**GEO 361 Final Project: Population of Ducks: Modeling and Interpolation**

**Ben Steves**

**GEO 361**

**Prof. Van Horn**

**May 5 2021**

**Introduction and Problem:**

It is important to keep track of animal populations in case they are on the decline or maybe overpopulating, and this applies to birds as well. Though breeding bird populations are usually slightly erratic, decreasing or increasing, duck population dropped quite a bit from 2018 to 2019. US Fish and Wildlife reported that Michigan Mallard populations dropped from about 250,000 to 179,000 from 2018 to 2019. The Michigan breeding duck population overall also fell from 452,000 to 333,000. (Waterfowl Population Status, FWS 48-53). Most differences are state-by-state, and the US did have an increase in Mallard population overall, but other species like the Blue Winged Teal and the Merganser fell in the US. Again, there is a lot of variability in short term trends, but it is always good to monitor populations in case there is something going on.

**Objectives:**

This report will look at populations of ducks in five Great Lakes states using spatial interpolation, in an attempt to predict where certain waterfowl species are more heavily located or seen by observers. I performed a few different interpolation methods all using data from 2009-2019 and tried to compare the conclusions of those with each other. Comparing these methods is another objective of this project. Lastly, I fit a GWR and created a few different models to see which was best of them, and tried to see if the Number of cars present on a route and excessive noise had any effect on the number of birds observed. This was done both partly in ArcPro as well as R. In R, I fit a regular regression, and in ArcGIS I fit a geographically-weighted regression. I did this to practice model assessment and apply it to what I’ve learned with interpolation.

**Methods:**

The data that I used comes from the North American Breeding Bird Survey from the Patuxent Wildlife Research Center, which has thousands of participants each year working on this dataset <https://www.sciencebase.gov/catalog/item/52b1dfa8e4b0d9b325230cd9>. There is bird observation data here dating as far back as 1967.

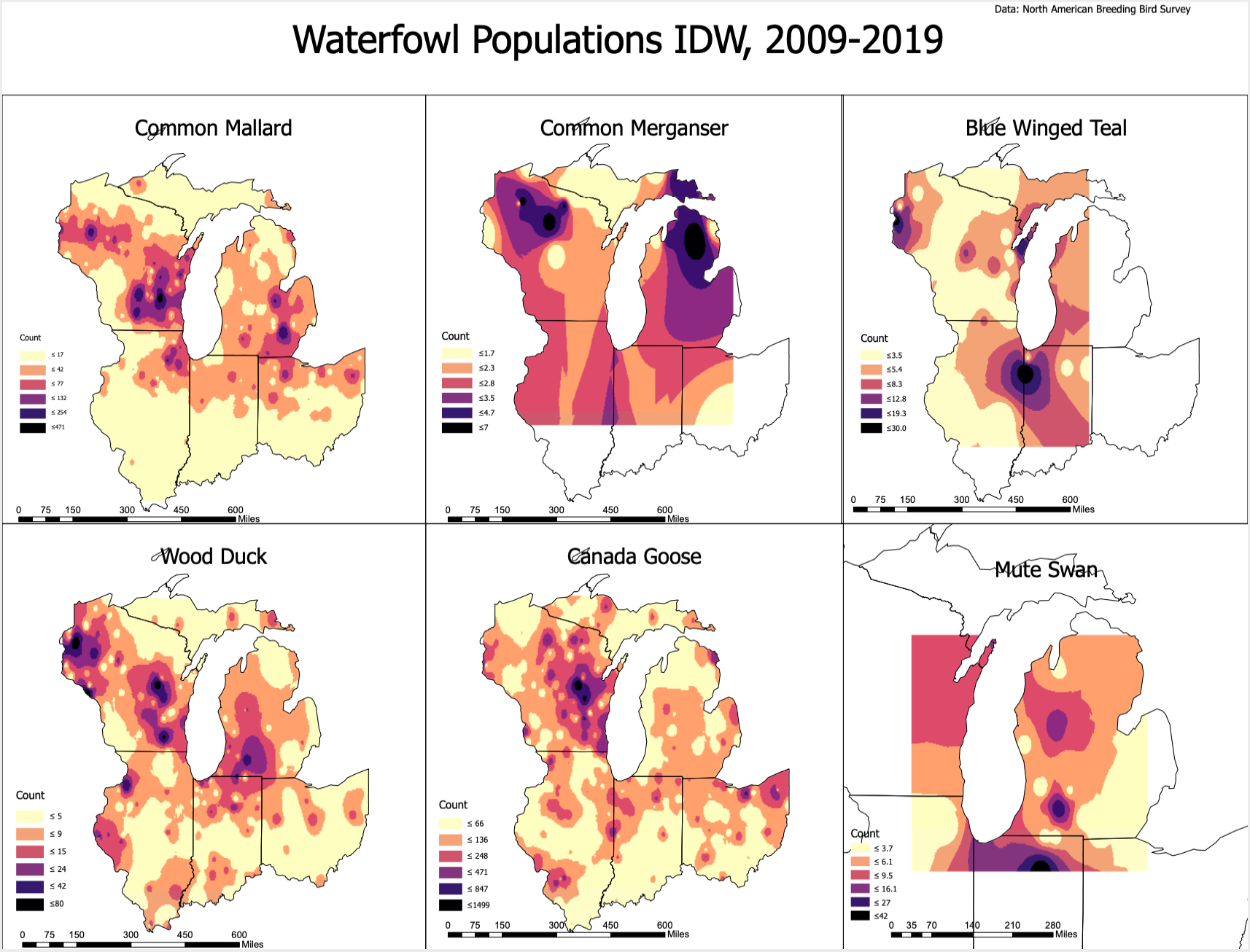
There was a decent amount of data prepping to do here. Included in the dataset were observations for each individual state, although not all states were needed as I wanted to limit it to a smaller, more centralized area of the US, this being the Great Lakes region. I used data from Michigan, Ohio, Indiana, Illinois, and Wisconsin, which I knew were all pretty common areas for many waterfowl species because of our Great Lakes. Included in the data are an extensive amount of different species from many different orders, like waterfowl, songbirds, shorebirds, birds of prey, etc. Of course, I was only interested in waterfowl, so I picked out six different species to look at in this region - Mallards, Common Mergansers, Blue-Winged Teals, Wood Ducks, Canada Geese, and Mute Swans. Route data was also included with XY coordinates, and this route data was easy to merge with the state observation data. I also filtered out anything before 2009.

