## Chapter2 Data Availability

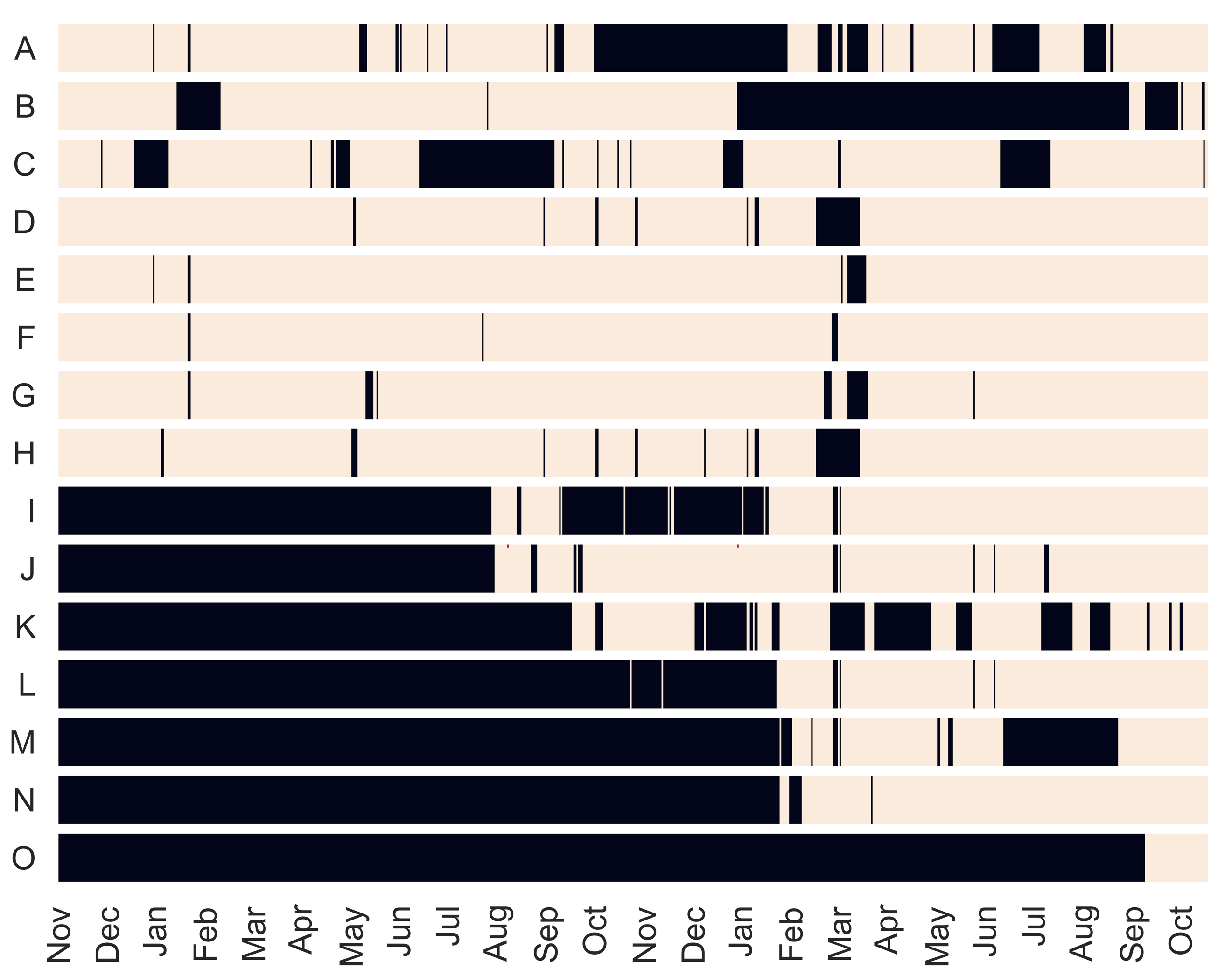
The data collected from these deployments were analysed to determine the variety of the sensors supported and the devices, the velocity of data arriving at the cloud infrastructure, as well as the variability of the data collected.

During this initial high-level analysis, it became apparent that the data collected from the IoT devices was not always delivered properly to the cloud.

As a first step towards understanding the availability of measurements:

Fig. X and Fig. X are included that depicts the availability of measurements on a daily basis for each site and each type separately.

In Fig. *X for site* each different device corresponds to a specific horizontal line, organized based on the site of deployment.

**

A single dark point signifies missing data for the specific on a specific day, while a white part signifies complete availability (i.e., according to the sensing rate of the specific sensor, see Tab. I). For those points remaining long-term dark from beginning, it is due to GAIA platform deployment not been settled at that period.

