MIT 805 Assignment 1

15091644 Jesse Parvess

Data Set

Weather measurements from sensors situated across the city of Sofia, the capital city of Bulgaria, were collected. This data set was collected by Luftdaten to generate a more representative set of weather measurements for the city.

Technical Aspects of Data Set

The data set contains one year's worth of weather information, collected from 2017-07-01 to 2018-08-01. Hence, the data on hand is approximately 4-3 years old.

The data consists of 9 columns with the following field names, data types and explanation:

Table 1: Sofia weather data set, field information.

Name	Data Type	Explanation
Unnamed: 0	Integer	Ordered event identifier
senor_id	Integer	Identifier of the weather sensor
location	Integer	Identifier of the sensor's location
lat	Float	Latitude of the sensor
lon	Float	Longitude of the sensor
timestamp	String	Datetime of recording
pressure	Float	Pressure recording from sensor (Pa)
temperature	Float	Temperature recording of sensor $(°C)$
humidity	Float	Humidity recording of sensor $(\%)$

When viewing the first few records the data is as follows:

	Unnamed: 0	sensor_id	location	lat	lon	timestamp	pressure	temperature	humidity
0	3	3642	1837	42.694	23.360	2017-08-01T00:00:02	95565.32	20.57	59.72
1	4	3102	1561	42.665	23.392	2017-08-01T00:00:05	95804.66	20.91	56.63
2	6	4475	2250	42.654	23.316	2017-08-01T00:00:06	95105.52	21.16	56.28

The data is currently stored in csv formatting and consists of 3.43 GB of information.

The 4 V's

Insights gained in the section, were obtained by taking a subset of the data (for example one month's worth of information) and validating the consistency of the finding on another randomly selected period in the set. This was done because processing the entire set at once, was too computationally expensive.

Veracity

The age of the data potentially makes it unrepresentative for use in 2021 and beyond. This assertion is only valid if the local weather conditions of the city have changed drastically. This would occur if the area has seen shifts in weather patterns from 2018. Therefore, using this information to infer assumptions about current weather conditions or making future weather predictions for the city (using this data) needs extreme caution.

However, for the data collected, one can conclude that it is representative of the weather for the period because:

1. 56 sensors were used across the entire city.
2. Sensors in very different locations make readings more representative of the entire area.
3. Multiple sensors occurred at some locations (hence one sensor can be used to validate the readings of those in its proximity).
4. The data contains one year's worth of information (encapsulating changes in season).

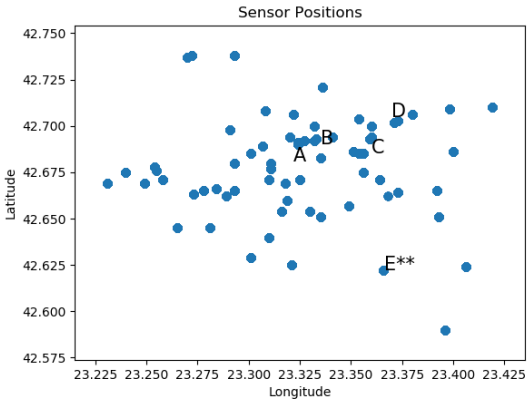

Figure 1: Sensor locations across the city of Sofia.

In figure 1, locations with multiple sensors were labelled A, B, C, D. Label E** indicates a non-proximal sensor, that was used for comparative purposes. Thus, when comparing humidity readings from the sensors at location A to location E** over the same period (that was randomly selected):

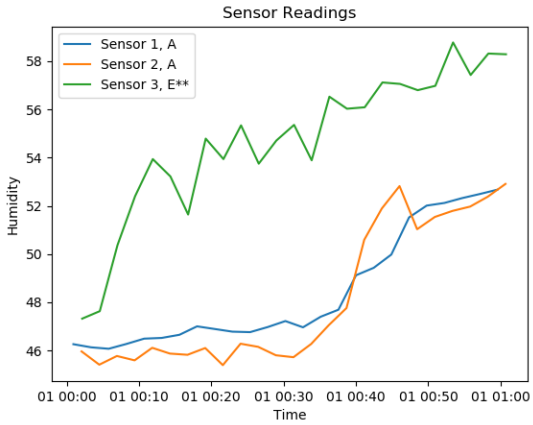

Figure 2: Humidity readings of sensors at location A and E**.

Figure 2 indicates, that as expected, sensors at the same location read similar values, when compared to sensors in completely different locations. Sensors in vastly different locations make a weather reading more representative. Likewise, sensors at the same location validate the readings of sensors in their proximity. If there is a severe disparity on a set of proximal sensors, one can distrust the readings coming from that set.