I then used an Inverse Distance Weighting (IDW) interpolation method to map populations of waterfowl across the Great lakes region. One challenge with the data was that routes were not always equally visited each year, and some routes had more data on bird observations than others. For example, route 1 in Ohio had only two years worth of data for Canada Geese (from 2009 to 2019), while route 3 had seven years worth of data for the same species. I initially used raw numbers in interpolation processes but changed to means of bird observations across those 11 years. The IDW map (shown later on) does show raw totals, and I was worried this would make the map show incorrect conclusions, but as it turns out there wasn’t a huge difference between maps using raw population and maps using mean population, at least for IDW. Later on, though, I ran into some trouble with another interpolation method in using raw counts and quickly redid my analysis to better represent the population of waterfowl. I carried out my IDW and created a map for each species that I was looking at. In the IDW, I used mainly the default settings and classified the data to six levels using natural breaks.

Next, I tried a Local Polynomial Interpolation (LPI) which is smoother over space than IDW overall and doesn’t completely cluster at route points with high populations. Instead, it makes a more gradual plane, which is maybe better representing where waterfowl are rather than specific local points where waterfowl were observed. It is here where using raw counts gave me different results, interestingly, so I calculated the mean number of birds seen for a given species at a given route, for all the years from 2009-2019. I figured this would be better in case 1) for some reason a route one particular year had way more waterfowl observations than usual, incorrectly weighting the raw total, and 2) in case one route had more observation data across the years 2009-2019 than another. Similarly to my IDW, I mapped the predictions of the six species and classified the prediction values up to 6.