This visualization clearly depicts that the stability of the IoT deployment. In almost all deployments there are values missing almost on a daily level.

Essentially these measurements are missing either because they were never reported to the cloud infrastructure or due to a failure occurring while they were processed and stored by the cloud services.

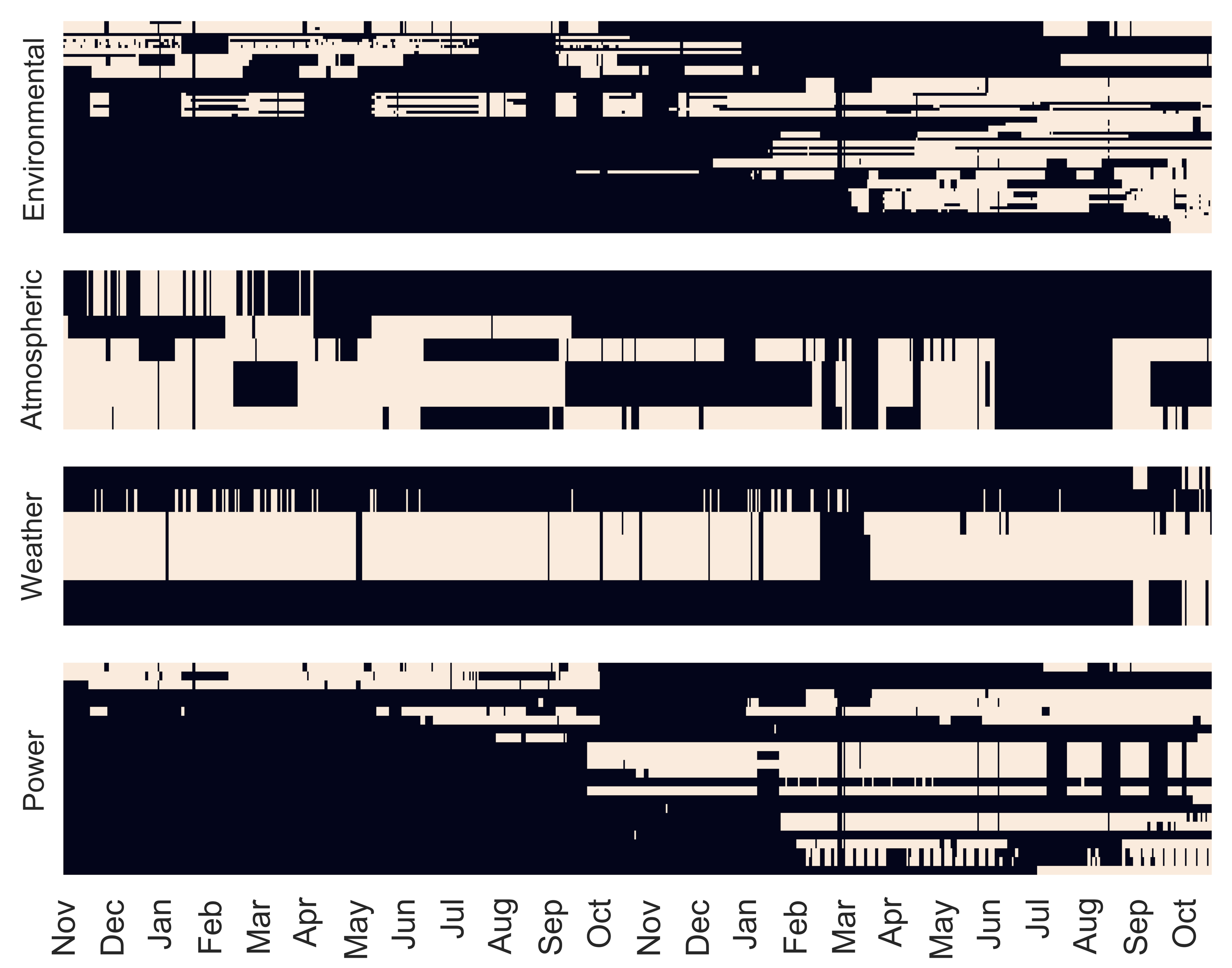
In the first case, network failures occur either due to a packet transmission error at the wireless network level (i.e., IEEE 802.15.4 or Wi-Fi), or due to a transmission error while an intermediate gateway transmitted the data to the cloud infrastructure over the Internet.

