Determining the Best Time Interval to Measure Home Ranges of Wildebeests in the Serengeti Desert

Project by Madi Arndt, Alyssa Keehan, Brianna Lee

Figure 1: Wildebeests in Serengeti National Park

Introduction:

This study involves researching the optimal time interval at which to measure convex hulls for wildebeests in the Serengeti National Park. Convex hulls output a polygon with the shortest perimeter that encloses a set of points. They are a great tool for animal movement analysis because they do not underestimate the habitat range of animals, they convert traveled points into a clear area, and they allow researchers to estimate an animal's home range. Convex hulls have the potential to be a very telling representation of an animal's home range, which is defined as the area over which an animal or group of animals regularly travels, but only if used appropriately. One factor that can determine how well a convex hull represents reality is the time interval that is used when creating a hull. For example, creating a convex hull based on one day's worth of movement would not be very representative of reality because it is too short and random of an interval. On the other hand, using a month interval would be overfitting and not representative because it takes too much movement into account that could fluctuate because of external factors that cannot be considered when analyzing a convex hull. Through experimenting with different time intervals for creating convex hulls, our group hopes to find the most fitting time interval for wildebeests in the Serengeti National Park.

Motivation:

After researching Wildebeests in the Serengeti National Park and discovering that they migrate, our group was surprised to see most of the animals in our dataset did not follow these extreme movement patterns. As a result of this realization, we decided to use convex hulls to observe their movement and home ranges. During this process we considered different time intervals and noticed longer intervals were too overfitting and short were too random. These different results were not representative of the wildebeests true home ranges and would have been extremely inaccurate if we had used them for our analysis. For our research and other research projects, using the wrong time frame can be misleading and can impact the accuracy of home ranges. This issue inspired us to shift our research towards finding the most appropriate time frame for creating and analyzing convex hulls. Through our research we want to determine which time frame produces convex hulls that realistically represent home ranges. The results of our research could be an important discovery for the field of movement analysis because the use of convex hulls is one that we have found to be used in multiple studies. Based on our results, we will discover a time interval that is ideal and that can be applied to animal movement studies to provide maximum accuracy. We hope the conclusions of our analysis can provide beneficial information for researchers, especially those studying wildebeests or the movement of similar animals.

Research Questions:

In this study we are investigating the question "Which time interval is the best suited for determining the home range for Wildebeests in the Serengeti National Park?" In further detail, at what time interval is the most self-similar convex hull produced as measured through the Jaccard Similarity Index?

Data:

For our project we utilized two different datasets to perform our analysis. The first dataset used for this project is the "Serengeti National Park Boundary." To provide context of this dataset and our study site, the Serengeti National Park is located in Southern Kenya and Northern Tanzania. We retrieved this dataset from serengetidata.weebly.com, which is a site that provides multiple datasets concerning Serengeti-related ecological and environmental data. This polygon shapefile was used to identify the national park of interest for our project and acted as a tool to identify and filter which wildebeests in our analysis were located within this study site.

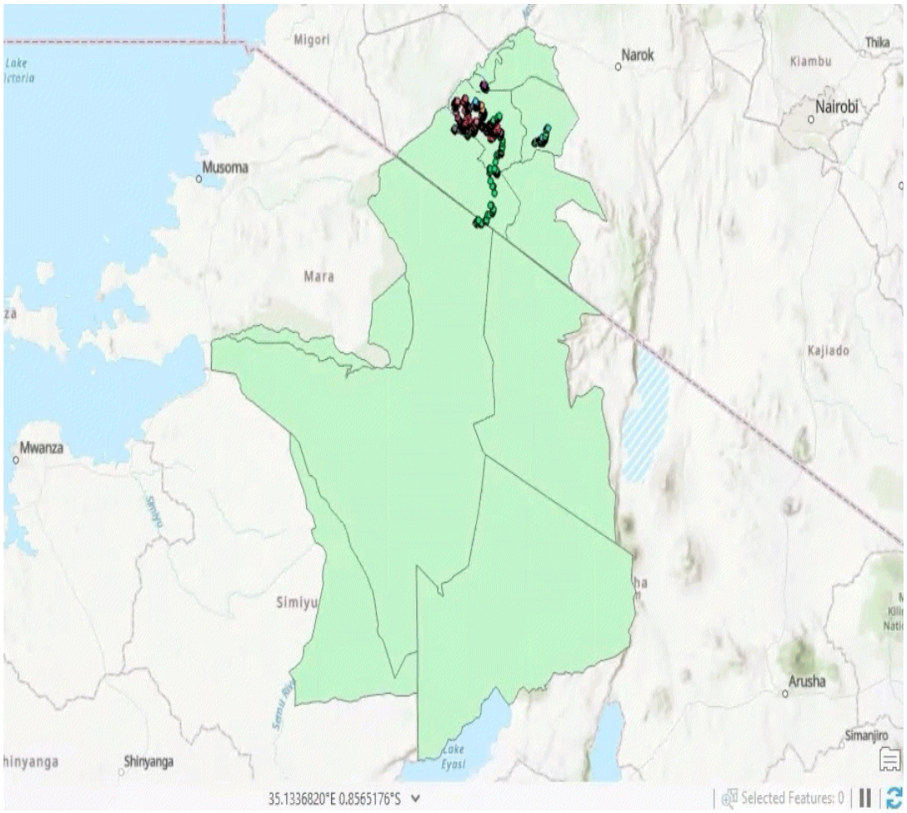


Figure 2: “Serengeti National Park Boundary” dataset

The second dataset we used is the “White-Bearded Wildebeest in Kenya” dataset. This vector point shapefile comes from the website Movebank.org, which is an online platform where researchers can post their animal movement- related data. This dataset contains 36 wildebeests and their specific movement coordinates. The time frame upon which this data was collected was a 2-3 year time period from May of 2010 to January of 2013. These coordinates were collected by a Lotek Wildcell GPS collar. These collars tracked the wildebeests and recorded their location multiple times each hour, providing plenty of detailed information for our group to use. This dataset included information on the tag identification for the wildebeest to differentiate each wildebeest by an unique number. For our project, we chose to study five wildebeest within the Serengeti National Park boundary from the time period June 2011 through August 2011.

