UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

A Public Transit Network Analysis with Data Processing

Gerar Francis Quispe Torres

Trabalho de Conclusão de Curso - MBA em Inteligência Artificial e Big Data

ICMC - UNIVERSIDADE DE SÃO PAULO

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: ______________________________

Gerar Francis Quispe Torres

A Public Transit Network Analysis with Data Processing

Final Paper submitted to the Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, in partial fulfillment of the requirements for the degree of Master of Business Administration in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence and Big Data

Advisor: Prof. Phd. Jean Roberto Ponciano

Original version

São Carlos
2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Q8p

Quispe Torres, Gerar Francis
A Public Transit Network Analysis with Data Processing / Gerar Francis Quispe Torres; orientador Jean Roberto Ponciano. -- São Carlos, 2024.
50 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Smart Public Transportation. 2. Trajectory Anomaly Detection. 3. Urban Transportation Efficiency Analysis. 4. Trajectory Descriptor. 5. GPS Data Processing. I. Ponciano, Jean Roberto, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

I want to dedicate this work to my family, especially to my mother, whose unwavering support has been a constant source of strength and inspiration. I also dedicate this achievement to my three brothers, whose wise advice and companionship have been essential in overcoming challenging times. Thanks to you, this journey has been more meaningful and fulfilling.

ACKNOWLEDGEMENTS

I want to thank all the staff that make this program possible, giving me the opportunity to be a part of it. I am deeply grateful for the full scholarship I received to join this program, which has been a crucial support in pursuing my academic goals.

I would like to express my sincere gratitude to my advisor, Professor Jean Roberto Ponciano, for his patience and invaluable advice throughout this journey. I also extend my thanks to Professor Solange Rezender, who runs the program, and to all the teachers in the program for their dedication and commitment to our learning.

Lastly, I want to thank my new colleagues who are part of this program, always willing to share their knowledge and answer questions in the chat. Your camaraderie and support have enriched my experience and made this journey even more rewarding.

Thank you all for being a part of this significant chapter in my life.

“ Let me tell you something you already know. The world ain’t all sunshine and rainbows. It’s a very mean and nasty place, and I don’t care how tough you are, it will beat you to your knees and keep you there permanently if you let it. You, me, or nobody is gonna hit as hard as life. ***But it ain’t about how hard you hit, it’s about how hard you can get hit and keep moving forward. How much you can take and keep moving forward.*** *That’s how winning is done! ”*

– Rocky Balboa

ABSTRACT

Quispe-Torres, G.F. **A Public Transit Network Analysis with Data Processing**. 2024. 50 p. Trabalho de conclusão de curso (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Urban public transportation faces growing challenges related to operational efficiency, particularly in the context of developing cities. This research analyzes public transportation networks, focusing on detecting anomalies in bus routes and extracting monthly efficiency-related insights. The main objective is to enhance urban mobility through data-driven approaches, addressing the challenges of managing large datasets generated by Smart Public Transportation (SPT) systems. Using the Try-Bus dataset, which includes extensive GPS records of bus locations, a comprehensive methodology was developed to achieve these aims. The study conducted two key analyses: anomaly detection to identify deviations from predefined routes, and efficiency analysis to evaluate the operational performance of bus fleets over time. The results indicate significant operational deviations and offer actionable insights for improving public transportation services in Cusco, Peru. This research contributes to the optimization of urban transportation through advanced data analysis and processing techniques, supporting decision-makers in enhancing service quality and efficiency.

Keywords: Smart Public Transportation. Trajectory Anomaly Detection. Efficiency Analysis.

