Assignment 2

Part 2: Visualisation Discussion

Introduction

Weather data for the town of Sofia was collected. The data set consists of the following:

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 $(\%)$

Here, for practical reasons, the temperature measurement is used to extract useful visualisations. The approaches discussed can easily be expanded to the humidity and pressure measurements.

It is assumed that these sensors were placed around the city by scientists who wish to monitor the local weather differences across the city. To ensure effective monitoring occurs, measurements are checked once per day to gain insights and detect issues.

These insights are the maximum, minimum, and mean registered temperature per sensor per day. This will give a detailed understanding of the daily temperature registered by the sensors. Additionally, because there are multiple sensors spread out spatially, understanding these measurements with respect to the locations of the sensors is important. Furthermore, it is important to check the validity of these measurements. Thus, visualisations that check the measurements of the closest sensors with respect to a reference sensor of interest is important. This is because if sensor is reading vastly different readings to its closest neighbours, it indicates a potential issue on that sensor (especially if its neighbours read similar readings). Furthermore, since data is registered every few seconds, understanding the progression of a measurement over time can allow for temporal patterns to be extracted for that day. This can be overlayed with the sensor's closest neighbours to gain insight about the temporal trend of a measurement across a local region.

Methodology

To simulate the above scenario a random day was chosen from the data set to develop various daily visualisations. Here the day of 2018-05-01 was considered.

Spatial Visualisations

The mean, maximum, and minimum temperatures on the day for each sensor was determined and overlayed on the positional data given (latitude and longitude):

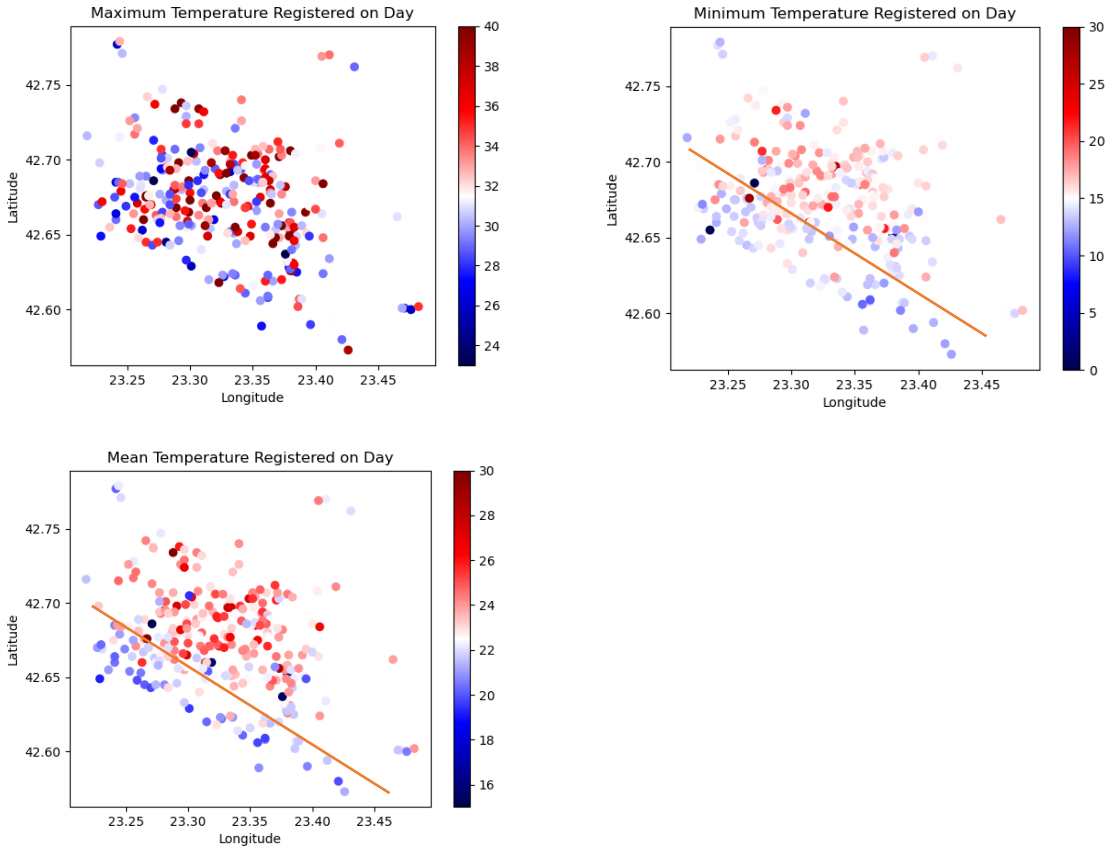


Figure 1: Maximum, Minimum and Mean Temperature Readings per Sensor on Day

Figure 1 indicates a clear difference between sensors located in the north of the city compared to the south of the city. Clearly the north-east of Sofia is on average warmer than its south-west region.

Looking at the maximum temperature scatter plot, it indicates that there are more uniform maximum readings across the city. Maximums on average were around $35°C$. Some maximum readings exceed $38°C$ and should be investigated as outliers.

Looking at the minimum temperature scatter plot, it indicates that there are less uniform minimum readings across the city (a clear north-westerly divide, shown with an orange line). Minimums on average were around $15°C$. Some minimum readings were as low as $5°C$ and should be investigated as outliers.

Looking at the average temperature scatter plot, it indicates that there are less uniform mean readings across the city (a clear north-westerly divide, shown with an orange line). Means on average were around $25°C$for the day. The clear spatial divide is confirmed by the mean temperature because it summarises a range of temperature values (and is therefore more temporally robust). This stands in contrast to the minimum temperature reading which may occur at a certain time of day when this pattern is prominent (say early morning)

An explanation for the higher northern temperatures could be due to the topography of the region:

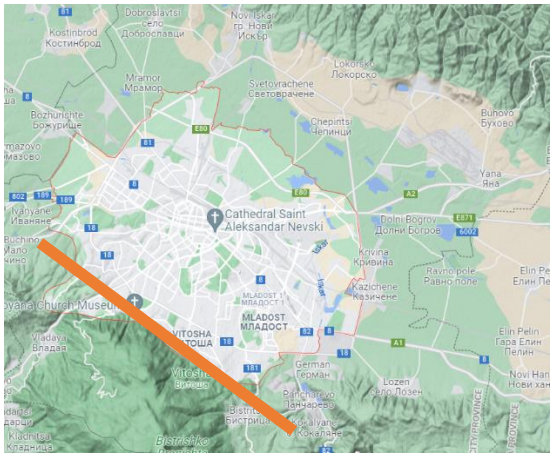


Figure 2: Topographical map of Sofia, Bulgaria

Figure 2 indicates an orange line showing the prominence of a mountain that runs in a north-westerly direction. Compare this to the lines in figure 1. Clearly the prominence of the mountain influences the temperature measurements. Here the mountain could influence wind flows and heating patterns that cause lower temperature measurements in sensors along this direction in the city.

Outlier Detection

To detect outliers among sensors a heatmap can be used to compare temperature readings across a single day and across time (across days) to detect sudden sensor degradation.