Analysis Methods:

For analysis, our project focuses on computing the Jaccard Similarity Index of the movements of our wildebeests during the time frame of June 2011 - August 2011. In short, the Jaccard Similarity Index is computed by dividing the size of the intersect area over the size of the union area. We did this calculation for every possible combination of dates for each of our five time intervals.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

In order to determine the best time interval using the Jaccard Similarity Index, we needed to look at what relative values prove to be the optimal choice for home ranges. Therefore, in our exploratory analysis, we will present multiple plots and graphs comparing the Jaccard Indexes across the different time intervals in addition to noting the time difference or number of time intervals between comparisons (i.e. one week apart versus 6 weeks apart). Our optimal value will be determined by crossing out the choices that have high variance and are underfitting or have high bias and are overfitting. This process enables us to choose the best time interval for our wildebeests that minimizes the mean squared error.

For our first analysis method, we incorporated regression analysis and time series using ggplot in R to compare the relationship between our computed Jaccard Indexes and the number of time intervals since the comparison. For further explanation on the number of time intervals or time difference, if we were working with the one week time interval and comparing the movements of one wildebeest from June 1 - June 7 and July 6- July 12, our number would be 4. Since the week of June 1 - June 7 is our starting point, counting up to the next 4 possible combinations would bring us to July 6 - July 12.

For this comparison, we expect the correlation to stay negative across all the different time intervals (i.e. the interval of one week, 10 days, two weeks, 18 days, and three weeks). This is because if we were to compare the movement areas of these wildebeests from a small difference in time, we expect a higher self-similarity than when looking at the movement areas with a longer difference in time. We typically wouldn't see a lot of intersection across the longer time differences (i.e. how many time intervals between comparison) since these animals move a lot. Something we will look for with these plots is a possible relationship change between the number of time intervals or time differences between comparison and the Jaccard Similarity Indexes. We expect the linear relationship and correlation between these two variables to get stronger as we increase the length of the time interval; which means that for longer time measurements, larger time differences produce inconsistent self-similarity values than for smaller time intervals. However to optimize the bias-variance trade-off, we are looking for the largest time interval with the weakest relationship. Having a weak relationship but a longer time interval would indicate more consistent Jaccard Similarity values nearly independent of the time differences in comparison.

For our second analysis method, we combined all the Jaccard Similarity Indexes for each of our five wildebeests and calculated the averages for each time interval. To visualize this, we used the geom_bar where the height of the bar represents the average Jaccard Similarity Index for each of the five time intervals we used. In order to help us determine the best time interval for a home range, we would analyze our bar chart and see where the average self-similarity index elbows off. To elaborate, we want to find the time interval that sort of levels off and then jumps significantly to a higher similarity value. This is because the elbow represents the optimal estimate as a trade-off for the self similarity. It allows us to choose the time interval that produces the least variance and bias while also minimizing the mean squared error.

Visualization Method:

For our project we used many tools in ArcGIS to create our maps. First we subsetted the wildebeest dataset to only contain the wildebeest inside the Serengeti National Park boundary. Among those wildebeest, we further separated the data to only contain wildebeest that were alive during our timeline of June 2011 to August 2011. We further divided the subsetted data into five unique tag identifiers to analyze the different time periods for each individual wildebeest. We made separate maps for each wildebeest and each of our time intervals - 1 week, 15 days, 2 weeks, 18 days, and 3 weeks. To conduct our analysis, we observed all of the possible combinations of the time intervals within our timeline of three months. For example, for the 1 week time interval we compared the first week of June to the second week of June, then the first week of June to the third week, and so on until all possible combinations were iterated. Our visualization analysis required us to use the minimum bounding tool, specifically the convex hull option within this tool, to create hulls for each of the intervals. Using the newly created convex hulls, we used the union tool to calculate the geometric union of the hulls for each possible combination. Then using the intersect tool, we calculated the geometric intersection of these convex hulls. Bringing together the two layers, we overlaid the intersect polygons over the union polygons to create a visualization of the Jaccard Similarity Index. The values calculated through the output of our intersects and unions were used for the calculation of the Jaccard Similarity Index, which also contributed to the analysis we conducted in R.

Regression and Time Series Analysis Results:

For our analysis, we decided to look at the regression plots for wildebeest 2836 because it was the most representative of our aggregated findings, as you will see in the bar charts in the next section. Like we mentioned in the methods section, we compared the number of time intervals between comparisons to the self similarity indexes to see if there are interesting results.

