DATA601-23X
GIS SPATIAL MOBILITY PROJECT

A SPATIAL ANALYSIS OF SMARTPHONE SMUGGLING IN BRAZILIAN PRISONS

PROJECT BY:

Junlin Jiang
Maritess Pagaduan
University of Canterbury
Master of Applied Data Science

PROJECT SUPERVISORS:

James Williams
University of Canterbury
Mathematics and Statistics

Vanessa Bastos
University of Canterbury
Earth and Environment

Zachary Todd
University of Canterbury
Mathematics and Statistics

ASTRACT / EXECUTIVE SUMMARY

Geospatial analysis has seen significant growth in recent years, finding applications in several fields such as earth science, commerce, urban planning, and healthcare. An active area of research within this discipline involves anomaly detection, focusing on identifying unexpected behaviors or patterns that deviate from the norm. This project explores the application of unsupervised learning techniques to discern anomalies in spatial mobility within major prisons in Rio de Janeiro and Sao Paulo, Brazil. The primary goal is to uncover the extent of illegal smartphone usage within confined areas, which poses potential security threats.

The datasets used in this project are large but have limited attributes. Due to privacy reasons, all movements outside the target boundary are generalized using random offset. We integrated the use of conventional and machine learning approaches to build more features from local information to apply in anomaly detection. Setting a conditional expected daily behavior using rule-based method and assessing its consistency and frequency inside the target location using statistics created movement characteristics that established pattern abnormalities. Applying K-Means and Agglomerative Hierarchical Clustering provided output clusters that learned from multiple inherent and engineered features.

The findings reveal irregular spatial patterns detected from devices within the prison vicinity during unconventional hours. The Gericino Complex shows heightened activity from April to July 2020, while CPD Pinheiros recorded increased movements in 2019 and the latter half of 2020. Usage typically surges after 8 p.m. and diminishes post-midnight at both penitentiaries. While clustering helps visually identify abnormal movements, the cluster quality output from both clustering models is moderately good with room for improvements. Future works using advanced modelling techniques can provide additional insights on precise and distinctive temporal and spatial dynamics within prison environments. These can help improve authority investigation and intervention strategies to address potential safety and security risks.

TABLE OF CONTENTS

1 INTRODUCTION

This report elaborates on an innovative data science project conducted within the framework of Geographic Information Systems (GIS). The project's objective is to explore and map the patterns of smartphone smuggling within the largest prisons in two major Brazilian cities, Rio de Janeiro, and São Paulo. The illegal use of smartphones has become a serious problem in the prisons of these two cities (Franklin, 2014). These devices are not only used for communication among inmates but may also be linked to criminal activities outside the prison walls (Mari, 2013). This project aims to identify critical smuggling patterns by analysing the spatial movement patterns of phones within the prisons.

1.1 Organisation and Domain

Geographic Information Systems (GIS) have emerged as a transformative tool for spatial data analysis and visualization, impacting fields as diverse as urban planning, environmental science, and security. GIS technology facilitates the integration of diverse types of data, overlaying them on geographic maps to provide insightful spatial analysis (Unwin, 1996). This project is a collaborative effort with the Department of Earth and Environmental Sciences at the University of Canterbury, emphasizing the use of GIS for advanced data analysis and visualization in research. One of the challenges presented by this field is the importance that should be placed on data privacy and security, and project members will need to avoid identifying or associating any individuals in the spatial datasets used with the project.

1.2 Goals, Activities, and Research Questions

The primary objective of this project is to use Geographic Information Systems and data analysis methods to reveal the extent and patterns of illegal smartphone usage in the largest prisons in São Paulo and Rio de Janeiro. This involves a comprehensive process of data extraction, filtering, storage, analysis, and visualization.

There are two core key questions that we need to address on this project:

1. What are the spatial patterns and characteristics of mobile device usage within the prisons of São Paulo and Rio de Janeiro?

2. How have the patterns of smartphone usage within these facilities evolved over time?

Understanding these patterns is crucial for assessing the extent of smartphone smuggling and enhancing prison security, which is a critical component of strengthening overall prison security. By identifying and comprehending these patterns, the project aims to facilitate the government in formulating more effective strategies for managing and safeguarding the prison environment.

1.3 Constraints

This project operates under a set of stringent constraints, primarily revolving around data security and ethical considerations. The use of Near Mobility dataset, a comprehensive set of anonymized mobile phone usage data, necessitates rigorous adherence to privacy standards. All data processing steps should be securely managed, and access should be restricted to the project team. Ethical considerations also play a significant role, and any form of data misuse or re-identification is prohibited.

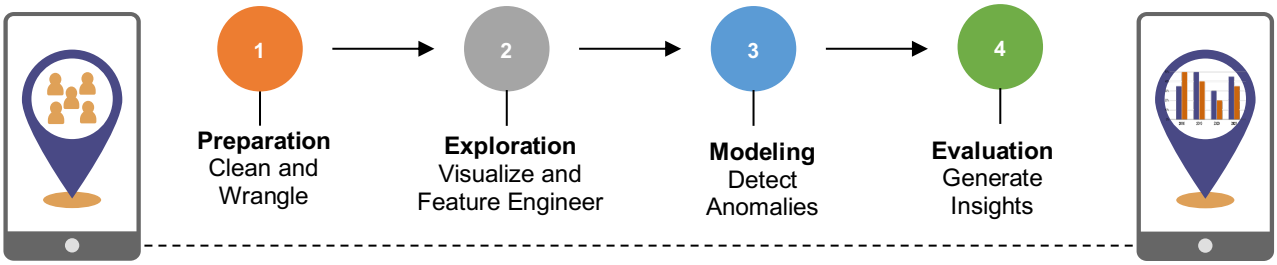
From a technical perspective, the main challenges faced by the project were the size and quality of the data. The vast size of the dataset has resulted in data corruption at the end after the transfer process, necessitating not only data retransfer but also increasing time costs. In particular, the São Paulo dataset, once uncompressed, reaches to 1.8 terabytes in size. Despite efforts at code optimization and the use of multi-threading techniques, data processing and analysis still require a substantial amount of time.

This scale of data processing places an extremely high demand on computational resources and imposes pressure on the project's timeline. The increase in processing time could affect the overall project schedule and delivery of results. In addition, team members must also address technical issues such as data integrity and memory management throughout the process.

2 METHOD

The overarching framework outlining the project's execution is presented in Figure 1.

Figure 1
Project Workflow



2.1 Data Retrieval and Cleansing

Handling large amount of data presents challenges, particularly when faced with resource limitations. Effectively managing large files requires ample RAM to serve as short-term memory for immediate data storage and retrieval by the CPU. Adequate storage capacity is equally vital for preserving both the original data and derived outputs, though it often comes at a cost. The process of collecting and ingesting large files can be time-consuming and

resource intensive. To address the slow transmission and communication of substantial datasets across systems or networks, employing efficient compression and transfer protocols becomes essential. Moreover, due to the considerable demand for computational power and memory, finding the right balance when designing and managing infrastructure for parallel and distributed tasks can be tricky. With human mobility collection, data increases exponentially over time. Careful consideration must be given to effective methods of managing pre-processing, extraction, and storage.

From the raw dataset, the focus is narrowed to the neighborhoods around the Gericino Penitentiary Complex and CPD Pinheiros in Rio de Janeiro and Sao Paulo, Brazil, respectively, from May 2019 to January 2021. Three sequential pipeline jobs deliver the required data:

1) Apply an optimized extraction method that will efficiently access a compressed file format and know the features of these two vast geographical datasets.
2) Retrieve all unique mobile HashIDs that entered the jail perimeter from the collected information.
3) Separate all data point movements from these devices and save them for future operation.

Less than a quarter of the original data is retained from the output of this workflow, which becomes the baseline for deriving new tables that will be used to gain insights.

2.1.1 Handling Large Datasets

The whole dataset for each city is provided as a separate gzip file due to its large size. Data stored in a compressed format typically requires less storage space, while uncompressed data occupies a substantial amount of disk space. Certain libraries, such as Spark, Pandas, and Dask, possess the capability to read compressed data on-the-fly, offering storage efficiency. However, handling compressed files can pose a challenge for parallel processing, as each compressed file is required to be loaded and read in memory. On the contrary, uncompressed formats necessitate in-memory loading, which can significantly consume RAM resources and may incur additional overhead costs such as backup and replication costs.

The choice between compressed and uncompressed formats involves a trade-off between storage efficiency and processing complexity, considering the task requirements. For our project, the latter is more relevant to leverage parallel processing strategies by distributing different tasks among several workers, resulting in an increase in productivity. This option required additional storage allocation as the unzipped file, written to standard text format using the Bash command ‘zcat,’ exceeded the original storage allocation of 2TB from the network server.

