Exploring MoveBank data and the movement characteristics of barnacle geese

Jacob Peterson

February, 12, 2020

Document Overview:

This document was originally used to create my final project for the ABE 651 (*Environmental Informatics*) class I took while at Purdue University. I am now repurposing it to use as an example in my data science portfolio I am putting together to host on my Github.

The purpose of this class project was meant to demonstrate our ability to find and manage a publicly available dataset, perform exploratory graphical analysis, run a data quality checking sequence, and perform a simple statistical analysis on the data. The project was meant to be open ended so that students could use data they had specific interests in and I chose to explore animal movement data from the data repository site MoveBank.

A repository with the data and code used to create this pdf can be found here.

Abstract

With this project, I explored the offerings of an animal movement data repository and created a Rmarkdown document that contained the code used to clean, format, and explore the data available at MoveBank. I chose to use data depicting movements of barnacle geese in Northern Europe and the North Atlantic within three separate sub-populations. With this data, I examined the regional differences in displacement between the summer and winter ranges as well as differences in summer home range size using one-way ANOVA's. Using post-hoc Tukey tests, I discovered that there was a significant difference in seasonal displacement between the Barents Sea and Greenland sub-populations. However, there was no significant variation between summer home range size in the sub-populations.

Dataset Description:

Dataset Source

The data was downloaded from MoveBank and consists of movement data of 44 barnacle geese (*Branta leucopsis*) individuals in Northern Europe from April 2006 to June 2011. This included data from three specific study regions in Greenland, Svalbard, and Barents Sea (Cabot 2014, Griffin 2014, Jeugd et al. 2014). In total, there were 52,443 total records spread among 44 individuals. This data was originally collected for and used in two other published studies (Shariatinajafabadi et al. 2014, Kölzsch et al. 2015). However, it is not likely that all the possible questions have been asked using the dataset. Open access datasets like these present opportunities to answer new ecological questions with existing datasets.

Dataset Format and Metadata

This data was downloaded in csv format and includes the following data:

Useful data included in all study regions as part of the original datasets:

- **event-id** unique identifier for each movement event
- **timestamp** the date and time the event was recorded in UTC
- **location-long** longitude in decimal degrees (Datum: WGS84)
- **location-lat** latitude in decimal degrees (Datum: WGS84)
- **individual-local-identifier** unique identifier for individual

Created as part of exploratory analysis using existing variables:

- **hours** hours between timesteps
- **dist** distance traveled during a timestep (in meters)
- **dist_hour** distance traveled divided by the hours between timesteps (m/hr)
- **month** month data was collected in
- **year** year data was collected in
- **monthFloor** month and year combination that the data was collected in

Graphical Analysis:

Graphical Analysis Methods

As not a lot of information was included in the original data set, I created a few new columns for the dataset as noted in the list above. This included splitting the data into movement tracts for each individual and calculating time per step and distance moved per step, as well as splitting the data up into different time periods for ease during graphical analysis. First, I identified the distribution of individuals across the study regions (Fig. 1). I then mapped the points to see where exactly the birds from the different study regions were located and look for spatial outliers (Fig. 2). Next I explored the distribution of movement records among individuals and study regions (Fig. 3). This would help identify individuals that had a lot less relocations as well as the variation in the number of relocations gathered between study regions. I also plotted the distribution of the amount of time between consecutive points for each individual (Fig. 4). Large gaps in time between successive points would make the data less useful and such gaps should be removed in further analyses. I was hoping to get fairly continuous data of year round movement in order to calculate both summer and winter home range sizes and estimate differences in step length during migration and within home ranges. To check for this, I used boxplots to plot the standardized distance over time by each month in the entire time series (Fig. 5). This not only helped identify possible variability at certain times of the year, but also time periods that only had data from certain study areas. The patterns identifying the distribution of points across time and study regions were summarized best in the final two graphs that plot the count of points occurring in the years of the study and across individual months (Figs. 6 and 7 respectively). Because there is a noticeable lack of points recorded during the winter months (Fig. 7), I will only be estimating home ranges in the summer breeding months.