RESUMO

Quispe-Torres, G.F. **A Public Transit Network Analysis with Data Processing**. 2024. 50 p. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

O transporte público urbano enfrenta desafios crescentes relacionados à eficiência operacional, especialmente no contexto de cidades em desenvolvimento. Esta pesquisa analisa as redes de transporte público, com foco na detecção de anomalias nas rotas de ônibus e na extração de insights mensais relacionados à eficiência. O principal objetivo é melhorar a mobilidade urbana por meio de abordagens baseadas em dados, abordando os desafios de gerenciar grandes conjuntos de dados gerados pelos sistemas de Transporte Público Inteligente (TPI). Utilizando o conjunto de dados Try-Bus, que inclui extensos registros de GPS das localizações dos ônibus, foi desenvolvida uma metodologia abrangente para alcançar esses objetivos. O estudo conduziu duas análises principais: detecção de anomalias para identificar desvios das rotas pré-definidas e análise de eficiência para avaliar o desempenho operacional das frotas de ônibus ao longo do tempo. Os resultados indicam desvios operacionais significativos e oferecem insights práticos para melhorar os serviços de transporte público em Cusco, Perú. Esta pesquisa contribui para a otimização do transporte urbano por meio de técnicas avançadas de análise e processamento de dados, apoiando os tomadores de decisão na melhoria da qualidade e da eficiência dos serviços.

Palavras-chave: Transporte Público Inteligente. Detecção de Trajectórias Anomalas. Análise de Eficiência.

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

ACM	Association for Computing Machinery
AFC	Automated Fare Collection
AP	Affinity Propagation
APC	Automatic Passenger Counter
ASYU	Innovation in Intelligent and Machine Learning Conference
AUC	Area Under the Curve
AVL	Automatic Vehicle Location
CLEI	Centro Latinoamericano de Estudios en Informática
DEA	Data Envelopment Analysis
EDA	Exploratory Data Analysis
ETL	Extract, Transform, Load
EU	European Union
GPS	Global Positioning System
ICMC	Institute of Mathematics and Computer Sciences
ITS	Intelligent Transportation System
KNN	K-Nearest Neighbors
SPT	Smart Public transportation
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos

CONTENTS

1 INTRODUCTION

1.1 Motivation and Justification

We are motivated to conduct this study by the fact that Smart public transportation systems (SPT) are transforming urban mobility thanks to the affordability of electronic devices that generate data for service improvements. These technology-driven systems aim to enhance efficiency, safety, and passenger comfort. Features like real-time vehicle tracking, contactless payment, data collection, traffic light prioritization, and incentives for electric or hybrid vehicles are all part of the SPT toolbox. As a growing trend, smart public transportation is attracting investment worldwide from cities seeking to modernize their transit networks.

The need for our study arises from the fact that automating anomaly detection, efficiency analysis, and mobility pattern analysis tasks significantly facilitates the supervision and management of tasks in a company's bus fleet. This information is constantly being generated and is growing in volume, but due to the lack of processing, valuable knowledge-rich information is being lost.

1.2 The problem

Abundant data processing poses a significant challenge due to the constraints in resources for both processing and extracting truthful and relevant knowledge. The sheer volume of data and the complexity of extracting meaningful insights make it a demanding task.

Automating a process necessitates the establishment of well-established methodologies to ensure their infallibility within a real-world application that encounters numerous contingencies. Robust and reliable automation requires rigorous testing and validation to guarantee its effectiveness in handling diverse scenarios and unexpected situations.

Oversight tasks for business owners can be tedious and inefficient when not equipped with a Smart Public Transportation (SPT) system. The lack of real-time data and automated insights makes it difficult to effectively monitor and manage operations, leading to time-consuming manual tasks and potential inefficiencies.

1.3 Objectives

The main objective of our study focuses on performing an analysis of a public transportation network, this includes an efficiency analysis, an analysis of mobility patterns, and an anomaly detection analysis on bus routes. In achieving these objectives, we identify peak times and areas of greatest and least demand for the public transportation service of the city of Cusco in Perú.

1.3.1 Specific objectives

The specific objectives of the present study are:

1. An anomaly detection analysis, this anomaly detection will serve to detect when a bus deviates from its route automatically. With this information, problems can be detected within the routes that the buses follow.

2. An analysis of the efficiency of a public transport network. This objective aims to identify bottlenecks, routes, and city areas with lower and higher demand. The factors that influence the variability of travel time are obtained.

