Ethical Considerations

This project is ethically significant because of the global threat of climate change. We have to consider that climate change will have vastly disproportionate impacts across different populations. Weather prediction is ethically significant because people depend on weather prediction for things like crop planting and disease prevention. We will be analyzing data from the ISD (International Surface Database), which is used to build global climate models. We are interested in whether these climate models are accurate globally, or if they have a U.S. or global north bias. To do this, we will analyze the global distribution of weather stations.

Historically, there have been large gaps in Africa and South America with this reporting system. Because NOAA uses data from climate stations to build their climate models, there may be serious ramifications in the climate models if the stations aren't evenly distributed. New station identifiers were implemented in January 2013, but we don't know what impact these will have. As the attached image shows, there are still significant gaps in where stations are located, and swaths where there are no stations. When assessing bias in what locations are chosen, we have to keep in mind that certain locations may be more or less accessible. There are also a variety of ways that we collect data from locations, including balloons, aircrafts, computers, and satellites. It is beyond the scope of this class to determine the accuracy and effectiveness of each of these weather data-collecting methods.

Additionally, the data we collected was administered by NOAA, which is a U.S.federal government entity. The access of data and administration of NOAA changes with the presidential administration, and is impacted by various climate and

fossil fuel lobbies. However, NOAA uses strict “Information Quality Guidelines” to conduct quality control on its data and minimize institutional biases. Of course, one of the things we are interested in for this project is NOAA’s institutional bias - we want to see how a U.S. governmental organization is biased in the weather stations it collects data from to build its climate models.

While cleaning the data, we removed data that was listed as N/A for temperature or station name. Further research might delve into stations that weren’t recorded or named, and if there’s a pattern within these. We chose to use clustering, which offers limited bias because it is an unsupervised learning tool.

We use maps in many of our diagrams that present the data. When we consider the way that maps are typically presented, they generally have North America on top and South America on the bottom. There isn’t a scientific reason most maps are built this way--after all, Earth is a spinning sphere with no true “top” or “bottom.” We can also look at pieces which argue for the ethical importance of mapping. We should also consider the spacing, and how this might lead to overlapping dots for stations, which could make the map harder to read.

Write Up

In a warming climate, weather predictions matter and data about weather matter because climate change will bring large food shortages and natural disasters. Weather prediction is important for making planting decisions, disease protection, and climate resiliency. Weather stations are distributed across the globe, and track data like precipitation, temperature, visibility, and pressure. In our project, we wanted to observe

bias and variability in weather stations. We analyzed this data cartographically using ArcGIS (Geographical Information System), which is a powerful mapping tool that allows the user to manipulate data cartographically to draw conclusions, and has broad applications in many fields (including environmental science).

The data is from the NOAA Integrated Surface Database, and is hourly access data from stations around the world. Our steps included finding the database, getting the database from AWS, and cleaning the data. We dealt with numeric versus character data, and tried to align these so everything ran smoothly. For example, the different station names were originally parsed as numeric, because each station is identified by a string of numbers. We also had to deal with the missing station names, which were sometimes marked by a series of Xs and other times marked by a string of 0s and 9s. The missing temperatures were marked as either a string of 9s or N/A. Because each of the recordings also dealt with dates, we used the “lubridate” package to effectively deal with this. We went through different station numbers and names, and parsed these as numeric. We also collected data from the WorldBank, to find information on the populations and relative wealth level of countries (and compare it to the number of weather stations).

