Snowfall Data Analysis and Visualization

CS226 BIG-DATA MANAGEMENT - Fall 2022

Group Name : Earth.ai
Group No. : 10

Sanjana Senthilkumar
Ritesh Singh
Puneet Singhania
Shadhrush Swaroop
Rigved Patil

UC RIVERSIDE

1. Background

- Large amounts of weather data collected from various weather stations around the world give us information on snowfall.

- This data is critical to obtain meaningful insights that can reduce damage from freak weather events like polar vortex and snowstorms.

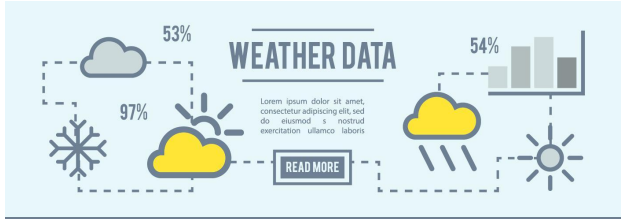


Fig 1.1 Pictorial depiction of weather data.

- Visualizing these insights is essential for people to interpret and take actions in response to disasters.

2. Motivation

- Snowfall and snow storms can cause loss of property and life.

- Frozen ice on the pathways and roads can lead to reduced friction between the vehicle and the road, poor visibility, and stranding of vehicles while commuting.

- An analytical insight into historical snowfall patterns can help in improving preparedness and help understand the rate of change of snowfall.

Fig 2.1 Snow mower working due to excessive snowfall

3. Problem Statement

“To help improve preparedness by visualizing the rate of change of snowfall using historical recorded snowfall data for easy interpretation.”

Data Visualization

4. Dataset

- “The Integrated Surface Dataset (ISD)” is a dataset that has surface weather observations from over **35,000 stations** around.

- The dataset is taken from the National Centres for Environmental Information, a part of the National Oceanic & Atmospheric Administration.

- It includes readings of **temperature, dew point, wind speed**, etc. We will be focussing on precipitation readings from thousands of weather stations.

NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

Fig 4.1 Logo of NOAA NCEI

Fig 4.2 Scan to view our data set in detail

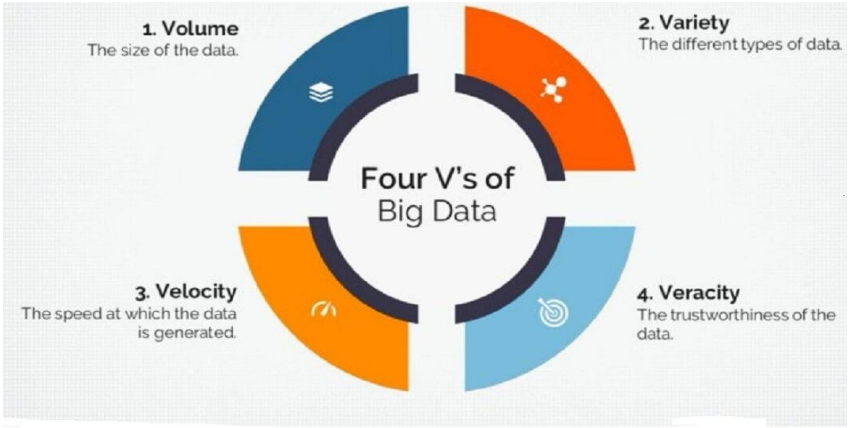
5. Data and what we did!

- Data cleaning using Python.
- Filtering out irrelevant columns in the data set.
- Replaced missing values with NaN.
- Matching station numbers and its location data on different csv files using an inner-merge.

Fig 2.1 Data preparation Steps

6. RELEVANCE TO BIG DATA?

- **Volume** - The size of datasets related to snowfall and its factors justify the big data's volume aspect.
- **Velocity** - The dataset for this problem is increasing every day, thus it justifies the velocity aspect of Big Data.
- **Veracity** - The accuracy of the data is guaranteed as we are taking the datasets from trusted sources.
- **Variety** - For the analysis of Snowfall at any particular location and time, we are considering different aspects such as latitude, longitude, temperature, pressure, wind speed, snow depth etc. This demonstrates variety in the data.