1.4 Text organization

Our text is organized as come next, on Chapter 2 is describing concepts that help us to better understand our study since these concepts will be mentioned as we go through the text reading, the following explanation aims to elucidate these concepts straightforwardly.

Relevant studies associated with public transit networks are presented and discussed in Chapter 3, the review begins with the topic of Anomaly Detection in public transport networks. It then emphasizes on study works which process GPS data and transform this in trajectories.

Subsequently, we discussed efficiency and pattern analysis in the transit network found in the scholarly articles. The concepts are involved in GPS data processing modeled as trajectories. Chapter 4 will begin presenting the methodology used in this study and a description of all methods involved in it. We added illustrative graphics for better explanation.

Experiments and results are detailed in Chapter 5. It includes graphics created based on the analysis, the validation methodology used in our study, and some limits of the method used in this study. The results are presented as qualitative and corroborated with real-world information. Chapter 6 concludes the study with three sub-topics the applicability of our study in other cities, the contribution to public transport analysis, and reflections on challenges encountered.

Finally, some final remarks and directions for future study works are in Chapter 7.

2 BACKGROUND

2.1 Intelligent transportation systems (ITS)

Intelligent transportation systems represent an advancement in transportation infrastructure, aiming to provide innovative services that cover various modes of transport and traffic management. Their primary function lies in empowering users with improved information, enabling them to make safer, more coordinated, and strategically optimized use of transportation networks. Overall, ITS aims to make transportation safer and more efficient(Verma *et al.*, 2024).

These systems leverage up-to-date technologies to achieve this objective. Examples include automated emergency service notification upon detection of an accident, the implementation of camera systems for enforcing traffic laws, and the deployment of dynamic signage that adjusts speed limits based on the most common conditions (Gordon, 2016).

The European Union defines ITS as applying information and communication technologies to improve road transportation. This includes infrastructure, vehicles, users, traffic management, and mobility management. The ITS systems do not necessarily use machine learning models, some applications only use algorithms that can help intelligently(DIRECTIVE. . . , 2010).

ITS can help reduce congestion, improve mobility, save lives, and optimize the performance of existing infrastructure and road users. Its prime essence is to facilitate easier and more efficient data collection, data transfer, communications between the transport infrastructure users, and assessment of facilities (Dighe; Nikam; Markad, 2024).

2.2 Efficiency analysis in a public transport network

Efficiency analysis in a public transport network involves the measurement and assessment of the performance of a public transport system, considering various dimensions such as economic efficiency, technical efficiency, and allocative efficiency. The analysis helps evaluate the degree to which results on efficiency and economies depend on the type of function and specification form estimated, and whether the differences frequently reported can be attributed to other factors. The findings suggest that efficiency assessment methods yield similar mean efficiencies but different efficiency distributions, and a stronger result was reached for effectiveness where both mean scores and the distribution of scores vary significantly between methodologies. This finding suggests that peer group comparisons and performance assessments based only on averages may yield erroneous findings and lead to skewed policy recommendations, (Karlaftis; Tsamboulas, 2012).

The efficiency analysis in a public transport network also involves determining the relative efficiency of transportation routes from one point to another. This is typically done using Data Envelopment Analysis (DEA), a method that compares operating results and investments to achieve maximal business success. In the context of public transport, DEA can be used to determine the optimal transport route by minimizing input values such as transport costs, external costs, and transport time while maintaining the same output value, which is the traveled distance. By using DEA, decision-makers can identify the most efficient route to a preferred destination and make informed decisions about transportation planning and decision-making, (Vukić *et al.*, 2020).

2.3 Mobility patterns

Mobility patterns are the recurrent behaviors observed in travel within a city or region. These patterns can be analyzed based on various variables such as the origin and destination of trips, the mode of transport used, the time of day, the frequency and duration of journeys.

The analysis of these patterns is essential for the planning and management of public transport, as it allows to identify the needs of the population, improve the efficiency of the system and promote its responsible use.