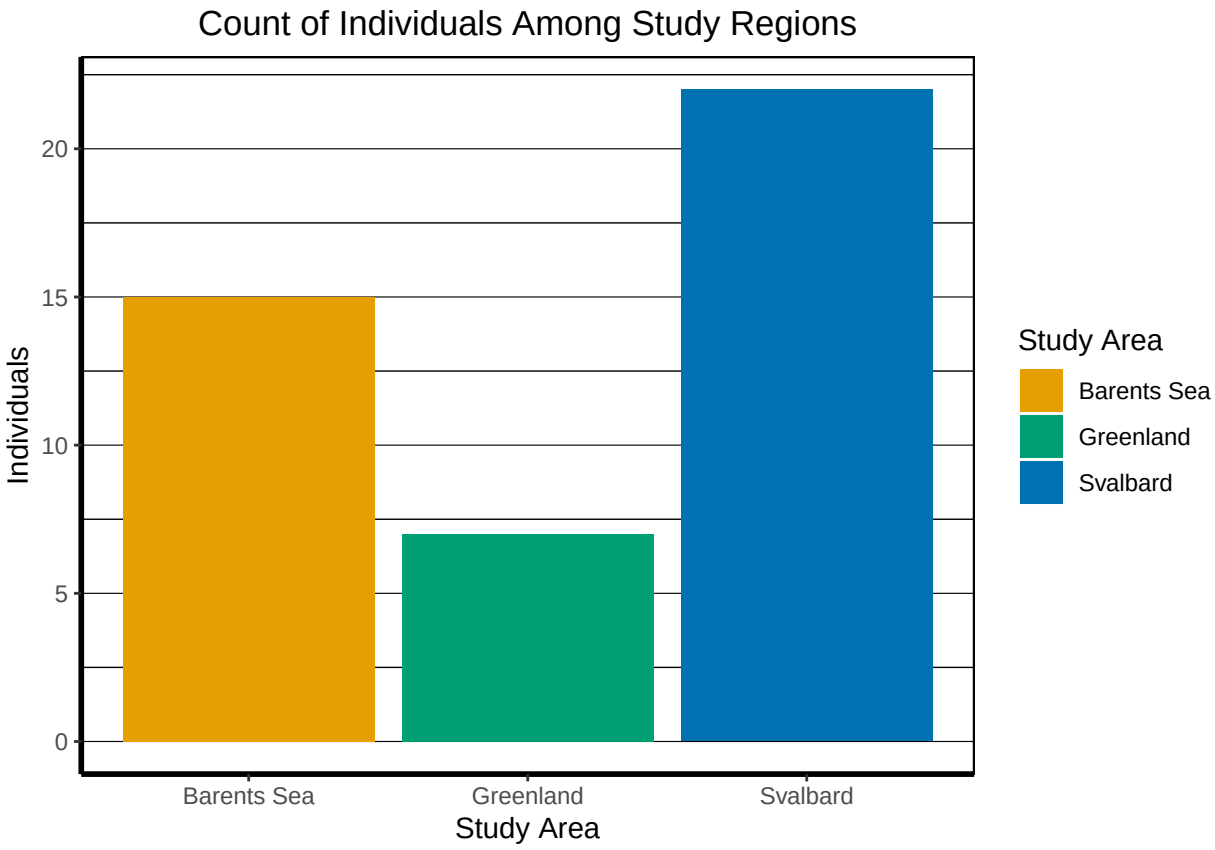

Figure 1: Distribution of individuals among study regions

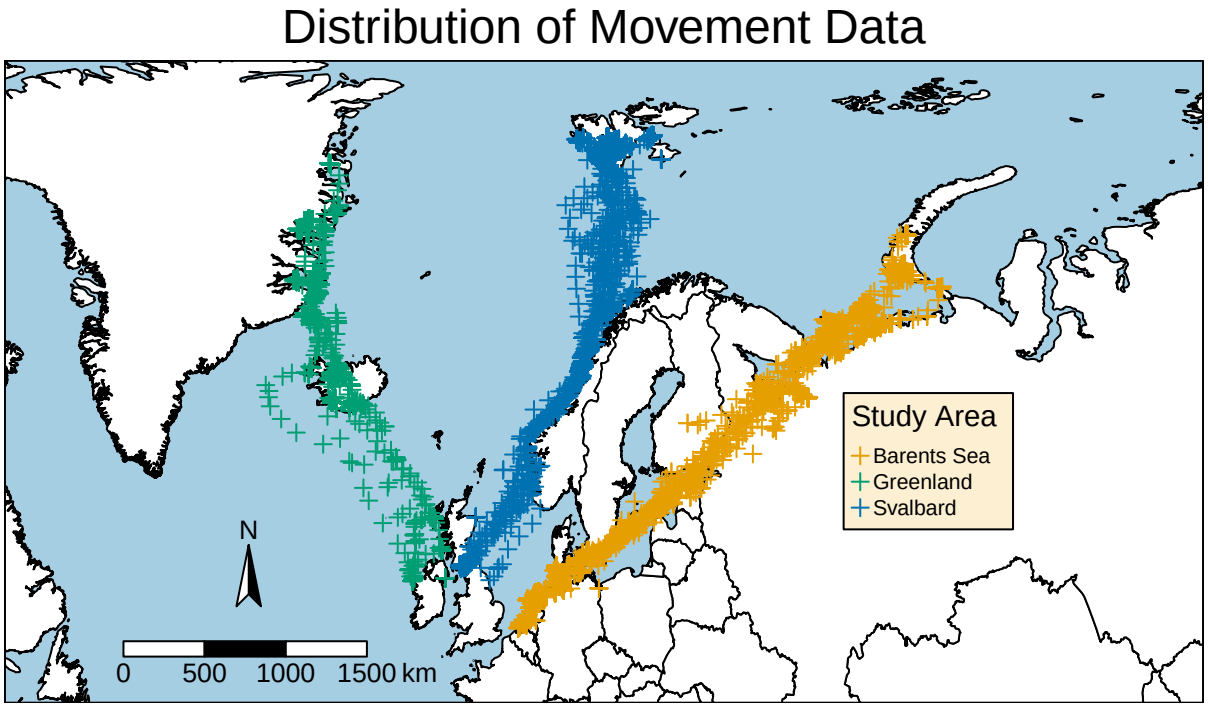

