LoRaWAN Data Analysis

Simone Zagaria – 2145389
Alessio Lani - 1857003

Smart environments – 2024/2025

1. Introduction

The Project

What?

- The objective of our project is to make a deep and meaningful analysis of the **LoRaWAN packets** collected by our gateway located in **Sapienza**

Why?

- 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**.

The Dataset

Our datasets consisted of two files containing **raw LoRaWAN packets**, in **CSV format,** one for **Uplink** and one for **Downlink.**
20+ Fields after decoding the packets, 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 (e.g., ADR, ACK).
  FCnt: Frame counter to prevent replay attacks.
  FPort: Port field; 0 indicates MAC commands, 1–223 for application-data.
  FRMPayload: Frame payload containing application data or MAC commands

Data Cleaning & Preprocessing

Formatting fields

- ‘**tmst**’ field was converted in a readable timestamp format
- **SF** and **Bandwidth** were parsed from the ‘**datr**’ field

Decoding packets

- All the packets were decoded using a **Base64 packet decoder**

Columns Selection

- We removed from both datasets columns that were either **not relevant**, contained **only NaN values**, or were replaced by **reformatted versions**.

2. Uplink Analyses

Uplink Overview

Total collected messages: **5’800’000**

Range of dates: **17 October 2024** to **15 May 2025**

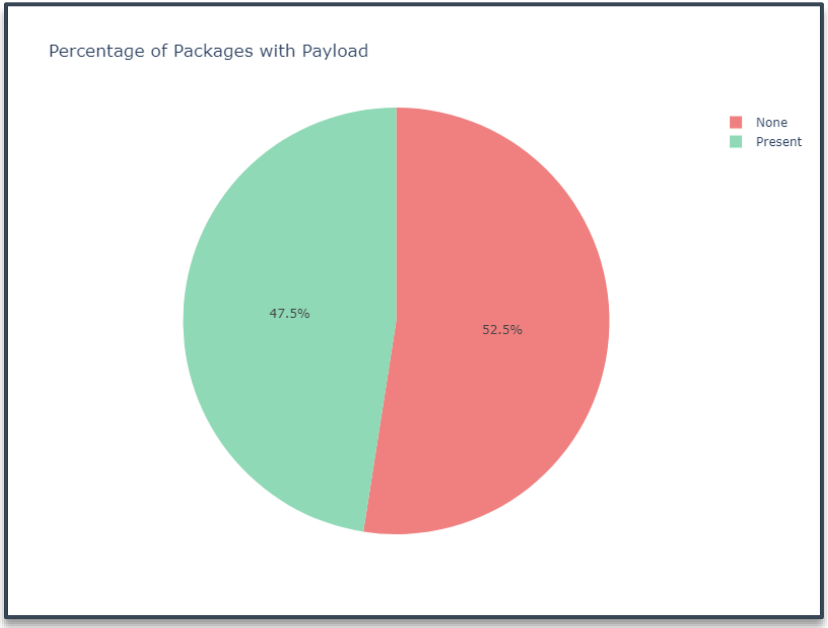
More than **150’000 unique devices** in uplink

Collected Packets vs. Usable Payloads

We noticed that most of the collected Payloads were **not Lora-Standard,** intercepted by the gateway

The chart illustrates the **Percentage of packets with data** and packets with unusable data (None).

This **large quantity of garbage packets** could indicate of numerous devices are transmitting on similar frequencies, polluting or cluttering the dataset.



Unique Devices Types and Frequencies

The number of **devices sending only garbage packets are only ~8%** of the total devices.

However, they appear to transmit at a significantly higher rate than the payload devices, **contributing disproportionately to the traffic and overall noise**.

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

Both the Payload and non-Payload Devices transmit at **similar frequencies**, and that could be the reason why they were intercepted by the gateway.

```
Min and Max frequencies: (867.1, 868.5)
```

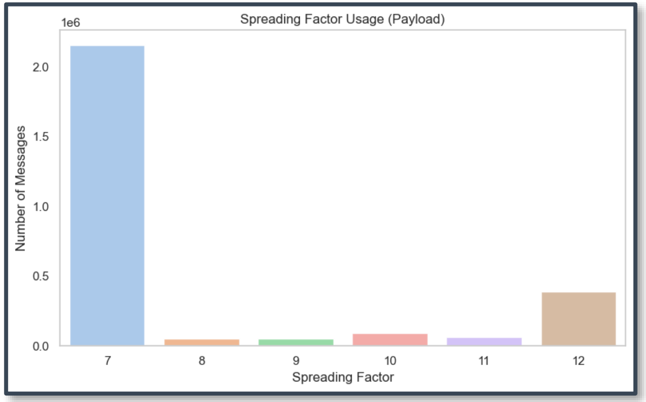
The transmission frequencies used in the dataset **fall within the typical LoRaWAN range**.

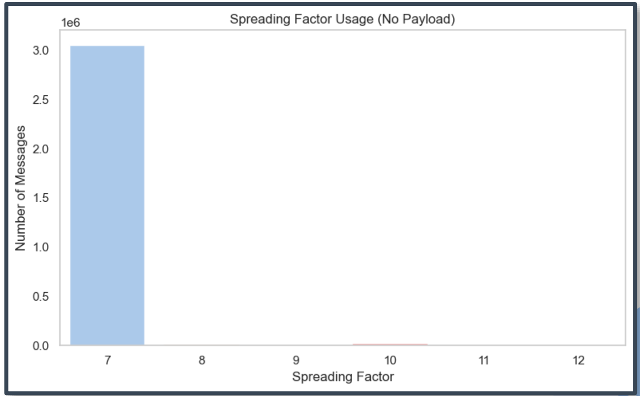
Spreading Factor Usage

The chart clearly shows that **Spreading Factor 7** is used in the vast majority of transmissions, while higher values like 8 to 12 are used very rarely.

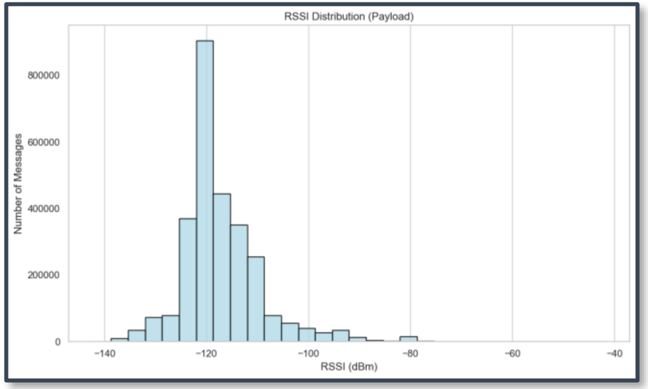
Garbage packets show an even with a more accentuated skewed distribution

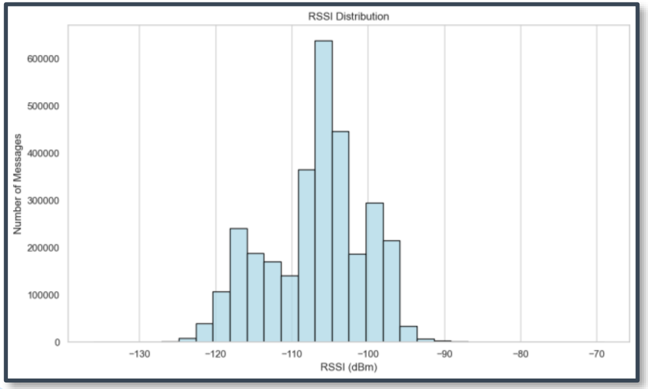
This suggests that devices were mostly operating in **good signal conditions**, likely near the gateway.





RSSI and used Frequencies





Next we also looked at The RSSI for the messages. In LoRa the acceptable RSSI value is **-30 dBm to -120 dBm**

Most RSSI values for payload-carrying messages are clustered around -125 dBm to -110 dBm, which is relatively weak, up to borderline **poor signal**.

In contrast, garbage packets tend to show better RSSI values, with their distribution concentrated between **-110 dBm and -100** dBm.

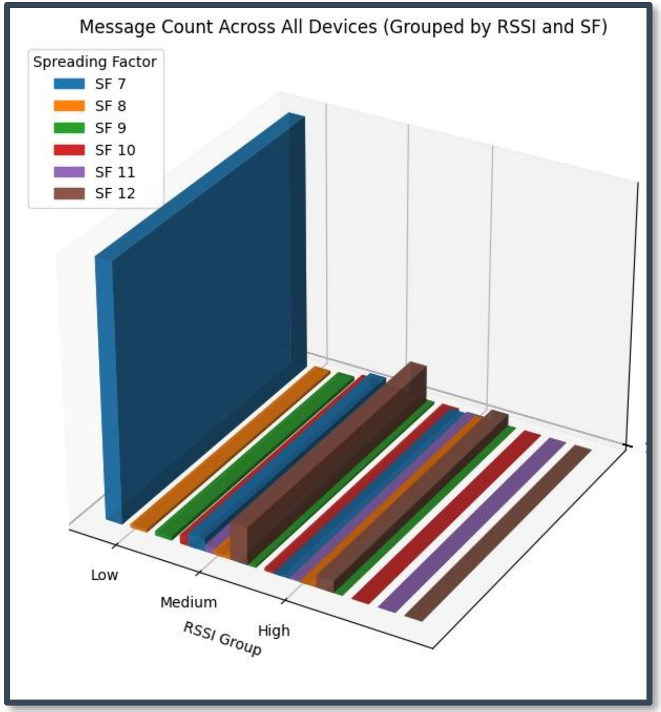
This indicates:

- **High interference**
- **Obstacles**
- Non Payload devices likely transmit from **shorter distances** than payload ones

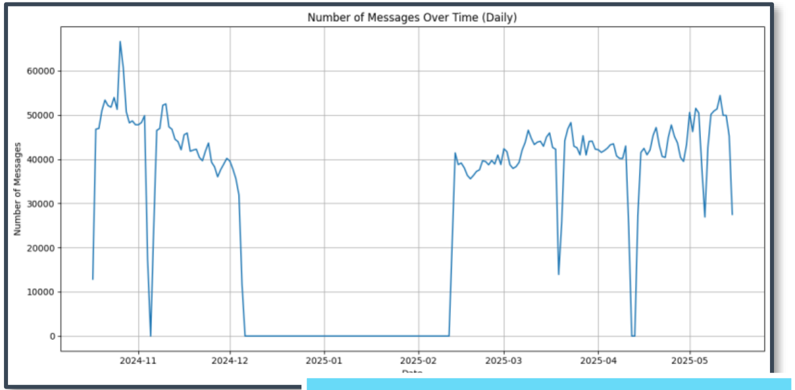
3D Plot

Next, we created 3 group of rssi (low, medium, high) and we confronted the **number of messages** and **spreading factor** to **rssi**

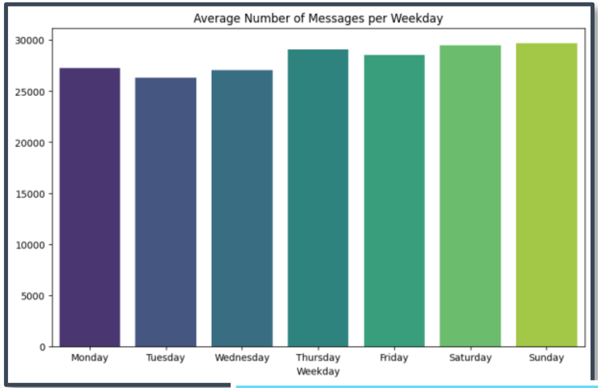
From the chart we observe, as expected, that **SF7** is used for **lower Rssi** and for a large number of messages, while higher values of rssi tend to be managed with higher values of Spreading Factor



Temporal Analysis



Number of messages Over time



Average of messages over the week days

- Data collection is consistent with occasional Spikes
- For approximately two months, from december 2024 to february 2025, the gateway was shut off or **data collection was not performed**.

- no significant temporal trend observed in the weekdays

Clustering

We clustered devices with similar properties, like **rssi, lsnr, size, SF and estimated distance**. We used the k-means algorithm and determined through the Elbow method that the **optimal number of cluster was 4**.

We also calculated the average number of messages per devicein each cluster. The results show a significant variation. Cluster 1 with 1.22 and Cluster 3 with over 110 messages

This suggests a number of devices that are much more active, possibly indicating different usage patterns or application types.

AVERAGE FEATURE VALUES BY CLUSTER

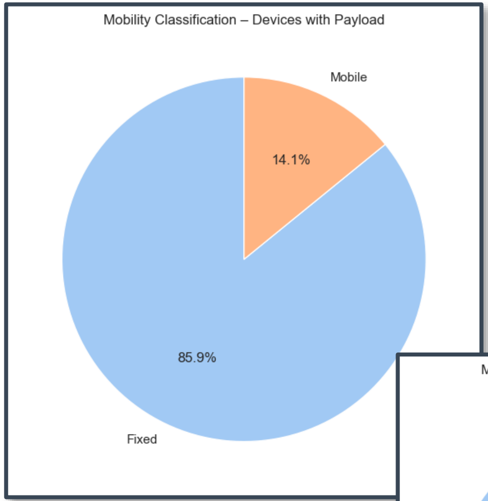
Cluster	RSSI	LSNR	Size	SF	Estimated Distance
0	-104.35	2.42	51.98	11.99	127.77
1	-92.57	-6.39	60.13	7.13	340.46
2	-126.69	-12.89	44.11	11.91	552.42
3	-97.53	4.85	12.57	9.98	83.50

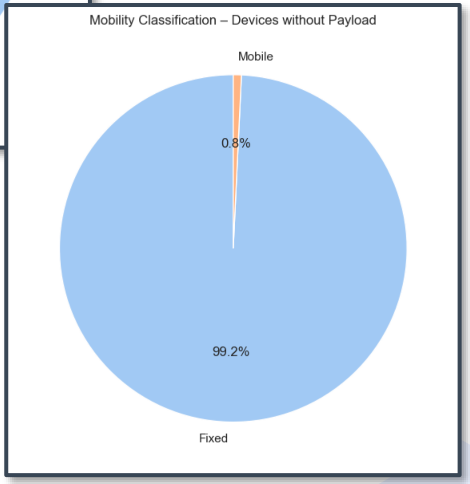
Mobility Estimation

For **Mobility Estimation** we used the **GM-RSSI technique**, which is based on the fact that RSSI usually tends to remain relatively stable over time.

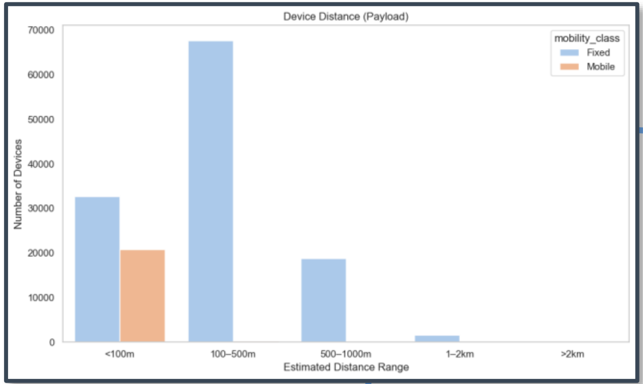
We notice that the two results differentiate strongly:

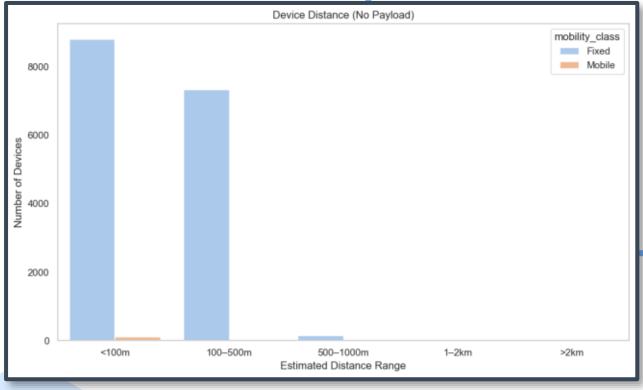
- Both resulted in having **more fixed devices than mobile**
- Non-Payload devices showed an even extremely **skewed trend** with nearly all devices estimated to be fixed





Distance Estimation





We made a rough approximation of the distance using the **RSSI-based distance estimation formula** $d = 10^{\frac{P-R}{10n}}$

Most Payload devices are **in the range of 1km** from the gateway

Most of the Non-payload devices tend to be within **100-200 meters** of the gateway

3. Downlink Analyses

Downlink Overview

Significantly **Lower total messages**: only **~500**

Same range of dates of the Uplink dataset

Same frequencies used by the Uplink (as expected)

Observations & Differences

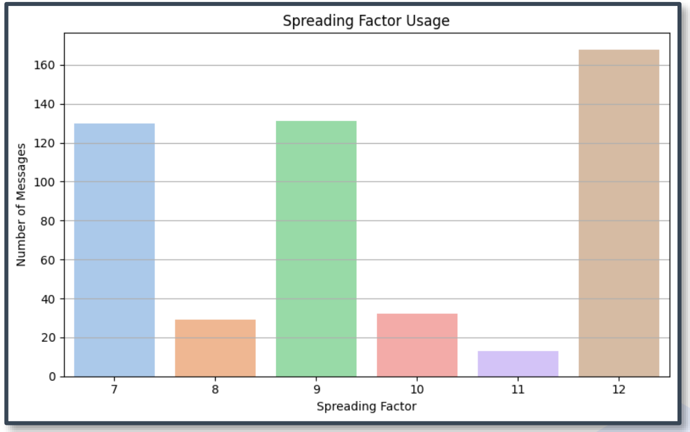
Only a small fraction of the devices in the uplink dataset actually received downlink messages

Given the low quantity of messages and the fact that so little devices receive downlink, reasonably **exclude the possibility that these devices are Class B**.

For the Spreading Factor, we observe a more even distribution, with a prevalent SF12

All Downlink messages have payload

201 Unique Devices



Class A or Class C?

For each message of each device in the downlink dataset we checked whether there was a corresponding uplink message from the same device within a **3-second time** window.

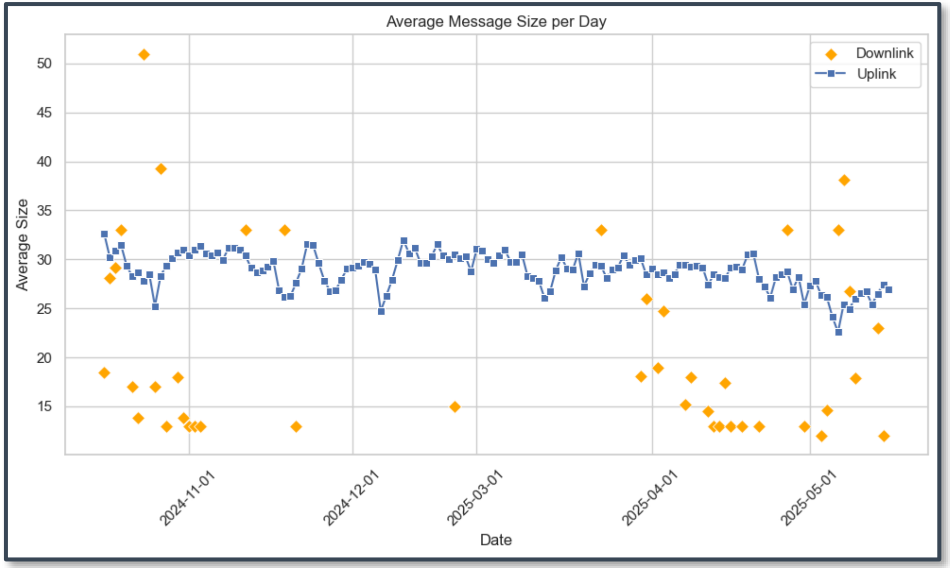
This is due to the fact that Class A devices can only receive downlink messages shortly after sending an uplink.

This analysis, supported by the fact that most of the devices have not received any downlink, resulted in the conclusion that **most of the devices being Class A**, with occasional devices of **Class C**, due to the fact that only few devices had isolated downlink**.**

Comparison of Average Message Size

We compared the average message size of Uplink and Downlink.

We observe that the **size of the Uplink packets** tends to be more **stable** whereas the **size of the downlink** is more **variable**.



4. Case Study: Mixed Payload Devices

Mixed Payload Devices

What?

we came across a group of devices with particular characteristics, specifically:

- An **abnormally high number of messages** compared tomost 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 haspayload and the other doesn't

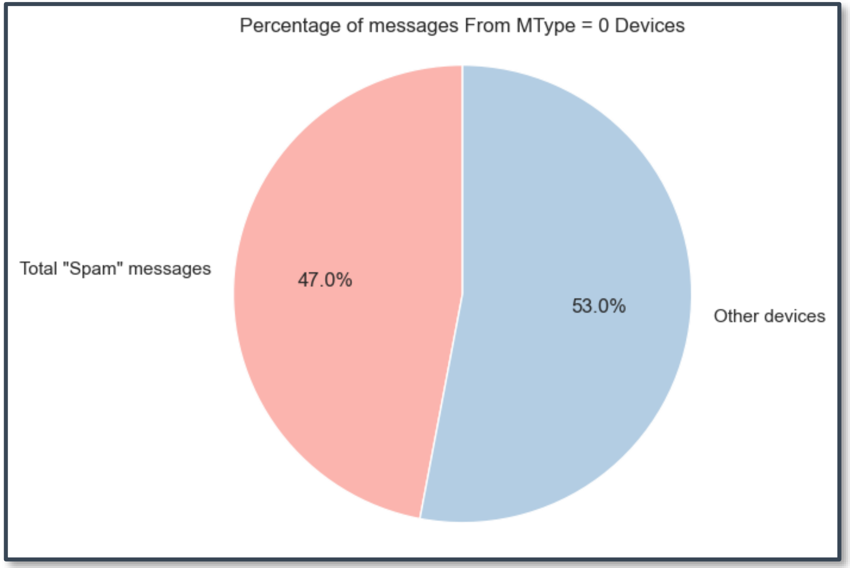
Relatively few:
Only **101 Devices**

This behaviour suggests continuous repeated attempts to associate the gateway that always get rejected

The Shocking Amount of Messages

We quantified their total number of messages toassess 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.**



5. Conclusions

Conclusions (Part 1)

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 Spreading Factor 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 lsnr and rssi distribution.

3) 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**.

Conclusions (Part 2)

4) We identified a group of devices that repeatedly **send the same Join Request to the gateway**—often in duplicate. These messages contribute further to the volume of unvaluable data

5) The majority of messages are **application-layer uplinks**, with only a small fraction serving other purposes. 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) Lastly, we observed **no significant Temporal trend observed** in the transmission patterns. However, there is a temporal gap in both the datasets of the duration of approximately two months where most likely the gateway was **shut off** or **data collection was not performed**

Thanks for the Attention!

Any questions?