Identifying the needs of the population: Knowing mobility patterns helps to better understand the needs of public transport users. This allows to determine where and when more investment is needed in infrastructure and services, such as the construction of new stations or the extension of the frequency of services at certain times(González; Anapolsky, 2022; Puebla *et al.*, 2020).

Improving the efficiency of public transport: Knowledge of mobility patterns allows to optimize routes, frequencies and schedules of public transport according to demand. This can improve the efficiency of the system, reduce travel times and reduce traffic congestion(Wu *et al.*, 2024; Lieshout; Dalmeijer, 2024; Borges *et al.*, 2024).

Promoting the use of public transport: Implementing policies and strategies that encourage the responsible use of public transport, such as lower fares or improvements in service quality, can help reduce traffic congestion and improve air quality(Guo *et al.*, 2020).

There are various tools for analyzing mobility patterns, such as mobility surveys, ticket sales data or sensors. In this study, GPS sensors will be used to collect data on the mobility patterns of the population.

3 RELATED WORKS

3.1 Anomaly detection in public transport networks

The study of Quispe-Torres *et al.* (2022) is about detecting anomalies in trajectories based on their morphological characteristics. The text proposes a two-stage methodology: (1) comparative analysis of the performance of two descriptors to group similar trajectories, and (2) detection of anomalies in the trajectories based on their similarities. The text defines Wavelet and Fourier transforms as trajectory descriptors for generating features and detecting anomalies. The text emphasizes measuring performance in describing the feature space of coefficients using unsupervised learning, specifically clustering techniques, to create subsets and identify irregular ones. The text demonstrates the implications and comparative analysis of their study through multiple experiments with synthetic and real data.

The study of Wang *et al.* (2016) proposes a method for detecting traffic anomalies, such as accidents or congestion, in real time using GPS data. The method involves mapping raw GPS trajectories to road segments and calculating traffic features, such as speed and flow, for each road segment in a given time slot. The traffic features used for anomaly detection include travel speed, traffic density, and object outflow. Travel speed refers to the average speed of moving objects on a road segment. A sudden decrease in travel speed may indicate a traffic anomaly, such as a traffic accident or congestion. Traffic density refers to the number of moving objects on a road segment within a given period. An increase in traffic density may indicate a traffic anomaly, such as a road closure or heavy traffic due to an event. Object outflow refers to the number of moving objects leaving a road segment within a given time. A sudden decrease in object outflow may indicate a traffic anomaly, such as a road closure or heavy congestion. These traffic features are calculated based on the GPS trajectories of moving objects, which are mapped to road segments using a map-matching algorithm. Anomaly candidates are then retrieved by checking the traffic flow acceleration based on their proposed method and examined further using the density change ratio or moving objects' outflow ratio to infer traffic anomalies. The effectiveness and efficiency of the proposed method were validated through extensive experimentation, showing that it can detect traffic anomalies efficiently and earlier.

3.2 Efficiency analysis in public transport network

The study Mazloumi, Currie and Rose (2010) focuses on the importance of the reliability of the transport service, which has been studied mainly from the perspective of passengers waiting at stops. The daily variability of travel time also deteriorates the reliability of the service, but it is not a well-investigated area in the literature due to the lack of complete data sets on travel times by bus. Explore nature and how travel time is distributed to different winds of departure time at different times of the day. The factors that cause the variability of public transport travel time are also explored using linear regression analysis in this study.

3.3 Pattern analysis in mobilities

The study of Özgün, Günay and Başaran (2021) focuses on how new technologies, such as Automated Fare Collection (AFC), Automatic Passenger Counter (APC), Automated Vehicle Location (AVL), and Geographic Positioning Systems (GPS)), have improved data, information collection and accelerated research in transit systems.

4 METHODOLOGY

The Figure 1 illustrates an academic pipeline that outlines the key phases of a research project focused on data analysis. This pipeline consists of five sequential and interconnected phases that organize the workflow within an academic and scientific context. Our study specifically focuses on phases 2, 3, and 4. Following, each phase is briefly described, emphasizing its role within the project.