Figure 2: Distribution of movement data across all study regions

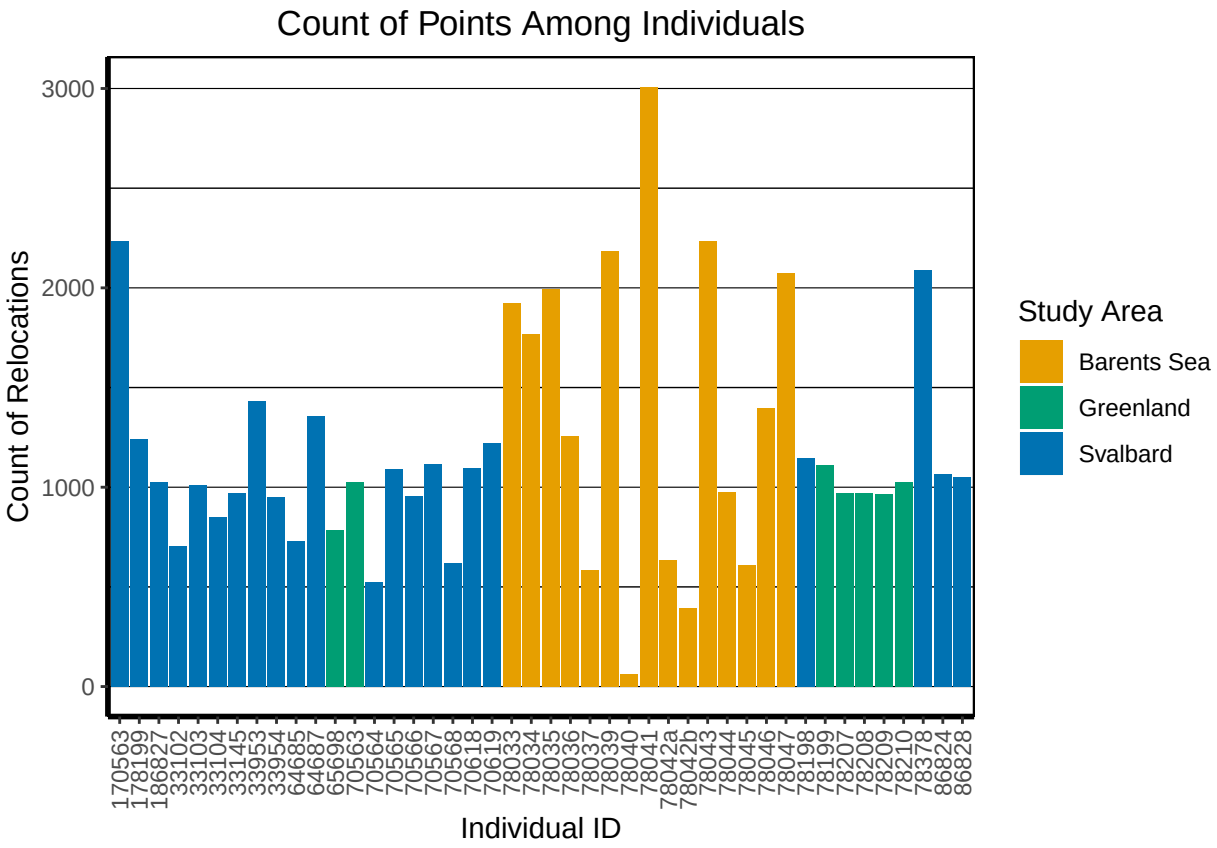

Figure 3: Distribution of movement relocations among individuals and study regions