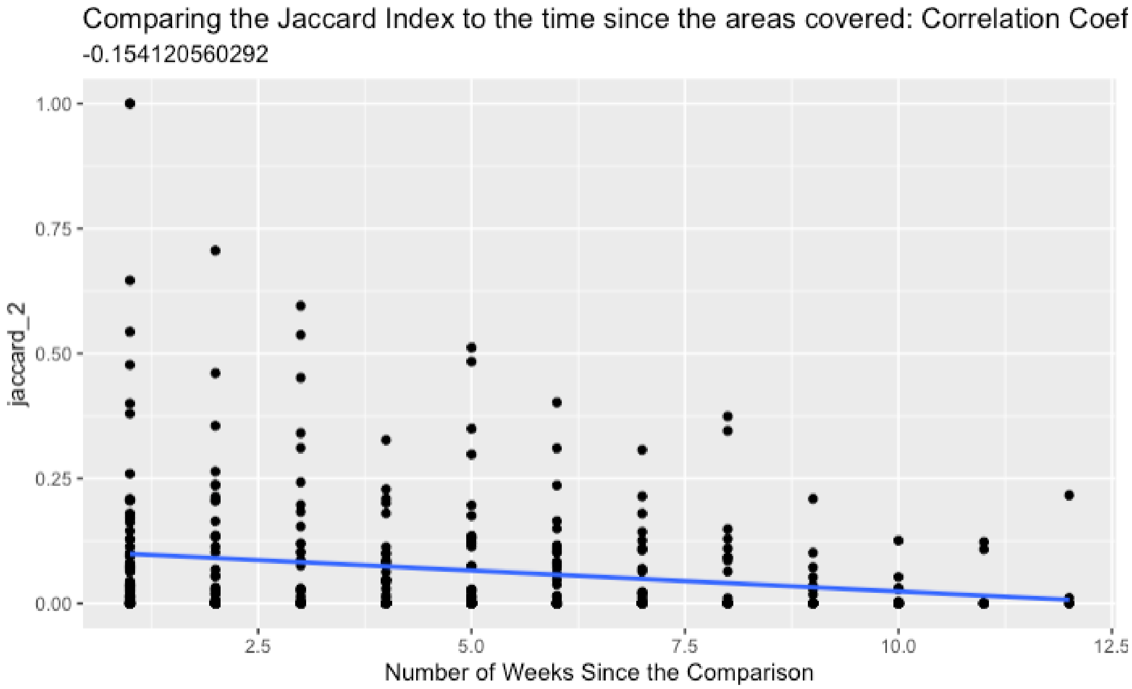


Figure 3: Comparing the Jaccard Index to the Time Since the Areas Covered for 1 Week Time Interval

For the aggregated data of the first time interval we used one week and got a correlation of about -0.15. This result maintained our expectations of a weak, negative relationship. This tells us that approximately 15% of the variability in the Jaccard Indexes can be explained by the time differences of comparisons. With a low correlation between the Jaccard Indexes and number of weeks between comparisons, this indicates that the time difference has little association with the Jaccard Similarity Index, especially since it is very low to begin with.

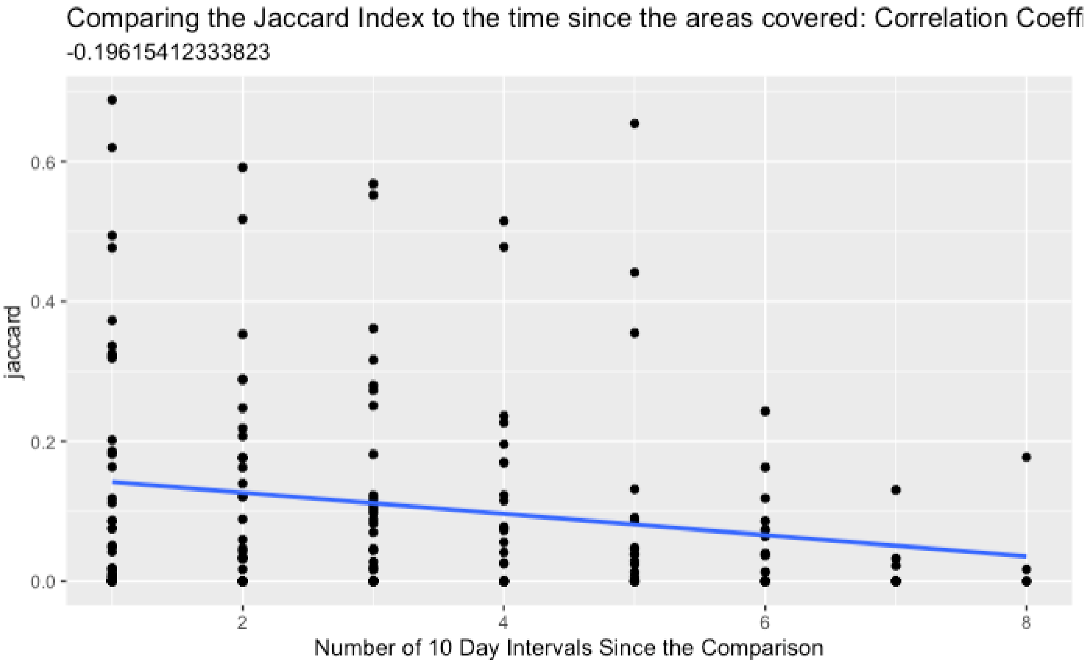


Figure 4: Comparing the Jaccard Index to the Time Since the Areas Covered for 10 Day Time Interval

For the second time interval we used 10 days and got a correlation of about -0.196. As we expected, the relationship between these two variables got stronger with approximately 19.6% of the variability in Jaccard Indexes being explained by the time differences of comparisons. Although the percentage increases just slightly, the number of 10 day intervals between comparison still has little association with the self-similarity values.

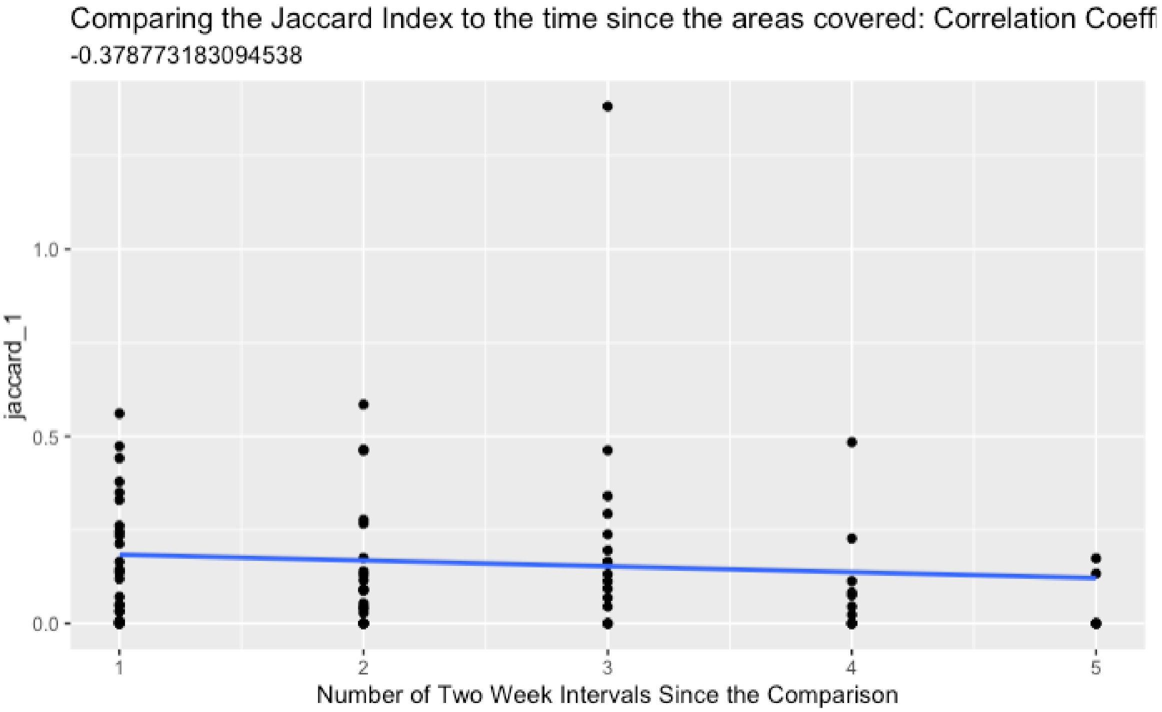


Figure 5: Comparing the Jaccard Index to the Time Since the Areas Covered for 18 Day Time Interval

The two week time interval presented a large increase in correlation with a value of 0.378. This presents that in our sample, approximately 37.8% of the variability in the Jaccard Similarity Indexes can be explained by the time differences of comparisons. At this point, the correlation is increasing as we increase the length of time intervals, and the relationship between the Jaccard Similarity Index and the time differences in comparison is getting stronger. This means that relatively, a linear relationship is developing between these two variables where higher Jaccard Indices correlate with smaller time differences and lower Jaccard Indexes with larger time differences.

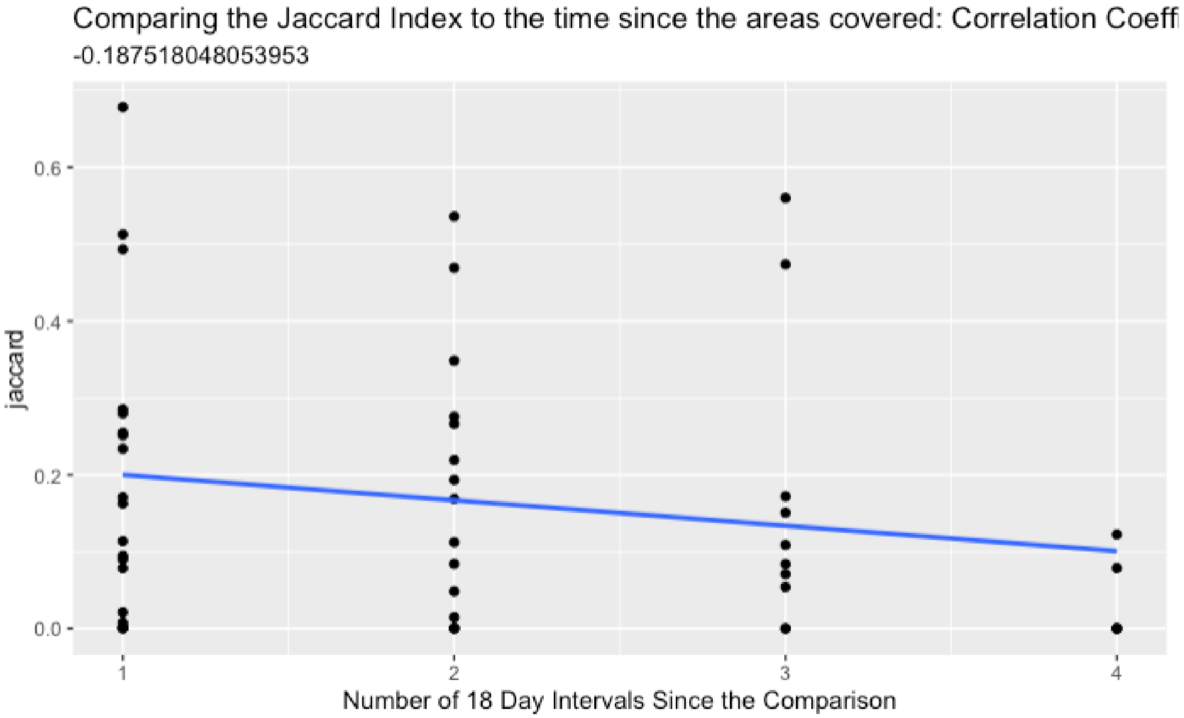


Figure 6: Comparing the Jaccard Index to the Time Since the Areas Covered for 18 Day Time Interval

Despite our expected results predicting a higher correlation as time interval length increases, the 18 day interval in fact decreased to a value of 0.187. In this case, approximately 18.7% of the variability in Jaccard Similarity can be explained by the time differences of comparisons. Since the correlation dropped quite a bit from being at about 38% in the next smaller time interval of 2 weeks, we see that the linear relationship is weakening and that there are constant Jaccard Indexes for any time difference. This is quite notable since it shows that the self-similarity is more constant even as the difference in time increases. As we mentioned before in the methods section, we want to see more constant values across the board like we do here because it shows more independence between the two variables thus making it a good candidate for our optimized time interval for the wildebeest's home range.