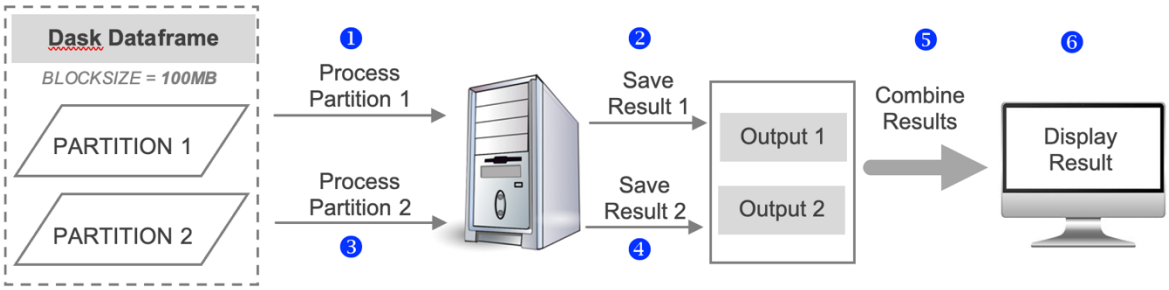
2.1.2 Dataframe Partitioning

Dask, a lighter alternative to Spark, is the primary library used for data extraction with a local client distribution. It works seamlessly with other numeric Python libraries like Pandas (*Comparison to Spark - Dask Documentation*, n.d.). The Dask Client and its interactive

dashboard facilitate advanced asynchronous parallelism, efficiently distributing tasks across computing infrastructure for reading and preprocessing files related to Rio and Sao Paulo (Schmitt, n.d.).

Figure 2

Dask divides large datasets into partitions and works on each partition independently.



Since we have two datasets, we wrapped each major task in a function for reusability. Figure 2 shows how Dask map partitions a dataframe and applies a function to each chunk of data for partition processing and output consolidation. There are two commonly used functions created to manage the unique HashID extraction within the penitentiary. In the first one, each partition undergoes pre-processing to extract all unique mobile hashed IDs associated with individuals who entered the vicinity of the prisons. The prison's boundary is defined using the 'feature_from_place' function from OSMNx, a Python module providing access to OpenStreetMap (OSM), the world's largest collaborative mapping project. Geopandas is then employed to convert each datapoint's coordinates into geospatial information relevant for mapping.

After extracting unique mobile IDs, the next step involves retrieving all associated movements into a dataframe. This process selects activities linked to the unique HashIDs from the original dataset, capturing both prison boundary movements and those occurring outside. It enhances dataset precision by removing irrelevant observations, resulting in a more streamlined file size for efficient analysis.

Given the complexity and resource demands of the process, it's recommended to save the output for documentation and future use, avoiding redundant work in case of notebook failure or errors. We used all three file types depending on dataframe size and nature. Table 1 provides a high-level comparison of these three file formats (Prajapati, 2023).

Table 1

Feature Comparison between CSV, TXT and Parquet

Feature	CSV	TXT	Parquet
Data Structure	Row-based (limited)	Typically row-based (limited	Columnar
File Size	Larger	Typically larger	Smaller
Read and Write Speed	Moderate	Moderate	Faster
Nested Structure	No	No	Yes
Support	Broad Support	Broad Support	Growing Support
Ease of Debugging	Easy	Easy	Moderate
Human Readability	Easy	Easy	Not human-readable

2.1.3 Cleaning and Transforming Data

Throughout the progress of this project, Pandas and Geopandas are utilized for managing derivative datasets, given their efficient handling of file sizes. Data undergoes crucial cleansing and transformation to rectify formatting errors, corrupted data, incompleteness, or duplication. Irrelevant columns like city and point ID are dropped, and duplicate rows are removed to maintain dataset integrity and accuracy, ensuring concise content.

With just five variable columns in the streamlined dataset (Table 2), schema reassignment is important in the data transformation process. This involves realigning column structures and attributes to match the desired schema, transitioning from their original string data type. This delivers and establishes a standardized and well-organized dataset, laying the groundwork for further analyses and operations.

Lastly, when working with spatial data that contains geometrical information like points, lines, or polygons, it is vital to convert a Pandas DataFrame to a Geopandas GeoDataFrame to support spatial operations and geometric data types, which is the heart of our case study. Geopandas is built on top of Pandas, extending its functionality to spatial data such as Shapefile or PostGIS databases. Both stores their data in tabular format, but one major difference between the two is the additional ‘geometry’ column, a GeoSeries, in the former. It also has more functionalities applicable to spatial processing and analysis. Similar APIs are Shapely and GDAL, which have comparable features to Geopandas.

Table 2
Sample extracted data from Sao Paulo and assigned schema

HashID (object)	Lat (float)	Lon (float)	Date (object)	Time (object)
2e7f2d117ea3a6ec66e443246f25efdc7d1ca7b9	-22.86453	-43.413139	12/4/19	8:18:13

2.2 Data Strategies

The soundness and completeness of the derived dataframe are analyzed to understand the shape and structure of the dataset before proceeding further to insight generation and visualization.

2.2.1 Near-Zero Variance

The absence of variability or data dispersion is the result of nearly-zero variance, also known as constant variables, which arise from uniform values in a dataset. An example is the city name column with only one unique variable. To enhance the dataset's efficiency, it is essential to exclude this column by dropping it.

2.2.2 Missing Values and Inaccurate Data Type

Handling missing values early on is crucial, as many algorithms cannot process such data conditions. Incomplete data capture, incorrect data entry, corrupted files, or a lack of data

input are common causes. Functions like 'isnull' or 'isna' help identify missing entries, while libraries like Missingno offer exploratory visualizations of missing data in various formats. Utilizing both methods allowed us to diagnose the datasets, confirming they were normal and complete.

To further validate data integrity, a function iterates through each column's input values, inspecting their alignment with the expected data type. This process returns a summary count of values and identifies any inaccurate formatting, confirming previous validation results. For missing data, strategies include removing records with missing values or using statistical methods for imputation or substitution to maintain dataset completeness (Shababuddin, 2020).

2.2.3 Outliers

Grubbs (1969) defined an outlier as a data point significantly different from others in a sample. Historical practices involves removing outliers for a smoother fit, but contemporary approaches recognize their potential significance as they often provide essential information (Boukerche et al., 2020, 1). In our case study, relevant outliers are HashIDs showing more movement within the prison boundary than outside. By employing conventional and data modelling techniques, including statistical exploration and anomaly detection, we can uncover unexpected patterns and behaviors associated with these gadgets.

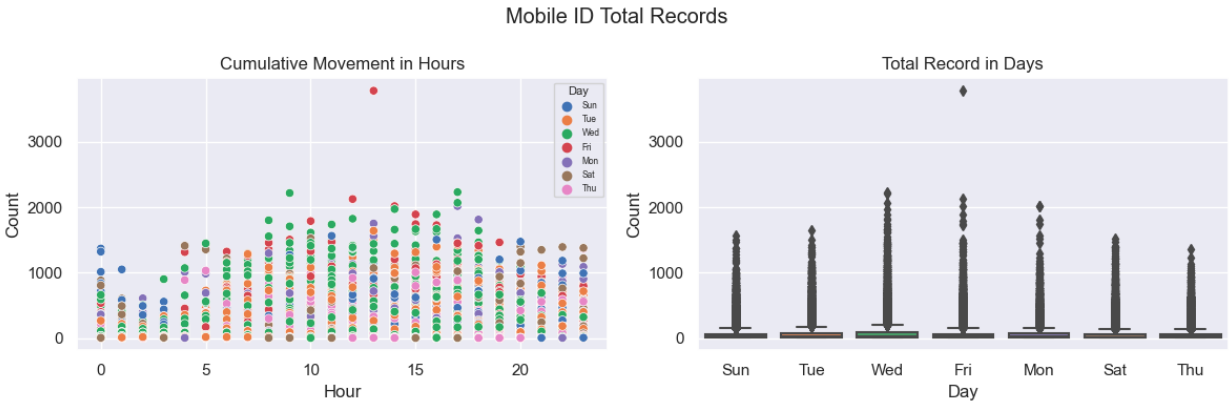
2.2.4 High Cardinality Nominals

When it comes to spatial information, the prevalence of high cardinality is attributed to the multitude of distinct mobile devices documenting location movements and the uniqueness and sparsity of each coordinate. While valuable for granular device tracking, handling numerous smartphones can be challenging and impractical. To address this, feature engineering is employed to classify points as inside or outside a specified area of interest, creating a new column 'Is_inside' with '1' for inside and '0' for outside. Additionally, the time column exhibits high cardinality due to numerous distinct values recording complete time formats across various points. To simplify this, a new attribute is created to extract the hour from the datetime, ranging from 0 to 23. Conversely, the point ID column, serving as a non-repetitive variable providing unique transaction IDs, is also deemed unnecessary and is consequently removed from the dataset.

2.2.5 Distribution Balance

Considering the data characteristics, imbalances are anticipated (Figure 3). Certain smartphones may have more location-sharing apps installed, affecting movement counts. Our focus is on mobile IDs with irregular movements inside prisons. Devices with regular patterns or frequent outdoor locations are excluded through label coding to streamline the dataset and address imbalances.

Figure 3
Cumulative Movement in Hour and Day per Hash ID
The aggregated count of mobile activities per HashID indicates more supplied record in the dataset by active users, an example of which is device outlier seen at 1p.m., Friday.



