Alexander Gustafson (algu6965)
Data Mining Project Milestone 2

Exploring Snow Depth Trends: Analysis, Prediction, and Insights Using Advanced Data Mining Techniques

Problem Statement and Motivation

The motivation behind this project is to find interesting patterns and long term trends regarding European Alp snow depth data collected between 1971 to 2019 to help our understanding of climate change's effects on snow depths. Secondly, the motivation is to use this knowledge to make predictions for future years' snow depths and snowfall amounts. The data mining process of the dataset will allow us to find these hidden patterns, which cannot be found with other methods due to the size of the dataset. Previous knowledge includes programming abilities and data mining concepts that will be crucial for mining the data, being able to understand and use the results, and to create visualizations to further help put the found patterns into an easy to understand visual. I will apply this previous knowledge to this project by using my programming and mathematical skills to dig through the data and find patterns. The interesting trends I hope to find from this project include how significantly climate change has lessened the snow depths in the region, when this began to occur, if there were timeframes where snow depths dropped the most, if there are any year over year increases, and what will the snow depths look like in the future.

Literature Survey

The previous work done on this data is from the paper "Observed snow depth trends in the European Alps: 1971 to 2019"[1] which describes where the data came from, what it contains, and their own analysis on the data. It came from the European Geosciences Union, which was able to receive the data from the six countries discussed in the data. This paper acknowledges that the data is more in depth in some countries than others, mainly due to some countries not having observation towers in many regions in the beginning. Most countries significantly increased the amount of observation areas several years to decades after when this study began. The study found that across all regions, the average mean snow depth was -8.4% per decade and that the maximum snow depth was down -5.6% per decade. It also notes that seasonal trends varied greatly across regions of similar elevations, which is an interesting pattern that I hope to replicate.

Most other work discussing the snow depth in the European Alps comes from another study that looked at trends over the last 600 years, which found an unprecedented decrease over that period of time. I hope to find similar patterns by looking at my more in depth data from a shorter time frame.

Proposed Work

There will be much to do in order to get the data ready for analysis due to how large the dataset is. For each of the six countries looked at, each one contains the data for every month over 40 years for over 100 cities. For example, the dataset for Germany contains all of the collective data over 40 years for 943 different towns and cities. To make the data more manageable, I will begin by finding the best methods of clustering. For one country, it would be ideal to group together regions with similar characteristics such as elevation. Instead of using each unique data point, the average value could be used for regions within certain predefined elevations such as 300-500, 500-750, etc. Similarly, since this

[1]Matiu, Michael, et al. "Observed Snow Depth Trends in the European Alps: 1971 to 2019." The Cryosphere, Copernicus GmbH, 18 Mar. 2021, tc.copernicus.org/articles/15/1343/2021/.

project aims to look at the Alps themselves and not their smaller, lower elevated towns, a threshold will be used to eliminate regions that fall below a certain elevation. One city mentioned is Absdorf, Austria with an elevation of 597 feet, which seems too low for its snow depths to be significant. The dataset contains many cities will elevations well over 5,000 feet which is more useful.This dataset comes with a file “meta_all.csv” which contains the name of the city as well as it’s coordinates, elevation, and snowfall and snow depths for the beginning and end of each year. This data can be merged with other datasets so that the elevation can be used to prune the data. In total, the meta_all file contains 2,979 unique cities, many of which likely do not fit in with the project.

The design of the ultimate dataset will be one dataset that contains the monthly observations for all regions above a certain elevation. The data will be inserted into a SQL table that can be used to query the data. The primary dataset will be made using pandas and will combine the monthly data from the 12 monthly datasets, and it will be merged with the meta_all dataset so that elevation will be included.

What sets this project apart from previous work done with the data is it will look at the data in a more broad sense to find patterns that cover the entire area. Previous work has looked primarily at each dataset separately and

Data Set

The dataset is a combination of several datasets that record snow depth measurements as well as many other relevant measurements within Austria, France, Germany, Italy, Slovenia, and Switzerland. is composed of 20 sub-data sets, with 7 representing daily measurements, 12 representing monthly measurements, and one representing the meta data. Each sub-data set represents a region indicated by the country and the area. For example, “data_monthly_AT_HZB” represents the country of Austria and within it are over 100 different cities where data was collected.

The data was not collected from one entity, but was instead collected individually from each country. For example, the data used for Austria was taken from the Austrian Hydrological Service while France was taken from the national weather service Météo-France. This means that consistency will be an issue as several countries did not begin collecting the data at the same time and include different measurements. Also, the number of weather stations varied greatly, with Germany doubling the number of stations in the late 70s yet they steadily decreased in number beginning in the early 2000s. However, in total the weather stations grew and reached their peak numbers in the 80s through the early 2000s.

Due to the variability of the data and when the data collection began, clustering will be used to group together regions that had similar timescales for data collection. Also, pruning will be used to cut down on the amount of regions looked at by removing those with too small of a time scale to find patterns in. Certain regions may be entirely eliminated if they are deemed to be too sporadic, inconsistent, or too late starting.

The measurements that remain consistent include HNsum, which gives the measurement of total snow depth (in centimeters). HSsum represents the mean snow depth, and HSmac measures the maximum snow depth. This project will focus on the monthly data, which explains where the maximum and mean snow depths come from and how they relate to the total snow depth. There are several gap filled values such as HSmean_gapfill and Frac_gapfilled that indicate how many observations had to be filled in from other nearby stations due to missing values. While this is important information, this project will likely not use columns like this and will instead trust that the values given are close enough to the true values to perform a good analysis on. Lastly, Month and Year remain consistent between data sets, representing the year that data was collected and the month indicated by 1-12, with 1 indicating January and 12 indicating December.

Evaluation Methods

Evaluation methods will include box plots for visualizations of the data for each year or decade, helping to recognize outliers and variability. K Means clustering will be used to group together years with similar snow depth patterns into clusters to identify distinct trends with similar patterns and behavior. When it comes to making predictions, decision trees will be used to identify important predictors of snow depth variability and to better classify years into different categories, perhaps on environmental factors and elevation. Bayesian classification will be used to compute the information gain on environmental variables such as temperature, snow fall, and elevation. This approach will assist in understanding the significance of each variable in predicting snow depth trends accurately.

Tools

Pandas has already been used to gain a general understanding of the dataset, and has been helpful in determining how large the dataset is. A jupyter notebook has been created on my local machine in VS Code, and the datasets have been imported in. Pandas has been used to load in the csv files where general commands have been used such as .nunique(), .max(), and .head() to determine what the datasets looks like and what they contain. With millions of data points across all of the csv files, pandas will be used more to combine the data into one general dataframe. Once this has been created, pandas will still be used for some general querying, however, it will be loaded into a SQL table so that I can make more specific commands that combine many features.

The use of SQL will be important due to the variability of when data was and was not collected throughout the dataset. An example of SQL's utility in this will be to "SELECT HSmean FROM data WHERE Elevation > 750;". Commands like this could be beneficial in getting a broad understanding of the data without the worry of missing yearly values. I can perform other queries that use "WHERE year BETWEEN 1971 AND 1980;" to extract all data where the year meets the requirements. Once this data is pulled, that is when pandas will be used.

Lastly, Scikit-learn will be used to make future predictions, GitHub will be used for version control, and the data mining textbook used in class will be referenced for methods and algorithms. Plotly and Seaborn will be used for visualizations including box-plots and histograms towards the end of the project.

Milestones

The first milestone is to clean the data, which I hope to have done by the end of week 10. The data cleaning will include pruning values and merging the datasets into one. Since the data came from many different sources this will likely be a lot of work and the tools listed previously will be used to aid in this. The second milestone is to begin data mining to find interesting patterns which will be done by the end of the spring break, around March 31st. Once this is done, I can begin a data analysis on the patterns found which will be done in early April, between weeks 11 and 12. The third milestone will be to use this newfound knowledge to create visualizations so that I can share the results with more people, as well as help myself gain a better understanding. Lastly, I will use the patterns found to train some machine learning models to make predictions on the next few decades which will be done by the end of week 13. Week 14 will be used to do any last minute cleaning to make the project presentable and I will practice presenting the data.