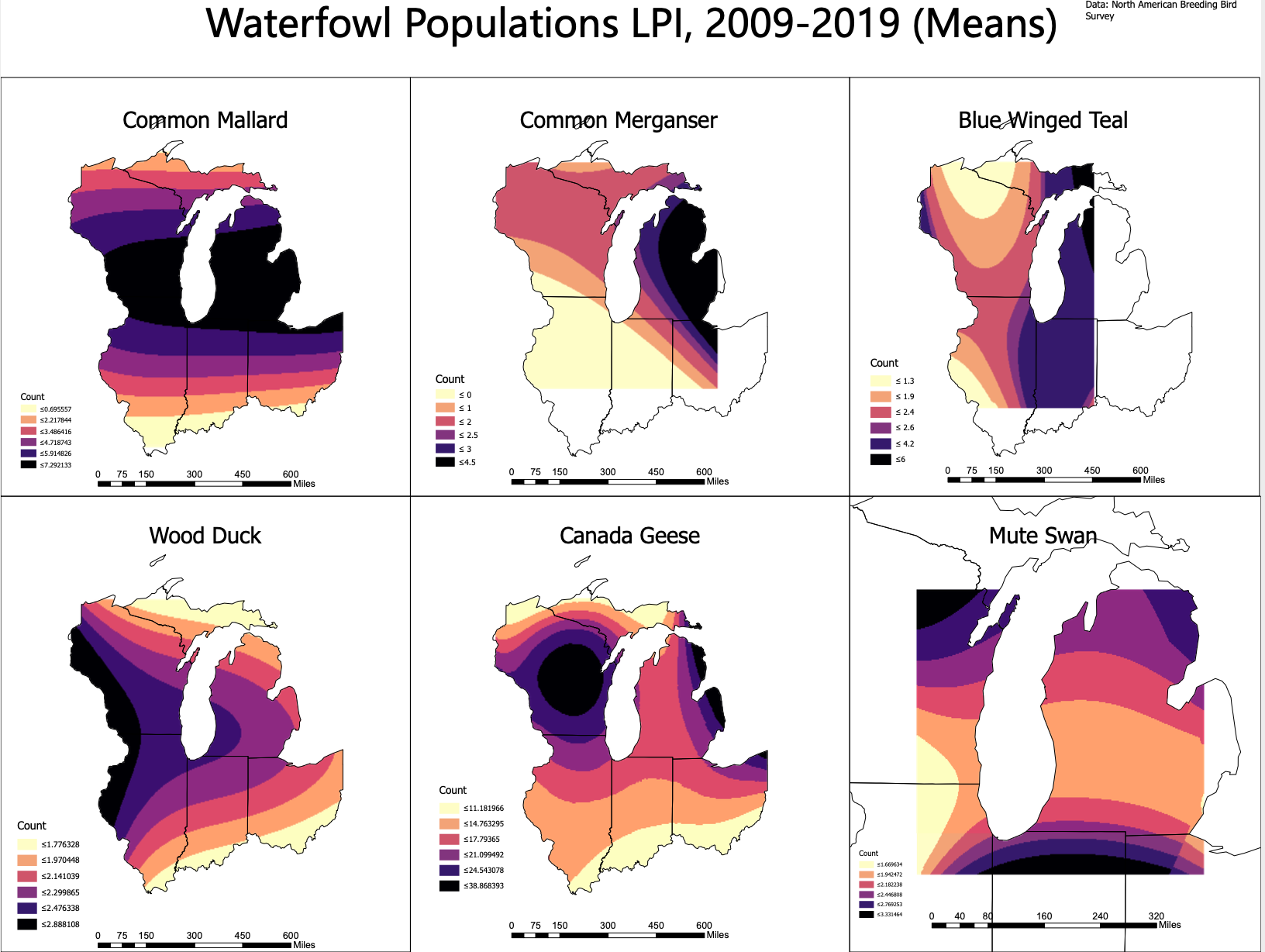
Lastly, I fit a Geographically Weighted Regression, and mapped predictions and standard errors. I created three different models with combinations of two predictor variables: the average number of cars observed at a route from 2009-2019, and the average number of stops with excessive noise observed at a route from the same time frame. One model contained both predictors, one contained only noise and one only total cars. I did a little bit of work in Rstudio as well and fit this same model as a non-geographically weighted model, mostly to better understand statistically if there was anything going on with the model I should be concerned about.

**Results and Conclusions**

IDW map:

The map above uses Inverse Distance Weighting to predict the number of birds seen for the six waterfowl species across space. Darker areas indicate larger observation totals. What should be noted again is that the results are showing, very specifically, the routes with the most observations for waterfowl species. It makes sense then to interpret the results as where waterfowl were seen, not exactly where they are located across space. Being a local interpolation method, it is making predictions very specific to this data and it would probably be too specific on new data. This is still useful to see where birds were spotted in an attempt to see which routes had more observations.

Mallards had highest densities in specific parts of Southern Wisconsin, Southeast and Central Michigan, and they are scattered all over Illinois, Indiana, and Ohio. Wood Ducks have similar densities except they were viewed more in West Michigan than East Michigan, and also were viewed a bit in Western Wisconsin. Canada Geese were viewed most in Wisconsin asd some pockets spread around as well. The other three species, being Teals, Mergansers, and Mute Swans, also have clusters in certain spots but they are a little bit less precise because there is less data on these species. For example, areas in black are larger in the Merganser map compared to the Mallard map because Merganser’s had a very limited number of observations. Specifically the Merganser, it may have been a more interesting bird to look at if I were looking at only Michigan, Wisconsin, and Ontario, as it’s not super common South of those areas in the breeding season. But it’s my favorite duck species, and I saw a bunch last summer in Charlevoix, so I wanted to include it purely because of that. Based on IDW, Michigan and Wisconsin overall have the highest number of predicted spotted birds, at least for the majority of species, as there are more patches of yellow in Ohio, Illinois, and Indiana.

