Analyzing the Effects of Weather

Erens, Sam
Department of Mathematics
Department of Statistics
University of Illinois, Champaign
Champaign, IL, United States
serens2@illinois.edu

Hasson, Emily
Department of Computer Science
Department of Statistics
University of Illinois, Champaign
Champaign, IL, United States
ehasson2@illinois.edu

Nelson, Lucas
Department of Economics
Department of Statistics
University of Illinois, Champaign
Champaign, IL, United States
lln2@illinois.edu

***Abstract*—In this paper, we address the challenges in collecting, cleaning, and analyzing gigabytes of weather related data. Using archives of**

***Index Terms*—spatiotemporal data, Apache Spark, model fitting**

I. Introduction

Each day, thousands of weather stations around the globe record and collect measurements of temperature, pressure, precipitation, and other factors involving the weather and climate. The National Oceanic and Atmospheric Administration is an American agency that forecasts weather and conducts scientific research about the climate and oceans. With history tracing back to the early 19th century, the NOAA has collected enormous amounts of atmospheric data since its inception. One of its largest collections of atmospheric data is the Global Surface Summary of the Day, which contains daily measurements dating back to 1929 and includes daily readings from over 9000 weather stations across the globe.

Weather prediction is one of the most classic and well-studied types of large data analysis problems, and its continued importance in research today stems from both its complexity and its practical significance to people and the economy. We wish to explore this problem by fitting models to historical weather data, while also exploring issues involved with gathering and analyzing large collections of data. Our analysis will contain models for predicting the occurrence of tornadoes as well as more specific regional analyses of severe weather occurrences.

II. Related Work

- Discuss published work that relates to your project.
- Emphasize how your approach will be similar or different from others.

We will be looking for a **minimum of 6** scholarly works cited.

Three citations must originate from either a scholarly journal or are pre-prints on arXiv. Consider searching for articles using Google Scholar: `https://scholar.google.com/`. From there, click the “cite” button to automatically generate the appropriate citation from MLA, APA, or BibTex. That said, there is a preference for using BibLaTeX or natbib to avoid any bibliography formatting errors or mixing styles.

III. Data

- What type of data is it (text/network/image/sound)?

  We are working primarily with weather data from various government agencies (NOAA, NCEI, etc). Most of the data is in the form of structured tables.

- Who collected the data?

  The data gathered for this project is provided by the National Oceanic and Atmospheric Administration (NOAA)[1], a government agency that collects an array of information pertaining to daily weather forecasts, severe storm warnings, and more climate-related instances.

- How large is the data’s size when uncompressed?
- How many records and features exist?
- Did you have to apply any pre-processing to the data?
  - List any preprocessing steps
  - e.g. Normalization, Tokenization, ...
- Were any special steps required?
- What kind of training/validation(dev)/test split is expected?
- Show examples (if possible) of the data.
  - For structured data, include the first 10 records.
  - For text data, include a few unstructured entries.
  - For image data, include a few images.
  - For sound data, include the wav graph of the music file.
  - And so on...

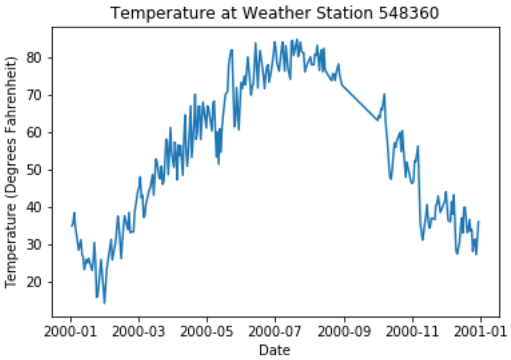
Please include a reference to where the data set can be found. **This does not count toward the minimum of 6 works cited.**

A. Preliminary Technical Details and Results

For points in this section, you must have at least one model fit and described within the progress report.

[1] This data is publicly available for anyone to use under the following terms provided by the Dataset Source

- Describe in detail the proposed model.
- Explain how it works in general (or specifically with your data).
- Show preliminary results in:
  - Summary tables:
    - Classification should highlight precision, recall, and accuracy metrics.
    - Regression should state the RMSE, MSE, or MAE.
  - Graphs

    The temperature at this weather station appears to peak in July at around 80 degrees Fahrenheit and drop to less than 20 degrees Fahrenheit in February. This suggests that this weather station is most likely located in the northern hemisphere.

    - Training and validation graphs based on the cost/loss function and evaluation metric (accuracy, F1, recall, RMSE, ...)
- Provide a timeline of the project to date and what the future work will entail.
  - Summarize completed and future work by using a Gantt chart to break down the tasks and due dates.
    - Break down each task by specifying **who** is **doing what** and **when** it will be done.
    - Avoid saying "Everyone in the group" is working on a single task.
- Code
  - Please provide code in a ZIP file or link to a GitHub repository. **Do <u>not</u> submit your data set!**
  - **If you have a private GitHub repository, please add @coatless.**

IV. Contributions

At the end of the progress report, please include a brief 1 - 2 sentence write-up of what each group member contributed to this stage of the project. Award each member with a percentage between 0 - 100 such that the sum of all percentages across group members will be equal to 100. **This section does not count toward the page limit.**

V. Grading

The project will be graded according to a rubric. There will be no possibility for resubmission. To avoid grading surprises, please speak with a member of course staff about your draft during Office Hours prior to turning it in.