7. Random Forest

- One of the reason why we chose this algorithm is because it runs efficiently on large datasets

- Random Forest Produces highly accurate results because it's learning rate is fast
- Random Forest Constructs multitudes of decision trees at the training time and uses a majority vote to make a final decision

- In our dataset there are multiple features which help us in predicting the Snowfall Intensity at different time, so random forest seemed to be a appropriate algorithm to start with

8. Gradient Boost Algorithm

- Stands for Extreme Gradient Boosting. The algorithm provides parallel tree boosting because of this, its one of the leading regression and classification algorithm.
- It's a supervised learning algorithm which attempts to predict a target variable in our case (SnowFall Intensity) by combining the estimates of a set of simpler, weaker models.
- We are using an maxItr value of 10.

Fig 8.1 General Regression

9. Relationship Diagram between the categories for the related work

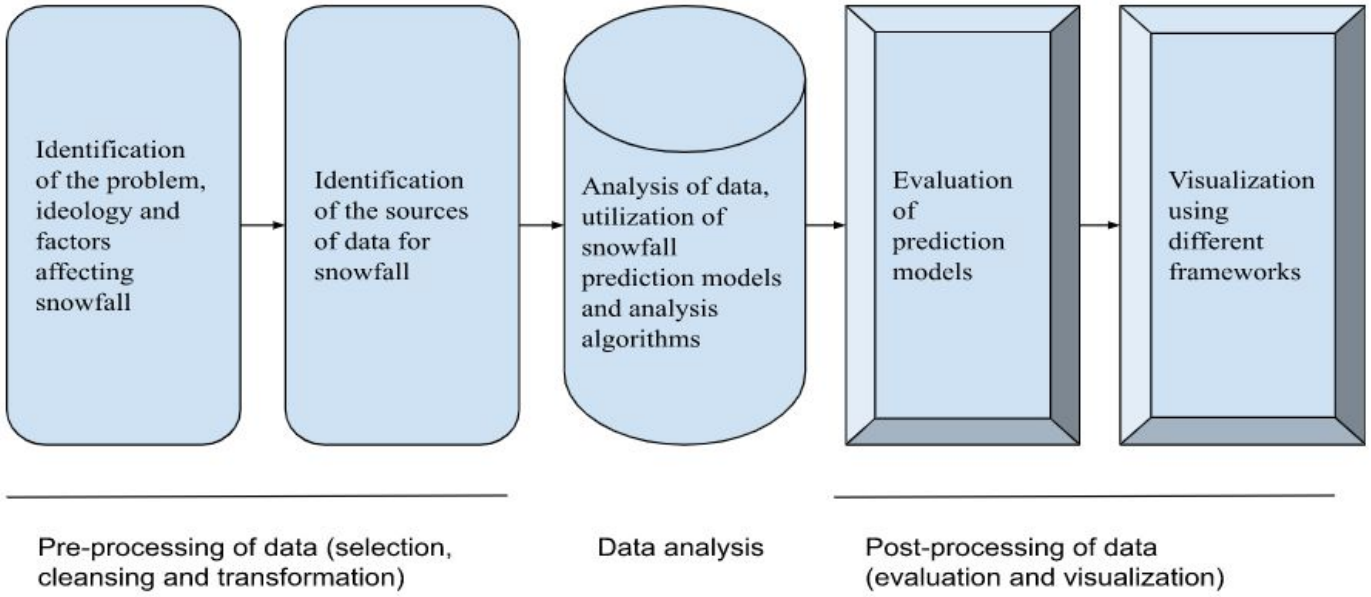


Fig 9.1 Project component overview and relationship

10. Spatial Big data

Spatial Big Data. Spatial Big Data **exceeds the capacity of commonly used spatial computing systems**.