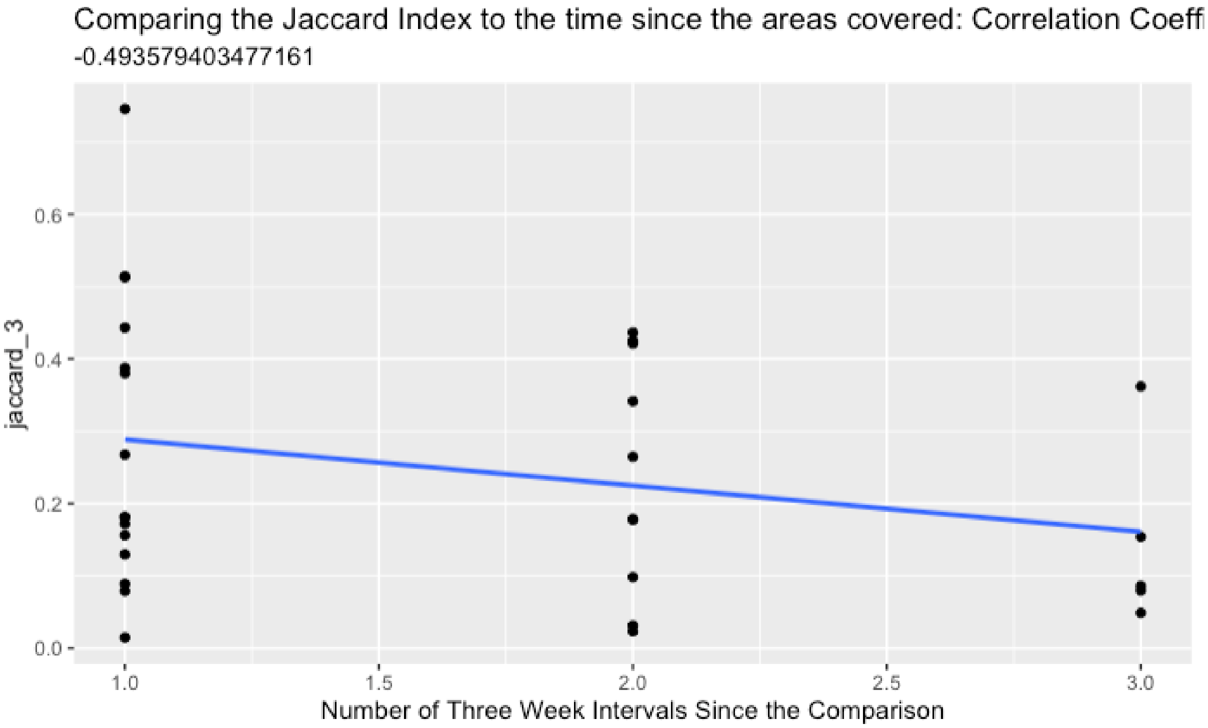


Figure 7: Comparing the Jaccard Index to the Time Since the Areas Covered for 3 Week Time Interval

Finally, for our last time interval--three weeks-- we see the correlation increased again to about 50% of the variability in the Jaccard Similarity being explained by the time differences of comparisons. Although this is a pretty high self-similarity value, the relationship between the Jaccard Indexes and the time differences seem to be too strong and may lead to high variance and overfitting. Again, the lack of independence between the two variables we are comparing and the non-optimized bias-variance tradeoff is what is keeping us from picking the three week interval as the most optimal.

Average Jaccard Self-Similarity Index Chart

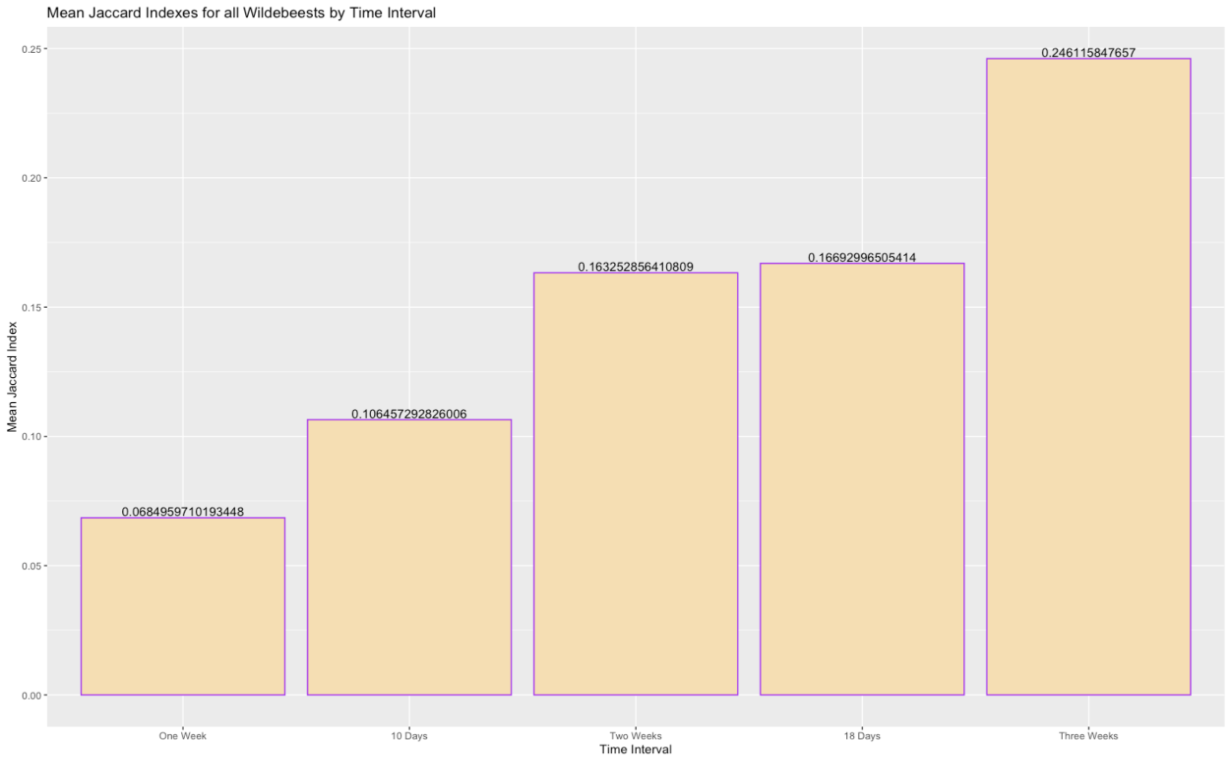
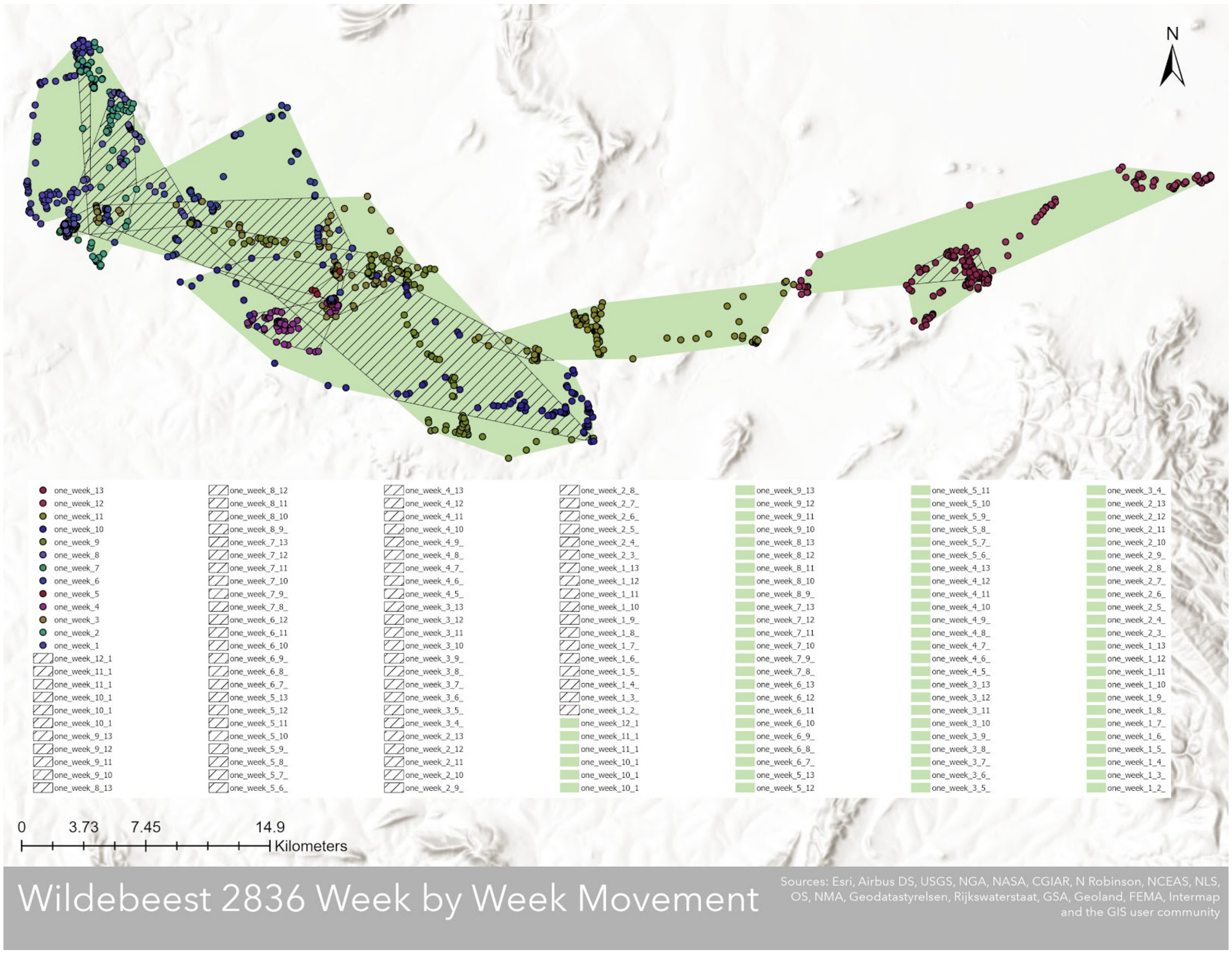


Figure 8: Average Jaccard Self-Similarity Index Chart for all Time Intervals

Above is the graph comparing the average Jaccard Indexes for each of the time frames we measured for all of our 5 wildebeests. Like we mentioned in the previous section highlighting our results from the regression analysis, the 18 day interval proved to be the optimal choice once again because it is shown as the elbow in this bar chart. As we mentioned before, we want to avoid having too much bias or too much variance in our self-similarity. It seems obvious to not choose the one week or 10 day comparison cause the average self similarity ended up being very low at approximately 0.068 and 0.106 respectively. However, the reason we did not choose the highest Jaccard, which ended up at the last interval of three weeks, is because it is too biased and overfits their ranges a bit. This leaves us with the two week and 18 Day time intervals which both have an average of about 0.16. However, since the elbow occurs directly at 18 Days our best option since it optimizes the bias-variance trade off and minimizes the mean square error.

Map Results of One Wildebeest and Notable Time Interval

For our spatial analysis, we decided to choose only one of our Wildebeests,number 2836, to present in our final project because we believe it represents the accumulated data the best. Below we will compare the results from the one week, 3 week, and 18 Day time periods to show what an underfitting, overfitting and optimal time frame for home ranges would be. We decided not to include our 10 days and 2 weeks measurements because they both showed to have similar qualities as the one week data being too underfitting and varied. The maps for both of these intervals did not highlight the high variance as much and our goal here is to present what bad and good choices would look like.

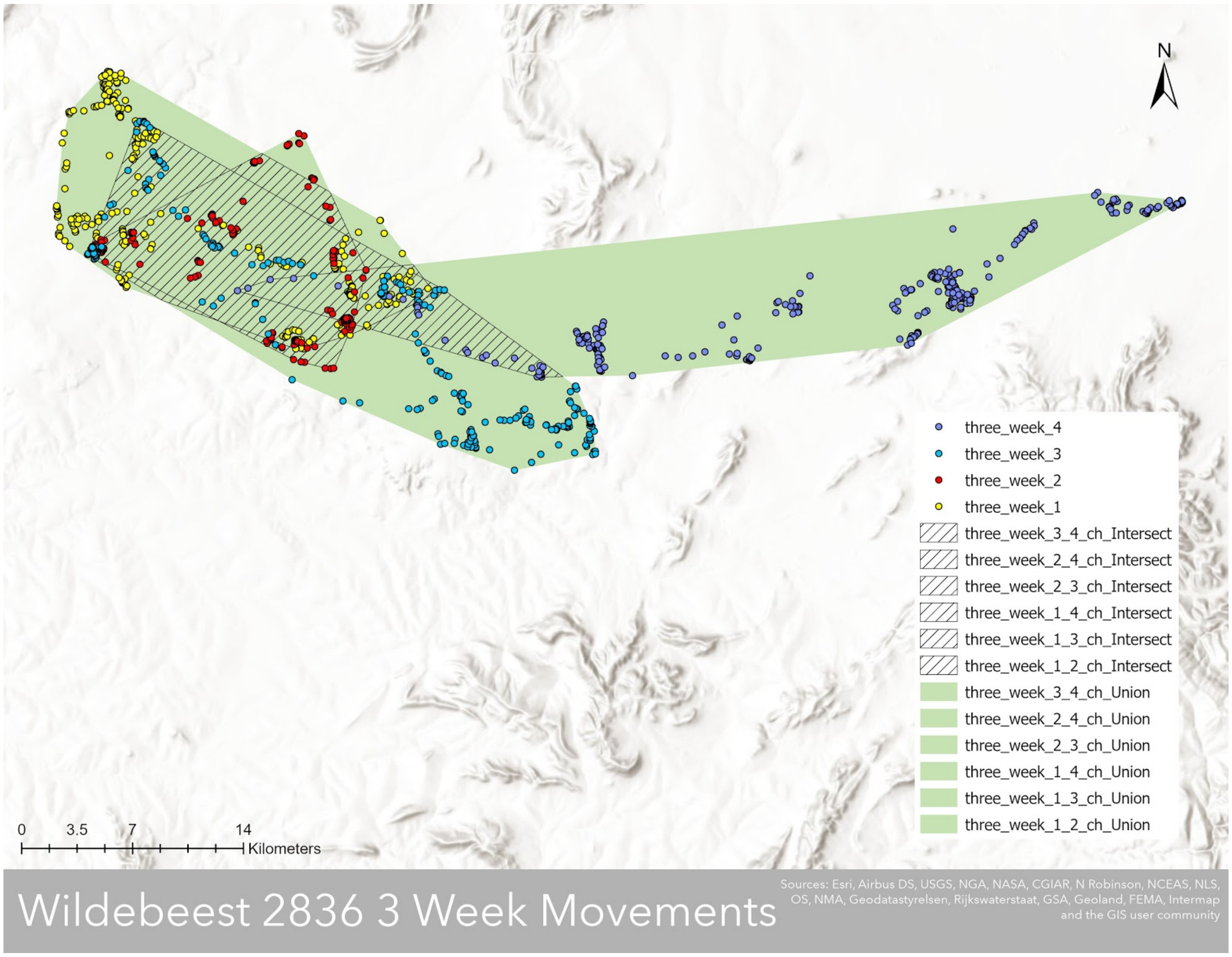


Wildebeest 2836 Week by Week Movement

Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community

Figure 9: Wildebeest 2836 Week by Week Movements from June 2011 - August 2011

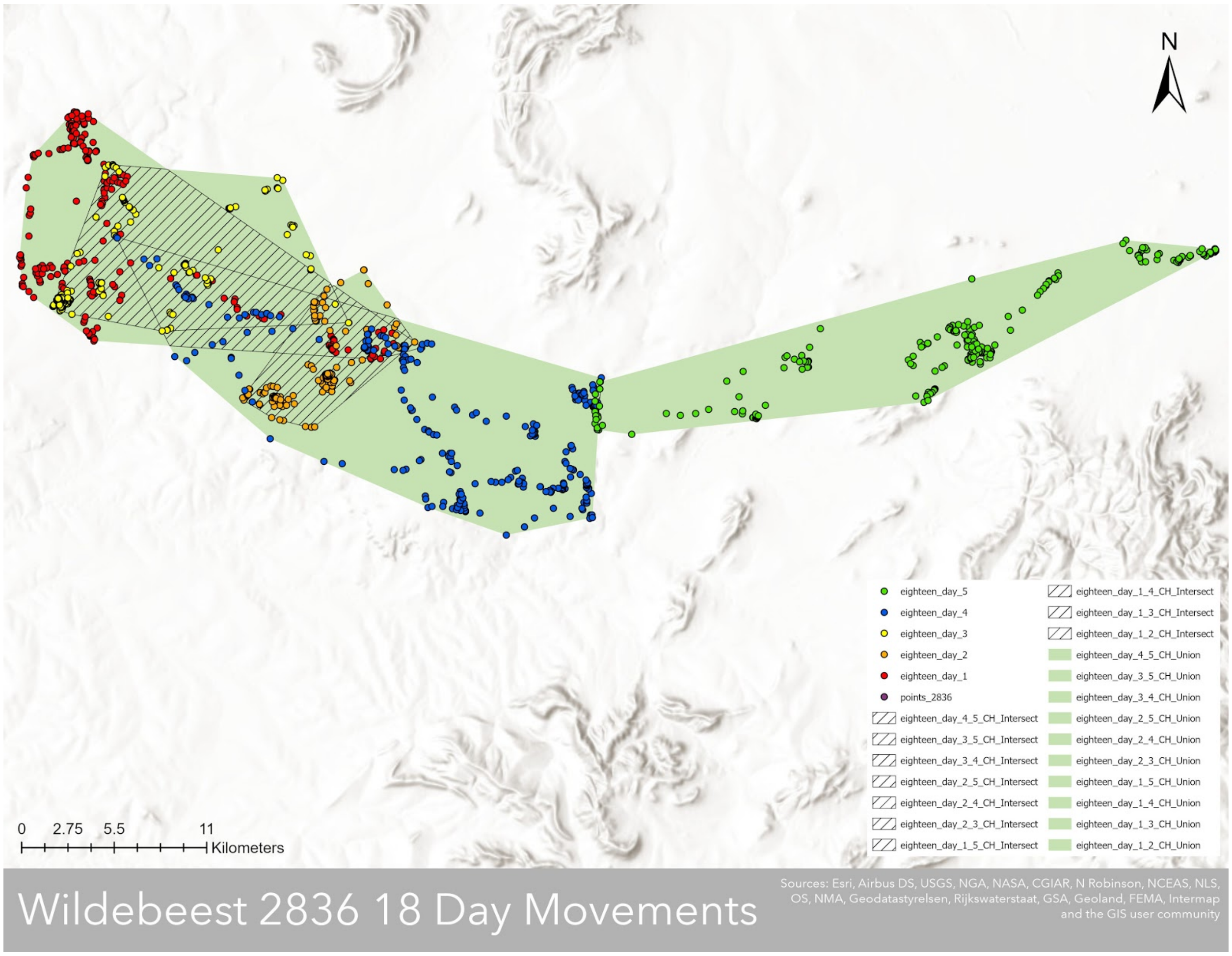
Above is the map of the week by week movements of Wildebeest 2836 from June 2011 - August 2011. For this time interval analysis, we had 78 total comparisons and an average Jaccard Similarity value of 0.03. The intersected area looks very spiky due to the underfitting and high variance of the animal's movements. Although it looks like the area covered is a lot, 50 of the combinations had an intersection of 0 meaning the areas of the compared movement time frames of the same interval did not overlap and share no similar qualities. Overall, what we can see in the map above are the intersections of time intervals that occurred very close, maybe 1-3 week intervals between one another, and not ones that were more spread out in time.



Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community

Figure 10: Wildebeest 2836 3 Week by 3 Week Movements from June 2011 - August 2011

Above is the map of the 3 week movements of Wildebeest 2836 from June 2011 - August 2011. For this time interval analysis, we had 6 total comparisons and an average Jaccard Similarity value of 0.25. Compared to the under-fitted intersection of the week-by-week movements, the three week movements show lack of variety and high bias due to overfitting the animal's range. Due to the interval providing more time and a higher chance of intersection between each of the time interval combinations, each combination in the three week intervals intersected. Because of this, we have larger, more squared-off hatched areas and a relatively large jaccard similarity index value, but this is not an optimal outcome due to too much bias.


N
0
2.75
5.5
11
Kilometers
eighteen_day_5
eighteen_day_4
eighteen_day_3
eighteen_day_2
eighteen_day_1
points_2836
eighteen_day_4_5_CH_Intersect
eighteen_day_3_5_CH_Intersect
eighteen_day_3_4_CH_Intersect
eighteen_day_2_5_CH_Intersect
eighteen_day_2_4_CH_Intersect
eighteen_day_2_3_CH_Intersect
eighteen_day_1_5_CH_Intersect
eighteen_day_1_4_CH_Intersect
eighteen_day_1_3_CH_Intersect
eighteen_day_1_2_CH_Intersect
eighteen_day_4_5_CH_Union
eighteen_day_3_5_CH_Union
eighteen_day_3_4_CH_Union
eighteen_day_2_5_CH_Union
eighteen_day_2_4_CH_Union
eighteen_day_2_3_CH_Union
eighteen_day_1_5_CH_Union
eighteen_day_1_4_CH_Union
eighteen_day_1_3_CH_Union
eighteen_day_1_2_CH_Union
Wildebeest 2836 18 Day Movements

Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community

Figure 11: Wildebeest 2836 18 Day Movements from June 2011 - August 2011

Above is the map of the 18 Day Movements for wildebeest 2836 from the time frame between June 2011 - August 2011. For this time interval analysis, we had 10 total combinations and an average Jaccard Similarity value of 0.11. Based on the results of our statistical analyses, we believe that the 18 Day movement is the most optimal time interval for estimating the home range because it is the best tradeoff between being too underfitting or too overfitting. Looking at the map above, the visualization proves the statistical analysis was correct because the intersection is not too spiky or too squared-off. Just like we expected, the proportion of combinations with no intersection is between that of the week by week and the three week movements at 30% of all the observations (3 combinations). Because of this, the hatched area becomes a bit more rounded off and more representative of the wildebeest's home range.

Conclusion

From our statistical and visual analysis , we concluded that the 18 day interval is the most ideal time interval for studying the home range of wildebeest in the Serengeti National Park. This is shown by the 18 day time period having a Jaccard Index of .187 compared to the two weeks that had a significantly higher Jaccard Index of .378. This drop in Jaccard Index concludes that the linear relationship between the Jaccard Index and time since the area covered is becoming less correlated and that the Jaccard index is becoming more similar for each time difference. With that, the 18 day interval is at the elbow of the graph which decreases the mean square error while being the most effective bias-variance trade. Similarly visualization through our maps, we can conclude that the 18 day interval is most ideal as the intersections are not as spiky or squared off, which allows for the time interval to not be too underfitting or overfitting. From all these analyses we can determine that the 18 day interval is the most ideal time interval for studying the home ranges of wildebeest in the Serengeti National Park. We hope our results prove to be helpful for future researchers looking to explore movement and home range analyses of wildebeests in this region.

Challenges, Gaps and Future Directions

Throughout this project we ran into multiple challenges. At first we were planning on using the Pompeiu Hausdorff distance which calculates the maximum distance between two representations of the same dimension. However, many of our convex hulls intersect, which violates the assumption of the Hausdorff distance that the two polygons in question are disjoint from each other. This eliminated the use of this mathematical distance measurement. If we had more time, we would have tried to explore different distance measurements that would be applicable to our data.

Another challenge we faced was data scarcity. In the White-Bearded Wildebeest dataset, it lacked information about the traits of the wildebeests. The dataset did not include information about the gender and age which would have been interesting information to do further analysis on. The male wildebeest are typically larger and usually reside at the edge of herds to protect the females in the middle of the herd during migration. This may affect the distance traveled by these male wildebeest compared to female wildebeest. Thus, further analysis would be helpful to explore if gender plays a role in determining the ideal time interval for studying wildebeest in Kenya. Along with that, as wildebeest age they travel less distance because they become weaker. Therefore in the future, we would want to conduct more analysis if given these variables about the wildebeest to see if these particular factors affect the ideal time interval for measuring home ranges.

If given more time, we would increase the timeline from just three months, June 2011 to August 2011, to a longer timeline. This would decrease the bias as more data points would be analyzed. This would also allow us to take into account the migration patterns of wildebeest when deciding the ideal time interval. However, due to time constraints we were not able to further expand our timeline. Therefore, in the future we would like to expand our timeline to about a year to incorporate the entirety of migration into our analysis.

In the future, we would also like to conduct the same analysis on different species of animals. The ideal time interval our results produced was an 18 day interval for the wildebeest in the Serengeti National Park, however, we would also want to reproduce this analysis on another animal in another location. Different animals have different behaviors that would change the mean distance they travel over time. Along with that, the contrasting location can introduce new variables that could affect the home range of the animals, such as vegetation and elevation. All these variables can influence the home range of the various animals and would want to test if the ideal time interval we found is representative of all animals.