4.1 Approach

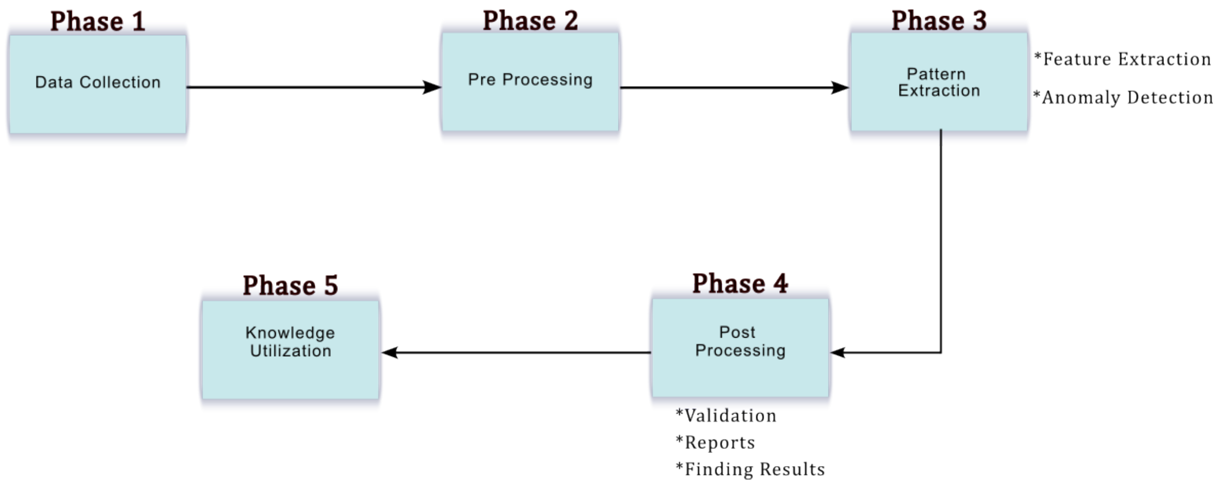


Figure 1 – Pipeline used in our study (Image of own creation).

1. Data Collection. This initial phase is fundamental and focuses on gathering the necessary data for analysis. For this study, data was obtained using GPS devices installed on buses, providing precise information about their movements and routes. Without this phase, progress would be impossible, as the analysis entirely depends on the availability of relevant data for processing.

2. Preprocessing. In this phase, the collected data undergoes rigorous treatment to ensure its quality and consistency. This includes noise removal and correction of inconsistencies. Additionally, the structure and format of the data are verified to ensure they are suitable for subsequent stages. This step is crucial in this study, as proper preprocessing significantly impacts the accuracy and relevance of the final results.

3. Pattern Extraction. This phase involves analyzing the processed data to identify patterns and extract key features that provide insights into the behavior of the

evaluated systems. Techniques such as feature extraction and anomaly detection are applied to uncover significant trends, relationships, or irregularities. In this study, this phase is central, as it employs statistical methods and machine learning algorithms to generate valuable information.

4. Postprocessing. After completing the analysis, this phase focuses on interpreting the results and presenting them in a structured and comprehensible manner. In this study reports and qualitative results are generated in this phase. Then key conclusions are drawn to validate the utility of the findings. This phase is essential, as it ensures that the results are effectively communicated, facilitating their interpretation and evaluating their impact within the study area.

5. Knowledge Utilization. The final phase aims to apply the obtained results in practical or theoretical contexts, maximizing their impact. Although this study does not directly address this phase, it suggests future work focusing on the implementation of practical solutions, the publication of relevant findings, or the design of strategies based on the analyses performed. The primary goal is to transform the analysis results into tangible benefits for industry, academia, or society, ensuring that the generated knowledge has a real-world impact.

4.1.1 Try-Bus Dataset

For our experiments, we used the Try-Bus Dataset, a collection of millions of records generated by GPS devices. This dataset was first introduced by Mamani (2018), whose work involved developing an Intelligent Transportation System (ITS) to monitor urban transport buses.

