LoRaWAN Data Analysis

Smart Environments Final Project

Simone Zagaria
Data Science
2145389

Alessio Lani
Data Science
1857003

***Abstract*—This project presents an in-depth analysis of LoRaWAN packets collected by a gateway located in Sapienza. Over 5 million uplink and around 500 downlink messages were examined. The analysis revealed a significant presence of "garbage packets" and anomalous devices that disproportionately contribute to network traffic. Clustering techniques, mobility estimation, and distance calculations were applied, and with a comparison between uplink and downlink data. Results indicate that most devices are fixed, Class A devices operating under weak but stable signal conditions.**

I. Introduction

The aim of our project is to make a deep and meaningful analysis on the LoRaWAN packets collected by our gateway located in Sapienza.

We chose this project because we want to understand how devices behave within a smart environment network, what kinds of packets are collected by our gateway, and how these devices interact with the network.

II. The Dataset

Our Datasets consisted of 2 files containing raw LoRaWAN packets, in CSV format, one for **uplink** and one for **downlink** messages. Each row represents a message received by a gateway (or sent by the gateway, in the case of downlink). The Uplink dataset is way larger than the downlink, containing more than *5 milion* rows, while the downlink contains only around 500. After decoding the packets, both contained more than 20 fields, the most important include:

- **tmst**: Arrival timestamp of the packet in seconds since the Unix epoch.
- **freq**: Frequency (in MHz) at which the packet was received.
- **datr**: Data rate, encodes both Spread Factor (SF) and Bandwidth.
- **rssi**: Received signal strength indicator (in dBm).
- **lsnr**: Signal-to-noise ratio (in dB). Provides insight into signal quality.
- **size**: Payload size (in bytes).
- **data**:
  - **DevAddr**: Unique device address within the network.
  - **FCtrl**: Frame control flags (ex. ACK).
  - **MType**: Indicates the message type (ex. JoinRequest, UnconfirmedDataUp)
  - **FPort**: Port field; 0 indicates MAC commands, 1-223 for application-data.
  - **FRMPayload**: Frame payload containing application data or MAC commands.

These fields, except for rssi and lsnr, which are contained only in the uplink, are common to both the datasets, and will be confronted with later analyses.

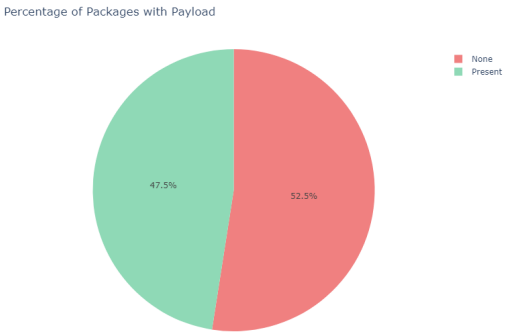
A. Preprocessing & Data Cleaning

Before conducting our analyses, we first preprocessed the datasets to remove columns that were either not relevant, contained only NaN values, had constant values across all rows, or were replaced by reformatted versions, in order to reduce memory usage and avoid cluttering the data. In particular, the columns 'time', 'tmms', 'chan', 'rfch', 'stat', 'codr' were dropped for the uplink dataset, and 'imme', 'tmms', 'rfch', 'rfch', 'fdev', 'prea' for the downlink.
Next, we converted the 'tmst' column from Unix time to a datetime format. The resulting timestamps revealed that the data spans from **17 October 2024** to **15 May 2025**, covering a total of **210 days**. Furthermore, we parsed the 'datr' field, containing the bandwidth (BW) and the spreading factor (SF) as a string like "SF7BW125" was formatted with the help of the string extract feature. Finally, as previously mentioned, a base64 LoRaWAN packet decoder was used to decode the 'data' field. This resulted in having available a number of valuable fields for analysis, including 'MType', 'DevAddr', and 'FPort'. However, we were not able to decode the payload content itself due to the lack of encryption keys.

III. Results

A. Uplink

When first looking at the cleaned dataset for the uplink, we immediately noticed that many of the payloads were missing. This suggested that the packets, although received on LoRa frequencies, were not actual LoRa transmissions. These so-called "garbage packets", as shown in the graph below



constitute a significant portion of the dataset, reaching up to 52% of the messages collected. This could indicate the presence of numerous devices near the gateway that are transmitting on similar frequencies, effectively polluting or cluttering the dataset. Such a scenario is plausible, especially considering that the gateway is located in a densely populated area within a highly urbanized city.