We used ArcGIS to create a visual analysis of weather station distribution, and used the aggregate function of the weather station coordinates to reverse geocode the number of weather stations in each country. We paired this mapping data with WorldBank data on 2020 population and wealth of the individual countries (which WorldBank either categorises as higher, upper middle, lower middle, or lower), to

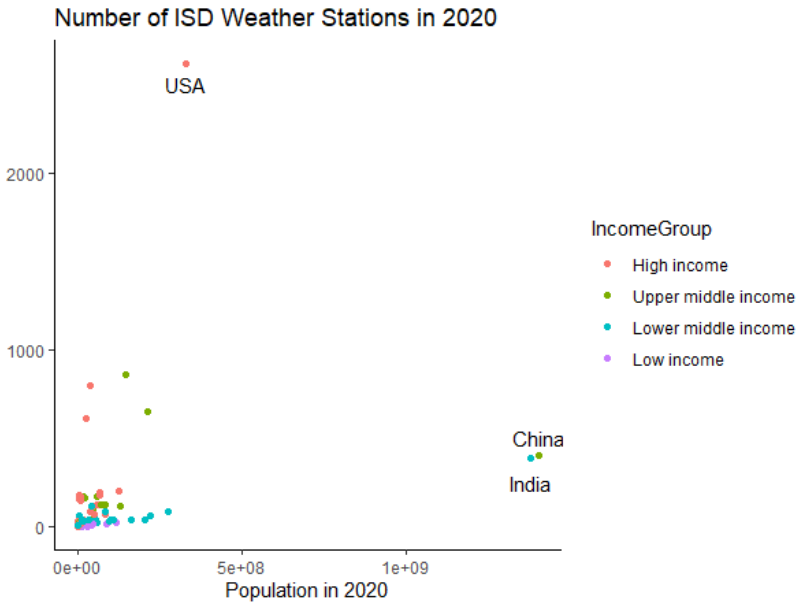
analyze the distribution of weather stations across richer countries, and whether they are distributed according to population.

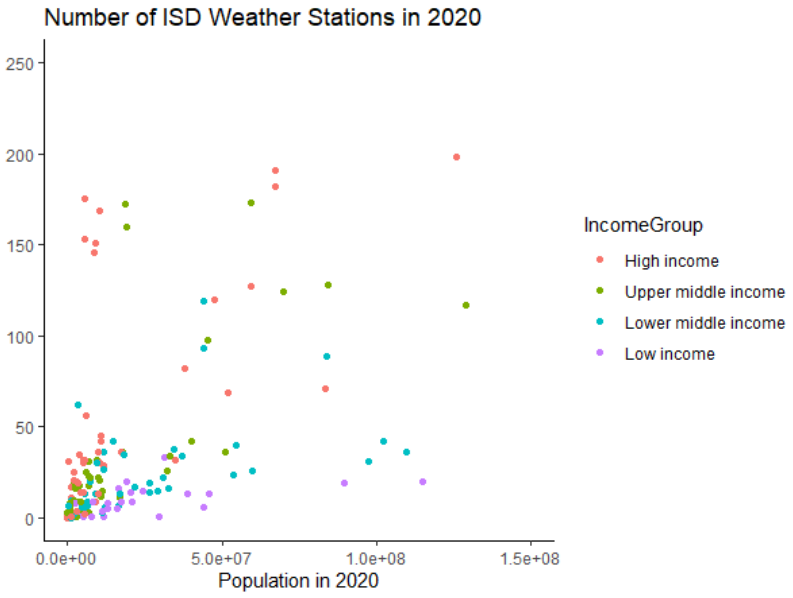
Based on visual analysis from the ArcGIS map and the comparisons with WorldBank data, it is clear that weather stations are unequally distributed between upper- and middle-income countries and low-income countries. From the zoomed in plot, we can see that no low-income country has over 50 weather stations; in comparison, wealthy countries such as the United States do not even fit on the zoomed map, with over 2500 stations. The mean number of stations for countries in each income group is as follows:

Income group	Mean # Weather Stations
High	112.207 +/- 360.969
Upper middle	67.135 +/- 156.570
Lower middle	31.453 +/- 56.072
Low	9.440 +/- 7.990

There is a clear trend in a decreasing mean number of weather stations as nations get less wealthy. Additionally, the standard deviation decreases with decrease in wealth, suggesting less variance in number of weather stations as countries get less wealthy. While there are some massive outliers (such as the United States and Russia) for the

wealthy countries, low income countries tend to uniformly have lower numbers of weather stations.