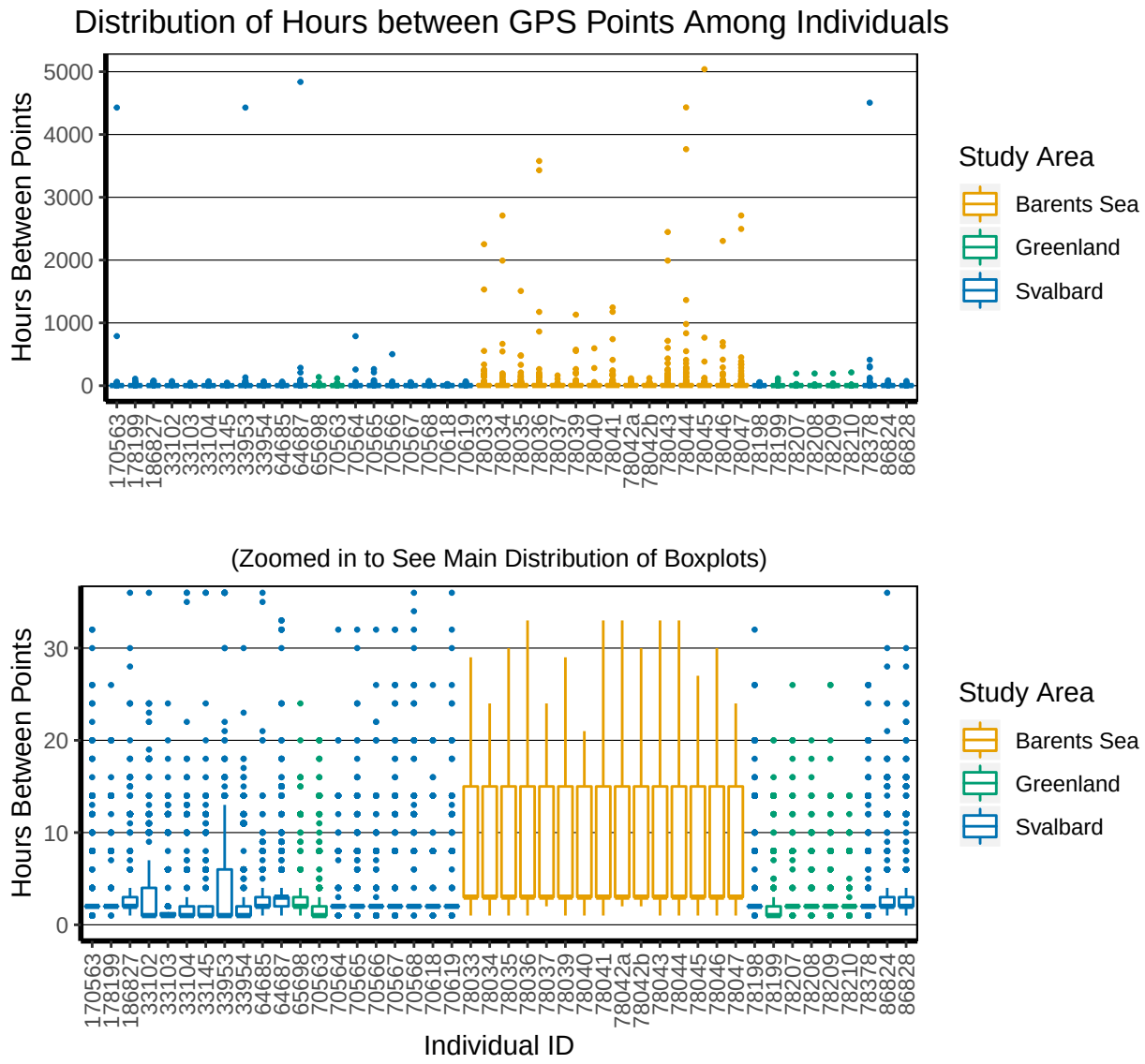

Figure 4: Distribution of the length of time between successive relocations on bird data by individuals and study regions

Figure 5: Distribution of distance traveled, standardized by an hour, during each movement across time