- ● Due to volume, variety and velocity. Spatial Big Data comes from many different sources.
- ● Eg. Satellites, drones, vehicles, geosocial networking services, mobile devices, cameras.

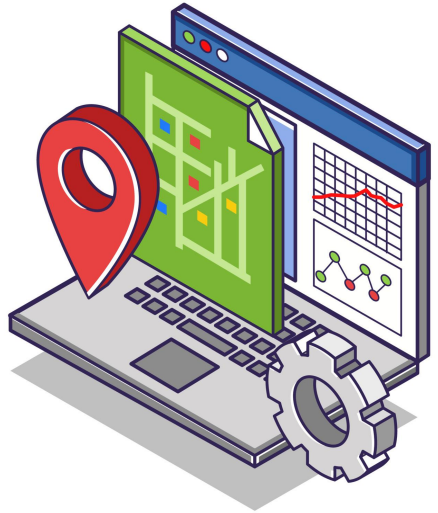


Fig 10.1 Image depiction location based spatial data

11. Evaluation

- Since we had access to a vast amount of weather information, we used this data to help our model predict the snowfall for a present date.

- Comparing this prediction to the already available data for that date, we calculated the accuracy of our prediction.

- We are going to use the following metrics to calculate the accuracy of our model:
  a. Root Mean Squared Error

12. Outcome

- Relevant data fields will be used to produce a pictorial and topographic representation of snowfall in a particular region.

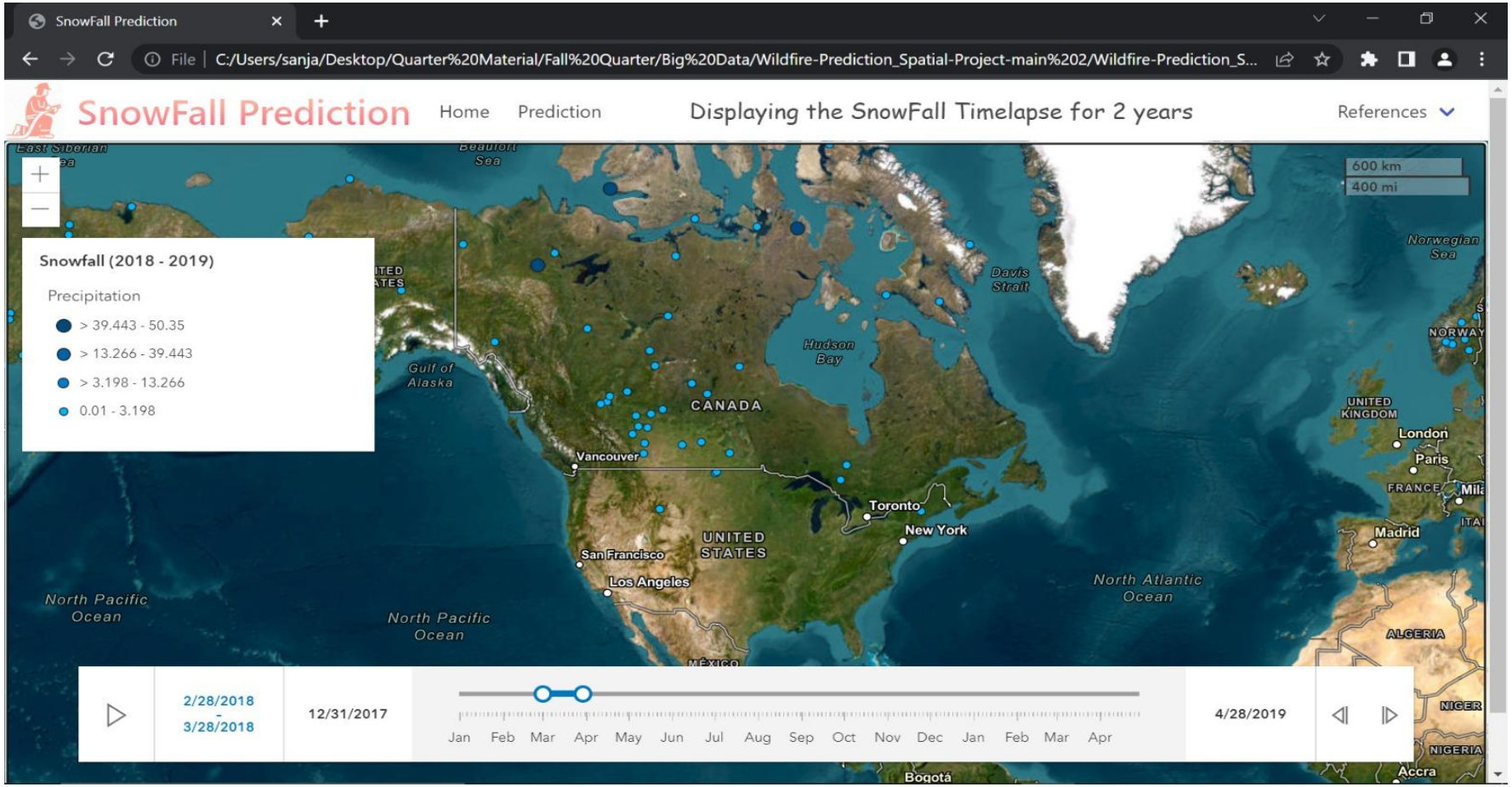
- Comparative analysis on different time frames to give insight into how snowfall patterns will change in real time with a interactive time-lapse.

