Challenges of Big Data: Predicting Occurrence of Storm Types per Historical Data (with LaTeX)

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 archived weather data.**

***Index Terms*—big data, spatiotemporal data, logistic regression**

I. Introduction

Emily's work goes here ...

II. Related Work

Weather forecasting has proven to be a difficult task, involving the ingestion of massive amounts of climate streaming data by models fit for big data with the purpose of summarizing observed or forecasting unobserved weather data. Numerous industries and professionals rely on the accuracy of these forecasts, stressing the importance of first accurately measuring various weather attributes across time.

- 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

For our project, we will be working with structured weather data provided by various weather-related government agencies[1], including the National Oceanic and Atmospheric Administration (NOAA) and National Climate and Environmental Institution (NCEI), gathered by each of the internationally distributed sensors from the beginning of 1901 to the end of 2019. Each of these sensors report basic logging information (e.g., datetime and station identification) - which can be merged with provided text files in the same archive that provide geographical information (latitude, longitude, named location) of each station - as well as climate data on a daily basis (e.g., mean temperature (F), dew point (F), natural disaster indicators, etc.). Although the entire archive spans 120 years, we will strictly be focusing on data generated beyond 1975, the year the NOAA decided to increase the number of sensors they deploy. Altogether, this sample of the original archives contains 23GB of data comprised of NN million observations with 23 attributes each.

Prior to conducting our analyses below, we did not perform any preprocessing to our data. However, this does not mean that the data was prepared ready for analysis. Rather, only after iterating over each year's series of '.gz' files were we able to prepare our dataset for analysis, including mapping a uniform null value to various missing value codes, expanding cluttered columns into multiple columns, and condensing each station's annual report to a single annual report. (Table 1 in the Appendix shows the first five and last five observations of the resulting dataset.) Given this, we will conduct k-fold cross validation on samples of our dataset below in our logistic regression models to allow hyperparameter selection to determine our best model.

[1]Archived dataset can be found on Kaggle: *NOAA Global Surface Summary of the Day* (see References)

IV. Preliminary Technical Details and Results

For points in this section, you must have at least one model fit 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).[2]

- 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.[3]

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

      There were no tornados in our dataset and the model predicted no tornados. Therefore, the misclassification rate is zero.[4]

    - Regression should state the RMSE, MSE, or MAE.
  - Graphs
    - 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 predictions 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 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 not submit your data set!**
  - **If you have a private GitHub repository, please add @coatless.**
    https://github.com/emilyhasson/gsod-analysis

[2]https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
[3]https://en.wikipedia.org/wiki/Logistic_regression
[4]https://github.com/emilyhasson/gsod-analysis/blob/main/gsod-analysis-part-1.ipynb

V. 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%

VI. 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.

VII. References

https://www.census.gov/
https://www.noaa.gov/
https://en.wikipedia.org/

VIII. 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