Furthermore, a random day was chosen and past weather information from online [1] was used to validate the data further:

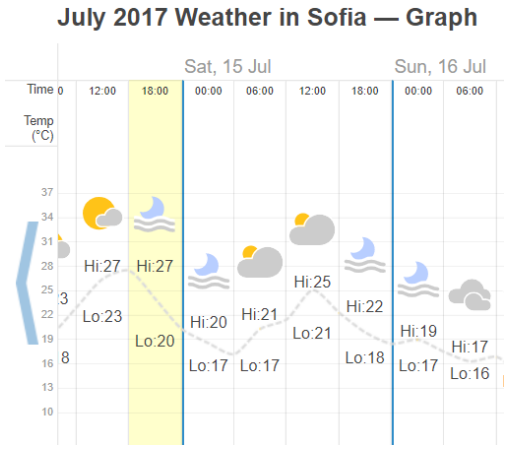

Figure 3: Weather readings for a randomly selected day, collected online.

Figure 3 indicates that July 15th, 2017, was used as a reference day for validation. Figure 3 shows that an average high of 23 $°C$ and average low of 17$°C$ occurred on the day. When comparing the distribution of temperature readings from the data on this day to figure 3:

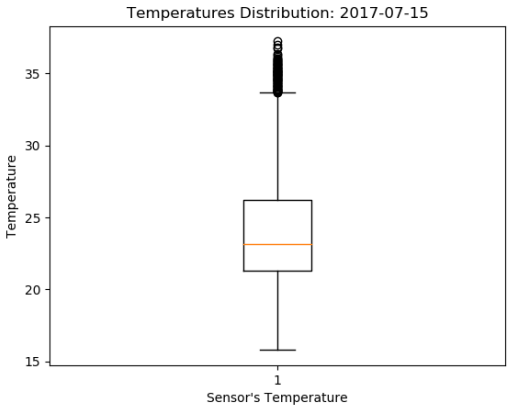

Figure 4: Temperature distribution from sensors on the 2017-07-15.

Figure 4 indicates that the mean temperature was approximately 23$°C$ with a minimum around 15$°C$. This closely corresponds to the weather recorded in figure 3.

From the above discussion we can conclude that the weather data collected from the sensors is representative and accurate for the period collected.

Velocity

To understand how often data is updated, the time difference between records per sensor was found. The following summarises the distribution of time intervals:

Table 2: Distribution of time between recordings per sensor.

Description	Value (Seconds)
Minimum	1.0
Maximum	1315841 (15 days)
Median	147
Mean	169.3
25th Percentile	147
75th Percentile	147

Table 2 indicates that most of the data is recorded every 147 seconds. The maximum value was investigated, and it was found the delay in recording was due to a dead sensor.

For 56 sensors, recording every 147 seconds, one would get roughly 33 000 records per day (if recorded at the same time). When looking at the data, an average of 21 000 records per day occurred (this is because of disparate recording times between sensors). Hence, 33 000 records per day represents a theoretical maximum load of records per day.

If this application were to be extended to a live scenario: The interval of recording makes detection of changes in weather conditions effortless, as weather does not perceptibly change over such a short period. Hence, information about current local weather conditions will always be reasonably up to date.

Variety

The data does not consist of a great variety of different types of data.

However, variety occurs in the type of measurements (temperature, humidity, and pressure) and the time it was collected (time of day and year). Variety is also introduced from the different locations and sensors used to collect the data.

Volume

3.43 GB of information collected over a period of 1 year, does not constitute massive quantities of data. It is not believed that this would not change drastically as sensors and different measurements are added to the system. Therefore, per year, it is estimated that approximately a maximum of 10 GB of information would be collected if additions to sensors and measurement types are made.

Predicted Corelations

Since weather events affect large areas, it is expected that readings from proximal sensors will be closely correlated. Additionally, even sensors that are not closely located should have a loose corelation in the temporal movements of recordings (i.e., if the temperatures increase in one part of the city over a period, it can be expected that temperatures should also rise in other parts over the same period, as weather events are delocalised).

Additionally, it is expected that there would be temporal correlations within a field. For instance, the temperature recorded 30 minutes ago would be a good approximation for the current temperature reading.

Furthermore, there would also be correlations between fields. For instance, a particular humidity level may be dictated directly by the associated temperature and pressure.

Conclusion

Weather measurements for Sofia from 2017-07-01 to 2018-08-01 represents a veracious set of data, that has a velocity frequent enough to make valuable weather updates. Additionally, it has some variety in measurement and a reasonably low volume.

Website Used

[1] https://www.timeanddate.com/weather/bulgaria/sofia/historic?month=7&year=2017