- Hotspots are mapped to visualize areas with highest density of snowfall.

13. Visualization

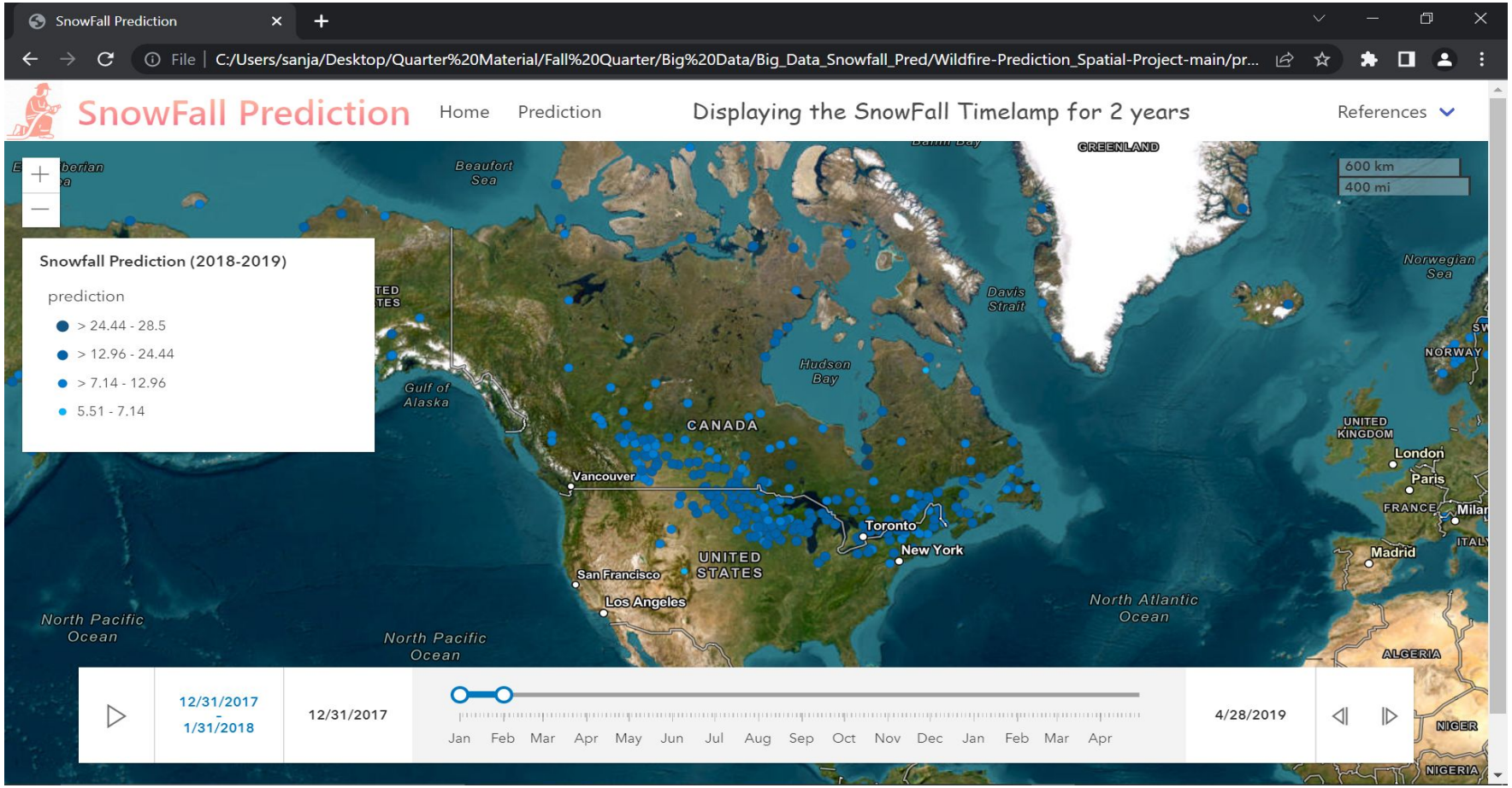
- We use ArcGIS Python API to upload the generated csv to ArcGIS online.
- We have utilized ArcGIS tool to generate the output map which shows the snowfall intensity over a period of time. We have provided an interactive time lapse for the snowfall prediction representation.
- We have used 4 different colors to display the snowfall intensity. The legend is provided in the attached image above.
- Precipitation here represents either snowfall or rainfall and if the temperature is below 22 Fahrenheit then it represents snowfall.