The Try-Bus dataset contains different public transport routes, corresponding to different transport companies within a city. This dataset also considers differences between cycles that buses complete during the day, since this dataset notes when a route cycle is completed, a cycle includes the outbound and return routes. The monitoring process carried out to create these datasets is constant, the GPS devices used operate twenty-four hours a day and the georeferenced data points are produced every 30 seconds.

Due to the large amount of monitoring data, a sampling process is needed. To do this, the data selection criteria are as follows: Only one transport company is chosen, in this case, we selected a transport company that has the biggest number of buses. We chose as the study date, data from 2019-04-01 to 2019-04-30 and the time between five and twenty-two hours, since the work schedule of the company, and also it is the schedule in which the buses are in motion because the company stops working outside these hours.

The filtered dataset consisting of a maximum of 1.5 million records, which were then transformed into 15,013 trajectories. The trajectories are necessary because we need to

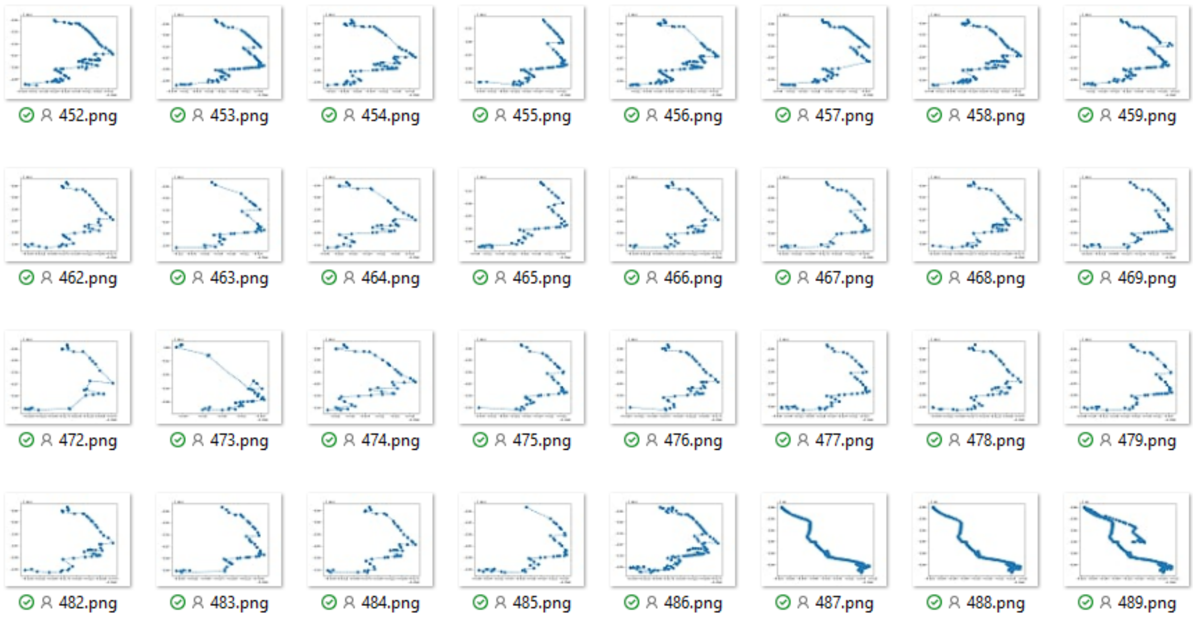


Figure 2 – First visualizations of the Try-Bus Dataset (Image of own creation).

measure the time taken by buses on their daily routes. The process of creating trajectories helps us to gather information that corresponds to a particular bus and not mix up information.

To build the trajectories, the following four attributes provided by the database were considered: the day, the bus license plate, the trip number, and the time at which the records were created. The trip number attribute allows us to indicate when the bus completed a cycle, including the outbound and return journeys in a trip segment. With the first three attributes mentioned, we created subsets of records, and then, using the time attribute, sorted them in each one. To build the trajectories, the chronological order of each record was used. Emphasizing again, it is necessary to group these attributes since any of these buses can make multiple trips in a single day, and it is necessary to distinguish them for a good analysis of the trips.