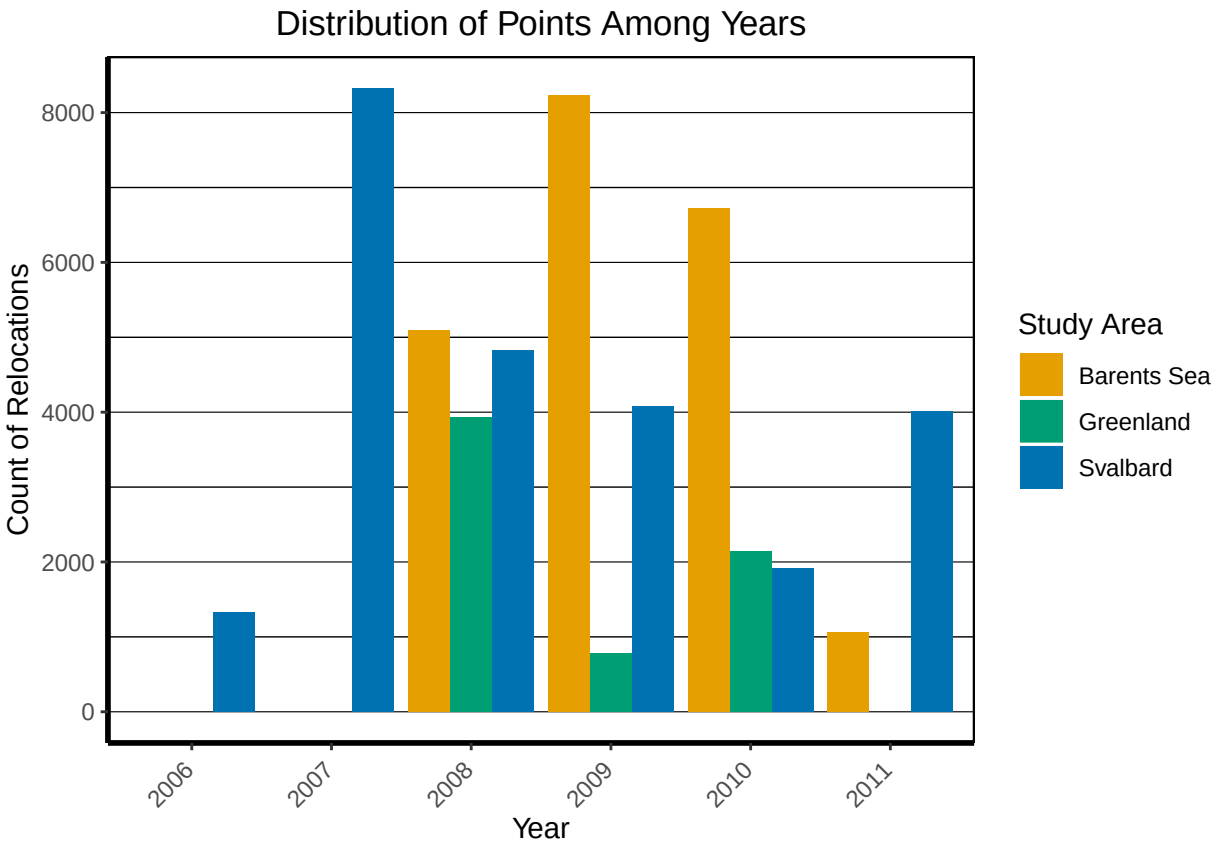

Figure 6: Count of points occurring in the different years of the studies

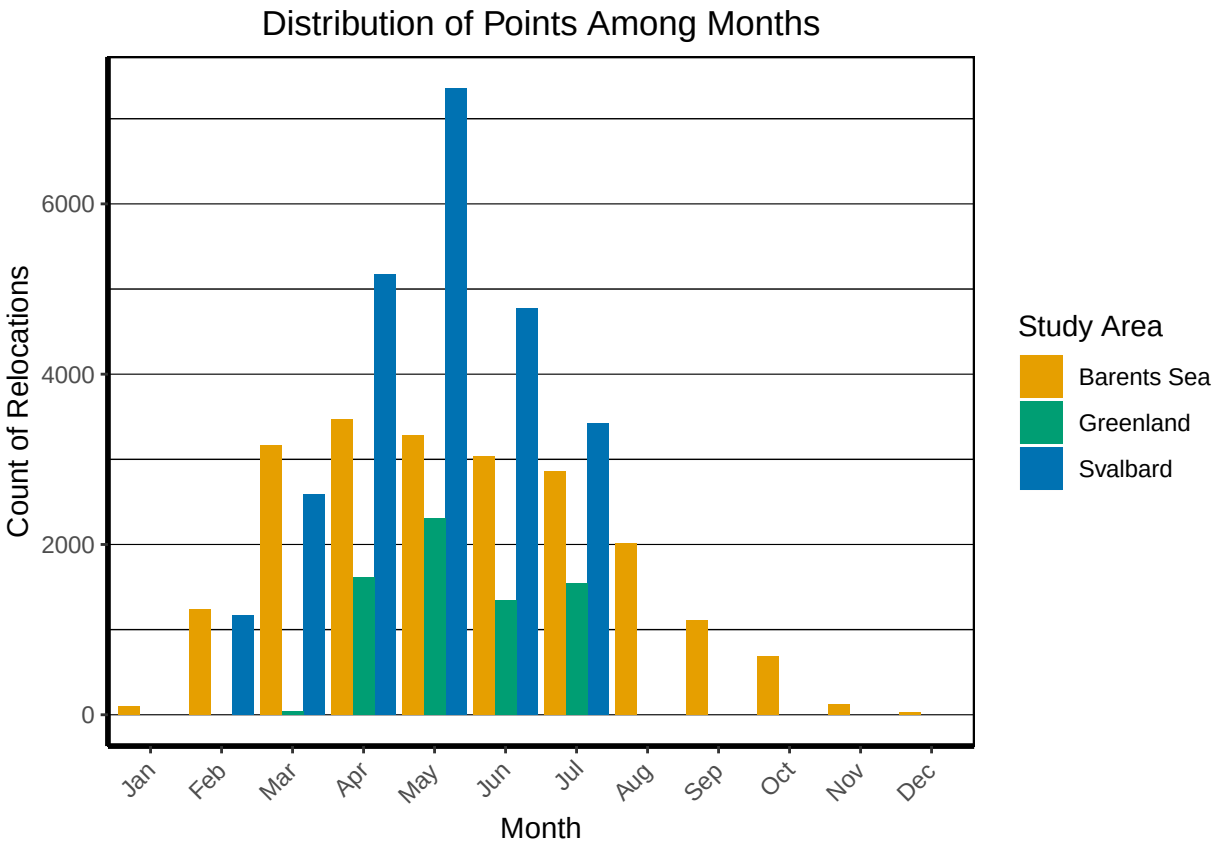

Figure 7: Count of points occurring in the different months of the year

Data Quality Checking

Data Quality Checking Methods

After completing the graphical data analysis, I found that the spatial data visually seemed to be consistent with no gross errors or GPS points way outside of the study areas (Fig. 2). This was likely influenced by the data quality checking standards required for submission to the MoveBank database. Despite this, we ran a nearest neighbor analysis to flag odd points. If a point was found to be greater than 75 km from its nearest point, it was flagged for closer inspection. This check ended up providing a few flagged points (Table 1), but after examining them individually, it was determined that they were just from individuals that made long movements in a timestep and weren't that close spatially to the other tagged individuals.

I also wanted to flag movements that spanned extremely large distances during a timestep. These locations may indicate GPS errors and I should consider their inclusion more closely. These flagged movements could also be an artifact of the GPS collar only logging points seasonally or in a case where there were blocks of missed satellite fixes. For this check I parsed the dataset for movements with a distance traveled greater than 100 km. This actually flagged 737 points, which would be a lot to go through. In order to cut down the number of points, I needed to consider the timestep length during those long movements. Moving 100 km over 1 hour is a lot different than moving 100 km over 24 hours. During the exploratory data analysis, I measured used hours as the unit for timesteps and created a distance per hour column in the dataframe as well. I then ran a similar check to flag movements greater than 100 km per hour which flagged a more reasonable 37 points. However, after doing some background research on barnacle geese, I did find evidence that while flying 100 km per hour is on the upper range of the species abilities, it is not unheard of. With this evidence in mind, I chose not to remove any of the flagged 37 locations.

During the graphical analysis, we also found that there was one individual in particular (individual ID 78040) that had a very low number of relocations, especially in comparison to other individuals in the study (Fig. 3). This could be indicative of either a failed GPS transmitter, or an individual that suffered from capture myopathy and died quickly. To avoid distorting our dataset with an individual that contributed very little data, all individuals with less than 100 points were removed from the dataset during this data quality check.

