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

- Describe the problem being solved and why it is important.
- Discuss your motivation for pursuing this problem.
- Clearly state what the features and model output will be.
  - Note that these components can be reused from the project proposal paragraph.

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)?

  For our project, we will be working with structured weather data provided by various government agencies (NOAA, NCEI, etc.). Although the range of all possible years covers 1901 to 2019, we will focus on more recent data - strictly speaking, 1975 to 2019, both due to the increase in quality and quantity that each of the NOAA's global stations provide.

- 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?

  Billions and billions of bytes (1.4 GB to be exact)

- How many records and features exist?

  Across the entire dataset, there exists millions records and 23 features, where each observation represents the daily observation of certain variables (e.g., datetime of record, mean temperature (F), indicators for natural disasters) for a given weather station.

- Did you have to apply any pre-processing to the data?

  Not yet...

- Were any special steps required?

  The archives of the data were stored in nested '.tar' folders, first segmented by year then segmented by station, and contained an array of non-ideal values that either contained two series of information in one series or were inconsistently alternating between missing value codes. By looping over each file, we extracted some series out into multiple series to discretize the data and preserve information and combatted inconsistent missing value codes using regular expression.

- What kind of training/validation(dev)/test split is expected?

- Show examples (if possible) of the data.

  An example of the final dataset can be found in the Appendix as Table 1.

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

  - 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 <u>at least one model fit</u> and described within the progress report.

- Describe in detail the proposed model.

  For the first phase of the project, we will fit a logistic regression model to the dataset to predict the presence or absence of a tornado at each weather station using most, if not all, of the other variables in the dataset as predictors. Currently, we have one model fit to the 1975 dataset that appears to be doing quite well; however, there is only one tornado recorded for 1975 so perhaps this is due to chance. We would like to expand our analysis to the years 1975 to 2019 and create separate logistic regression models for each of the four regions of the U.S. (Northeast, Midwest, West, and South).

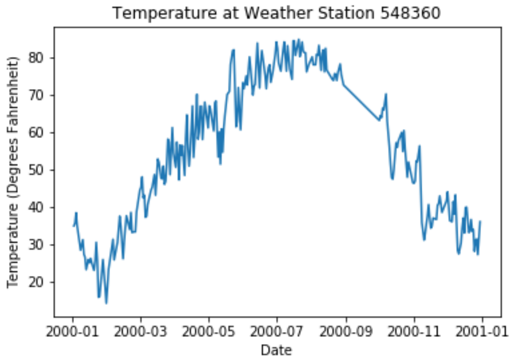
- Explain how it works in general (or specifically with your data).

  The model is quite simple in principle. As we learned in class, logistic regression uses numeric predictor variables to predict the outcome of a zero-one binary response variable. In this case, the response variable is the presence or absence of a tornado, but logistic regression models have also been used in the past to predict credit card default and detect malignancy in cancer patients. Things get a little bit more complicated when you are working with multiple predictor variables rather than a single predictor, but the underlying principle is the same.

- Show preliminary results in:
  - Summary tables:
    - Classification should highlight precision, recall, and accuracy metrics.

      There were zero tornados in our dataset, and the model predicted zero tornados. Therefore, the misclassification rate is zero.

    - Regression should state the RMSE, MSE, or MAE.
  - Graphs



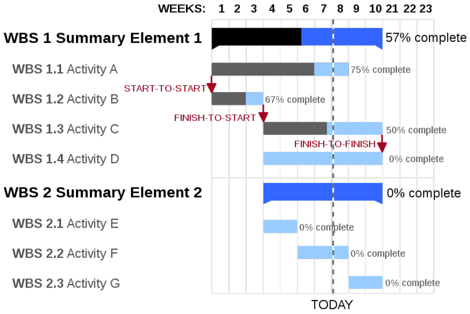
The temperature at this weather station appears to peak in July at around 80 degrees Fahrenheit and drop to less than 20 degrees Fahrenheit in February. This suggests that this weather station is most likely located in the northern hemisphere.[2]

    - 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.

  We hope to have data for the entire period from 1975 to 2019 by Friday, April 29th. We will then focus on developing separate models for each geographic region of the United States. We hope to be finished with the modeling portion of the project no later than Friday, May 6th so that we can spend the final week analyzing the results and editing the video presentation.

  - Summarize completed and future work by using a <u>Gantt chart</u> 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!**

[2] https://en.wikipedia.org/wiki/Northern_Hemisphere

- **If you have a private GitHub repository, please add @coatless.**

  https://github.com/emilyhasson/gsod-analysis

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.**

Lucas helped find and read in the data, even going so far as to write an automated Python script that pulls the data from the NOAA servers and then uploading it to Google Drive, our makeshift cloud storage solution. Lucas also created the initial LaTeX file for the report.

Sam made several contributions to the report, including the preliminary results section. He also began the initial modeling phase of the project by fitting a logistic regression model to the 1975 data.

Emily wrote the introduction section of the report. She also helped find the dataset and created a shared GitHub repository, which you can access at the following URL: https://github.com/emilyhasson/gsod-analysis

Sam: 33.3%
Lucas: 33.6%
Emily: 33.1%

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.

VI. REFERENCES

https://en.wikipedia.org/wiki/Northern_Hemisphere

VII. APPENDIX

Station	Datetime	Temperature	...	Precip Flag
548360	19750101	35.7	...	G
548360	19750101	35.7	...	G
548360	19750101	35.7	...	G
548360	19750101	35.7	...	G
...	...	...	...	...
548360	19750101	35.7	...	G
548360	19750101	35.7	...	G
548360	19750101	35.7	...	G
548360	19750101	35.7	...	G
548360	19750101	35.7	...	G

TABLE I
SAMPLE OF FINAL DATASET