Next, we want to assess how many unique devices were responsible for sending these messages, in order to better quantify their impact on the dataset. According to the analysis, it turned out that

TABLE I
DEVICE TYPE COUNTS

Device Type	Count
Devices with only payload	150550
Devices with only no-payload	16333
Devices with mixed traffic	101
Total unique devices	**166984**

the number of devices sending only garbage packets was significantly lower than the number of actual LoRa devices, representing only 8% of the total devices. However, these devices appeared to transmit at a much higher rate, meaning that, despite being fewer, they **contribute disproportionately** to the overall noise in the data. This suggests that, if their number were to grow, they could pose a greater issue than initially expected.

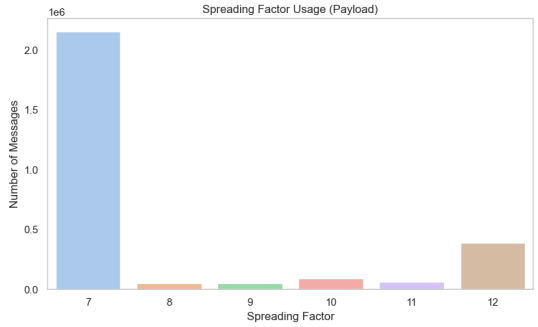
Next, made various analyses regarding the packets with payload. These were compared to the garbage packets to assess the possible similarities and differences between the two, but the main analyses focused on the payload packets. The maximum and minimum frequencies of the ”payload packets” are:

TABLE II
MEASURED FREQUENCIES

Measured Frequencies (MHz)	
Maximum frequency	867.1
Minimum frequency	868.5

which are considered normal frequencies for LoRa transmission, given the fact that LoRa typically transmits from frequency 863 MHz to 870 MHz. The garbage packets also transmitted at similar frequencies, and that could be the reason why they were intercepted by the gateway.

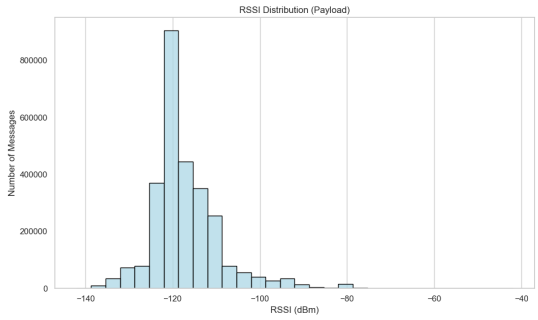
We then examined the distribution of the **Spreading factor** and **RSSI** *per message*, which could help identify the quality of the transmissions.



The results clearly show that Spreading Factor 7 (SF7) is used in the vast majority of transmissions, while higher values, such as SF8 through SF12, appear only rarely. This pattern suggests that most devices are communicating over *short distances* and maintain a *strong, stable connection with the gateway*. Interestingly, a similar distribution is observed among the garbage packets, and further accentuates the already unbalanced trend, with virtually all transmissions using SF7, and higher SF values being almost entirely absent. This suggests that these transmissions also originate from areas with good signal conditions, likely close to the gateway.

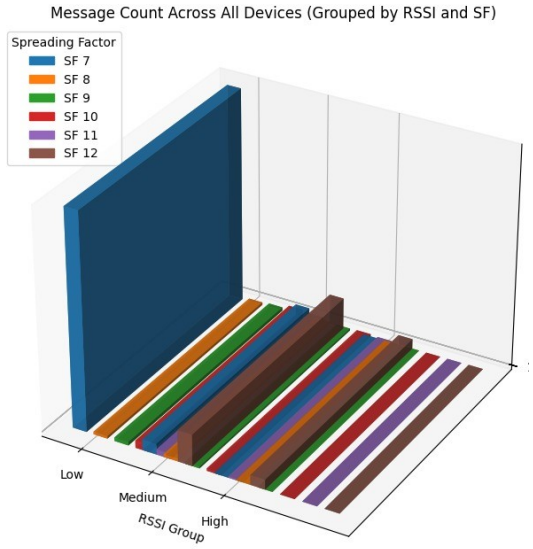
As for the **Bandwidth**, it’s always set to 125 kHz, which is the default and most commonly used setting in LoRaWAN networks.