Since the access to the measurements are done through the GAIA platform API, the reason for the missing information is completely unknown. However, as it will become evident in the following sections, specific data mining techniques can be used to overcome the problem of missing values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site ID | Alias | device number | resources number (without SITE) | start time | Outrage | Outlier |
| 144024 | A | 8 | 43 | 2015-OCT | 29.78% | 2.67% |
| 28843 | B | 9 | 56 | 2015-OCT | 41.26% | 2.84% |
| 144243 | C | 6 | 32 | 2015-OCT | 23.63% | 2.82% |
| 28850 | D | 9 | 50 | 2015-OCT | 5.33% | 1.45% |
| 144242 | E | 7 | 43 | 2015-OCT | 2.19% | 1.63% |
| 19640 | F | 11 | 54 | 2015-OCT | 0.96% | 3.36% |
| 27827 | G | 7 | 45 | 2015-OCT | 3.69% | 1.33% |
| 155849 | H | 6 | 27 | 2015-OCT | 6.01% | 1.31% |
| 155851 | I | 7 | 36 | 2016-SEP | 29.23% | 1.68% |
| 155076 | J | 34 | 103 | 2016-APR | 3.97% | 3.19% |
| 155865 | K | 5 | 26 | 2016-SEP | 37.87% | 1.70% |
| 155077 | L | 5 | 109 | 2016-OCT | 26.43% | 1.22% |
| 155877 | M | 5 | 24 | 2017-FEB | 33.09% | 2.67% |
| 157185 | N | 12 | 55 | 2017-FEB | 3.31% | 5.09% |
| 159705 | O | 4 | 22 | 2017-SEP | 0.00% | 13.22% |

*In Table II the different school sites are summarized in dicating the time when they were incorporated in the GAIA platform. For each school building the number of points of sensing (POS) are listed along with the total number of sensors deployed. The table reports the percentage of outages recorded for the particular site (reflecting the periods during which no measurements were received from the site, as a percentage from the point when the site was first incorporated in the GAIA platform) along with the total number of measurements received from this site is listed along with the percentage of values that have been identified as outliers. Remark that to avoid issues related the confidentiality of private data, the names of the school buildings have been ommitted.* The analysis reveals that certain buildings experience very often data outages.

At a second level the same analysis of data is repeated based on the device type of the sensor.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | POS | Sensor | Inactive | Outlier |
| Env | 101 | 505 | 14.62% | 7.76% |
| Libelium for outdoor weather | 7 | 56 | 19.56% | 6.29% |
| Synfield for outdoor weather | 7 | 28 | 20.25% | 0.95% |
| Electrical Power Consumption | 20 | 56 | 12.55% | 4.17% |

The sensors deployed in the GAIA platform are organized in four main categories:

(1) *classroom environmental comfort sensors*: sensors positioned within classrooms;

(2) *atmospheric sensors*: sensors positioned outdoors;

(3) *weather stations*: sensors positioned on rooftops of buildings;

(4) *power consumption meters*: sensors attached to the main breakout box of the buildings that measure energy consumption.

*For each device category, the percentage of outages and the percentage of outliers observed are reported in Table III.* In Fig. 3 the representation of the availability of data is depicted on a daily basis for each sensor separately organized by sensor category. Based on this visualization one observes that all sensors experience period loss of data. This may be justified by the wireless networking technology used to interconnect the sensors located in the classrooms as reported in [7]. Apparently the low-power, lossy nature of the networking technologies used result to a non-negligible data lost.

A second step is to examine the actual values received from the IoT devices.

It is very common in the relevant literature to deploy relatively low-cost devices that produce low-quality measurements or are not properly calibrated.

For this reason, we examined the values to identify possible outliers, that is observation points that are distant from the historic values.

Such observations may be due to transient errors occurring on the sensing equipment and should be excluded from the data set. The identification of outliers is based on the *interquartile range* (IQR) using the upper and lower quartiles Q3 (75th percentile) and Q1 (25th percentile).

The lower boundary is set to and the upper bounder is set to where The values are examined per sensor/site basis using a timed- window of size W. If a value is outside the boundaries [,] it is flagged as an outlier.

In the following sections, we replace it with the history minimum, or maximum or average value observed during the time window W.

After examining the values characterized as outliers two distinct cases were identified: (a) 0 values which were clearly sensor error rather than natural events (e.g. humidity of 0%, temperature dropping from ⇠ 20 to 0), and (b) drastic changes of power consumption (i.e., spikes or fast drops) that could not be justified by the daily school activities.

To overcome the fact that the IoT deployment (1) experiences outages on a regular basis and (2) at a significantly lower rate, sensors report values characterized as outliers a moving window average technique is used. The moving window all to (1) smooths out short-term fluctuations for the case of outliers and (2) fills-in missing values using a simple local algorithm that introduces historic values to fill in the missing data for the specific time period. The moving window average also helps to highlight longer-term trends on the sensor values. Fig. 4 depicts an example of the analysis conducted over a specific temperature sensor located in a classroom.