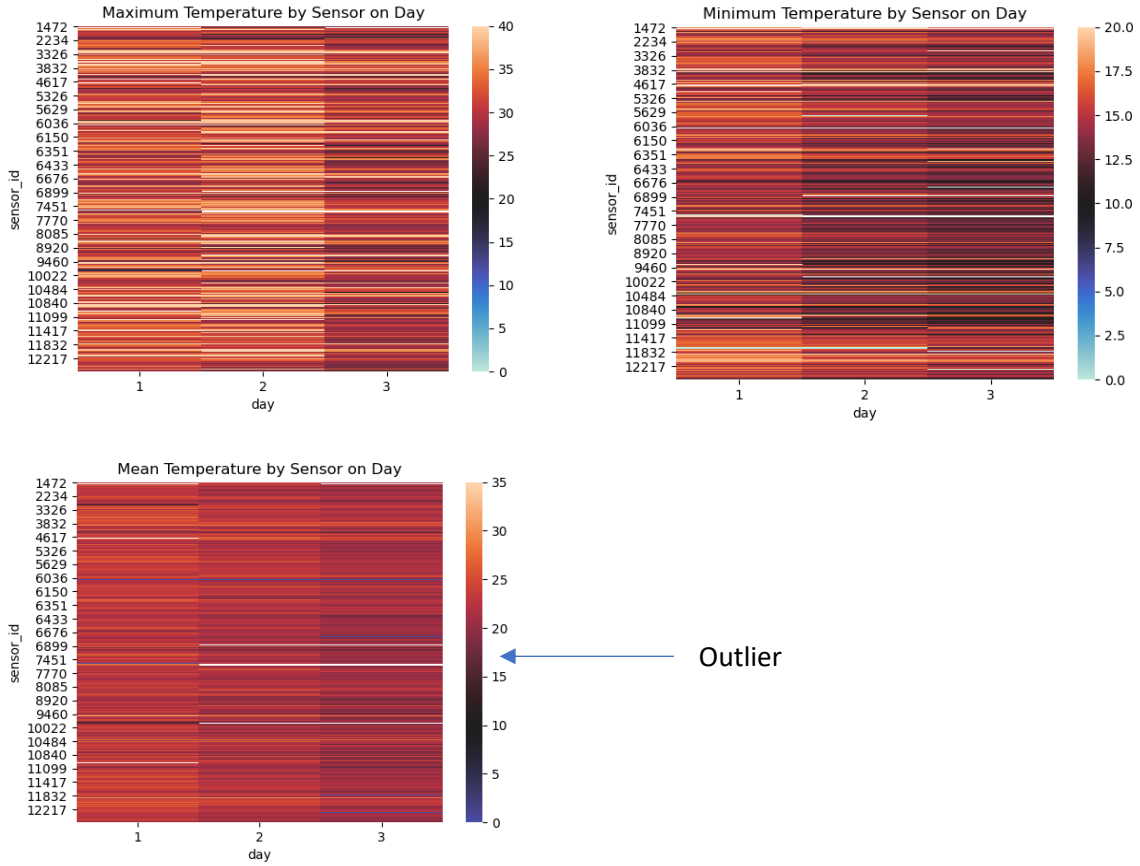


Figure 3: Temperature measurement heatmaps for 3 days

Figure 3 provides easy interpretation of the state of sensors in the system. For instance, what most readings are on a day. One can detect, in general, that from the 1st to the 3rd of May 2018 the temperature decreased (prominence of shaded blue readings). Additionally, by looking at the mean temperature heatmap, sensor 7451 suddenly started reading high values when clearly most temperature readings decreased. This potentially indicates an outlier (indicated by an arrow) and a faulty sensor.

To incorporate spatial information, the distribution of temperature readings of the closest sensors to a sensor of interest can be plotted. Vastly different temperature measurements of the reference sensor to its closest neighbours potentially indicates a sensor issue. Here three sensors (IDs: 1472, 1954, 1962) were chosen to demonstrate the algorithm:

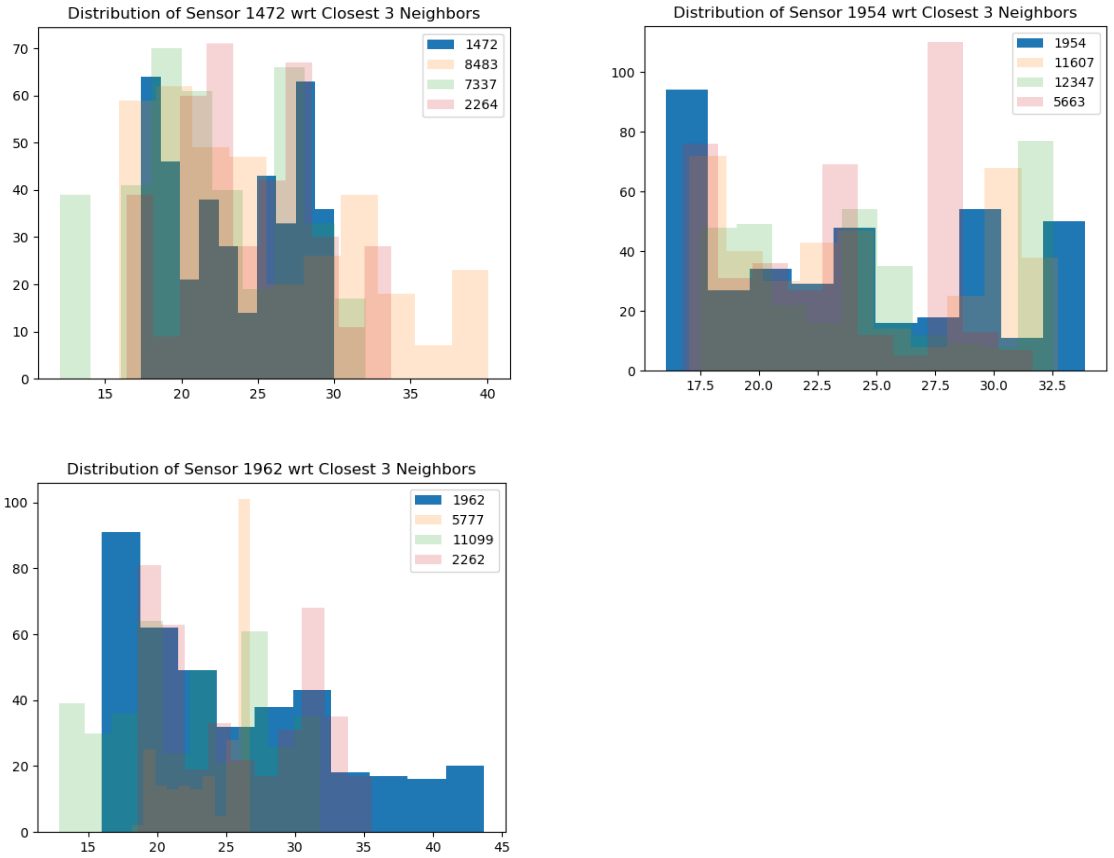


Figure 4: Distribution of a reference sensor's (blue) temperature measurements for a single day compared to its three closest neighbours (orange, pink and green)

Figure 4 indicates that all reference sensors (in blue) have similar distributions to the three closest neighbours to that sensor. There are some differences in measurement (at the extremes of measurement). This approach can be used to quickly ascertain if a sensor's measurements can be trusted by looking at the distribution of its neighbours.

If we compare the same sensors to their **furthest** neighbours, we find:

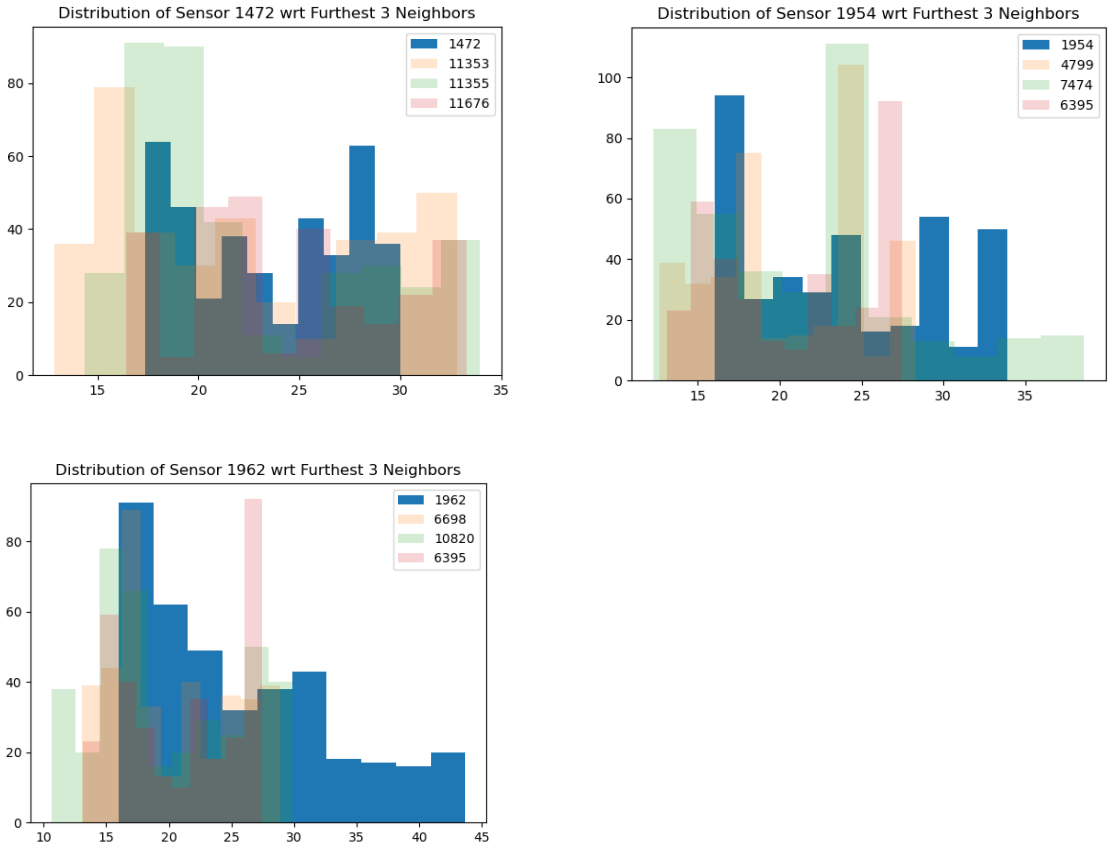


Figure 5: Distribution of a reference sensor's (blue) temperature measurements for a single day compared to its three furthest neighbours (orange, pink and green)

Figure 5 indicates that, when considering the reference sensor to its furthest sensor, very different distributions can be detected (see subfigure 3). This proves that the above method can be effectively used to tack outliers and sensor degradation when considering the reference sensor's closest neighbours (figure 4). This is because the furthest sensors should have different temperature distribution profiles.

Temporal Considerations

Since the data is updated every 20 minutes, and because weather changes with time, tracking the temporal movement of measurements is important to monitoring. Here, the above outlier approach can be employed where the closest sensors to a reference sensor are overlayed temporally. This allows for faulty measurements to be detected (if the reference sensor does follow similar measurements to its closest neighbours) and for the general temperature trend in a locality to be validated by several close sensors.

Here the reference sensor is plotted over time in blue, and its closest neighbours are overlayed over the same period of time:

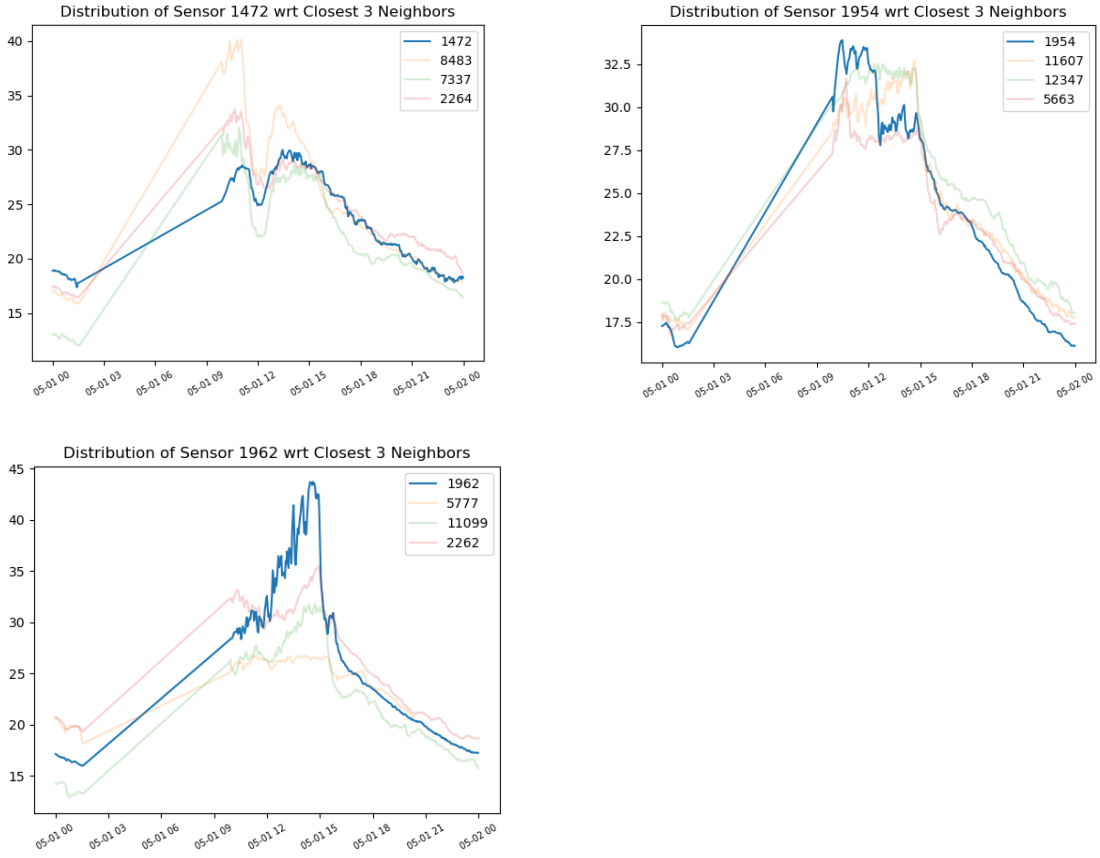
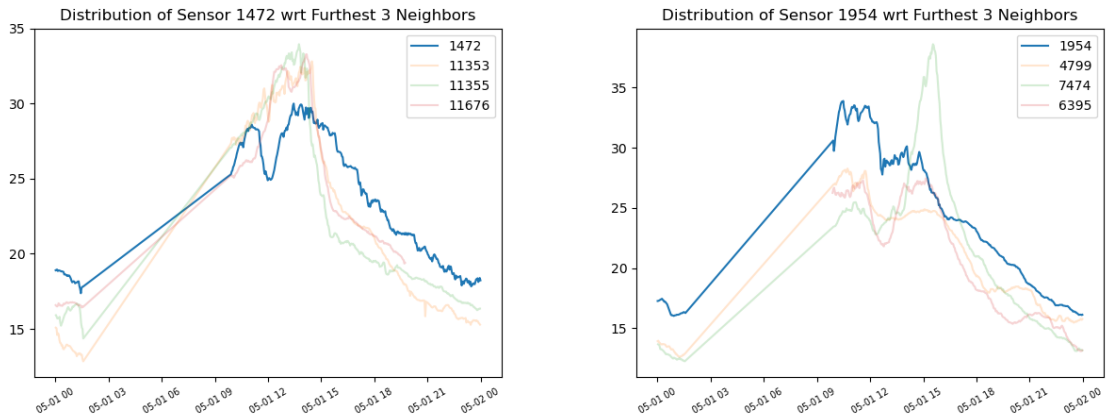


