Challenges of Big Data: Predicting Occurrence of Storm Types per Historical Data

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. Using historical weather data provided by NOAA, we apply an array of statistical models to classify the occurrence of a specific weather event. Using popular software of the field, our findings contribute to an immensely extensive body of literature of weather classification techniques.**

***Index Terms*—big data, spatiotemporal data, Apache Spark, logistic regression, k-means clustering**

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. To give an example of the economic importance of weather forecasting, the U.S. government spent over 5 billion USD on weather forecasting in 2009 [1]. Additionally, weather events caused over 27 billion USD in damages to property and crops in 2020 [2]. Furthermore, these numbers to not account for the loss of human life, which cannot be compared in value to property losses.

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 and fog as well as more specific regional analyses of severe weather occurrences. We will use Apache PySpark to handle the somewhat large quantity of data required for this analysis.

II. Related Work

Weather forecasting has proven to be a difficult task, both in retrospective analysis and unobserved forecasting, due to the massive amounts of data required for an analysis and possible repercussions of an inaccurate forecast. As a result, researchers have drawn from other fields to help manage the ingestion of massive, multidimensional data [4, 3, 5], relying on clustering and dimensionality reduction techniques to derive patterns and remove noise or outliers among less-computationally expensive subsets of weather data. These results have allowed others to design and refine models that adapt to the changing climate [6] and are more robust and statistically significant when presented with the problem of frequent climate changes.

However, even with the aforementioned advances in the field, weather forecasting remains a challenging issue that holds heavy implications in the case of an inaccurate or misleading forecast. Businesses and industries that predominantly operate outdoors rely on the accuracy of forecasts, both short-term and long-term, as it has a direct effect on their profits [1], allowing for these businesses to better manage costs of operation well before they could otherwise [2].

With these advances and corresponding challenges in mind, the field is set on adapting recognized ”big data” technologies that can allow for greater insights thanks to the added capabilities: digest data of **volume** and accelerated **velocity**, produce analyses of higher **value** and greater **variety**, and ensure the pipeline is backed by **veracity**. Such technological capabilities have allowed for meteorological researchers and data gathering stations to develop models that report and adjust as data is streaming in, making use of Apache Spark streaming [7] functionality. Furthermore, Apache Spark has been utilized to ingest data in parallel fashion, partitioning the previously mentioned streamed data into work streams that can be parallelized, operated on, and combined into a resulting database [8] that would have required greater monetary and computing resources previously.

Our analysis makes use of the previous findings, and we aim to add to this existing body of literature by demonstrating these difficulties in gathering weather data as well as composing a meaningful analysis from massive, multidimensional weather-related data.

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.

IV. Exploratory Data Analysis

Emily's work goes here ...

V. Methods

For our third model, we utilized the k-means clustering algorithm to determine the underlying cluster structure of the country's various climates. Although a relatively simple clustering algorithm, k-means identifies k clusters by minimizing the Euclidean distance between a point and its assigned centroid for all points in the passed data matrix. Our input matrix consists of each state's average temperature, dew point temperature, elevation, wind speed, and sea level pressure for each of the years in the range 2000 to 2019. Due to the drastically different standard deviations across all five attributes, we first normalized the attributes to that no one attribute would greatly influence the clustering results.

We evaluate our models using model-specific evaluation metrics. ... , whereas our clustering algorithm, since it is an unsupervised learning problem, will be evaluated by cluster cohesion and separation (e.g., silhouette score) as well as intuition.

VI. Results/Discussion

Considering our k-means model, we selected a model that identified four clusters. This decision was informed by our silhouette elbow plot fighere as well as intuition. Although our silhouette plot is shows similar values at $k = 2$ and $k = 4$ clusters, we preferred $k = 4$ clusters because the slight change in silhouette score allows us to define more distinct clusters that $k = 2$ clusters cannot convey. Given this, we constructed a series of boxplots grouped by predicted cluster and plotted them against various attributes used to cluster the observations in the first place fighere. We notice that there exists many relationships between the clusters as seen in tablehere and these relationships correspond to different geographical regions fighere.