Most of the inconsistencies identified within the dataset in the exploratory analysis seemed to be related with the temporal continuation within the dataset. The Barents Sea data had the most variation in the spread of hours between consecutive points (Fig. 4). While the other study regions had an average of 2 hour timesteps, Barents Sea data logged a lot data on different temporal scales. Patterns I noticed while searching through the data included a 5 hour period and a 19 hour period daily and a 3-3-3-15 hour schedule daily. This will limit some analyses that can be done with the Barents Sea data. Despite this, all regions had some large outliers in in times between GPS fixes that need to be flagged and dealt with. For this check, I flagged all points that had greater than 48 hours between the GPS fixes. This check will be useful when deciding which points to remove from the distance per hour check. I expected to get a lot of flags during this check and I will have to look at where these breaks occur and if there is a pattern there. GPS transmitters could be programmed to take less points during certain seasons or it could just be an error with the transmitter in communicating with the satellites. To use this data I will likely have to separate the data into bursts of movements around these long periods of time. After examining the flagged data, I resolved to remove all of the flagged points in this check as non-continuous bursts shouldn't impact results involving home range and spatial displacement analyses.

Final Data Analysis

Final Data Analysis Methods

For my analysis, I wanted to test for differences in the average displacement between winter and summer (breeding) sites, as well as regional differences in summer home range size among the sub-populations of barnacle geese. To calculate displacement distance I created a nearest neighbor matrix using the 'FNN' package in R for each individual across all of their relocations and recorded the max distance between any

Table 1: Table summarizing the results of the data quality checking

Quality Check	# of Individuals Flagged	# of Relocations Flagged	# of Relocations Removed
Spatially Isoloated Points	8	20	0
Long Distance	43	737	0
Long Distance by Hour	10	37	0
Low Data Individuals	1	63	63
Hours Between Steps	41	310	310
Totals	44	977	369

two points. I then ran a one-way ANOVA across the three sub-populations and used a post-hoc Tukey test to examine pairwise differences in displacement.