We can observe from the distribution of the lsnr that



RSSI measures the strength of the received signal at the gateway or end device. In LoRa the acceptable RSSI value is -30 dBm to -120 dBm. the value of -30 dBm shows a strong signal and -120 dBm shows a weak signal. In our case, the Most RSSI values for fot the payload device are clustered around -125 dBm to -110 dBm, which is relatively weak, up to borderline poor signal. In contrast, garbage packets tend to show stronger RSSI values, with their distribution concentrated between -110 dBm and -100 dBm. This indicates that these transmissions likely originate from devices much closer to the gateway. In general, these low rssi values could indicate strong interference and obstacles between the device and the gateway.

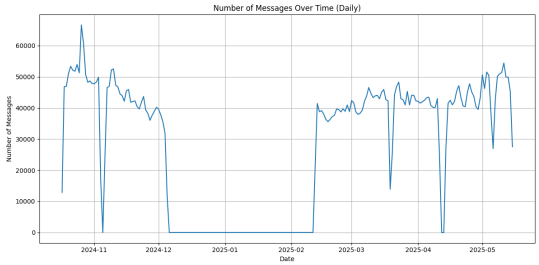
This is likely, given the densely populated and urbanized area the gateway is in.
Next, we created 3 RSSI levels, and we decided to compute the number of messages sent for each combination of spreading factor (SF) and RSSI level:



We can see that, as expected, the main SF used for the devices with the highest number of messages are 7 and in second position 12. We can also notice a trend that lower SF used lower values of RSSI and instead a higher SF tend to be associated to higher values of RSSI.

B. Temporal Analysis of Uplink

As we can see from the image there is not a clear trend, we observed there is a temporal gap in the dataset of the duration of approximately two months, from december 2024 to february 2025, where most likely the gateway was shut off or data collection was not performed. We also examined the average number of messages over the weekdays and did not detect any significant trend.



C. Clustering

In this analysis, we wanted to cluster devices with similar properties, like rssi, lsnr, size, SF and estimated distance (which will be computed in the next paragraphs); using the k-means algorithm through the elbow method we concluded that the optimal number of cluster was 4.

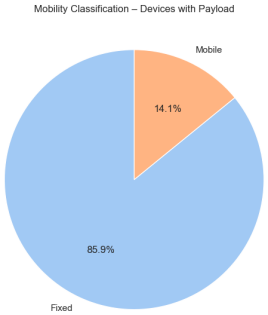
TABLE III
AVERAGE FEATURE VALUES BY CLUSTER

Cluster	RSSI	LSNR	Size	SF	Estimated Distance
0	-104.35	2.42	51.98	11.99	127.77
1	-92.57	-6.39	60.13	7.13	340.46
2	-126.69	-12.89	44.11	11.91	552.42
3	-97.53	4.85	12.57	9.98	83.50

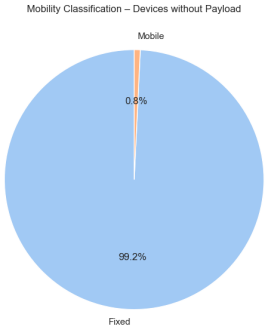
We also calculated the average number of messages per device in each cluster. The results show a significant variation: Cluster 0 has the lowest activity with just 1.22 messages per device on average, while Cluster 3 shows the highest communication frequency, with over 110 messages per device. This suggests that devices in Cluster 3 are much more active, possibly indicating different usage patterns or application types across clusters.

D. Mobility Estimation of Devices

Next, we wanted to estimate whether the devices were mobile or fixed, and then estimate their distance from the gateway. To distinguish between mobile and fixed devices in this LoRaWAN network, we analyze the behavior over time of the Received Signal Strength Indicator (RSSI). In fact, in fixed devices, RSSI usually tends to remain relatively stable over time, typically with a standard deviation of less than 2 dB. In contrast, mobile devices introduce more pronounced and irregular fluctuations, often exceeding 3 to 4 dB.
To capture these dynamics more reliably, we based our classification on deviations from the Geometric Mean rather than the arithmetic mean. This choice offers advantages such as being robust to outliers and being compatible with signal propagation. Based on this, we computed the **RSSI-GM score**, defined as the average absolute deviation from the geometric mean. We considered Mobile the devices that computed this metric resulted above a certain threshold. We also differentiated Devices that only sent meesages with payload to devices that sent only garbage packets:



We observe that only 15% of the devices that transmit payload are estimated to be mobile. This is an interesting result, as it differentiates strongly with the devices that transmitted only garbage packets, which display an extremely skewed trend:



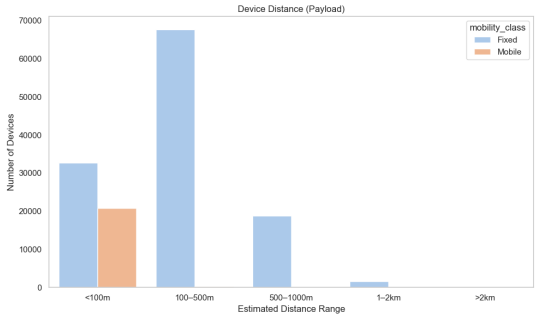
nearly all of them are classified as fixed. This results suggests that these devices exhibit very low signal fluctuation over time, likely due to their fixed position near the gateway or consistent transmission conditions. From this,we estimated the distance of the devices. To do this, we used the RSSI-based distance estimation formula:

$$d = 10^{\frac{(P-R)}{10n}} \tag{1}$$

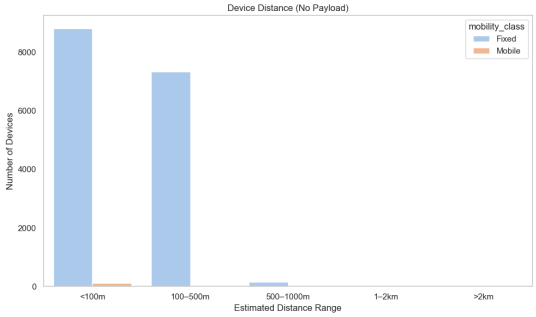
Where:

- P: RSSI at 1 meter (as a reference value)
- R: measured RSSI from the dataset
- n: path loss exponent, typically from 2 to 4 for LoRaWAN, depending on the environment type

For this dataset, we assume a reference value of $P = -40dBm$, which is common for LoRa at 1 meter with typical antennas. For n, we choose 3.2, an high path loss due to the higher urban interference, since typically in higher urbanized area values from 2.7 to 3.5 are chosen. This method provides only a rough approximation due to the influence of environmental factors such as buildings, obstacles, and interference, but it is useful to have a general idea of the patterns of distance within the datasets.
It resulted in:



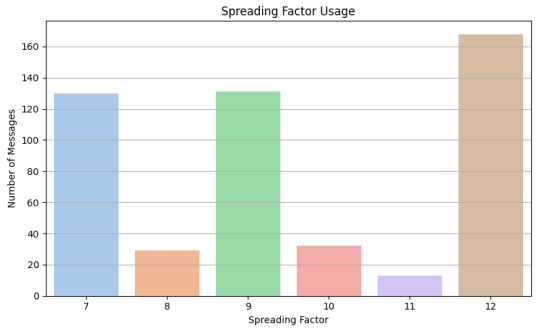
We observe that, according to the estimation, most devices are in the range of 1km from the gateway, while the most being in the $100 - 500$ meters range.



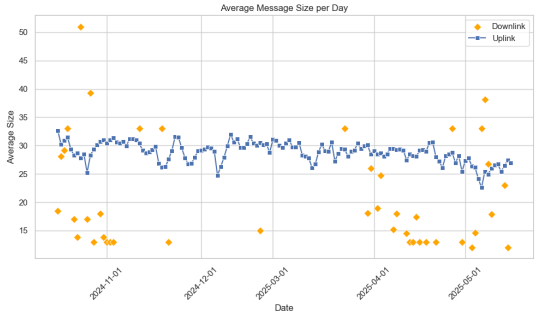
For the results regarding the garbage packets, the graph reveals a moderately different distribution for the distances. In fact, while the majority of devices still fall within the 100–500 meter range, this distribution has a large number of very close devices in the range of 100 meters, while only less than 10 devices are estimated to be above 500 meters. This suggest that spatially in general, the garbage devices tend to be closer to the gateway than the LoRa devices.

E. Analysis of Downlink and Comparison with Uplink

We immediately observed that the number of downlink messages is significantly lower than that of uplink messages, and their transmission is far more sporadic, even if the range of dates almost corresponds with the uplink. As expected, unlike uplink messages, all downlink packets contain a payload. In fact, there are no empty payloads in the downlink dataset.
Counting the number of device addresses in the dataset we notice that there are only **208** unique devices that the gateway has transmitted to. This confirms that only a small fraction of the devices in the uplink dataset actually received downlink messages, while the vast majority were limited to sending data without receiving any. Given the low volume of downlink traffic, we can also reasonably exclude the possibility that these devices are *Class B*. Class B devices require regular beacon transmissions to remain synchronized with the network, and the absence of frequent downlink messages makes this scenario unlikely. Regarding the frequencies used for downlink transmissions, we found no anomalies. They fall within the standard LoRaWAN frequency range, specifically between 867.1 MHz and 869.5 MHz.
Starting from the Spreading factor distribution, we start to notice interesting differences.



We notice a shift from a prevalent SF 7 towards higher Spreading factor, in fact we observe a prevalent spreading factor of 12 in the downlink. Next, calculating the average size of the packets over time and comparing it to the uplink shows that



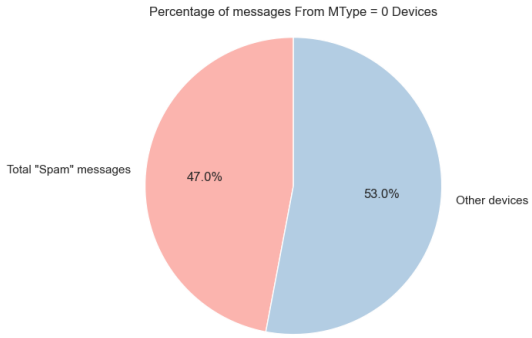
the size of the Uplink packets tends to be more stable whereas the size of the downlink is more variable. Next, we want to estimate the device class, therefore, we analyzed all downlink messages with MType = 3 or 5 (Unconfirmed and Confirmed Data Down, respectively). For each of these messages, we checked whether there was a corresponding uplink message from the same device within a 3-second time window. Since Class A devices can only receive downlink messages shortly after sending an uplink, if there is at least one message for that device that doesn't satisfy this condition, then we can be sure that, that device is not a class A.
This analysis, supported by the fact that most of the devices have not received any downlink, resulted in most of the devices being **Class A**, with occasional devices of **Class C**, due to the fact that only few devices had isolated downlink.

F. The Stange Case of Mixed Payload Devices

Observing the various devices, and from the results of the Clustering Analyses, we came across a group of devices with particular characteristics, specifically:

- An abnormally high number of messages compared to most other devices
- All messages with MType = 0 (0 = Join Request), 6 or 7 (Reserved for Future Use)
- they usually send packets in pairs, one message has payload and the other doesn't

These devices have never been accepted by the gateway, given the lack of downlink MType = 1 (Join Request Accept) for any of those devices, and this suggests repeated and continuous attempts. As stated previously in the Table 1, the number of devices with mixed traffic is relatively small, just 101. However, upon noticing the volume of traffic they generated, we quantified their total number of messages to assess their overall impact on the total traffic.



The results were striking: these 101 devices amount to a number of messages that is almost equivalent to half of the entire uplink dataset. This indicates that mixed-traffic devices may pose an even greater threat to network efficiency than those that consistently don't transmit payloads.

IV. Conclusions

We conducted an in-depth analysis of both uplink and downlink packets handled by the gateway. The **key conclusions** we have drawn are as follows:

1) Most of the devices transmitting to the gateway appear to be LoRaWAN **Class A and C devices**. They are likely **fixed in position** and show a **relatively consistent signal strength over time**.
2) The frequent use of **Spreding Factor of 7** suggests a stable connection, however, the signal quality is likely affected by the **dense urban environment**, which causes some degradation, as seen in the rssi distribution.
3) While the majority of devices transmit standard LoRa data, a **small subset of devices generates a disproportionately large number of non-LoRa-standard packets**. These packets are unusable and essentially flood the gateway. These devices are also estimated to be stationary and located within **200 meters** of the gateway, indicating local interference.
4) In addition to those sending unusable data, we identified **a group of devices that repeatedly send the same Join Request to the gateway**, always in duplicate.
5) The majority of messages are application-layer uplinks. Most devices only transmit data and **do not receive any downlink**.
6) Both uplink and downlink packets follow standard LoRaWAN configurations in terms of frequency and bandwidth, with no anomalies detected in these parameters.
7) Finally, there is **no significant temporal trend observed in the transmission patterns**. However, we observed there is a **temporal gap** of the duration of approximately two months, where most likely the gateway was shut off or data collection was not performed.

Acknowledgment

We would like to thank Prof. Cuomo for her availability and advices throughout the project, and Mr. Spadaccino for his supervision and for providing the base64 packet decoder.