References

[1] Jain H, Jain R (2017). *Big data in weather forecasting: applications and challenges.* In: 2017 International conference on big data analytics and computational intelligence (ICBDAC). pp 138–142

[2] Reddy PC, Babu AS (2017). *Survey on weather prediction using big data analystics.* In: 2017 Second international conference on electrical, computer and communication technologies (ICECCT). pp 1–6

[3] Mittal S, Sangwan OP (2019). *Big data analytics using data mining techniques: a survey.* In: Advanced informatics for computing research, Singapore. Springer Singapore, pp 264–273

[4] Sahasrabuddhe DV, Jamsandekar P (2015). *Data structure for representation of big data of weather forecasting: a review.* Int J Comput Sci Trends Technol IJCST 3(6):48–56

[5] Hassani H, Silva ES (2015). *Forecasting with big data: a review.* Ann Data Sci 2(1):5–19

[6] Vannitsem S et al (2021). *Statistical postprocessing for weather forecasts: review, challenges, and avenues in a big data world.* Bull Am Meteorol Soc 102(3):E681–E699

[7] Jayanthi D, Sumathi G (2017). *Weather data analysis using spark—an in-memory computing framework.* In: 2017 Innovations in power and advanced computing technologies (i-PACT). pp 1–5

[8] Palamuttam R et al (2015). *SciSpark: Applying in-memory distributed computing to weather event detection and tracking.* In: 2015 IEEE International conference on big data (big data). pp 2020–2026

VII. Contributions

VIII. Appendix

Station	Datetime	Temperature	...	Precip Flag
725130	19750101	34.8	...	G
725130	19750102	31.4	...	G
725130	19750103	26.0	...	G
725130	19750104	34.8	...	G
...	...	...	...	...
702310	20191227	-16.4	...	I
702310	20191228	-5.7	...	G
702310	20191229	0.6	...	G
702310	20191230	-20.8	...	A
702310	20191231	-16.3	...	A

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

A. *Figures*

Figure 1. Display of silhouette scores for k-means clustering algorithm.

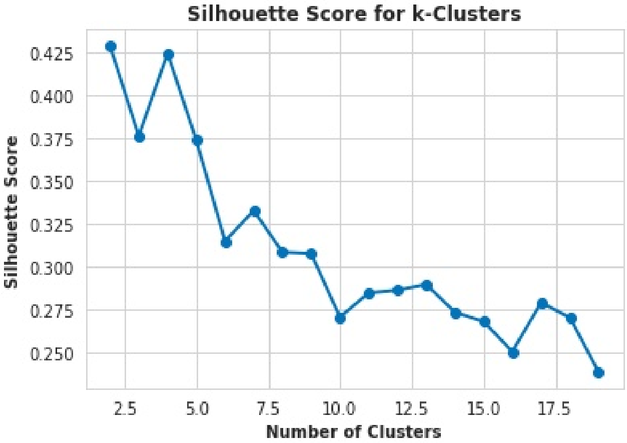

Figure 2. Grouped boxplot of standardized cluster attributes.

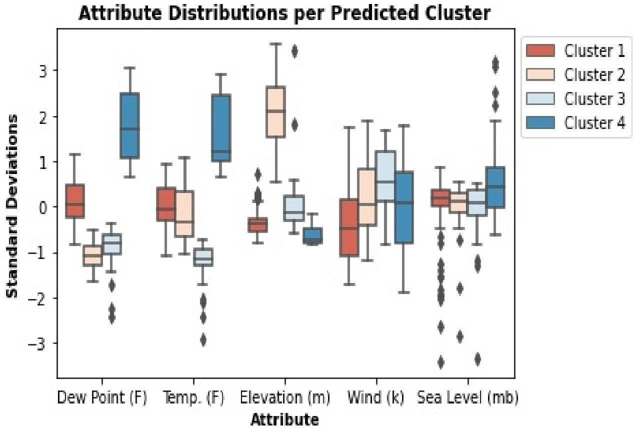

Figure 3. Mapping of k-means predicted clusters to United States.