“DansMaRue”

ANOMALY REPORTING CROWD SOURCING SYSTEM IN PARIS 2016 – 2022

Exploiting data for better city management

Marcia Fernanda
OTALORA FLORES

https://github.com/marciafof/dansmarue_rncp

Contents

INTRODUCTION

Engaging the public and non-experts in the search for solutions is no longer seen as just a trend but an objective of public decision-making. In recent years the term *"crowdsourcing"* has become widespread among the topic of sustainable development of cities (Certomà et al., 2015). It can be described as the practice of obtaining information or input from a large group of people typically through online platforms or mobile apps. The gathered information can include feedback, ideas, and data from a variety of actors like residents, businesses and visitors across different aspects of urban life such as transportation, public services, environmental issues, and urban planning.

From a city management perspective, engaging public participation in the design and implementation of solutions is necessary to implement effective policies and to achieve a balance between environmental protection measures, social cohesion, and the provision of democracy (Liao et al., 2019). At the same time, the widespread availability and use of personal ICTs, coupled with easy access to the internet, can result in the generation of vast amounts of data through crowdsourcing (Niu & Silva, 2020). Leveraging this data is essential for enabling decision-makers to make well-informed decisions.

Within the pool of crowdsourcing services, public service systems for monitoring non-emergency civic issues are one of the most matured examples of crowdsourcing for city management. These tools are part of collective monitoring services and shared management processes with operational goals (Mericskay, 2021). They work through collecting reported data on *anomalies,* which refer to issues in public spaces like street potholes, fallen trees, etc. The information is obtained by the public user and uploaded through the service's platform (website, mobile application) and then transmitted to the responsible public service department. Some of the most known examples include the **FixMyStreet** initiative in the UK and the **Street Bump** initiative in the city of Boston that focused on road surface conditions.

While the aforementioned examples are considered successful at improving public monitoring capabilities, they can also bring along challenges managing the reported anomalies and interpreting the collected data. For instance, Mericskay (2021) noted that the immediate nature of the information requires the reassessment of the service responsible for handling the anomaly. Additionally, data analytics raises questions about how to utilize retrieved data for diagnosing and prognosticating anomalies. Extracting trends and generalized characteristics of the city from these anomalies is essential to understand and improve public services.

This project focuses on the latter point and presents the use case of the data obtained through the anomaly reporting system **DansMaRue**, developed by the Ville de Paris in the city of Paris, France. The service was first launched in 2012 by the municipality of Paris, with the goal of using the citizens' participation in the identification of anomalies (graffities, etc.) in public spaces within the city of Paris. Its

popularity has exploded in the past five years, going from around 90 000 reports in 2017 to almost 900 000 in 2022, despite the COVID crisis in 2020.

The report is divided in three sections: Section 1 is dedicated to describing the context of the problem and defining the objective of this project. Section 2 describes the datasets used in this analysis. Section 3 addresses the methodology used in data preparation and processing. Section 4 shows the results and insights obtained in the exploratory data analysis (EDA) and finally Section 5 presents our conclusions.

CONTEXT AND OBJECTIVE

DansMaRue is a well-established crowdsourcing service with over 1000 records of anomalies a day. The collected data, which includes geographical position and imagery (photos), is currently only used as a collective monitoring tool. However, given the volume of anomalies registered each year, could we be missing important insights?

This project has the objective of exploring additional ways to utilize the collected data to prevent the loss of valuable information. By doing so, the goal is to develop a framework to identify key parameters in the data that could improve the efficiency of city management.

The datasets that were collected encompassed the most recent half of the application's existence, spanning from 2016 to 2022.

DATA BASE

DansMaRue SERVICE

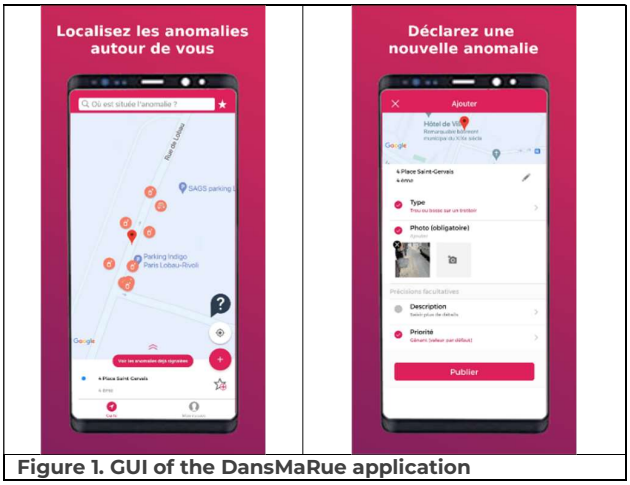


Figure 1. GUI of the DansMaRue application

Open to public in 2013 and inspired by the FixMyStreet project, the initiative was led by the Direction de la Propreté et de l'Eau à la Mairie de Paris as a collective monitoring tool. The application allows residents to report issues related to cleanliness, safety, and maintenance in public spaces. The app enables users to take photos of problems such as broken street lamps, illegal dumping, or damaged sidewalks, and report them to the appropriate city department. The information submitted through the app is then processed by the city's technical services and action is taken to address the issue. The app alsos provides users with real-time updates on the status of their reported issues.

DATA SOURCES

Core datasets

The principal source of data used in this work is the open data catalogue of the Ville de Paris accessible through the platform Open Data Soft. The platform offers the possibility of using its API through https://parisdata.opendatasoft.com/api/v2/console or downloading databases directly from the website. The following databases were obtained through this source:

- Historical records of **DansMaRue** application for the years 2016, 2017, 2018, 2019, and 2020.

Column Name	Description
id_dmr	Id unique obtained from raw database
type	Category of anomaly (FR)
soustype	Subcategory of anomaly (FR)
adresse	Address of the reported anomalie
code_postal	Postal code of the anomaly detected
numero	No description available
prefixe	A pour Iphone (Apple), G pour smartphone (Google), B pour Backoffice, S pour Paris.fr
intervenant	Name of the service in charge of fixing the anomaly if possible
conseilquartier	Neighbourhood of the anomaly
lon	Longitude in WGS 84
lat	Latitude in WGS 84
date_input	Date of the received report
subcategory	Subcategory (clean)
extrainfo	If extra information in the subcategory column

- Historical records of **DansMaRue** application for the years 2021 and 2022. The tables contained two additional columns and a different format of columns. Hence its separation from historical data from years before.

Column Name	Description
code_postal	Postal code of the anomaly detected
etat	Type of current state of the report treament

- Geographical information of each district (“Arrondissement”)

Official geographical delimitation of the districts or “Arrondissements” in the city of Paris. The data was downloaded as a geojson file and worked to be presented as a table.

Column Name	Description
n_sq_ar	Identifiant séquentiel de l'arrondissement
c_ar	Numéro d'arrondissement
c_arinsee	c_arinsee
l_ar	Nom de l'arrondissement
l_aroff	Nom officiel de l'arrondissement
n_sq_co	n_sq_co
Surface	surface
perimetre	perimetre
geom_x_y	geo_point_2d
geom	geo_shape

- Geographical information of each quarter ("Conseil de Quartier")

Official geographical delimitation of the ensemble of "conseils de quartier", here on referred to as "quartier", within each district. The data was downloaded as a geojson file and worked to be presented as a table.

Column Name	Description
NO_CONSQRT	Numéro du Conseil de Quartier
NOM_QUART	Nom du Conseil de Quartier
NAR	Arrondissement
NSQ_CA	*No description available*
AREA	*No description available*
Geometry X Y	geo_point_2d
Geometry	geo_shape

Complementary datasets

To enrich our analysis we downloaded data from other official and non-official sources.

- Population and surface data for each arrondissement

Additionally, the population data for each arrondissement was obtained through the National Institute of statistics, INSEE portal, Comparatuer de territoires. The data is not tabular but it was reconstructed to a tabular format.

Column Name	Description
Population en 2019	Number of population
Densité de la population (nombre d'habitants au km²) en 2019	Number of population per km² in 2019
Superficie en 2019, en km²	Area in km²
Variation de la population : taux annuel moyen entre 2013 et 2019, en %	Variation in percentage of the population between 2013 and 2019
dont variation due au solde naturel : taux annuel moyen entre 2013 et 2019, en %	Variation due to natural balance
dont variation due au solde apparent des entrées sorties : taux annuel moyen entre 2013 et 2019, en %	Variation due to entries and leaves
Nombre de ménages en 2019	Number of households

- Amenities and commercial points over the city of Paris

OpenStreetMap (OSM) is a free, open geographic database updated and maintained by a community of volunteers via open collaboration. The level of detail of the available geographical data makes it a good alternative when official data is not available or open. Two map features of interest were collected within the city of Paris for further exploration and link to the detection of anomalies.

- The amenity tag is the top-level tag describing useful and important facilities for visitors and residents, such as toilets, telephones, banks, pharmacies, prisons and schools.
- The shop tag is used as a place of business that has stocked goods for sale.

Column Name	Description
id_	Numéro du Conseil de Quartier
Lat	Nom du Conseil de Quartier
lon	Arrondissement
category	Tag name (In this case amenity or shop)
name	If name available we collected it
geometry	

The ensemble of datasets were collected and restructure and formatted to create the final working database. The methodology is explain the following section.

METHODOLOGY

DATA COLLECTION

The heterogeneity in the sources of data used in this work requires the development of ad hoc scripts. Since the geographical aspect is the predominant feature that will be used as the linking point between the majority of the datasets, this demands to be familiar with Geographical Information System (GIS) programming tools.

The collection process was divided into steps:

1. Streaming data from the API of interest.
2. If data not available in API it was downloaded manually through the official website. For example, the historical data of **DansMaRue** is only available in direct download.
3. Save the information locally as text, json or to our local SQL Server in its raw version.

The data was parsed properly into a tabular form using Python and using SQL for the creation of tables within the server. The list of scripts linked to this stage are:

- Ville de Paris Datasets: **`scripts/get_raw_data_parisville.py`**
- OSM data : **`scripts/get_osm.py`**

DATA PROCESSING

Data cleaning

Once the data collection is processed we continue to the inspection, formatting and cleaning of each dataset. For this purpose I created a python script for the cleaning of the **DansMaRue** data, its formatting and finally sending it to the SQL Server. This can be found as **`scripts/data_cleaning.py`**.

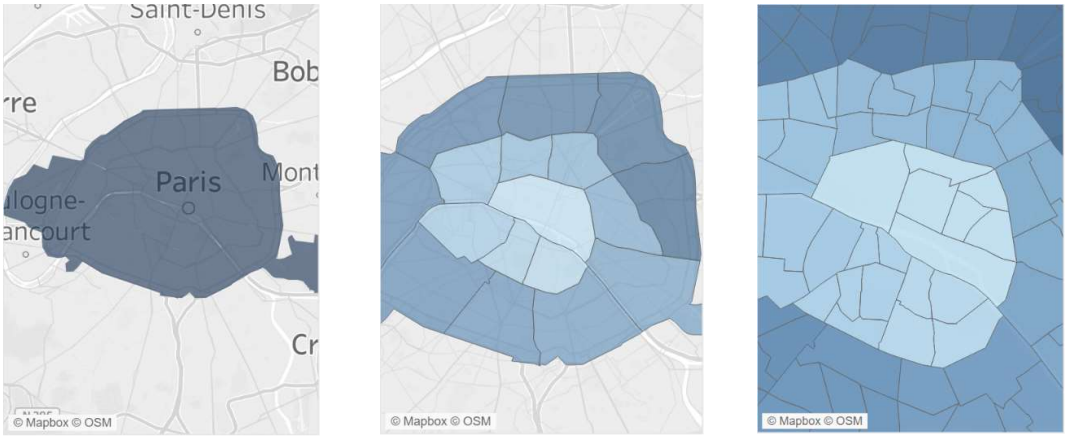
The general steps in the cleaning process can be summarized as following:

- Parsing files correctly: choosing right delimitator and encoding
- Standarizing columns names
- Verification and standardization of columns names
- Parsing datetype columns
- Checking duplicates
- Checking NaN values
- Dropping unnecessary columns
- Additionally, the geographical data was checked for values outside of the geographical boundaries of interest (Paris).

Overall, the datasets collected were “clean” with very low percentage of NaN values (<5%) and duplicates allowing the dropping of rows without any impact in the quality of the data. Note that the detection of outliers was not included in this part. This is because at this stage, the dataset did not have any measurable features (only categorical).

Data harmonization

One of the main challenges of this dataset was the choice of the Level Of Detail (LOD) set for the analysis. Although the granularity of the main dataset was adequate to understand local phenomena, our goal was to provide insights at a level which can be useful for decision-making. In this case, we decided to perform the analysis at 3 levels shown in Figure 2:



a. Level 0 **b. Level 1** **c. Level 2**

Figure 2. Examples of Levels of detail (LOD)

The first level considers the entire city of Paris. At this level we expect to get the overall picture on the number of reports and main issues at a city level. Secondly, we zoom in to the district, or “Arrondissement” level. This is the working level of the main municipality authorities. Finally, the quartier level divides the districts into “neighborhood councils” that are structures that involve city residents in municipal management.

The harmonization of the dataset at these three LODs was the main time consuming task during the data preparation. Modifications in the official delimitations at the district level and quartiers were done in the time period of the study. To avoid inconsistencies, the latest official administrative division for both quartier and district was backpropagated to every dataset. The technical actions in this process include the use of a specific library for the treatment of geospatial data name Shapely.

Additional features

The information collected from the anomaly reports are for the most part categorical. However, datasets collected from the year 2021 onwards contain information on the status category and the date of the reported status change. Although no clear description of these fields is provided, we assume that it corresponds to the time it took for the anomaly to change its status, e.g. from “NEW” to “RESOLVED”.

This feature is calculated when the data is available and store under the column:

deltadays_signal_etat

After the data cleaning and harmonization, the tables were sent to the SQL server and the final database structure was created.

DATA EXPLORING WITH SQL

The collected datasets were formatted into a tabular format. This format has the advantage of being easy to organize and making the interpretation and analysis a lot simpler for the human eye. Under the context of our problem, tabular data is a good choice because it will allow us flexibility when it comes to manipulating different sources of information and verifying its consistency.

Why SQL?

The vast availability of data analytics tools often blurs the benefits of classical languages like SQL. Pandas is one of the most popular tools nowadays thanks to the raising popularity of Python programming language. But since it is based on SQL, is it really better?

Both SQL and Pandas are confirmed tools for data analysis, meaning that there is vast community of users behind them. Each has its own set of advantages and disadvantages. Personally, the intuitiveness aspect of Pandas wins over SQL, however, when dealing with large databases SQL is an optimal choice. Some of the main advantages of SQL are:

- **Speed**: SQL is typically faster than Pandas when working with large datasets. This is because SQL is designed to work efficiently with databases, while Pandas is designed for more general data manipulation tasks. SQL can quickly filter and aggregate data using indexes, while Pandas needs to load the entire dataset into memory.
- **Scalability**: SQL is designed to work with large, distributed databases, making it more scalable than Pandas. SQL can easily handle datasets that are too large to fit into memory, while Pandas may struggle with these types of datasets. The possibility of working with cloud sourced data servers is definitely a win for SQL.
- **Data security**: SQL provides robust data security features that allow you to control who can access and modify your data. You can set up user accounts and permissions, and audit data access and changes. Pandas does not have built-in data security features.
- **Data consistency**: SQL is designed to enforce data consistency by using constraints, such as primary keys, foreign keys, and unique constraints. These constraints prevent you from entering invalid data into your database. Pandas does not have built-in data consistency features.

In summary, SQL has several advantages over Pandas when it comes to working with large datasets, scalability, data security, data consistency, and integration with databases. However, Pandas is still a powerful tool for data analysis, particularly for smaller datasets that can be loaded into memory. The choice between SQL and Pandas will depend on the specific requirements of your analysis.

BUILDING MAIN DATABASE

After the process of data cleaning, we proceeded with the creation of the playfield in SQL for the EDA. The ensemble of queries can be found in the **`scripts`** folder.

```
create database if not exists dansmarue;
USE dansmarue;
```

```
-- JOIN TABLES FROM EACH YEAR
 CREATE TABLE dmr_all
    SELECT * FROM dmr_2016_clean
    UNION ALL
    SELECT * FROM dmr_2017_clean
    UNION ALL
    SELECT * FROM dmr_2018_clean
    UNION ALL
    SELECT * FROM dmr_2019_clean
    UNION ALL
    SELECT * FROM dmr_2020_clean
    UNION ALL
    SELECT * FROM dmr_2021_clean
    UNION ALL
    SELECT * FROM dmr_2022_clean
    ;
ALTER TABLE dmr_all
ADD PRIMARY KEY (id_dmr);
```

SQL QUERIES

```
/* ---------------------- OVERALL VIEW *------------------ */
    -- GET THE TOTAL COUNT OF REPORTS
SELECT count(*) as count
from dmr_all
ORDER by count DESC ;

    -- GET THE AVERAGE OF TOTAL COUNT OF REPORTS PER WEEK
SELECT count_p_week.year, AVG(count_p_week.count)
FROM
    (SELECT extract(year from date_input) as year, extract(week from
date_input) as week, count(*) as count
    from dmr_all
    group by year, week
    ORDER by count DESC) as count_p_week
group by year
;

   -- GET THE COUNT PER YEAR  COMPARE TO PREVIOUS YEAR IN PERC
WITH yearly_count as (
 SELECT
   EXTRACT(year from date_input) as year,
```

```
   COUNT(*) as count
 FROM dmr_all
 GROUP BY year
)
    select
      *,
      (yearly_lag.count - yearly_lag.count_previous_year) /
yearly_lag.count_previous_year * 100 as perc_diff
    from (
            SELECT yearly_count.year, yearly_count.count,
                    LAG(count) OVER ( ORDER BY year ) AS count_previous_year
            from yearly_count
            ) as yearly_lag
    GROUP BY year;


-- GET THE COUNT PER YEAR PER MONTH AND COMPARE TO PREVIOUS MONTH
WITH yearly_monthly_count as (
 SELECT
   extract(year from date_input) as year,
   extract(month from date_input) as month,
   COUNT(*) as count
 FROM dmr_all
 GROUP BY year, month
)
    SELECT yearly_monthly_count.year, yearly_monthly_count.month,
yearly_monthly_count.count,
            LAG(count) OVER ( ORDER BY year , month) AS count_previous_month
    from yearly_monthly_count;

-- GET THE COUNT PER MONTH AND COMPARE TO PREVIOUS MONTH
WITH monthly_count as (
 SELECT
   extract(month from date_input) as month,
   COUNT(*) as count
 FROM dmr_all
 GROUP BY  month
)
    SELECT monthly_count.month, monthly_count.count,
            LAG(count) OVER ( ORDER BY month) AS count_previous_month
    from monthly_count;
```

```
-- GET THE COUNT PER ARRONDISSEMENT
SELECT code_postal_id, COUNT(*) as count
FROM dmr_all
GROUP BY code_postal_id
ORDER BY count DESC;
```

```sql
 -- GET THE COUNT OF CATEGORY
SELECT category_FR, count(*) as count
from dmr_all
group by category_FR
ORDER by count DESC ;
```

```sql
 -- GET THE COUNT OF CATEGORY PER YEAR
SELECT category_FR, EXTRACT(YEAR FROM date_input) AS year, count(*) as count
from dmr_all
group by category_FR, year
ORDER by count DESC ;
```

```sql
/* ---------------------- JOINING WITH OTHER FIELDS *------------------ */

-- GET COLUMNS FROM shop_georef
  select * from shop_georef;
 /* id_, lat, lon, category, subcategory, name, geometry, quartier_id,
code_postal_id */
 -- SHOPS GROUP BY quartier_id
 SELECT category, quartier_id, code_postal_id, COUNT(id_) as count_shops
 FROM shop_georef
 GROUP BY category, quartier_id, code_postal_id;

WITH count_by_quartier AS (
    SELECT code_postal_id, quartier_id, COUNT(*) as count
    FROM dmr_all
    GROUP BY code_postal_id, quartier_id)

    SELECT *
    FROM count_by_quartier
    LEFT JOIN
            (SELECT category, quartier_id, code_postal_id, COUNT(id_) as
count_shops
             FROM shop_georef
             GROUP BY category, quartier_id, code_postal_id) shops_by_quartier
    ON count_by_quartier.quartier_id = shops_by_quartier.quartier_id

;
```

FINAL ERD

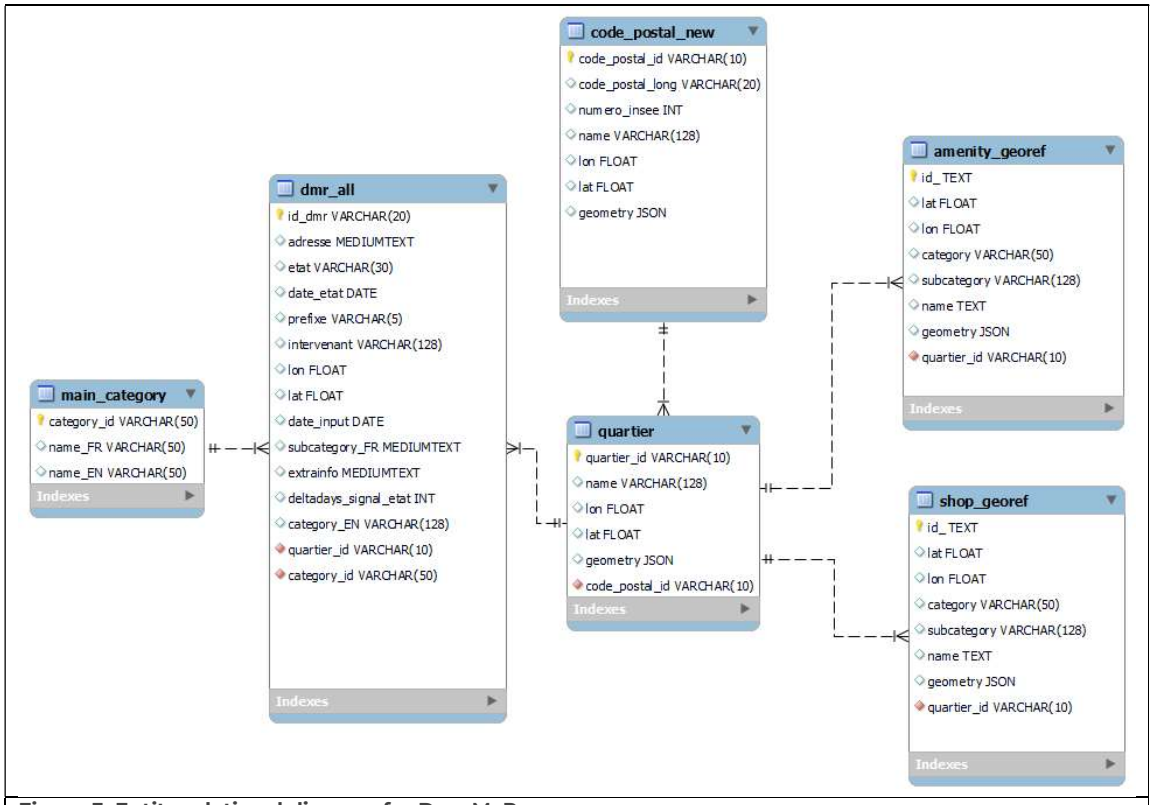


Figure 3. Entity relational diagram for DansMaRue use case

The final schema contains 6 tables.

RESULTS OF EXPLORATORY DATA ANALYSIS (EDA)

For the EDA section was divided between SQL, Python and Tableau.

1. SQL
   Queries were processed to obtain overview of aggregated fields.
2. Python
   a. Manipulate aggregated fields obtained from SQL query
   b. Additional EDA like identification of outliers and timeseries analysis in detail
   c. Format output for Tableau
3. Tableau
   a. Visualization
   b. Dashboard creation

The first part of the EDA focuses on the evolution of the number of reported anomalies in the past 6 years, including the period concerning the COVID pandemic.

How many anomalies and when do they occur?

A total of 3 253 435 reports were collected by the **DansMaRue** service between the year 2016 and 2022.

Although available for over 10 years, the service showed an important increase in the number of reports starting from the year 2018, as shown in Figure 4 , but shows and stabilization of the number of anomaly reports in the last 2 years reaching approximately 800 000 annual records with a decrease of reports between 2021 and 2022 of -0.5%.

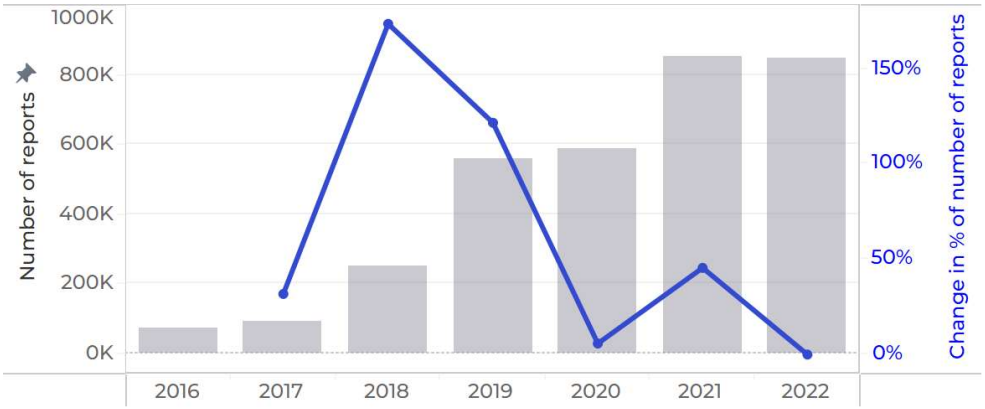


Figure 4. Number of annual reports (grey bar) and year by year change in percentage (blue line)

Looking into the distribution of reports by month (Figure 5) we can see that there is a lower percentage of anomalies being reported in the winter season between the months of January and Mars. The months with highest number of anomalies reported are June through October, with the exception of august. This is explained by period of holiday season in France where many locals leave town.

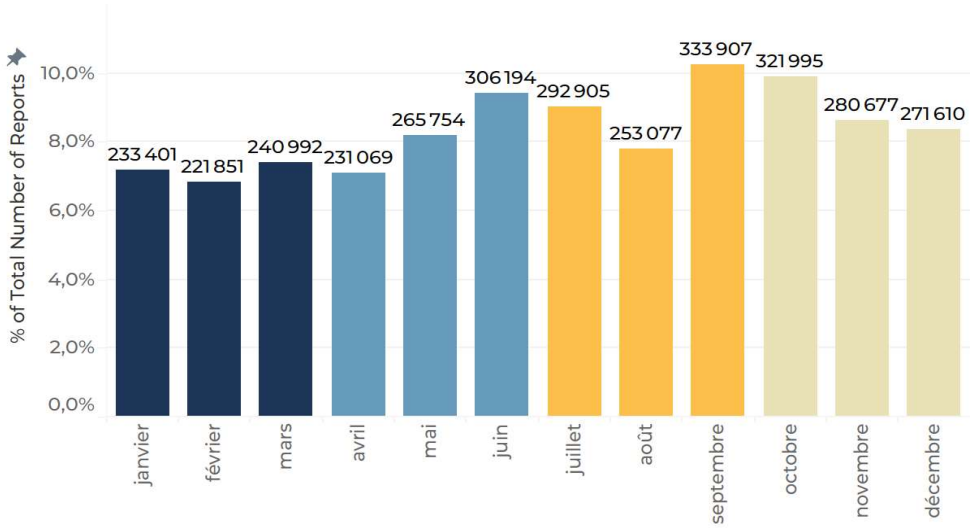


Figure 5. Percentage of the total number of reports per month, colored by season of the year

Figure 6 shows the number of anomaly reports in terms of day of the week. Monday is the day with the highest number of reporting, with a decreasing trend for the rest of week. Nevertheless, the day to day variation during the weekday is fairly constant with each day between Monday through Friday accounting between 14-15% of the total data.

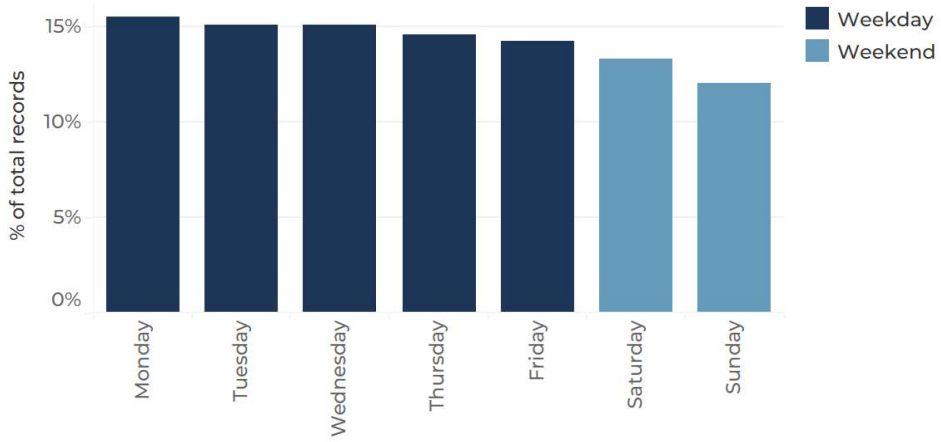


Figure 6. Distribution of total number of reports per week of the day

What type of anomalies?

We are also interested in knowing which are the most common type of issues in the area of study. Figure 7 shows the results at a city level considering the last 6 years of data. We can observe that the top four categories of anomalies are occupied by abandoned objects, graffities, cleanliness related issues and vehicle related problems.

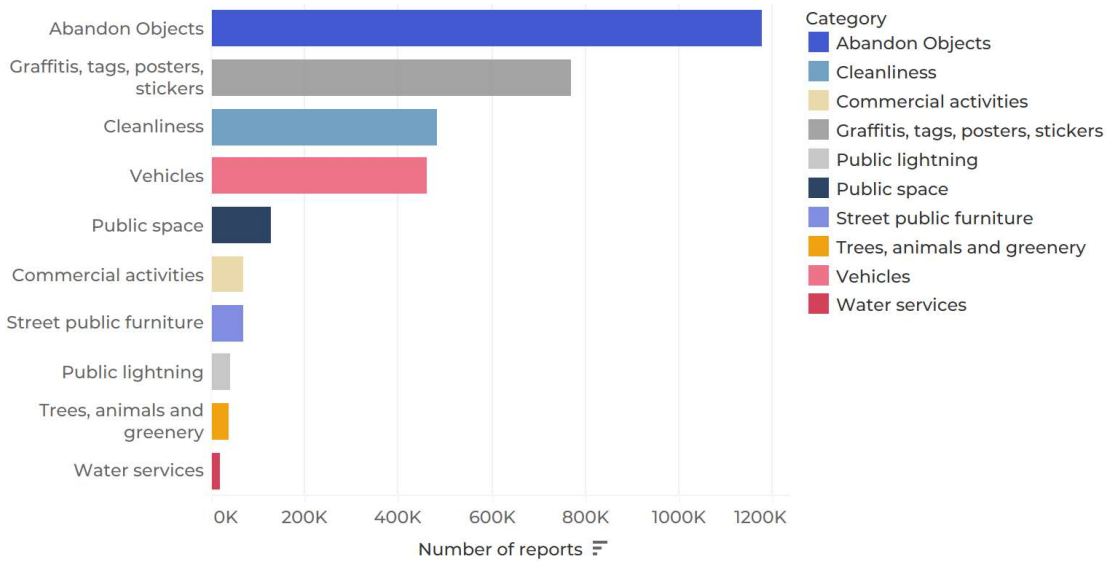


Figure 7. Number of reports for the period between 2016 and 2022

A deeper look in the year by year data (Figure 8) shows that the Vehicles related incidents have increased in volume of reports, as well as the graffities related anomalies.

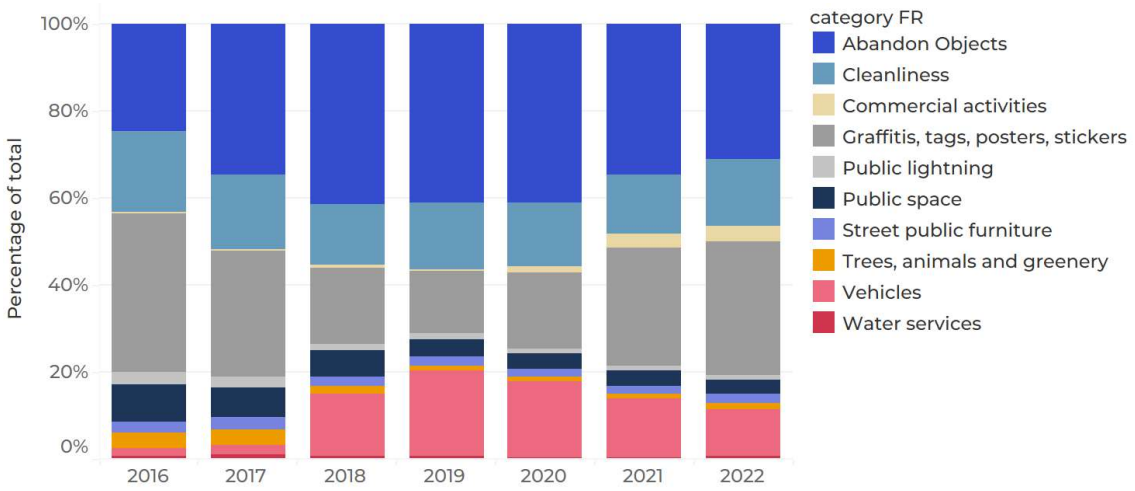


Figure 8. Distribution of the category of anomaly per year

Where are the anomalies happening?

So far we have looked into how the data is distributed temporarily, however, the main aspect of this dataset is its richness in terms of spatial features. The results in this report focus on the 2 level de detail, i.e. district.

Figure 9, shows the percentage of reports per district for the period between 2016 and 2022. The calculation uses the total number of records associated with each of the 20 districts in Paris. As explained in the Data Processing section, the data was adjusted to group the 1st, 2nd , 3rd and 4th districts into the PARIS CENTRE district.

The top 5 districts accounted for 50 % of the total cumulative reports a shown in Figure 9. The highest percentage of anomalies are located in the 18th district. It is followed by the 15th , 10th, 20th and PARIS CENTRE which share similar number of reports. On the contrary, the bottom 5 districts account for less than 10% of the data.

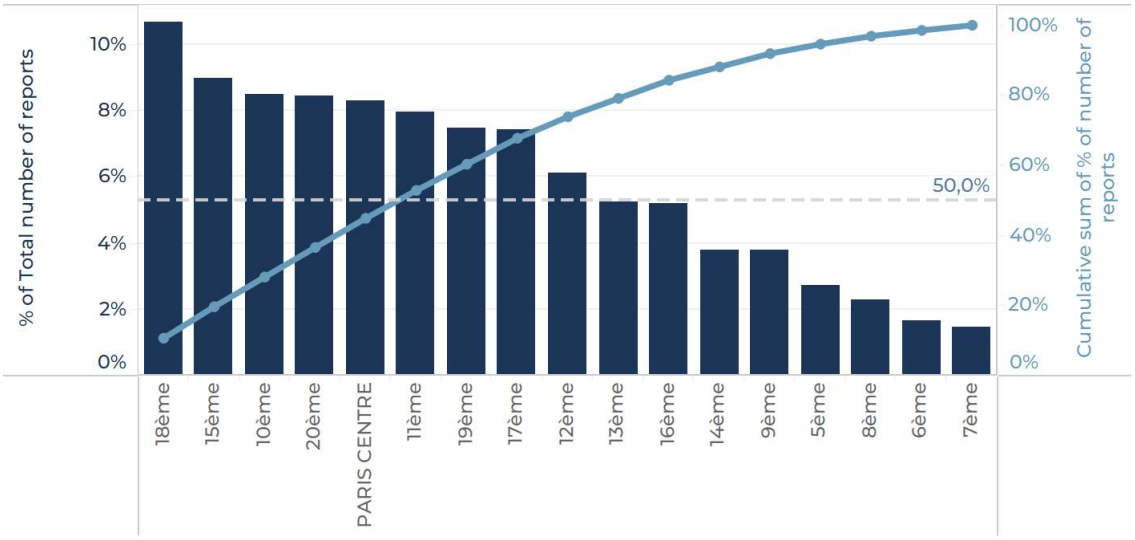
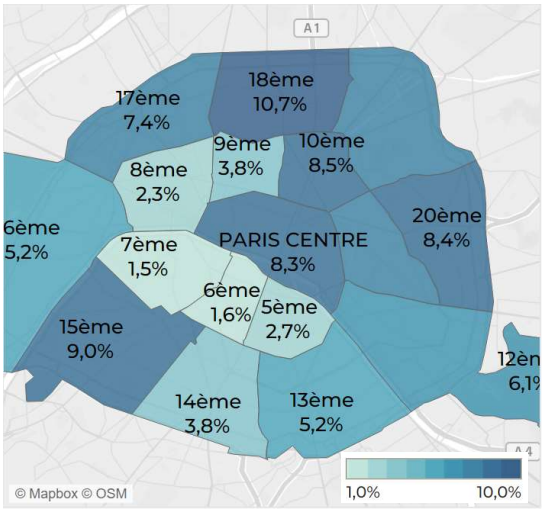


Figure 9. Percentage of number of reports per district over the total sum between 2016 and 2022 and cumulative sum

The interpretation of these results require further considerations. Firstly, the absolute number of reports over each district is not normalize by the size or population per district. Figure 12 shows the map of Paris colored by the percentage of the total number of anomalies for each district. Here, the smallest districts are also part of the bottom 5 in number of anomalies. Figure 13, shows the same result but considering the number of anomalies divided by the surface of each district.



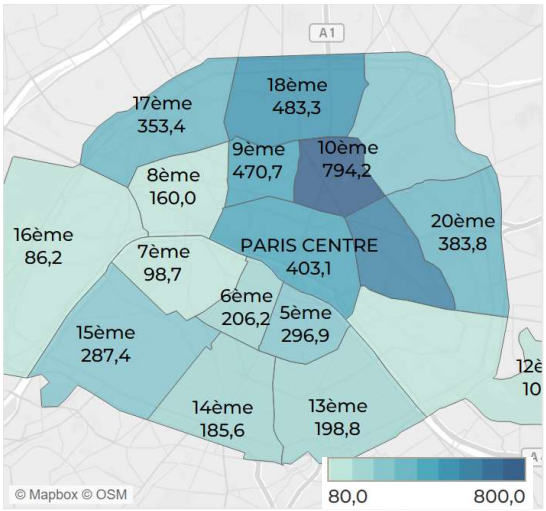


Figure 10. Percentage of number of anomalies by district for the period 2016-2022

Figure 11. Average number of anomalies per km2

These results showcase the importance of normalization when doing comparative analysis. The average number of anomalies by km^2 show a different picture in terms of most impacted districts by anomalies. The 10^{th} district, takes over the first place in the ranking, followed by the 11^{th} , 18^{th} and 9^{th}. In particular the 10^{th} district shows values

twice as high as other districts in the top 5. Further analysis in this area is recommended.

What are the main problems in my district?

Besides the distribution of the reports, we are interested in the categorical distribution of the anomalies over each LOD. At the district level we summarize the information in a colored table (Table 1) . Each row is the data divided by the principal category of the anomaly used by the DansMaRue service. The districts are arranged according to the average density of anomalies by km^2.

Table 1. Percentage of reported anomalies by category for the period 2016 to 2022

	Abandon Objects	Cleanliness	Commercial activities	Graffitis, tags, posters, stickers	Public lightning	Public space	Street public furniture	Trees, animals and greenery	Vehicles	Water services
10ème	31,7%	12,6%	3,0%	33,8%	0,7%	2,6%	1,4%	0,7%	13,1%	0,4%
11ème	36,2%	9,9%	2,6%	34,8%	0,9%	2,7%	1,7%	0,9%	10,0%	0,4%
18ème	42,0%	15,2%	2,3%	24,3%	1,1%	2,7%	1,5%	0,8%	9,4%	0,5%
9ème	24,0%	14,0%	4,8%	30,0%	0,7%	3,7%	1,9%	0,7%	19,7%	0,5%
PARIS CENTRE	15,8%	12,6%	5,0%	38,6%	1,3%	4,8%	2,8%	1,2%	17,4%	0,6%
20ème	48,9%	15,1%	1,0%	16,7%	0,8%	2,9%	2,1%	1,2%	10,9%	0,5%
17ème	40,1%	21,5%	2,0%	11,9%	0,9%	3,6%	1,7%	1,3%	16,3%	0,8%
19ème	40,7%	17,3%	0,6%	21,8%	1,0%	3,1%	2,3%	1,2%	11,6%	0,4%
5ème	18,8%	8,4%	2,3%	38,0%	1,4%	4,9%	3,1%	0,9%	21,5%	0,6%
15ème	53,9%	12,0%	0,9%	12,0%	1,5%	4,4%	1,7%	1,3%	11,8%	0,5%
6ème	28,2%	10,1%	4,5%	30,6%	1,0%	4,9%	2,6%	0,8%	16,8%	0,6%
13ème	40,2%	15,2%	0,8%	21,8%	1,6%	4,0%	2,5%	1,8%	11,7%	0,5%
14ème	27,4%	20,3%	1,4%	20,0%	1,9%	4,7%	2,4%	1,6%	19,7%	0,6%
8ème	14,6%	23,3%	2,7%	15,3%	2,1%	7,6%	2,6%	1,2%	30,0%	0,7%
12ème	38,3%	12,7%	1,3%	23,0%	1,4%	4,8%	3,0%	1,6%	13,2%	0,7%
7ème	15,3%	22,2%	2,3%	24,8%	2,4%	8,4%	3,3%	1,9%	18,9%	0,6%
16ème	37,0%	18,8%	1,5%	9,8%	2,5%	6,5%	2,2%	1,7%	19,3%	0,7%

The top 3 districts share show that the main issues belong to the Abandoned objects and Graffities related categories. PARIS CENTRE and the 5th district show that their main source of anomalies are graffities. Whereas the 8th district main issue is vehicle related anomalies.

A clear distribution of the top anomaly category for each district is shown in Figure 12. From this representation we can observe that the districts with a larger surface are mostly affected by the category of Abandoned objects. Districts within the inner areas of Paris are mostly concerned by the Graffities category. The 8th district is the “outlier” within the group as it is the only one primarily affected by Vehicle related issues.

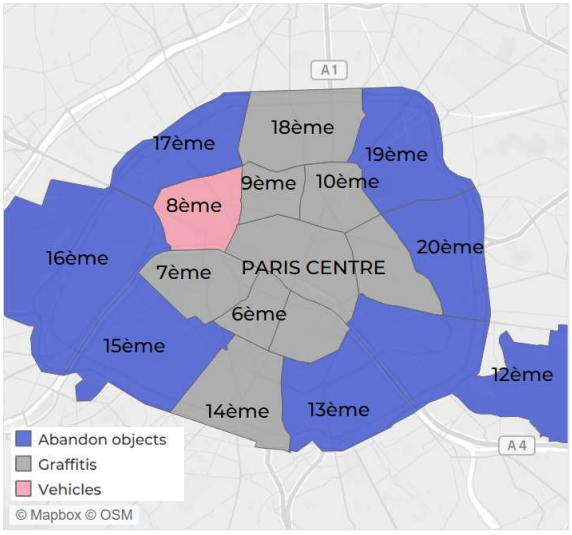


Figure 12. Principal category of anomaly for period 2016-2022

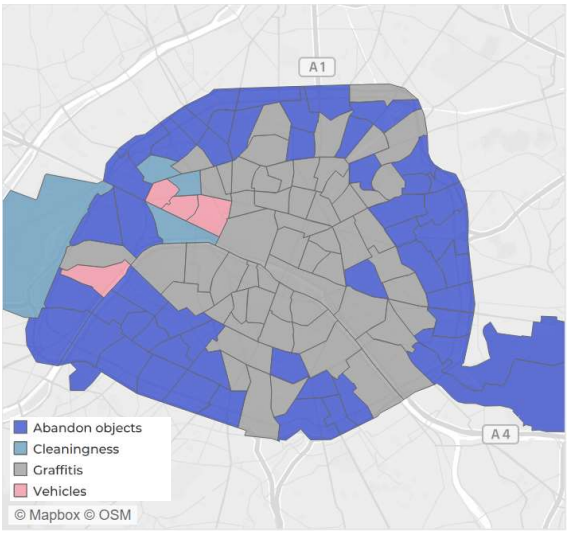


Figure 13. Principal category of anomaly by quartier for period 2016-2022

Zooming into the quartier level we can observe that not all quartier are affected by the main category of its district and the cleanliness category also impacts quartiers within the 8th and 16th district.

CONCLUSIONS AND RECOMMENDATIONS

Based on the exploratory analysis in this study, we were able to get insights on the distribution and trends of the main categories of anomalies from DansMaRue service. Firstly, the number of anomalies reported annually has been stabled over the past two years and it represents around 800 000 points per year, or approximately 16 000 anomalies per week. The majority of these records are transmitted to the responsible department. Thus, understanding the distribution and trends of the frequency of these anomalies is a key information for city management.

One of the insights obtained from the seasonal trends is the increase in frequency in the "warmer" months. Additional insights on the seasonal trend on subcategories could be useful to narrow down anomalies at a quartier level.

The main categories affecting the city of Paris are the "Abandoned objects" and the "Graffity" related anomalies. Considering the data obtained for the period 2016 through 2022, the districts on the outer ring of the city are affected by a higher percentage of "Abandoned objects" while the inner city shows a high percentage on the "Graffity" category.

Overall the insights presented in this work are the beginning of a longer process of introspection of data but they present the general picture of the anomaly reporting over the city of Paris. With the crossing of alternative data, which is a work in progress, we are aiming at a better understanding of the link between anomalies and the city's features.

All the scripts used in this project are available in the following github portfolio:

https://github.com/marciafof/dansmarue_rncp

BIBLIOGRAPHY

Certomà, C., Corsini, F., & Rizzi, F. (2015). Crowdsourcing urban sustainability. Data, people and technologies in participatory governance. *Futures*, *74*, 93–106. https://doi.org/10.1016/j.futures.2014.11.006

Liao, P., Wan, Y., Tang, P., Wu, C., Hu, Y., & Zhang, S. (2019). Applying crowdsourcing techniques in urban planning: A bibliometric analysis of research and practice prospects. *Cities*, *94*, 33–43. https://doi.org/10.1016/j.cities.2019.05.024

Mericskay, B. (2021). *Le crowdsourcing urbain comme nouvelle forme d'engagement citoyen : Etude de cas autour du service de signalement d'anomalies DansMaRue de la ville de Paris*. https://hal.science/hal-03170659

Niu, H., & Silva, E. A. (2020). Crowdsourced Data Mining for Urban Activity: Review of Data Sources, Applications, and Methods. *Journal of Urban Planning and Development*, *146*(2). https://doi.org/10.1061/(asce)up.1943-5444.0000566