14. Result



UC RIVERSIDE

14. Result



UC RIVERSIDE

15. Conclusion

- The results obtained from this project were able to provide a comparative analysis of the historical patterns of snowfall data using big data technologies and we are visualizing it using ArcGIS.
- We effectively utilized big data technologies to process huge datasets and built a prediction model.
- Our hope is that this can be utilized by governments bodies to better prepare themselves for weather disasters.

16. References

Ideology and factors influencing snowfall and its patterns

[1]Daniel Eisenberg and Kenneth E. Warner. 2005. Effects of Snowfalls on Motor Vehicle Collisions, Injuries, and Fatalities. American Journal of Public Health 95, 1 (January 2005), 120–124.

Snowfall data set

[2]Jamie L. Dyer and Thomas L. Mote. 2006. Spatial variability and trends in observed snow depth over North America. Geophysical Research Letters 33, 16 (2006).

Snowfall prediction models and algorithms

[3]Moon-Soo Song, Hong-Sik Yun, Jae-Joon Lee, and Sang-Guk Yum. 2022. NHESSD - A Comparative Analysis of Machine Learning Algorithms for Snowfall Prediction Models in South Korea. *NHESSD - A Comparative Analysis of Machine Learning Algorithms for Snowfall Prediction Models in South Korea*. Retrieved from https://nhess.copernicus.org/preprints/nhess-2022-118/

Evaluation of ML models

[4]2017. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives - ScienceDirect*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0957417417303457?fr=RR-2&ref=pdf_download&rr=768ae2de9c6e3149

Visualization techniques used for our ML models

[5]Ajibade, Samuel & Adediran, Anthonia. (2016). An Overview of Big Data Visualization Techniques in Data Mining. 4. 105-113.

Thank You

UC RIVERSIDE