The Figure 2 shows the first visualizations of the Try-Bus Dataset, created from records containing latitude and longitude attributes. This information was used to construct trajectories representing movements in a geographical space. Each of these trajectories was plotted using *Python's Matplotlib* library, enabling a detailed visual assessment of their characteristics.

The image displays multiple trajectories with diverse shapes and patterns. Some exhibit inaccuracies, evidenced by scattered points or different trajectories, while others show a high density of points, which could indicate a larger dataset or longer trajectory durations. The variability in shapes suggests that the dataset includes different patterns, possibly related to distinct routes or system dynamics.

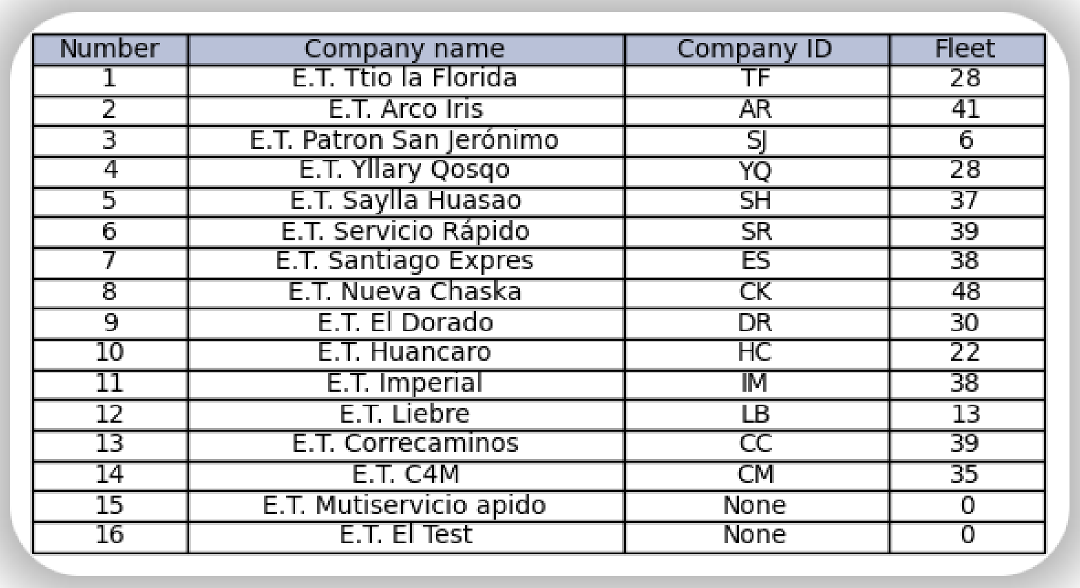
Each sub-image has a unique identifier, such as "476.png", which is used to individually identify each created trajectory.

Finally, once we built the trajectories, in some cases was necessary to eliminate the noisy points generated by some GPS device errors. For this purpose, we performed two steps: we considered only the points whose difference in the latitude and longitude plane is less than one unit and we did not consider trajectories with less than twenty points. The criteria are the following: a) This difference in units was taken because buses cannot move more than one unit of measurement in these planes in such a short time, and b) a route cannot be defined by so few points (twenty points) in this generated database.

4.1.2 Data exploration

We started the data exploration using the read_sql_table and read_sql_query functions of the *pandas.io* library in order to connect to a postgresql database from a notebook. With this exploration we were able to detect the amount of data available in each table, arriving to the conclusion that there are two tables which contain most of the data, both tables assigned to the omnibus monitoring.

The first file to be analyzed had a weight of 1074KB, which caused the filling of a RAM memory of 16GB, this demonstrated the difficulty of processing this amount of data at once.