For home range analysis, I only focused on the summer breeding season as there was a lack of data during the winter months (Fig. 7). In order to focus on relocations of birds residing in their summer range (not still migrating) across all sub-populations, I subset the data to only include relocations occurring in June and July. I then used the 'adehabitatHR' package in R to calculate the 95% kernel density home range for each individual using their points from the selected time period. Similarly to the displacement analysis, I used a one-way ANOVAs using the sub-populations as the grouping factor to examine regional differences in home range size.

Final Data Analysis Results

When examining the regional differences between winter and breeding sites using max displacement distance, I found that the only significant regional difference was between the Greenland and Barents Sea populations ($p = 0.02$; Fig. 8). Displacement distances ranged from 918 to 3325 km, with the most variation occurring in the Barents Sea subpopulation.

When comparing summer home range sizes, I ended up having to remove another individual from the Barents Sea sup-population because it lacked at least 100 points during the summer breeding months. The ANOVA found no significant variation between regions ($p > 0.05$; Fig. 9). The largest variation in home range size occurred in the Svalbard region, while the least occurred in the Greenland subpopulation. The home range size varied across study regions from 1 square m up to 56 square km. This variation is likely caused by the difference between individuals that were always on the nest when a GPS location was taken and individuals that did not breed that summer.

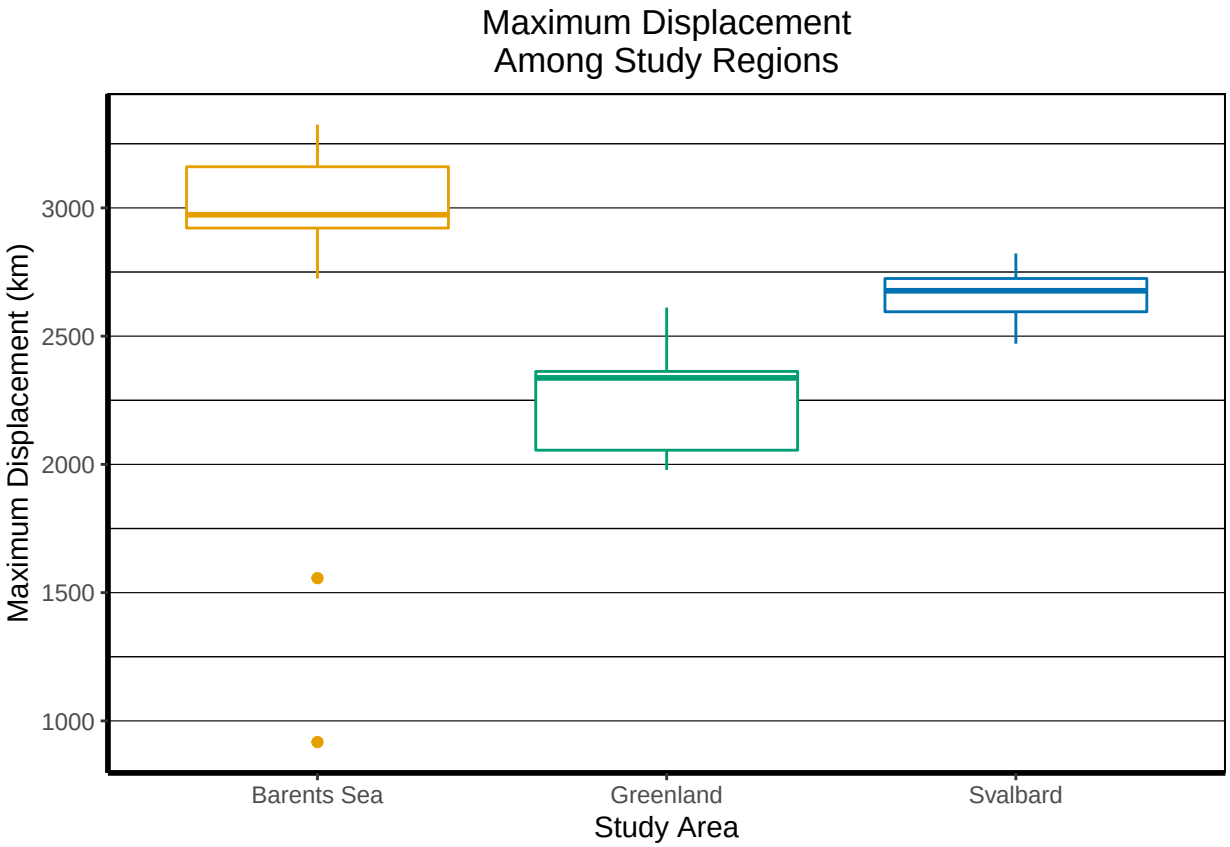

Figure 8: Distribution of maximum displacement distance (km) in barnacle geese between winter and summer ranges across breeding sub-populations