LPI Map:

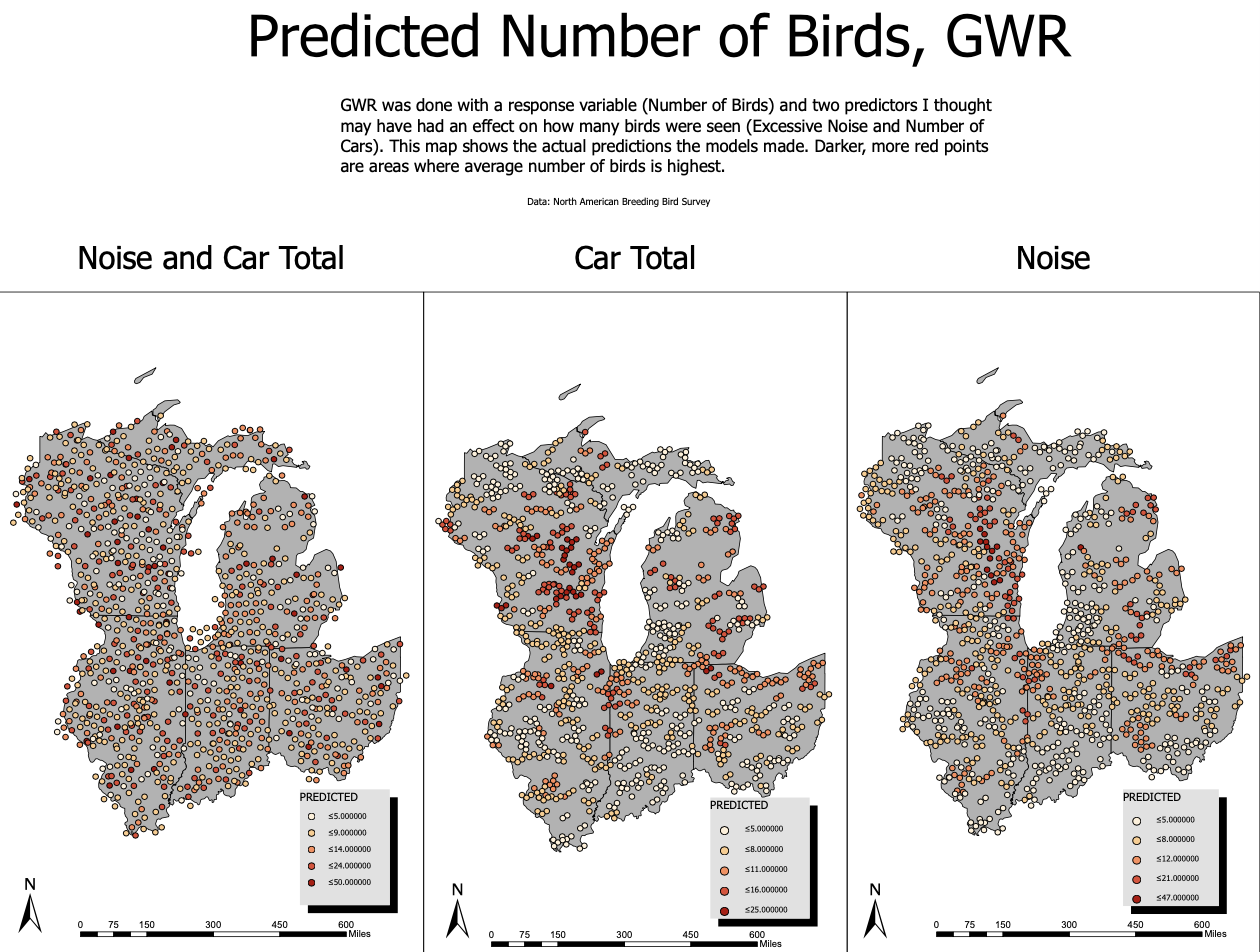
Here is the LPI map, which is much smoother and less local. This probably better predicts where waterfowl are, and matches up better with bird abundance maps. Counts are lower overall because this measures the mean number of birds seen at a route, averaged from all of the years looked at. Again, I did this in case certain routes were sampled more. This also makes the results more accurate.

Similar to the IDW map, darker areas have larger predicted numbers of birds for a region. The Mallard’s largest predictions are in Michigan and Southern Wisconsin. Mergansers are predicted to be most common in the Lake Huron region. Teals have large densities in Michigan and Indiana. Wood ducks seem to be more common in the Western portion of the Great Lakes region. Wisconsin also contains the highest densities for Canada Geese as well. The Mute Swan is a little harder to interpret, and it may be that in the data, the points on the edges of the geography had the highest number of birds, and this would probably be more interesting on a larger geography.

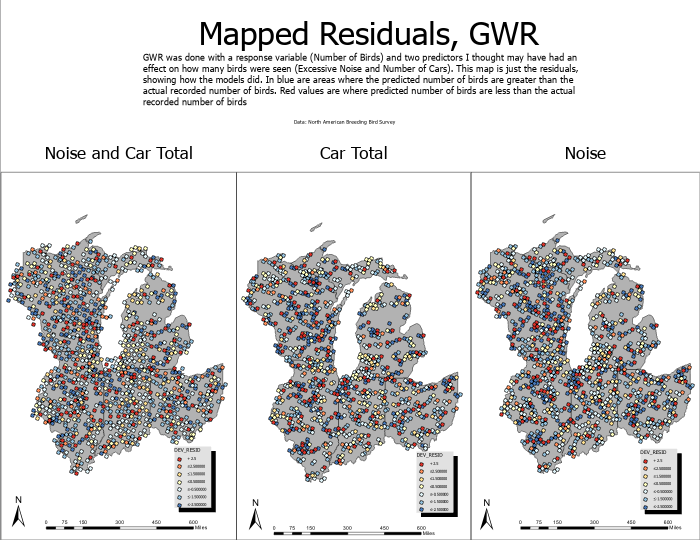
GWR:

Lastly, I carried out Geographically Weighted Regression with two predictors, the number of stops with excessive noise on a path, and the number of cars on a specific path. Each route for a given year had data on 50 different stops along a route, and on each stop, the number of cars was recorded along with a 1 or a 0 if there was excessive noise, with 1 being yes and 0 being no. I added up the totals for each stop using excel formulas. For example, if through 50 stops there were 7 with excessive noise, then the total number of stops with excessive noise for that route in that given year would be 7. Sums were similar for the number of cars as well. Then I averaged these totals across the years 2009-2019 for each route. Illustrating another example, say a path had data for three years in that span, and the car totals were 23, 33, and 26. I would just take the average of those to get 27.33 cars for a route, averaged across all of the years. I did this for the noise variable as well.

I then fit three different GWR models in ArcPro. The first model was with both predictors. The second had one predictor, being the total number of cars, and the third had one predictor, noise. After fitting the models, I compared their predictions using ArcPro. Below are the predictions of the models, mapped out.

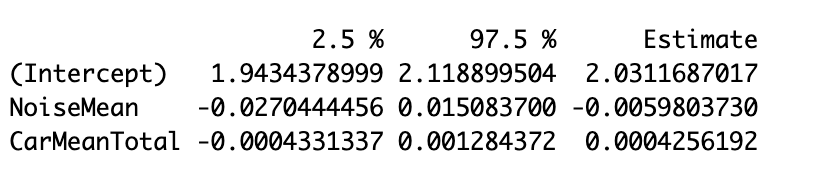


Because of overlapping points, I used the disperse markers tool to cluster them around their location. The models with only one prediction had less randomness in the predictions for number of ducks overall, with larger clusters in Wisconsin and Michigan, similar to other maps, except this one took into account all six species. In comparing the models using AIC, the model with both predictors had an AIC of 9093, the model with just car total had an AIC of 9647, and the model with just noise count had an AIC of 9381. These were all from summary tables in R. AIC values are best at comparing models, and the numbers don’t mean anything by themselves, they only mean something when compared to other models. Lower AIC’s indicate better fits and better model performance, so the model with both predictors, being the model with the lowest AIC, makes the better predictions overall.

Also shown below is the residuals map.

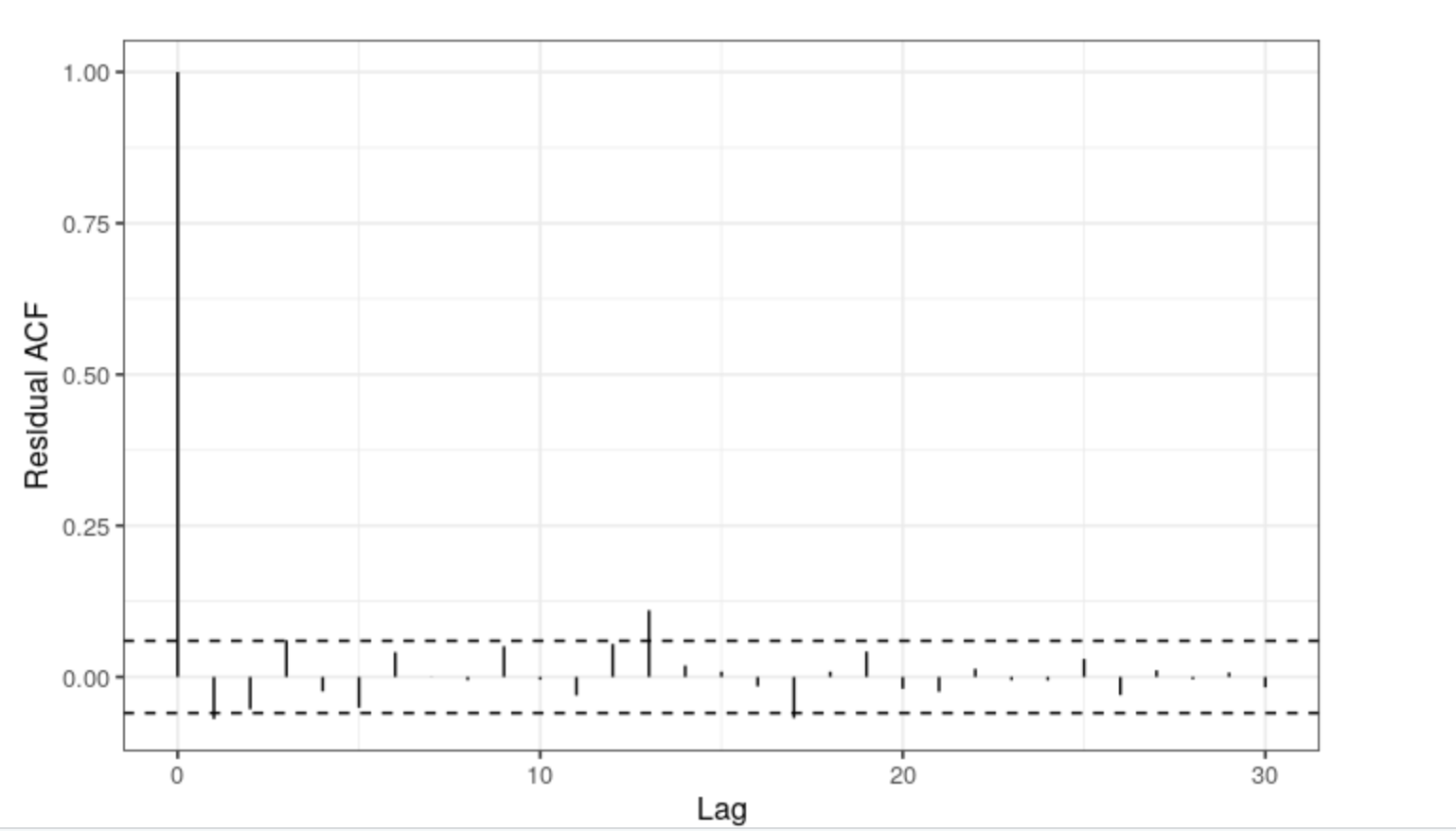
There is some slight blur because this map had some trouble exporting as a pdf, but to reiterate what the explanatory text says about these values, blue points are where the predicted number of birds are larger than the observed number of birds. On the other hand, red points are where the predicted number of birds were less than the observed number of birds. The scale for the blue values went up a lot higher than 2.5, so there is maybe some evidence that there is some overprediction in this model. There does seem to be more values in the middle in the model with both predictors on the left, which might be indicative of its better AIC compared to the other two models.