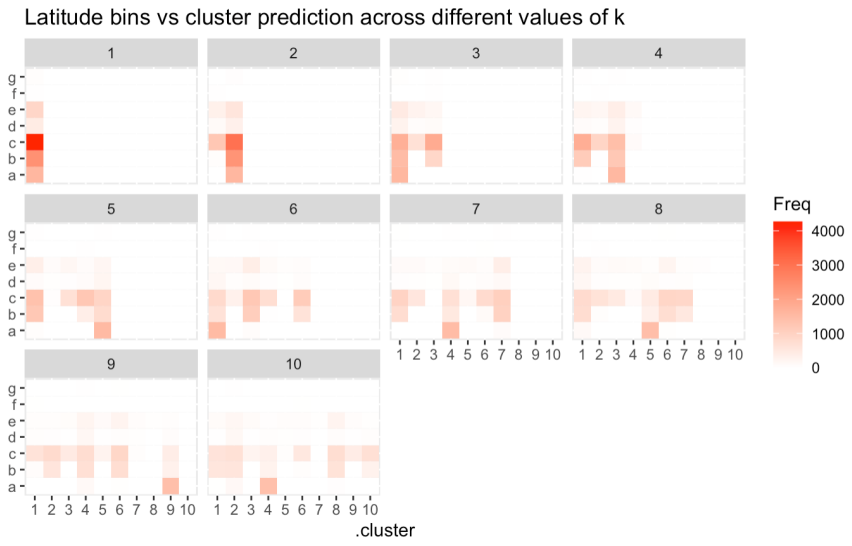
One major limiting factor during our project was computing power. With hundreds of thousands of stations and hourly observations, it took around a day to run sections of our code. This also led us to reduce the amount of stations we tracked, and the time

period we tracked them during. Because these stations come from around the globe, they are fairly generalizable. However, we should consider the time period we took these samples during. For example, if one area was going through monsoon season or another area was going through a drought, then this would have impacted their weather measurements during this time.

Another limitation with the WorldBank data is that a country with a large population could be a small, highly dense area that does not need as many weather stations as a less populous, but more geographically diverse country. Therefore, the number of weather stations according to population is not a perfect measure of whether individuals are “covered” by a weather station. Additionally, wealth of a nation is a broad measurement that does not account for wealth inequalities within the citizens, and so flattens our understanding of station distribution on a national level.

Our look at clustering also revealed areas for further data analysis. The clustering was a multi-stage process. The coding process is broken up into the cleaning of the data, the creation of a categorical variable to compare clusters, and the clustering process itself. For cleaning, the temperature data was originally documented with hour, day and month. Since we want to examine variance across all 12 months, we tried for each station to only focus on a single day of the month. We then averaged the temperatures over the course of the day to deal with the multiple measurements per day. While originally focusing on the 15th of each month, we found that many stations did not have measurements for the 15th for all 12 months. We used the 17th day of the month to substitute in for the missing 15th day. By making sure that each station had 12 measurements, we knew that we weren’t double counting the 15th and 17th. We found

annual variance for each station based on the 12 measurements. For the creation of the categorical variable we binned the absolute value of the latitude into bins of range 15. This was the categorical variable lat_bin which we tried to predict for using clustering on the variable of annual variance. We then applied the clustering algorithm on the data for the annual variance variable and observed whether those clusters isolated the bins of absolute value latitude. When we clustered, we expected to see clustering by distance from the equator. Our code for clustering involved looking at the absolute distance by longitude. We clustered on the Annual Temperature Variance for values k Between 1 and 10, and we then compared these clusters to divisions. The results did not cluster as we expected. This could be for a variety of reasons--perhaps we didn't have enough data, perhaps there aren't equally distributed data points, perhaps climate change is already impacting weather variance in stations far from the equator.



While mapping often claims to be unbiased, in reality the map-maker holds a lot of power in the construction of geographical narratives. Reverse geo-coding has an element of bias because the coders had to make a decision on where to classify borders, which is dependent on politics. Also, we had to choose the bins for the number of weather stations to color-code the map according to station distribution, which was heavily influenced by outlier nations with extremely high numbers of weather stations. For example, all countries with more than 253 stations are colored dark blue, but the United States has over 2500 stations! Therefore, these bins are highly influenced by outliers.

To conduct this project, we used AWS (Amazon Web Services) to connect with the data and share it between team members. We also used GIS (Geographic Information Systems) to look at weather station distribution. Both of these are outside of the scope of the class.

Works Cited

ABC News. 2020. "'It's a Gimmick': There's More to Your World Map than Meets the Eye," August 1, 2020. https://www.abc.net.au/news/2020-08-02/theres-no-such-thing-as-upside-down-world-map-racist/12495868.

"Data for High Income, Middle Income, Low Income | Data." n.d. Accessed December 14, 2021. https://data.worldbank.org/?locations=XD-XP-XM.

"Data Search | National Centers for Environmental Information (NCEI)." n.d. Accessed December 12, 2021. https://www.ncei.noaa.gov/access/search/data-search/global-hourly.

David Kretch and Adam Banker (2021). paws: Amazon Web Services Development Kit. R package version 0.1.12. https://CRAN.R-project.org/package=paws

Hadley Wickham, Jeroen Ooms, Kirill Müller (2021). RPostgres: Rcpp interface to PostgreSQL.

Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr

"Index of /Data/Global-Hourly/Access/2020." n.d. Accessed December 12, 2021. https://www.ncei.noaa.gov/data/global-hourly/access/2020/.

"Information Quality Guidelines." 2002. Accessed 12 December, 2021. https://www.noaa.gov/organization/information-technology/policy-oversight/information-quality/information-quality-guidelines.

Jeroen Ooms and Hadley Wickham (2021). rcurl: a Modern and Flexible Web Client for R. R package version 4.3.2. https://CRAN.R-project.org/package=curl

Koch, Tom. 2020. “2. Ethics, Geography, and Mapping: The Failure of the Simple.” In *Ethics in Everyday Places*. https://covid-19.mitpress.mit.edu/pub/8jw8qhr2/release/1.

Laet, Joseph Dumit and Marianne de. 2014. “Curves to Bodies:The Material Life of Graphs.” In *Routledge Handbook of Science, Technology, and Society*. Routledge.

“Most World Maps Show North at the Top. But It Doesn’t Have to Be That Way - ABC News.” n.d. Accessed December 12, 2021. https://www.abc.net.au/news/2020-08-02/theres-no-such-thing-as-upside-down-world-map-racist/12495868.

“Population, Total.” n.d. Accessed 12 December, 2021. https://data.worldbank.org/indicator/SP.POP.TOTL.

Smith, Adam, Neal Lott, and Russ Vose. 2011. “The Integrated Surface Database: Recent Developments and Partnerships.” *Bulletin of the American Meteorological Society* 92 (6): 704–8.

Steve Dutky and Martin Maechler (2021). bitops: Bitwise Operations. R package version 1.0.7. https://CRAN.R-project.org/package=bitops

Thomas J. Leeper, Boettiger Carl, Andrew Martin, Mark Thompson, Tyler Hunt, Steven Akins, Bao Nguyen, Thierry Onkelinx, Andrii Degtiarov, Dhruv Aggarwal, Alyssa Columbus, and Simon Urbanek (2021). aws.s3: ‘AWS S3’ Client Package. R package version 0.3.21.

https://CRAN.R-project.org/package=aws.s3

Thomas J. Leeper, Jonathan Stott, and Mike Kaminsky (2020). Aws.signature: Amazon Web Services Request Signatures. R package version 0.6.0. https://CRAN.R-project.org/package=aws.signature.

“The Ethical Mapping Guidelines: How Not to Map – EthicalGEO.” n.d. Accessed December 12, 2021. https://ethicalgeo.org/the-ethical-mapping-guidelines-how-not-to-map/.