Figure 6: Distribution of a reference sensor's (blue) temperature measurements for a single day compared to its three closest neighbours (orange, pink and green) over time.

Figure 6 indicates that all the closest sensors to a reference sensor follow a similar trend (if not in magnitude then in general profile). All three subfigures represent the same period of time, and thus show the variability of measurement profiles across localities.

Additionally, a period (indicated by a straight gradient from one measurement to the next) indicates that these sensors did not register information and were off. This indicates another form of error detection in the system.

If we compare the same sensors to their **furthest** neighbours, we find:



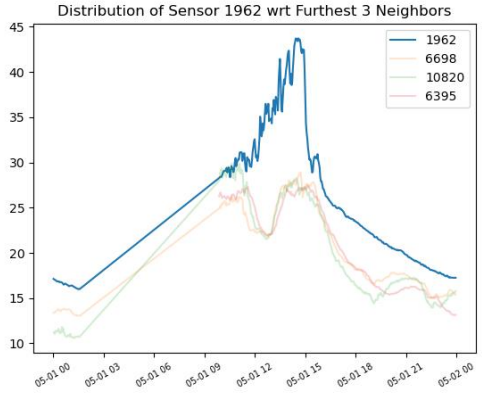


Figure 7: Distribution of a reference sensor's (blue) temperature measurements for a single day compared to its three furthest neighbours (orange, pink and green) over time.

Figure 7 indicates that all the furthest sensors to a reference sensor follow a different trend (in magnitude and in general profile). This indicates that above method is an effective means to detect sensor degradation or faulty readings.

Conclusion

The above approach only considered a single parameter (temperature), this can be expanded to the humidity and pressure measurements (other **varieties** of data in the set). The approach was only focused on daily tracking, such that a user could find:

1. Spatial differences in measurements (the **volume** of data is summarised to daily statistics)
2. Outliers in sensors (the **veracity** of the data is measured in the visualisations)
3. Change in time in sensor measurements (the **velocity** and change in time is accounted for)

The approach was heavily focused on outlier detection and sensor degradation since the purpose of these sensors is for scientific study. Thus, the researchers could know when to service the system and the veracity of the data they collected can be upheld.

To improve the approach, one can summarise the data over larger periods of time (like for a month in a year) and gather summary statistics.