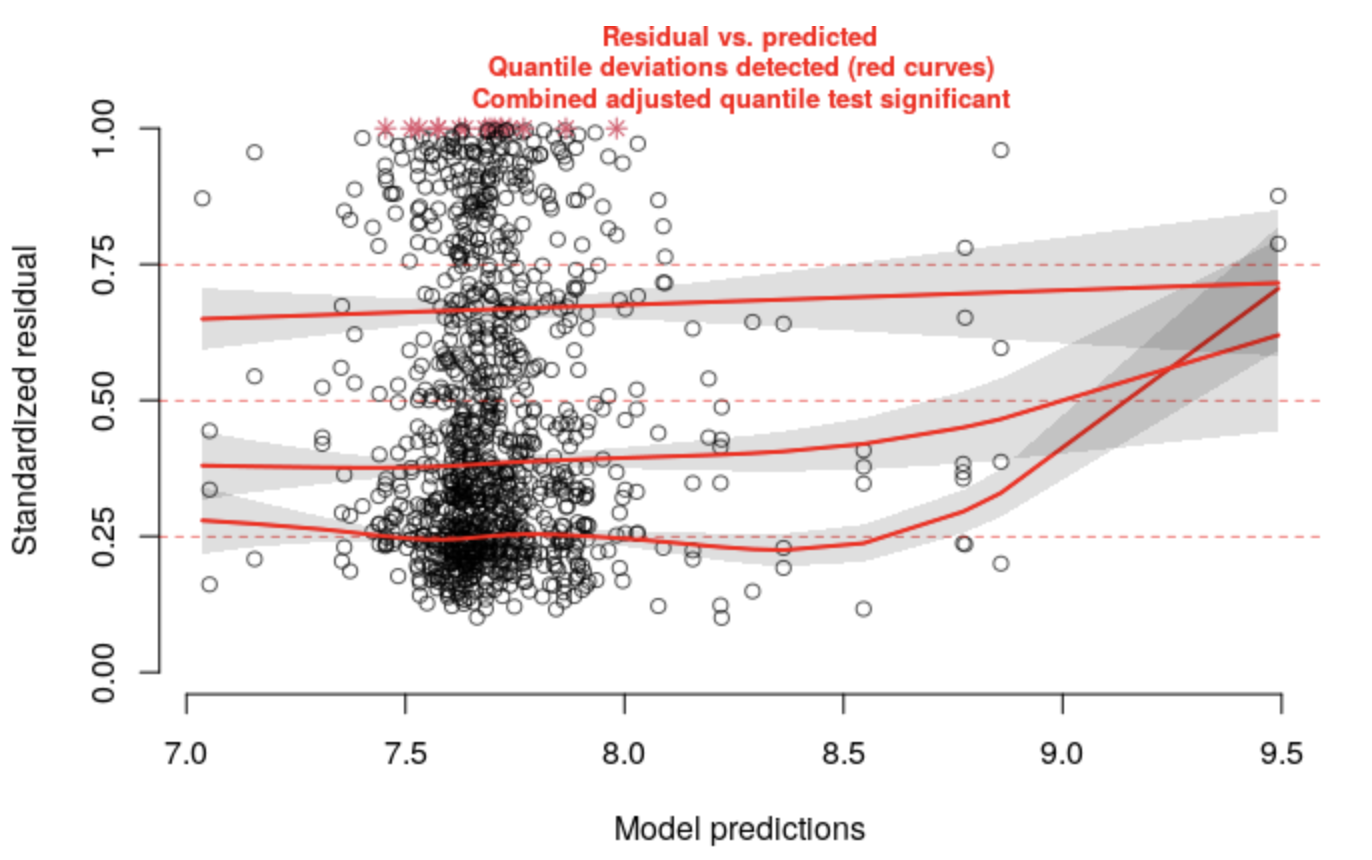
I also fit this model as a standard regression in R. Unlike ArcPro, in R I can fit the model with a negative binomial link function instead of poisson, as negative binomial is usually better in its results, so I tried that instead. Both poisson and negative binomial functions are standard for count data or only positive data. I was able to do some more digging on the effects of these variables. Before getting into model diagnostics, the model predicted that neither variable had much impact on the number of birds seen. Below is a table for a 95% confidence interval in the model’s predictions on the effects of these variables.