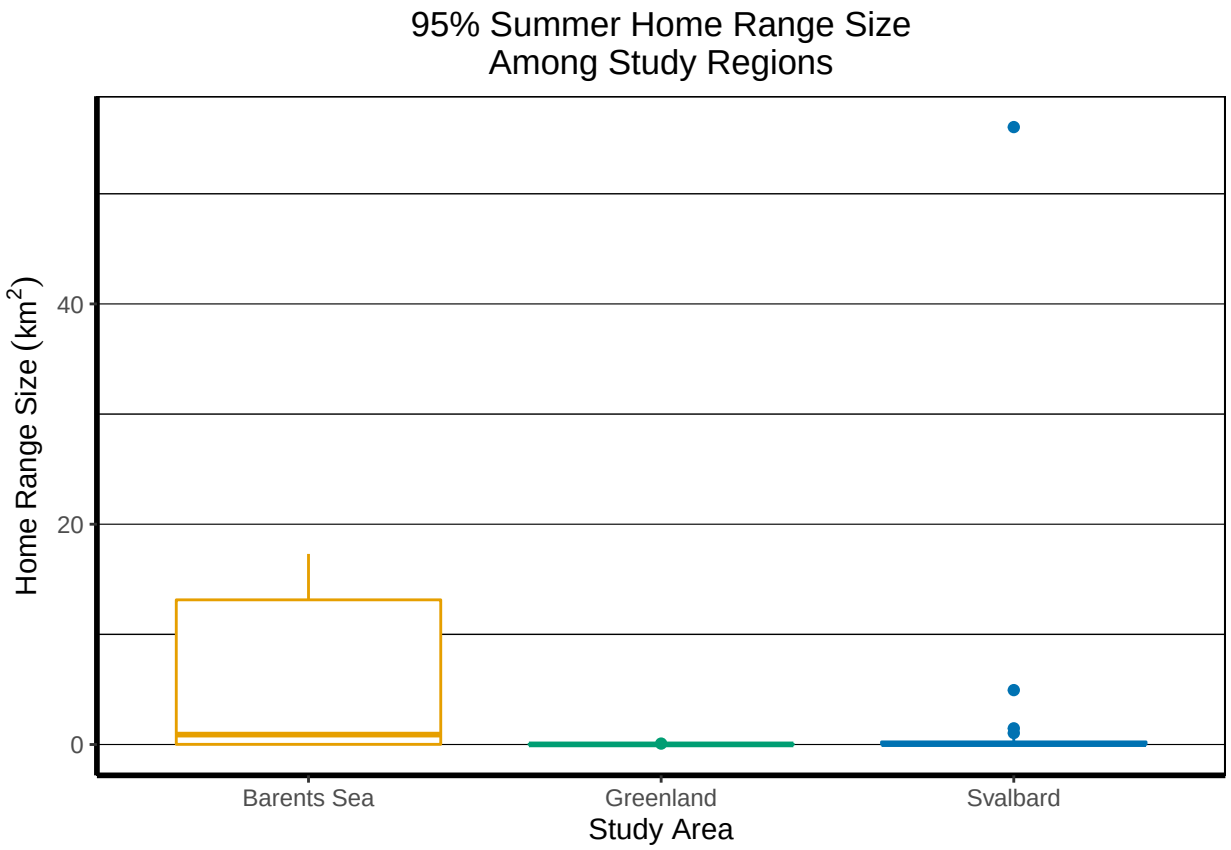

Figure 9: Distribution of summer home range size (km^2) in barnacle geese across breeding sub-populations

Works Cited

Cabot, D. 2014. Data from: Forecasting spring from afar? Timing of migration and predictability of phenology along different migration routes of an avian herbivore [greenland data]. DOI: doi:10.5441/001/1.5d3f0664. Movebank data repository.

Griffin, L. 2014. Data from: Forecasting spring from afar? Timing of migration and predictability of phenology along different migration routes of an avian herbivore [svalbard data]. DOI: doi:10.5441/001/1.5k6b1364. Movebank data repository.

Jeugd, H. van der, K. Oosterbeek, B. Ens, J. Shamoun-Baranes, and K. Exo. 2014. Data from: Forecasting spring from afar? Timing of migration and predictability of phenology along different migration routes of an avian herbivore [barents sea data]. DOI: doi:10.5441/001/1.ps244r11. Movebank data repository.

Kölzsch, A., S. Bauer, R. de Boer, L. Griffin, D. Cabot, K. M. Exo, H. P. van der Jeugd, and B. A. Nolet. 2015. Forecasting spring from afar? Timing of migration and predictability of phenology along different migration routes of an avian herbivore. Journal of Animal Ecology 84:272–283. DOI: 10.1111/1365-2656.12281.

Shariatinajafabadi, M., T. Wang, A. K. Skidmore, A. G. Toxopeus, A. Kölzsch, B. A. Nolet, K. M. Exo, L. Griffin, J. Stahl, and D. Cabot. 2014. Migratory herbivorous waterfowl track satellite-derived green wave index. PLoS ONE 9:1–11. DOI: 10.1371/journal.pone.0108331.