Number	Company name	Company ID	Fleet
1	E.T. Ttio la Florida	TF	28
2	E.T. Arco Iris	AR	41
3	E.T. Patron San Jerónimo	SJ	6
4	E.T. Yllary Qosqo	YQ	28
5	E.T. Saylla Huasao	SH	37
6	E.T. Servicio Rápido	SR	39
7	E.T. Santiago Expres	ES	38
8	E.T. Nueva Chaska	CK	48
9	E.T. El Dorado	DR	30
10	E.T. Huancaro	HC	22
11	E.T. Imperial	IM	38
12	E.T. Liebre	LB	13
13	E.T. Correcaminos	CC	39
14	E.T. C4M	CM	35
15	E.T. Mutiservicio apido	None	0
16	E.T. El Test	None	0

Figure 3 – The number of elements of each fleet.

With the data collected we found the number of elements of each fleet of buses, the Figure 3 shows these results, we can see that the companies *Nueva Chaska* and *Arco Iris* have more data, since they have the largest fleets of buses.

With the size information of the buses fleets, it was decided to perform a data segmentation process by month and by bus, choosing among the months in which complete data per month is available and first analyzing the companies with the largest number of buses.

Eliminating points of rest sectors. To avoid unusable data, it is defined the ends of routes of points, these points are generated once a bus arrives at its final destination where buses have to wait a certain time to return to transit on the next route, for our analysis these points do not generate information. This discarding is performed to reduce the size of the unused data and improve the analysis of routes.

Two approaches were tested for the elimination of these points, the first consists of creating a quadrilateral in the resting zones and the second consists of creating a circumference taking as centroid the point with the highest frequency as well as a radius that covers the resting zone.

This segmentation is done to improve the route analysis process, since both the outbound and return routes are linked for each bus, this is because the GPS devices are constantly on.

The Figure 4 shows the resting area defined with the rectangle approach, the points that are eliminated within this resting area are also shown, many of these points are

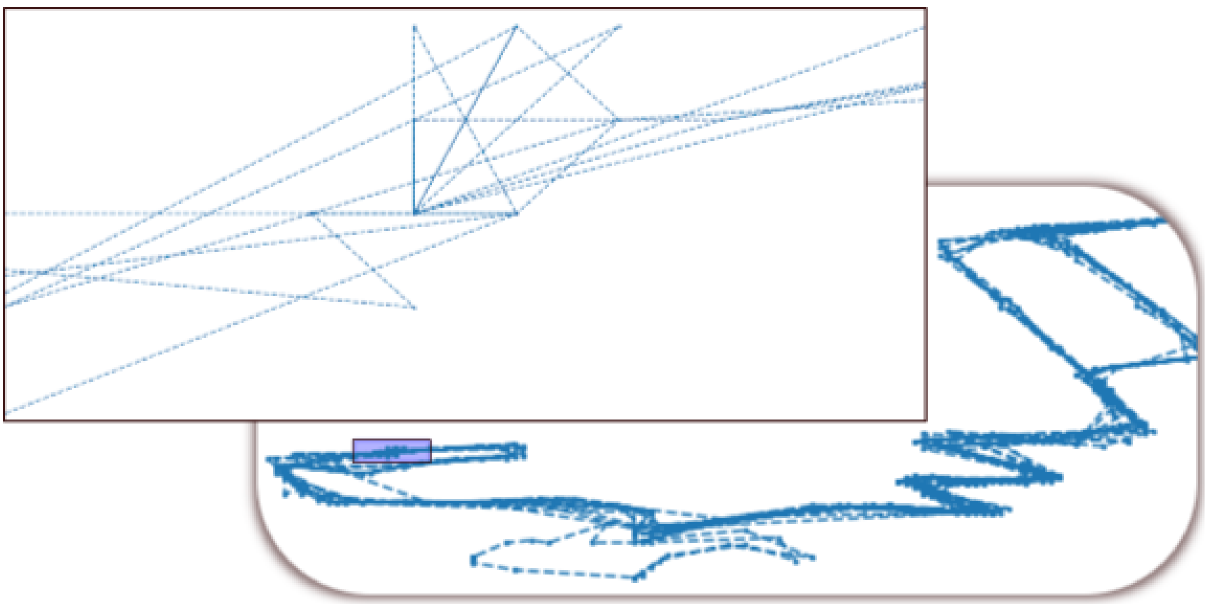


Figure 4 – