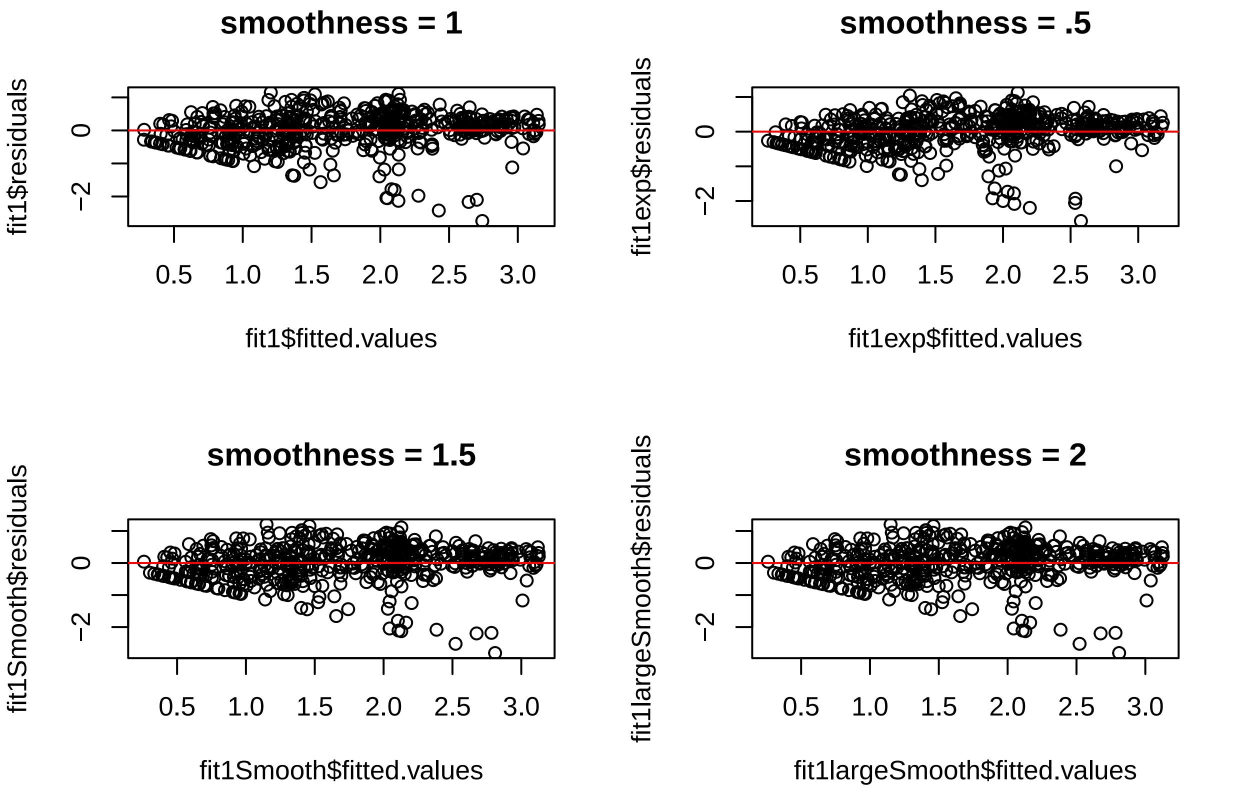
We try a $\nu = 2$ below:

```
## lnProfileLike.FULL
##          -573.5707
```

```
## lnProfileREML.FULL
##          -579.3514
```

The above MLEs show that $\nu = 1.5$ results in the best model (when purely considering MLEs). It should be noted that all of the MLEs for $\nu = 0.5, 1, 1.5, 2$ are very close, however when splitting hairs as we do here a $\nu = 1.5$ results in the best model since it has the highest MLE values.

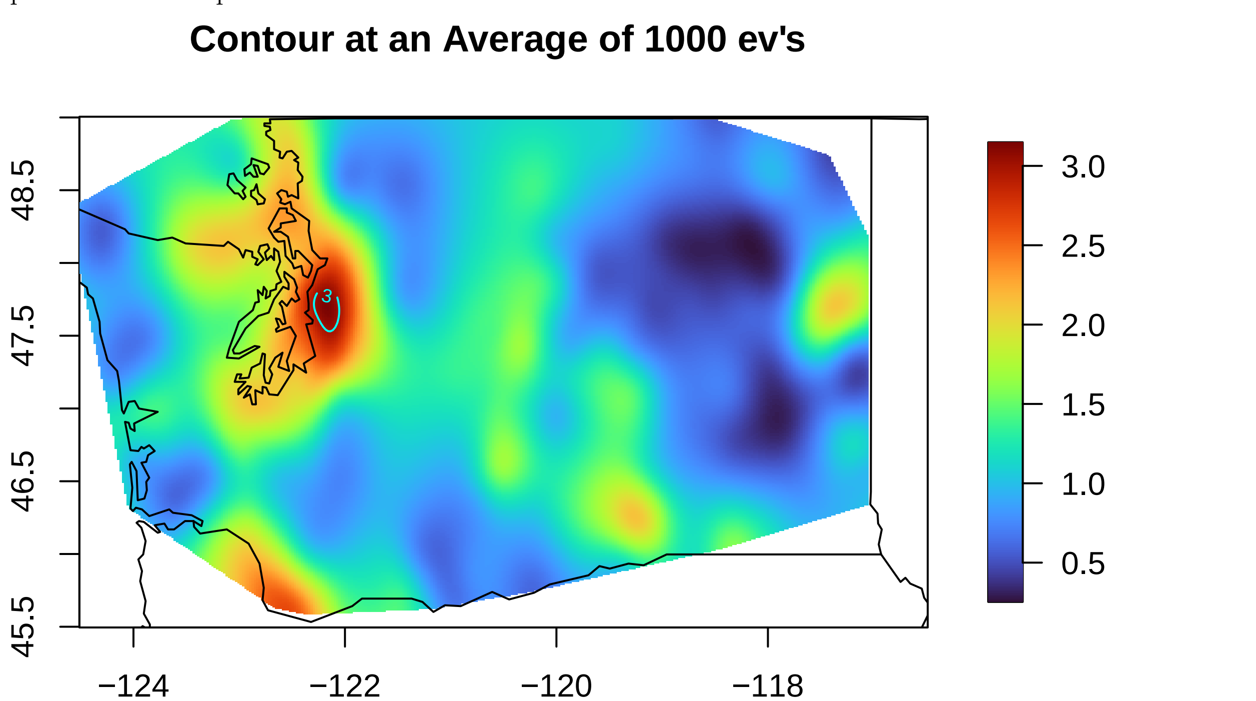
To further show this the below is a residuals vs fitted plot for all of the spatial models using $\nu = 0.5, 1, 1.5, 2$.



All of these fitted versus residual plots look very similar, however as shown in the MLE values we will continue with a $\nu = 1.5$ for our final model.

Creating Predictions

We predict the surface using the model with $\nu = 1.5$. We also create a 360x360 grid to predict over. This predicted surface plot is below:

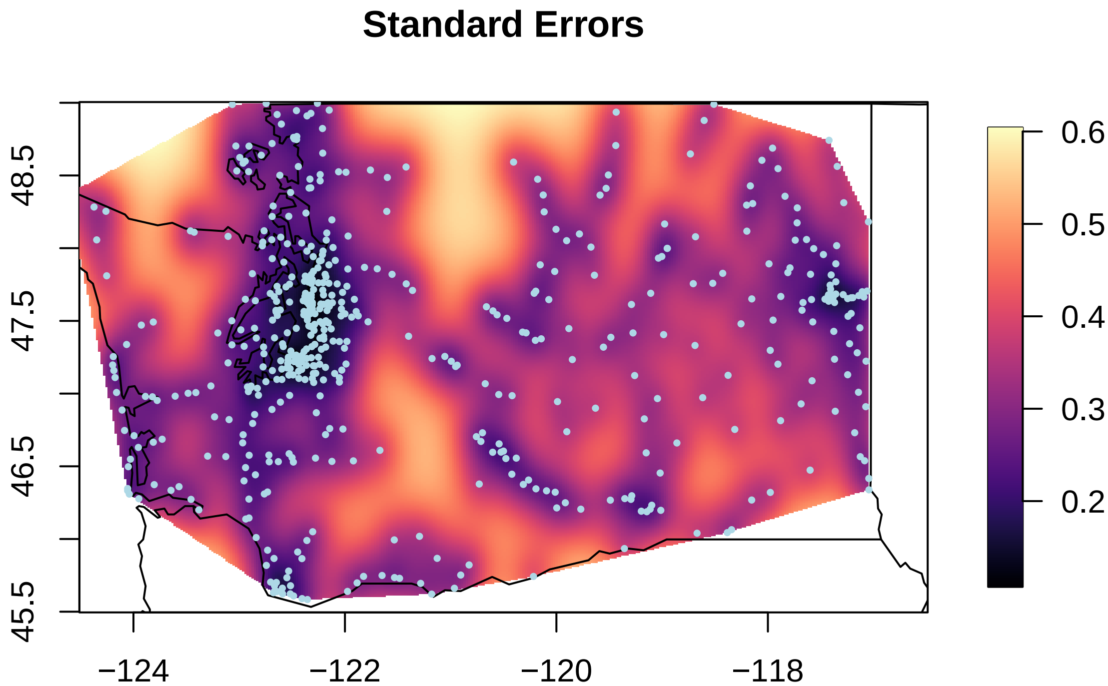


The spatial prediction plot above matches our assumptions. In this the spatial process model using $\nu = 1.5$ predicts higher amount of electric vehicles in places with a larger population (especially in Seattle and down towards Portland). Rural areas show less predicted amounts of electric vehicles.

The contour line in the plot above is around an average of $1,000$ electric vehicles. As seen throughout this report but this line is strictly around Seattle, again showing the importance of large cities and the prescense of electric vehicles.

Below we predict the surface and this time plot the standard errors. The light blue points are locations where there is a data point.

Standard Errors of the predictions. Light blue points is where we have our data.

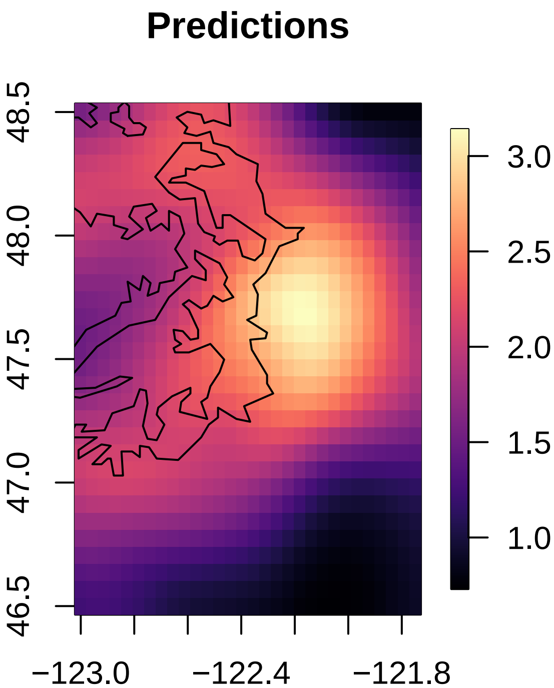


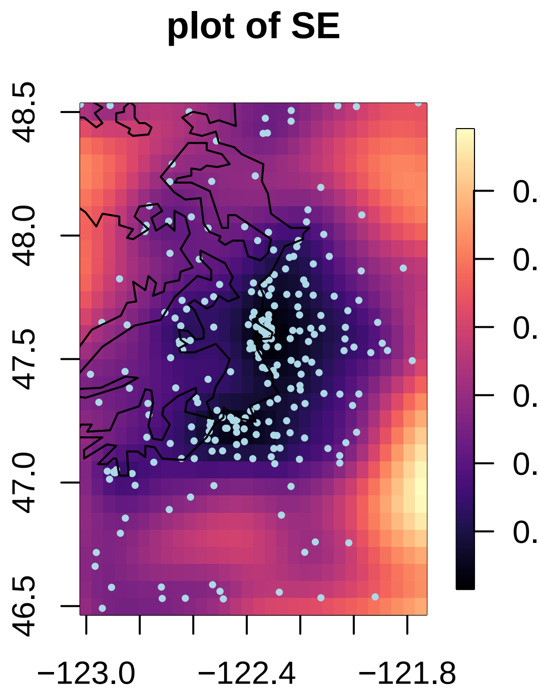
The above shows that our spatial predictions has larger standard errors in places with no associated data points, and lower standard errors in areas where we have more associated data points. This is to be expected and makes sense.

There also appears to be more data points in more urban areas (particularly Seattle and down towards Portland), again showing that there are more electric vehicles in densely populated areas.

To go further with Seattle as a whole we focus on just Seattle below:

The below plots of the predicted surface as well as standard error over just Seattle paint a similar picture. In the city center there is almost no error and the highest prediction. There is also the most amount of data points in the city center.





Results and Conclusion

The spatial analysis revealed distinct patterns in the distribution of electric vehicles across the State of Washington. It was uncovered that urban areas, especially Seattle, have a high density of EVs compared to rural regions. This aligns with our assumptions that urban areas have a higher concentration of EVs than rural areas.

The finalized model utilized a smoothness of $\nu = 1.5$. The *Big 3* from this model show a $\tau = 0.6092537$, $\sigma^2 = 0.3432394$ and aRange $= 0.1867752$. When considering lambda we see $\lambda = 0.6092537$.

A $\tau = 0.6092537$ value indicates a moderate level of spatial variance in the model at short distances.

A $\sigma^2 = 0.3432394$ value indicates that there is a moderate amount of unexplained variability in the data.

An aRange $= 0.1867752$ value indicates that points are correlated at short distances. This means that the spatial process has a local scale structure instead of broad scale structure. This makes sense given what is seen in the graphs above.

A $\lambda = 0.6092537$ value indicates that we have moderately smooth spatial variation in our data.

In conclusion, the spatial analysis confirms the notion that urban areas in the State of Washington have a higher density of electric vehicles. The reasons for this likely include higher population and better existing infrastructure for electric vehicles. These findings are important in the deployment of strategic planning of EV charging stations.

Possible Directions or Issues for Future Analysis

Next steps include applying the model to a wider geographic area so that it is not just limited to Washington State. THe model is applicable and can be easily applied to other areas so by applying our model to a wider range of an area we can better strategize in the planning/deployment of EV charging stations.

Additional further analysis could be including temporal analysis in our data and seeing how the distribution of EVs in relation to the expansion of charging infrastructure and public policy changes.