2.3 Data Ethics

The surge in volume, precision advancement, and accessibility of spatial data enable various sectors to identify trends, movements, and interactions of human and geographical systems at many levels and in various settings such as transportation, urban planning, and healthcare. However, there are risk factors associated with sharing location information, especially when it comes to issues like data quality, data privacy, and ethical issues.

Data quality challenges arise from uncertainty in collected information, leading to errors, inconsistencies, or incompleteness. Metadata communication, including origin, methods, and quality indicators, is vital for data interpretation. Near Intelligence's thorough metadata documentation ensures reliability, guiding our data utilization.

Unlike open data, where ownership and usage policies are unclear, the Near Mobility dataset is sourced from a reputable provider with established legal frameworks to ensure transparency, rights, and control, preventing unauthorized data exploitation. Security measures are crucial for data consumers like us, with password protection mandated for data storage and processing. Access is restricted to the project team, and transmission, disclosure, or sharing through third-party services is prohibited.

We prioritize data confidentiality and refrain from attempting to re-identify mobile HashIDs or associating information with people within the scope of our study because of ethical considerations. To maintain data privacy, the geodata underwent transformations by adding random offset to latitude and longitude of datapoints outside the target boundary, compromising accuracy but minimizing disclosure risk. Additionally, all data is deleted upon project completion to adhere to ethical guidelines and data protection protocols.

2.4 Data Visualization and Modelling

In the absence of labelled datasets, anomaly detection relies on unsupervised algorithms trained on the entire dataset. We employ three approaches: rule-based, statistical analysis, and machine learning modelling to identify movement irregularities within the prison. Spatial analysis tools and visual inspection using maps, charts, or graphs are utilized to examine

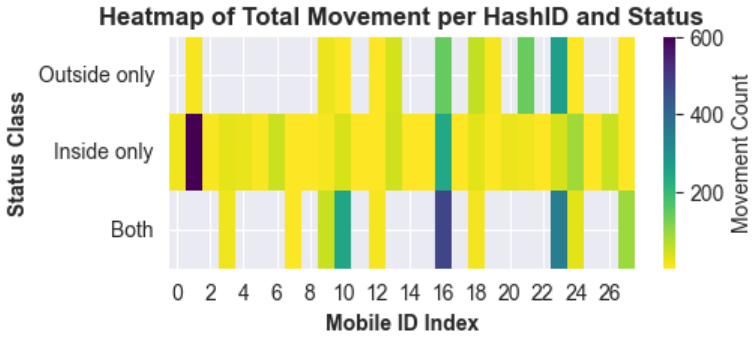
irregular patterns and movements. For a list of Python libraries and tools used in the project, please refer to Appendix 1.

2.4.1 Rule-Based Approach

Given the small features of the dataset, finding anomalies with basic information—like inside or outside a prison—might not establish movement characteristics alone. By specifying the daily anticipated movement, an additional layer of conditions is added to provide the sequence of movements with further context and highlight outliers.

Feature engineering using a multi-classification method generates a new attribute for the daily transaction categorization of each unique HashID. A conditional algorithm is applied based on the values in the 'Is_inside' column. Devices with instances only from outside the prison are labelled as "0," representing movements beyond the prison perimeter. Conversely, devices recording transactions only inside the prison are labelled as "1," considered outliers due to their unexpected pattern of remaining within a restricted area. Devices with regular movements both inside and outside the prison on the same day are labelled as "2." This approach allows for a visual assessment of the dispersion and concentration of presumed outliers using scatterplots or mapping techniques. Figure 4 is a subset matrix of target profiles and their aggregated count per status type.

Figure 4
Degree of Total Movements by Status from HashIDs with 'Inside Only" Count > 0. Higher inside only movements presents irregular activities indicating possible anomaly.

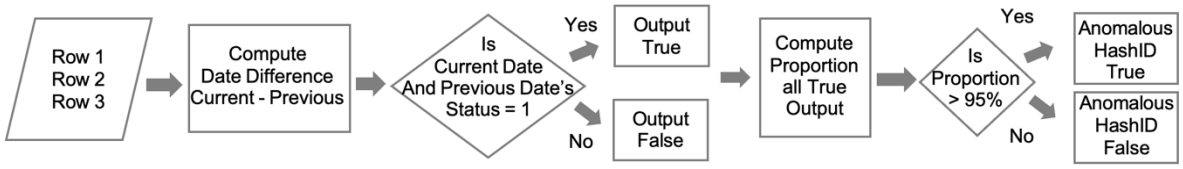


2.4.2 Statistical Analysis

Another approach to detect anomalies is by employing statistical methods, a conventional and straightforward technique. Finding collective anomalies is key to distinguishing if a mobile device is categorized as unauthorized. With spatial features, descriptive statistics focusing on frequency and consistency analysis in combination with the output of the rule-based approach are implemented.

To calculate daily movement, we concentrate on mobile HashIDs with unusual entry records (Status=1) at the target boundary, aggregating their movements by unique HashID and date.

Figure 5
Workflow Diagram for Consistency Analysis



In Figure 5, consistency is validated by comparing current and previous dates, considering a status of 1 (inside only) and zero if the status is 0 or 2. A boolean expression then checks for consecutive dates by verifying if the status is 1 and if the previous date has a similar status, and computes the proportion of consistent counts. If both conditions are met, it's labelled 'True'; otherwise, 'False.' This helps estimate the possibility of HashIDs being anomalous devices.

An additional layer of validation involves examining the reactivation behavior of device owners. By filtering out devices consistently detected inside the prison for extended periods, we capture the first instance of reactivation to compare usage patterns during normal operating hours and beyond. This approach offers insights into the rigor of inspection procedures and provides an estimate of potential smuggled phones.

Another method involves examining the frequency of visits inside the prison by counting how often each HashID appears in the dataset. Similar to the consistency analysis, frequency is calculated for HashIDs with points labelled as status 1. The total movement count per status class is calculated, and the frequency for label 1 is divided by the overall movement count. A conditional check is then performed to ensure that the frequency of status 1 exceeds the combined count of status 0 and 2, and that the proportion of status 1 surpasses a predefined threshold. These conditions help identify instances where status 1 is predominant, suggesting potential irregularities.

2.4.3 Machine Learning Modelling

Handling unlabeled data is one key strength of clustering. In this project, we employed the two commonly used clustering techniques, K-Means and Hierarchical Clustering. The SKLearn library is primarily used to perform cluster assignments for both types.

To augment the features and factor in a time series analogy, hour and day columns are transformed into ordinal values using label encoder. Next, using the mathematical functions sine and cosine, these two attributes are converted into cyclic features, an important step for finding distance and height. The frequency ratio of each status type is also feature-engineered as an added attribute. The data is further preprocessed by standardizing the features to provide a balanced scale with the employment of ‘StandardScaler’.

Using the simple and popular K-Means method, each point in our dataset is inferred using input vectors without any outcome guide. The functionality of this method is explained using the Table 3 left algorithm (Pati et al., 2020, 26).

Table 3
Basic Algorithms of K-Means and Hierarchical Clustering Methods

Algorithm : Simple K-Means Clustering	**Simple Agglomerative Hierarchical Clustering**
Input: X: A database of n data items {x1, x2, x3,..., xn} k: the number of clusters **Output**: A set of k clusters 1 Arbitrarily select k data items as initial cluster centers. 2 Repeat 3 Assign each data item to the cluster center to which the data item is closest. 4 Update the cluster center. 5 Until 6 No change in allotment.	**Input:** X: A database of n data items {x1, x2, x3,..., xn} k: the number of clusters **Output:** A set of k clusters 1 Start by treating each data point as a single cluster. 2 Merge the two closest clusters into a single cluster. 3 Repeat step 2 until there are k clusters remaining.

We begin by utilizing the elbow method, a graphical technique employed to determine the optimal number of clusters (K) in k-means clustering. This method relies on inertia, also known as Within-Cluster Sum of Squares (WSS), which calculates the sum of squared distances between each data point and its centroid across different values of K. The K value that results in the least WSS represents the optimal number of clusters.

To further validate the chosen cluster count, we calculate the silhouette score, ranging from –1 to 1. This score measures the similarity of each point to its cluster compared to other clusters, with higher values indicating better clustering.

Once we have determined the suitable dataset and optimal cluster count, we execute the clustering algorithm. This process involves iteratively assigning each data point to its nearest cluster centroid and recalculating the cluster centers based on the newly combined clusters until the maximum number of iterations is reached.

In contrast to k-means clustering, hierarchical clustering focuses on organizing data into dendrograms, visual representations of the hierarchical relationships between clusters. We specifically apply agglomerative clustering, a bottom-up approach where each data point initially forms its own cluster. Larger clusters are then created by iteratively merging the closest clusters until all data points belong to a single large cluster.

After obtaining the clusters, we assess the quality of the clustering using the silhouette score. Finally, we visualize the output of both clustering models using scatterplots for a clearer perspective of the data distribution and cluster formation.