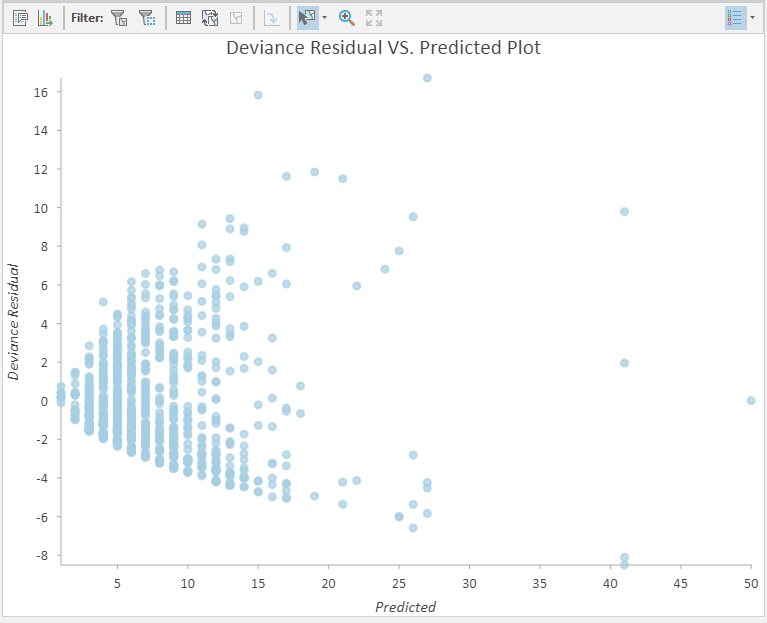
Both predictors were modeled as slopes, so anything heavily positive indicates a positive relationship between that variable and birds seen, and anything heavily negative indicates a negative relationship. Looking at the CI output above, slopes for both variables are close to 0, indicating that neither excessive noise or number of cars really had an effect on whether birds were seen. This is maybe a little surprising to me, as I figured more noise and more cars would lead to less waterfowl being observed, but it turns out it didn’t matter a lot. In the Rmarkdown document that I shared this is evident in the plots as well.

It makes sense, though, to double check to make sure the model is working the way we want it to. In STAT 245 with professor DeRuiter, we learned about assessing the accuracy of models first by making sure there aren't any issues with how the model is working. There are a series of tests we used to assess the diagnostics of a model. The first of these is a test of residual independence. We don’t want there to be heavy autocorrelation in residuals, because it would indicate data collections weren’t done independently. In time series data, sometimes certain values can depend on the value before it, and so on. The plot below is good, though, showing there isn’t any autocorrelation in the residuals. Usually, if there are, lag bars will all be above or below the dotted black lines. Only a few lag bars cross the lines, so there doesn’t appear to be any autocorrelation in the residuals.



One other test that is important in determining the accuracy of the model is error variance. Residuals should a) be mostly trendless or uniform and b) should have pretty similar variance across predicted values of y. This is where I ran into problems, both in ArcPro and in R.

The plot above shows the standardized residuals on the y axis and the predictions of the model on the x axis. If you look closely, there is a heavy cluster of points lower on the y-axis at residuals close to 0.25, and smaller clusters above at residuals of higher values. If this plot were passable, you’d see no red anywhere, with all of the lines being black and also have a slope of 0.

ArcPro gave me similar errors, in all of the models, including the model with both predictors that AIC said was a good model. Below is a residual vs predicted plot for that model obtained from the GWR output. 

What is shown above can happen a lot when examining models. The trumpeting shape here, also known as heteroscedasticity, shows that our results may not be reliable. Higher predicted numbers of birds have more extreme residual lengths than smaller ones, when it should be about equal across all predicted values. Another translation of this is that higher predicted waterfowl counts were much different from their actual observed values, with some being way higher and some being way lower than expected. I believe this is because there are a lot less values with a higher predicted number of ducks, and there are much larger clusters at values in the 0-10 range. If I had more time, I would have probably tried to standardize all of the variables to see if that made a difference. I don’t know if I am a skilled enough data scientist yet to totally be able to fix this issue anyway, as there are some solutions to this that I haven’t fully learned about yet. In the future, I would really like to learn more about the capabilities that R has in GWR as well.

What this means, though, is that there is some possibility that for GWR, predictions aren’t 100% accurate and shouldn’t be fully trusted. While I don’t have any fully-accurate conclusions here, it was interesting to try a GWR and learn how it differs from regular regression analysis. I do think there is some issue with how the data is transformed that is causing this though.

The light conclusions here are that neither noise nor number of cars have much effect on the number of birds seen from a statistical perspective. From a spatial perspective, neither predictor together accurately predicts where larger counts of birds were found, but both models with just one predictor do seem to show some spatial trends in where more birds were observed more often. But, results may not be completely reliable as there were roadblocks in some of the statistical testing of the model, hence these are merely “light” conclusions and should be taken more seriously with an improved set of models.

References:

Data: <https://www.sciencebase.gov/catalog/item/52b1dfa8e4b0d9b325230cd9> - many thanks to the thousands of individuals who take part in this annual survey.

Dooley, Joshua, et al. “Waterfowl Population Status, 2019.” 19 Aug. 2019, pp. 48–53.

<https://www.fws.gov/migratorybirds/pdf/surveys-and-data/Population-status/Waterfowl/WaterfowlPopulationStatusReport19.pdf> article mentioned from above.