3 DATA

Smartphones, whether on iOS or Android, collect geospatial human data movement via installed apps, capturing latitude and longitude coordinates. Users grant location sharing consent upon app installation, enabling data gathering through WIFI, GPS, Bluetooth, or

cellular hardware. This data is transformed by aggregators into an advertising identification (e.g., IDFA/AAID), assigning a unique device identifier.

The dataset employed in this study is the Near Mobility dataset, encompassing human movement patterns derived from opt-in shared locational data acquired through smartphone applications. This dataset was sourced from Near Intelligence Pte. Limited, a renowned company specializing in privacy-centric data intelligence known for compiling a large-scale database of high-quality mobile location data.

3.1 Descriptive Statistics

Billions of individuals, knowingly or unconsciously, share their temporal and geographic movements through smartphones, contributing to a global forecast of approximately 7.41 billion mobile users in 2024 (Taylor, 2023). This surge has led to interdisciplinary studies on human mobility, leveraging vast amounts of data from participating devices. Anonymized location data from these devices is a valuable academic resource, enabling analysis of human behavior across cultural, public health, social equality, and urbanization dimensions.

Specific to our case study, the focus is directed towards major penitentiaries in Rio de Janeiro and Sao Paulo, Brazil, using the Near Human Movement Dataset, which captures spatial information spanning from May 2019 to January 2021. The former consists of 2.87 billion lines extracted from a 128GB compressed file, while the latter contains 19.27 billion lines from a 40GB compressed file, both saved in gzip format. Uncompressed, they are 480GB and 1.9TB respectively, which can be challenging due to their large size and resource constraints. The former is received with the trailing portion of the file in a damaged condition, but most of the material remains intact, while the data compression quality of the latter is excellent without any noticeable issues.

The Near User database, integral to our study, incorporates the following key feature variables originally captured in string format and transformed to their appropriate types:

Table 4
Rio de Janeiro and Sao Paulo's Attribute Summary

Feature	Description	Format
HashID	Anonymized or Hashed Mobile Device Identifiers (Apple IDFA or Android AAID), each representing a distinct user. Unique mobile HashID for Rio is 1,668, while 15,743 for Sao Paulo.	object
City	City name where the recorded movement took place.	object
PointID	Indicators using Common Evening Location (CEL) and Common Daytime Location (CDL). CEL approximates frequent mobile positions during non-work hours (between 6 p.m. and 8 a.m. and weekends), while CDL identifies movement during work hours (between 8 a.m. and 6 p.m. and weekdays)	integer
Date and Time	Records when the information was collected.	datetime
Lat and Lon	Coordinates are derived from location Software Development Kits (SDKs) and mobile advertising.	float

The raw datasets are restructured using data wrangling to produce a refined dataframe that concentrates more on pertinent information, increasing computing efficiency and reducing dimensionality. After the removal of unnecessary columns and the extraction of unique

HashIDs from transactions that were found to have at least one entry in the prison area, the text file size significantly decreased from 484GB to 1.2GB with 12,131,247 rows for Rio de Janeiro and from 1.9TB to 32GB with 24,939,993 rows for Sao Paulo.

3.2 Missing Data

In geospatial tool extraction, such as GPS and sensor data recording, it is possible to come across values that are missing. But with the use of the Near Intelligence Mobility dataset, the information that has been obtained is pre-processed, filtered, and compared to reputable resources to produce an output that is statistically sound (*Derivation of Common Evening and Common Daytime Locations*, n.d., 2). Upon our validation of data missingness, the outcome returned zero.

3.3 Outliers

The presence of outliers, data points that significantly deviate from the general pattern or distribution, is observed within the major penitentiaries in Rio de Janeiro and Sao Paulo. These outliers manifested unusual or erratic movements, which may be caused by a technical error, software glitch, or human behavior. The natural and expected flow of movement in a day for staff and visitors to the penitentiary is to enter the prison facility and exit within the operation time or vice versa. Mobile pings with a sequential daily record inside the vicinity may only be detected as anomalous devices. HashIDs with more transactions inside than outside may also be identified as abnormal. Due to the lack of labelled anomaly outlier data, it can be challenging to draw a definite conclusion against devices that exhibited erratic activities within the area of interest.

3.4 Geodata Distribution

A potential disparity in point distribution may occur if some device IDs are disproportionately represented while others are less represented in the collection. Since anomalies are often uncommon compared to normal data points, it can be difficult to spot them because of the limited exposure to such occurrences. Table 5 displays the percentage of mobility by location:

Table 5

Mobility Count of Unique HashIDs with Prison Entry

Dataset:	Rio		
Total:	12,131,247		
Breakdown:			
Label	**Name**	**Count**	**Ratio**
0	Outside	11,965,975	0.9864
1	Inside	165,272	0.0136

Dataset:	CPD Pinheiros		
Total:	319,994,063		
Breakdown:			
Label	**Name**	**Count**	**Ratio**
0	Outside	319,882,834	0.9997
1	Inside	111,229	0.0003

This could be due to the prevalence of specific smartphone models, software apps installed, or user demographics and behaviors. Consistency in data recording can also influence imbalances. For instance, if certain time periods or days have significantly more data points than others, it could introduce an imbalance. This implies that the frequency of data collection varied at different time periods. Another cause of the increase in records is user

activity. Some users may use their smartphones more actively, leading to a higher frequency of recordings. If certain devices have a higher number of installed applications that actively collect location data, it can result in more frequent recordings.

This may be attributed to smartphone types, software apps, user demographics, and behaviors. Data recording consistency might also affect imbalances. If certain days or time periods contain more data points than others, it might lead to disparity. This suggests that the frequency of data gathering varies over time. The record count also increases due to user engagement. Some smartphone users have a higher frequency due to active use. Devices with more location-collecting apps may also track more often.

3.5 Near-Zero-Variance and High Cardinality

The provided datasets consist of a limited number of features. The 'Place' feature serves as a single unique identifier for the location of coordinates, while 'Point ID' comprises unique identifiers, making it highly cardinal. Both place and point ID are considered redundant and can be excluded. However, 'Time' is represented in a 24-hour format with precision to the nearest second, while 'Latitude' and 'Longitude' are recorded in decimal degrees, resulting in a wide range of distinct combinations. Although all three are highly diverse, they are essential attributes in the analysis.

4 RESULTS

This report encapsulates the comprehensive findings from extensive research into the patterns of smartphone smuggling within prison facilities, initially focusing on Rio de Janeiro before extending the analysis to São Paulo. The investigations in this study employed three methods: rule-based method, statistical method, and data modelling using machine learning. These three methods aim to reveal various aspects of spatial and temporal patterns associated with the use of smuggled smartphones within the confines of a prison environment.

4.1 Explanation

4.1.1 Explanation of the Rule-Based Method

Firstly, this report will delve into the rule-based method, using a systematic approach to classify mobile device activities based on predefined rules. This approach helps provide an initial understanding of the data and lays the foundation for more sophisticated statistical and machine learning techniques.

For the rule-based approach, this study initially created a key indicator called "Status" to classify and label the activity patterns of each mobile device (represented by its hash ID) within a given date. This classification is based on the device's movement within and outside the prison boundaries. The definition of the "Status" variable is as follows:

Value 2: Indicates that the device has records both inside and outside the prison on the same day. This suggests that the mobile device associated with this Hash ID has crossed the prison boundary at least once within a 24-hour period. This may indicate legitimate entry and exit or potential smuggling activity.

Value 1: Indicates that the device was detected only inside the prison on that day. This situation may warrant further investigation, as it could imply that the device is being kept inside the prison, possibly for clandestine use by inmates.

value 0: Indicates that the device was only detected outside the prison on that day. Usually this is the normal case, as most devices should only be active outside the prison.

Table 6

Rio Multi-Class: Prison Ingress and Egress Status

Dataset:	Rio		
Total:	12,131,247		
Breakdown:			
Label	**Name**	**Count**	**Ratio**
0	Outside	11,238,861	0.926
1	Inside	827,018	0.068
2	Both	65,368	0.005

Dataset	CPD Pinheiros		
Total:	319,994,063		
Breakdown:			
Label	**Name**	**Count**	**Ratio**
0	Outside	315,253,605	0.985
1	Inside	4,385	0.000
2	Both	4,736,073	0.015

As shown in Table 6, these ratios indicate that the majority of mobile activity occurs outside the prison (Class 0), and this high ratio suggests that most mobile activity takes place outside the prison perimeter, which is expected in any regular scenario. However, there are noteworthy instances of internal activity (Class 1) and cross-boundary movement (Class 2) that may be related to smartphone smuggling. Classes 1 and 2, although small in scale, are critical to this study as they can indicate patterns that are relevant to the core issues being investigated.

Figure 5

Scatter Plot of Actual vs. Generalized Locations Beyond Normal Work & Prison Visit Time (Top: Gericino; Bottom: CPD Pinheiros | Left: Actual; Right: Generalized)

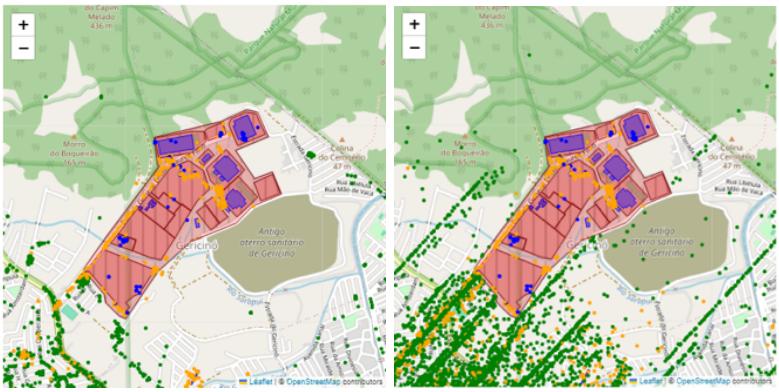


Figure 5 illustrates the locations where mobile devices were detected, categorised according to their position relative to inside and outside the prison boundaries. Patterns of flagging provide initial insight into areas of high activity that may be associated with smuggling behaviour or unauthorised use within the prison.

Figure 6
Points Inside the Boundary and their Concentration (Left: Gericino; Right: CPD Pinheiros)

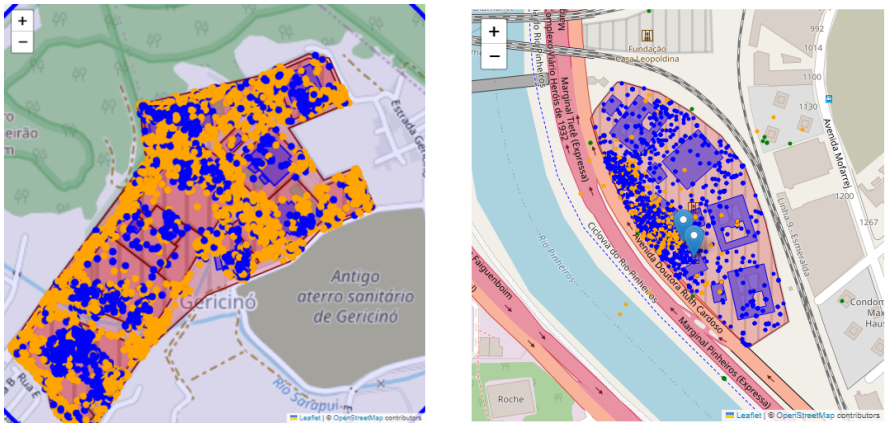


Figure 6 zooms in on points within the prison boundary, with colours distinguishing between states – blue indicating state 1 and orange indicating state 2. This provides a detailed view of device activity, aiding in pinpointing specific locations within the prison that may require further monitoring or investigation.

These visualizations are derived from a rule-based classification system designed to identify patterns in the movement of mobile devices within and outside the prison area. Systematic classification based on device location data at different times reveals potential areas of concern that warrant additional scrutiny. These represent only the first stage of a comprehensive analysis process. Further statistical methods and machine learning models will be applied to refine these findings and extract actionable insights.

4.1.2 Explanation of the Statistical Method

In the ongoing investigation into mobile phone smuggling in prisons, statistical methods are a key component of the data exploration process. Unlike rule-based methods, which rely on

predetermined criteria, statistical methods utilise the inherent characteristics of the data to uncover patterns, anomalies, and relationships that may not be immediately obvious.

The primary objective of statistical methods is to identify potential anomalous patterns that may indicate smuggling activities. This is achieved by examining the frequency, timing, and location of mobile device activity within the prison, with a particular focus on devices that remain within the prison boundaries (State 1). Through this approach, the aim of the study is to quantify the extent of abnormal behaviour and identify specific HashIDs that exhibit irregular patterns warranting further investigation.

Figure 7

Gericino: Time Gap Between Consecutive Records for Each HashID (Status 1)

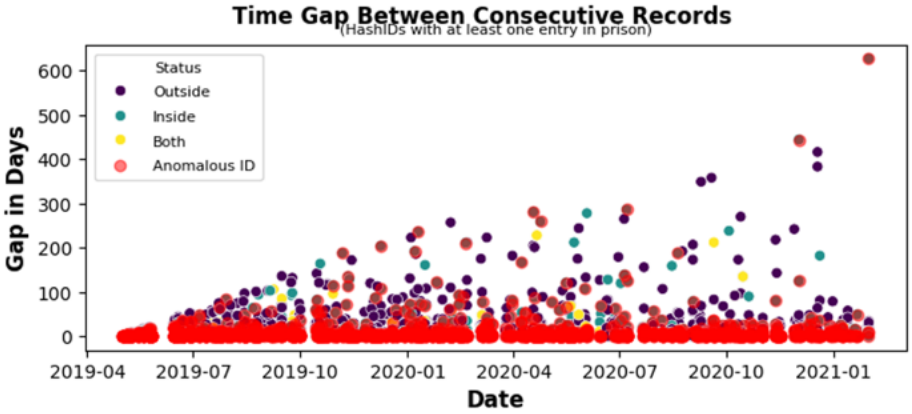


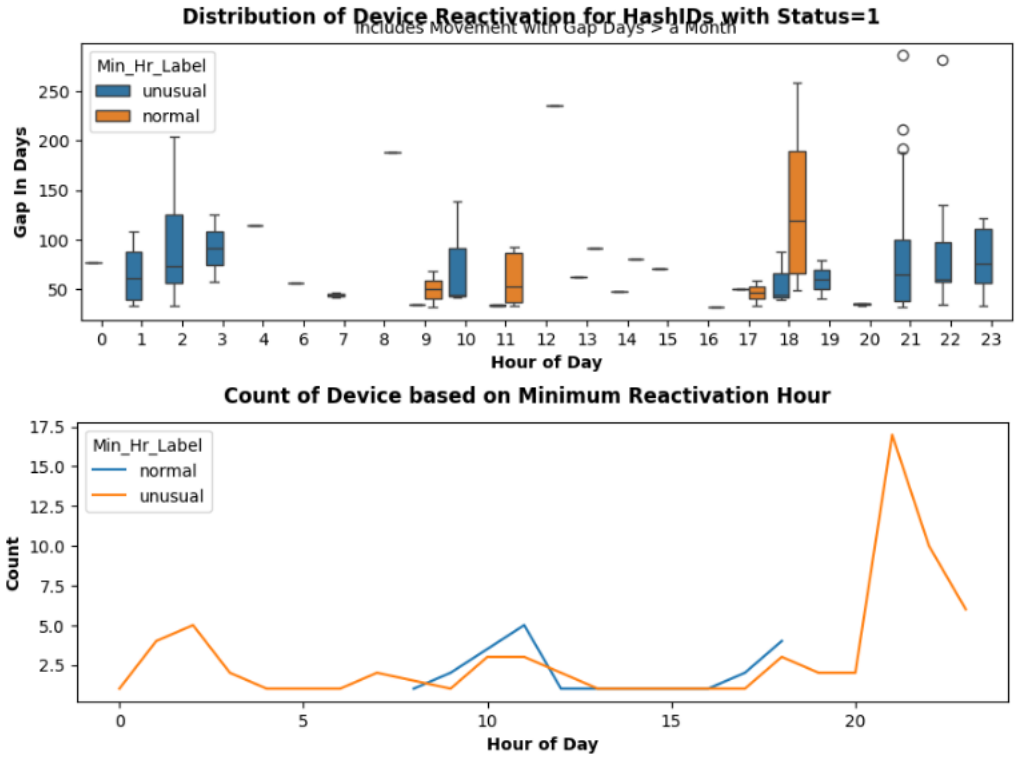
Figure 7 illustrates the time gaps between consecutive records of mobile devices, with a specific focus on those that have been consistently detected inside the prison (Status 1). The variation in time gaps across different HashIDs can provide insights into the temporal patterns of device usage within the prison. Continuous detections (marked in red) indicate persistence and may be associated with the illegal storage of devices within the prison over time.

Moreover, in the statistical analysis section, both a histogram and a line chart are created. The histogram (shown in Appendix 4) displays the distribution of the anomaly ratio, which measures the extent to which individual mobile devices are continuously detected inside the prison over time (Status 1). A higher proportion of anomalies indicates a greater likelihood of abnormal behavior. Proportions skewed towards higher values suggest a subset of devices with highly suspicious activity.

The line chart (displayed in Appendix 5) visualizes the frequency of each status for every HashID, capturing the frequency at which each device is detected in different statuses. HashIDs exhibiting abnormal behavior (those with a higher proportion of Status 1) are distinctly marked. A high frequency of Status 1 detections, especially when it exceeds the combined frequency of Statuses 0 and 2, marks a HashID as potentially anomalous.

Figure 8

Gericino: Device Reactivation after Extended Inactivity



To gain deeper insight into the usage of smuggled devices within the penitentiaries, we examined the reactivation behavior of devices within Gericino, focusing on those consistently detected inside but with extended periods of inactivity. Our observation from the boxplot revealed that some smartphones reactivated not only during unusual hours but also during normal operating times. The line plot below illustrates the number of devices attempting illicit usage per hour. Interestingly, about a quarter of the total device count was used during daytime on normal weekdays.

The statistical method specifically focuses on identifying and describing characteristics related to the time and frequency of device activity that deviates from the expected patterns, with a particular emphasis on transactions associated with HashIDs that have at least one record indicating their presence in prison (status 1). This may include prolonged periods of activity within the prison that are undetectable externally, or frequent transitions between internal and external states, both of which may indicate unauthorized usage or smuggling of mobile devices.

4.1.3 Explanation of Machine Learning Method

The application of clustering techniques is crucial in the analytical process of detecting anomalous patterns within spatial data. These techniques allow for the examination of unlabelled data and facilitate the identification of groups or clusters based on the inherent similarities among the data points. Clustering algorithms are particularly adept at uncovering hidden structures within data, making them invaluable for elucidating the spatial distribution of potentially smuggled phones.

In this context, the study applied two clustering methods, namely K-means clustering and Agglomerative Hierarchical Clustering (AHC), and compared them to discern potential spatial distribution patterns indicative of smartphone smuggling within prison boundaries.

Figure 9

Geospatial Distribution of Mobile Device Clusters Identified by K-Means Clustering (Left: Gericino; Right: CPD Pinheiros)

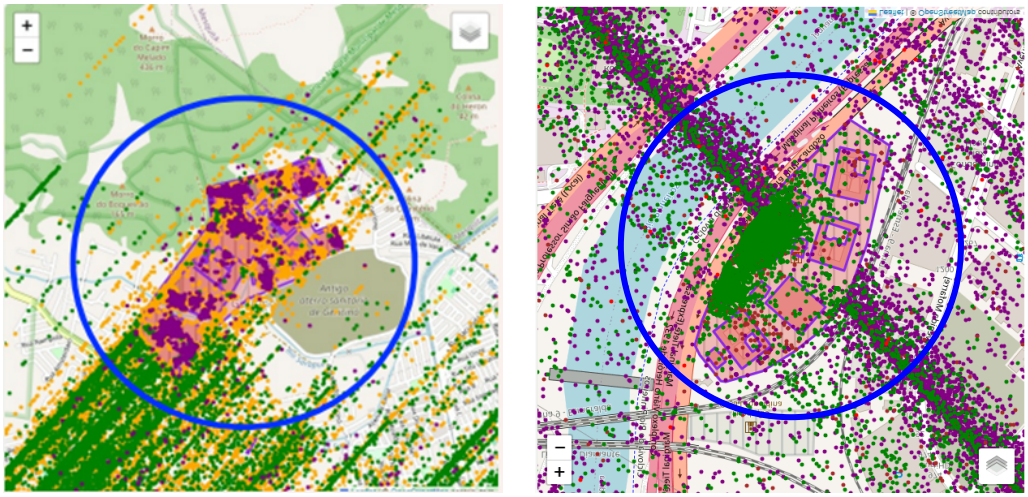


Figure 9 illustrates the spatial distribution of mobile data points near the Gericino Complex and CPD Pinheiros prisons, classified using the k-means clustering algorithm. In this visualization, different clusters are represented by different colours, highlighting clusters of data points with similar features.

Specifically, in the Gericino Complex, as shown in the left graph, the purple cluster stands out significantly, containing a large number of internal points and a few external points. Similarly, for CPD Pinheiros in São Paulo, the green cluster is noteworthy, also consisting of mostly internal prison points and a small portion of external points. This indicates potential aggregations of anomalous data points. These clusters are predominantly concentrated in specific areas of the prisons, suggesting potential hotspots of smartphone smuggling. The clear delineation of these clusters supports the hypothesis that these points are not part of regular traffic flows around the prison areas but rather indicative of clustered anomalous activities potentially associated with smuggling patterns.

Figure 10

Geospatial Distribution of Mobile Device Clusters Identified by Hierarchical Clustering (Left: Gericino; Right: CPD Pinheiros)

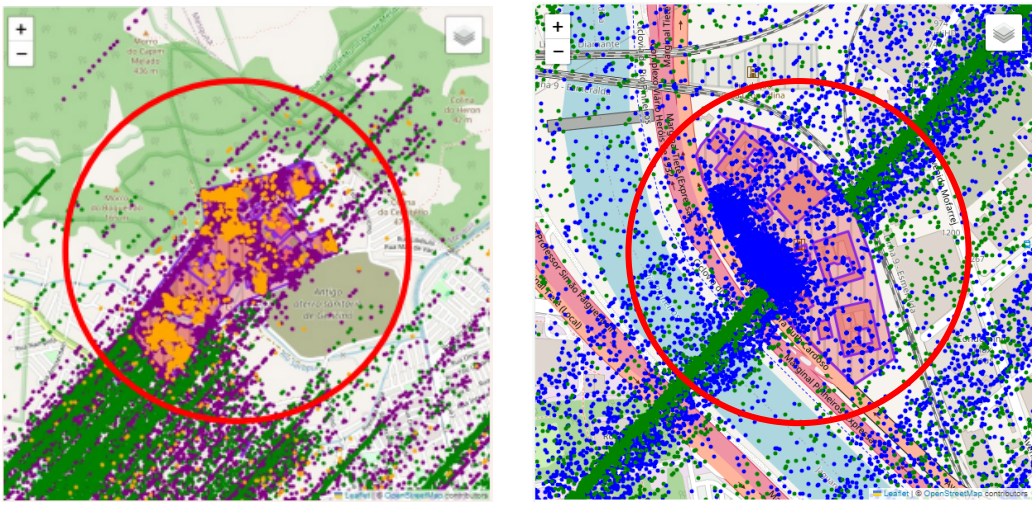


Figure 10 displays the results of the agglomerative hierarchical clustering method regarding the Gericino Complex and CPD Pinheiros prisons. Similar to the k-means results, clustering is color-coded to distinguish the groups identified by the algorithm. Although the number of clusters in hierarchical clustering differs from that of k-means clustering, the spatial distribution patterns of anomalous points in hierarchical clustering appear similar to those of k-means, indicating the presence of anomalous activities potentially related to smartphone smuggling within the prisons. The hierarchical clustering results further reinforce the findings of k-means clustering, providing an additional layer of validation for suspicious areas of interest.

Both k-means and agglomerative hierarchical clustering are powerful tools for identifying patterns in unlabelled data. When applied to geographic spatial data of mobile devices within prison boundaries, these methods can provide insights into spatial patterns indicative of smartphone smuggling. The clustering results offer valuable clues for focused investigative efforts to intercept and prevent such activities from continuing.

In the report, these visualizations and analyses constitute crucial components in understanding the complex dynamics of prohibited smartphone usage within prison facilities. The identified clusters are not merely abstract representations of data points; they represent real-world impacts that may contribute to strengthening prison security measures and protocols.

4.2 Model Selection and Performance Justification

4.2.1 Best Model Determination

The study has developed two machine learning clustering algorithms, namely Agglomerative Hierarchical Clustering (AHC) and K-means clustering, to discern the spatial distribution patterns potentially indicative of smartphone smuggling within prison boundaries. The performance of these models was primarily evaluated using the silhouette coefficient, a metric that measures how similar an object is to its own cluster compared to other clusters.

For the Gericino Complex in Rio de Janeiro, the silhouette coefficient for K-means clustering was recorded at 0.404, while the AHC yielded a silhouette coefficient of 0.362. In contrast,

for CPD Pinheiros in São Paulo, K-means clustering achieved a silhouette coefficient of 0.395, and AHC presented a silhouette coefficient of 0.408.

4.2.2 Justification of Model Choice

Given the observed silhouette coefficients, the K-means clustering model demonstrated marginally superior performance for the Gericino Complex, as evidenced by the higher silhouette score. This suggests that, for this particular prison, K-means clustering was more effective in identifying well-defined and separated clusters of mobile device activity, which is crucial for pinpointing potential smuggling activity. The distinct clustering of data points, particularly in the purple cluster, highlights areas of concentrated activity that could correspond to the loci of unauthorized use.

Conversely, AHC outperformed K-means clustering for CPD Pinheiros, as indicated by the silhouette coefficient. The slight superiority of AHC in this case implies a more cohesive and separated clustering, which is beneficial in identifying subtle patterns of smuggling behaviour potentially overseen in K-means clustering. The AHC's ability to produce a nuanced breakdown and its hierarchical nature may capture the intricate structures of data which are significant for complex environments like CPD Pinheiros.

4.3 Suitability of Chosen Machine Learning Algorithms

4.3.1 Alignment with Data Characteristics

The project’s data, characterized by its spatial-temporal nature and lack of predefined labels, necessitates an analytical approach that can intuitively discern patterns without supervision. The chosen machine learning algorithms, K-means clustering, and Agglomerative Hierarchical Clustering (AHC), align well with these data attributes due to their proficiency in unsupervised learning tasks.

K-means clustering is particularly suited for large datasets like those encountered in this project. It offers a computationally efficient solution to segment mobile device data into distinct groups based on their spatial attributes. This algorithm excels when the clusters tend to be spherical and evenly sized, which aligns with the distribution pattern of mobile devices within the open spaces of a prison complex. Its simplicity in concept and implementation makes K-means an accessible and potent tool for identifying conspicuous aggregations of data points that may signify unauthorized mobile device use.

On the other hand, AHC is well-matched for the project due to its ability to construct a hierarchy of clusters, which is invaluable for understanding the layered structure of mobile device activity within a confined environment. Unlike K-means, AHC does not require a predetermined number of clusters, allowing it to reveal the natural groupings within the data (Murtagh & Contreras, 2017). This characteristic is particularly advantageous when dealing with the potentially complex and subtle patterns of smartphone smuggling in prisons, where the distinction between normal and anomalous behaviour can be nuanced.

4.3.2 Problem-Specific Advantages

One of the research objectives is to detect spatial patterns of mobile device usage within prison boundaries. Unsupervised learning algorithms are particularly apposite for scenarios where potential patterns or structures within the data are not pre-defined. Given the protean and multifarious nature of contraband smartphone distribution networks within correctional settings, these algorithms are invaluable as they facilitate the emergence of previously unrecognized patterns. This inherent capacity to unveil latent structures within the data renders unsupervised learning methodologies an apt selection for this investigative task.

K-means clustering is adept at quickly isolating areas of dense activity, which are potential hotspots for smuggling (Ahmed, Seraj, & Islam, 2020). Its effectiveness is enhanced by the relative homogeneity of the prison environment, where deviations from the norm are more easily detectable against the backdrop of expected behaviour patterns.

Hierarchical clustering approach is beneficial in discerning not just the existence of smuggling activity but also its organizational structure within the prison. It can identify both broad patterns and intricate sub-patterns, offering a comprehensive view of the data that can inform more strategic and targeted interventions.

4.3.3 Consideration of Ethical and Privacy Concerns

In addition to their analytical strengths, both K-means and AHC inherently respect the privacy concerns intrinsic to the project. As unsupervised methods, they do not rely on, nor do they reveal, any individualized information. Instead, they process and analyse data in aggregate form, which is crucial in maintaining the anonymity of the individuals involved. This ensures that the project adheres to ethical standards and privacy regulations while still achieving its objectives.

In the end, the selection of K-means and AHC for this project is underpinned by their compatibility with the data's spatial-temporal nature, their ability to identify patterns within unlabelled datasets, and their congruence with the problem’s requirements. These algorithms provide a balance of simplicity and depth in analysis, facilitating the identification of smuggling patterns while upholding the necessary ethical and privacy standards.

5 CONCLUSION

This study embarked on an exploratory journey, utilizing Geographic Information Systems (GIS) and a suite of analytical methodologies to reveal the clandestine patterns of smartphone usage within the confines of major penitentiaries in Rio de Janeiro and Sao Paulo. Through the meticulous analysis of geospatial data, the project aimed to illuminate the extent of illegal smartphone smuggling, a pressing issue that poses significant security threats within these facilities.

Employing a robust methodological framework that combined rule-based analysis, statistical methods, and machine learning algorithms, the research provided a multifaceted view of the spatial and temporal dynamics of smartphone usage within prison environments. The rule-

based approach laid the groundwork by classifying mobile device activities, revealing substantial movements both within and across prison boundaries. This initial classification highlighted areas of concern, pointing towards potential illicit activities that warranted deeper investigation.

The statistical analysis further delved into the behavioral patterns of these devices, focusing on the frequency and consistency of their presence within prison boundaries. This phase of the study quantified the extent of anomalies, identifying specific devices that exhibited irregular patterns indicative of smuggling activities. The application of machine learning techniques, specifically K-Means and Agglomerative Hierarchical Clustering, refined these insights by revealing the spatial distribution of potentially smuggled phones. These clustering algorithms not only highlighted hotspots of illicit activity but also underscored the complexities and nuances of smartphone smuggling within these settings.

One of the pivotal findings of this research is the identification of irregular spatial patterns and heightened activity periods within the prisons, especially during unconventional hours. These insights point to the sophisticated nature of the smuggling operations and the utilization of smartphones within these facilities. Despite the challenges posed by data privacy and the enormity of the datasets, the study successfully navigated through these complexities to provide actionable intelligence on smartphone smuggling patterns.

Furthermore, the research emphasized the importance of ethical considerations and data privacy, ensuring that the methodologies and analyses respected the anonymity and security of the data subjects. This ethical approach not only reinforced the validity of the findings but also highlighted the potential of GIS and unsupervised learning techniques in addressing complex security issues while adhering to stringent ethical standards.

In conclusion, this project not only achieved its objectives of mapping and understanding the patterns of illegal smartphone usage within major Brazilian prisons but also contributed valuable insights into the application of GIS and data science in security and policy-making. The findings underscore the need for continued investigation and the development of targeted strategies to mitigate the security risks associated with smartphone smuggling. As the project move forward, the methodologies and insights gleaned from this study will undoubtedly serve as a cornerstone for future research and interventions aimed at enhancing prison security and thwarting illicit communications within these facilities.

6 FUTURE WORK

The exploration of illegal smartphone usage patterns within the prisons of Rio de Janeiro and São Paulo paves the way to strengthen and broaden the scope of this research. With the application of Geographic Information Systems (GIS) and a combination of rule-based, statistical, and machine learning methodologies, this study has provided valuable insights into a complex issue facing correctional facilities. Looking ahead, there are several areas where future work can extend and deepen the impact of this research.

Enhancing the analytical framework with more sophisticated machine learning models presents a promising direction for future research. Deep learning techniques, for instance, could offer improved accuracy and granularity in detecting anomalous behaviour within spatial data. These models may reveal subtle patterns and relationships that traditional methods might overlook, thus providing a more nuanced understanding of the dynamics of smartphone smuggling.

Expanding the geographic scope of this study is also essential. The methodologies and insights derived from the Brazilian context could be applied to correctional facilities worldwide, taking into account regional variations in smartphone usage and smuggling strategies.

Moreover, the research framework utilized in this project has potential applications beyond the realm of prison security. Similar geospatial analysis techniques could be employed in other fields requiring the monitoring and analysis of spatial-temporal data, such as wildlife conservation, border security, and public health. By adapting the methods developed in this study to new contexts, researchers can address a wide range of societal challenges.

In conclusion, future work in this project should not only refine and expand the analytical techniques employed but also explore new domains and applications of this research. Collaborative efforts across disciplines will be crucial in driving these advancements, ensuring that the potential of GIS and machine learning in addressing security and societal issues is fully realized. Through continued innovation and application, the work initiated by this study can contribute to more effective strategies for managing prison environments and beyond, enhancing safety and security on a broader scale.

7 ACKNOWLEDGEMENT

We extend our heartfelt gratitude to our project supervisors, the UC MADS staff, family, and friends for their unwavering support throughout this project's development. Special thanks to James Williams, Vanessa Bastos, and Zac Todd for their invaluable insights, practical suggestions, and guidance during times of uncertainty. We also express our gratitude to Steve Gourdie for his assistance with data extraction and technical support, even during the holidays. To our families and friends who provided constant encouragement and inspiration, we are deeply thankful for your unwavering support.

8 REFERENCES

2.3. Clustering — scikit-learn 1.4.0 documentation. (n.d.). Scikit-learn. Retrieved February 12, 2024, from https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

About Our Data. (n.d.). Knowledge Base. Retrieved January 10, 2024, from https://knowledge.near.com/about-our-data

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics, 9*(8), 1295. https://doi.org/10.3390/electronics9081295

Audibert, J. (2022, May 30). Unsupervised anomaly detection in time-series. Retrieved January 15, 2024, from https://theses.hal.science/tel-03681871/document

Babitz, K. (n.d.). *Introduction to k-Means Clustering with scikit-learn in Python*. DataCamp. Retrieved January 10, 2024, from https://www.datacamp.com/tutorial/k-means-clustering-python

Bancroft, T. (n.d.). *Chapter 1 Introduction to GIS | Intro to GIS and Spatial Analysis*. mgimond.github.io. Retrieved December 1, 2023, from https://mgimond.github.io/Spatial/introGIS.html

Boukerche, A., Zheng, L., & Alfandi, O. (2020, June 12). Outlier Detection: Methods, Models, and Classification. *ACM Computing Surveys*, *Volume 53*(Issue 3 Article No.: 55), pp 1–37. https://doi.org/10.1145/3381028

Casado, L., & Londoño, E. (2020, November 30). *As Coronavirus Strikes Prisons, Hundreds of Thousands Are Released (Published 2020)*. The New York Times. Retrieved February 15, 2024, from https://www.nytimes.com/2020/04/26/world/americas/coronavirus-brazil-prisons.html

ChatGPT. On code debugging, code queries on Pandas to Dask conversion, content proofreading. *https://chat.openai.com/*

Comparison to Spark — Dask documentation. (n.d.). Dask documentation. Retrieved January 15, 2024, from https://docs.dask.org/en/latest/spark.html

Derivation of Common Evening and Common Daytime Locations. (n.d.). Knowledge Base. https://knowledge.near.com/understanding-mobile-location-data-1

Franklin, J. (2014, April 15). Cell to Cell: How Smuggled Mobile Phones Are Rewiring Brazil's Prisons. *VICE.* https://www.vice.com/en/article/4x3wjw/cell-to-cell-how-smuggled-mobile-phones-are-rewiring-brazils-prisons

Grubbs, F. E. (1968, February). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *Vol. 11*(No. 1), pp. 1-21. https://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657

Guide, S. (2020, October 17). *Handling Missing Data In Large Datasets | by Md Shababuddin*. Medium. Retrieved January 23, 2024, from

https://medium.com/@shababuddinmd/handling-missing-data-in-large-datasets-cfe3a493f61a

Hirst, T. (2018, June 29). *Working With OpenStreetMap Roads Data Using osmnx – OUseful.Info, the blog…*. OUseful.Info, the blog... Retrieved November 22, 2023, from https://blog.ouseful.info/2018/06/29/working-with-openstreetmap-roads-data-using-osmnx/

Mari, A. (2013, August 12). São Paulo to spend R$1bi to block cell phones in prisons. *ZDNet.* https://www.zdnet.com/article/sao-paulo-to-spend-r1bi-to-block-cell-phones-in-prisons/

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7*(6), e1219. https://doi.org/10.1002/widm.1219

Natarajan, M. (2023, October 16). *Deciphering Optimal Clusters: Elbow Method vs. Silhouette Method | by Megha Natarajan*. Medium. Retrieved February 15, 2024, from https://medium.com/@megha.natarajan/deciphering-optimal-clusters-elbow-method-vs-silhouette-method-7e311c604201

Pandas and Geopandas -modules — GeoPython - AutoGIS 1 documentation. (n.d.). Automating GIS Processes. Retrieved February 8, 2024, from https://automating-gis-processes.github.io/2016/Lesson2-overview-pandas-geopandas.html

Pati, B., Panigrahi, C. R., Buyya, R., & Li, K.-C. (Eds.). (2020). *Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume 1*. Springer Nature Singapore. https://doi.org/10.1007/978-981-15-1081-6

Patro, P. (n.d.). *Time Series Anomaly Detection*. Kaggle. Retrieved February 15, 2024, from https://www.kaggle.com/code/pratyushpatro/time-series-anomaly-detection

Plotting with Folium — GeoPandas 0+untagged.50.g9a9f097.dirty documentation. (n.d.). GeoPandas. Retrieved November 22, 2023, from https://geopandas.org/en/stable/gallery/plotting_with_folium.html

Prajapati, S. (2023, June 13). *The Perfect File Format Unveiled: Parquet vs. CSV*. (2023, June 13). LinkedIn. Retrieved January 16, 2024, from https://www.linkedin.com/pulse/perfect-file-format-unveiled-parquet-vs-csv-shailendra-prajapati

Schmitt, M. (n.d.). *Understanding Dask Architecture: Client, Scheduler, Workers*. Data Revenue. Retrieved January 15, 2024, from https://www.datarevenue.com/en-blog/understanding-dask-architecture-client-scheduler-workers

Unwin, D. J. (1996). GIS, spatial analysis and spatial statistics. *Progress in Human Geography, 20(*4), 540-551. https://doi.org/10.1177/030913259602000408

Visitar preso do Sistema Penitenciário Federal. (DEPEN). (n.d.). Governo Federal. Retrieved November 24, 2023, from https://www.gov.br/pt-br/servicos/visitar-preso-do-sistema-penitenciario-federal

9 APPENDICES

Appendix 1:

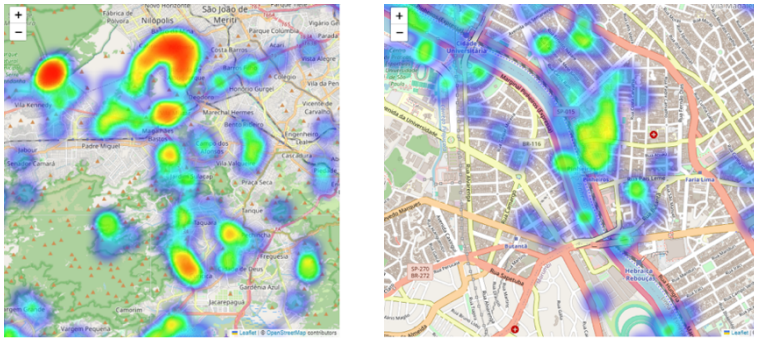
List of Tools and Libraries used for project coding

Preparation		Exploration		Model & Evaluation
For Data Extraction • Dask Client • Dask DataFrame	For Data Wrangling • Pandas • Geopandas • Shapely • MissingNo & Random	For Map Visualization • Folium • OSMNx	For Graphs & Charts • Seaborn • Matplotlib • Holoview	Scikit Learn

Appendix 2:

Heatmapping 1M subset of Mobile Records

Left: Gericino
Right: CPD Pinheiros



Appendix 3:

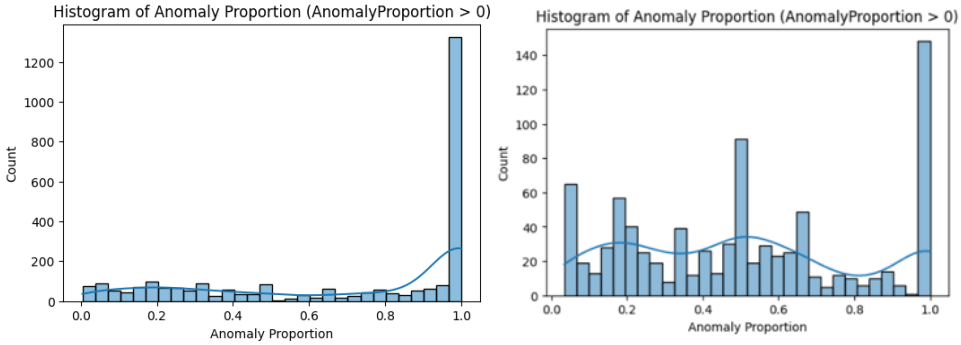
CPD Pinheiros: Subset of Movements labeled as "Inside Only" and "Outside Only"

Left: Actual Location
Right: Generalized

Appendix 4:

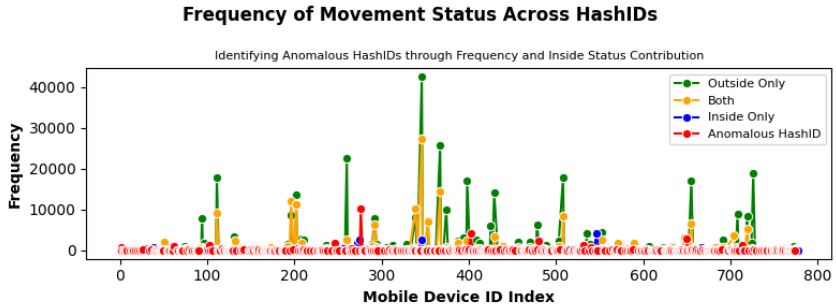
Histogram of Anomaly Proportion (AnomalyProportion > 0) based on Gap in Days Analysis. A significant number of profiles are deemed anomalous.

Left: Gericino
Right: CPD Pinheiros



Appendix 5:

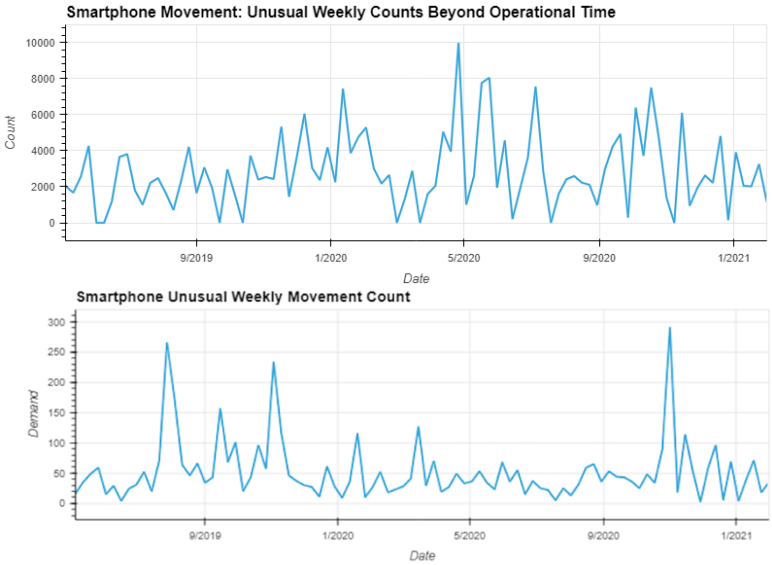
Status Frequency Over HashIDs focused on HashIDs with Status 1 in Gericino Complex
It contains the total records pings per status for each HashIDs. Points overlayed with red marks are deemed anomalous due to high proportion of staying inside the vicinity.



Appendix 6:

Timeseries of Total Movement by Target HashIDs beyond Operating Time
Total recorded pings for target mobile device outside of the normal visiting and working time in prison.

Top: Gericino
Bottom: CPD Pinheiros



Appendix 7:

Timeseries of Total Hourly Movement by Day of Target HashIDs Beyond Operating Time
Total recorded capturing only 6pm up to 8am on weekdays and the whole weekend.

Top: Gericino
Bottom: CPD Pinheiros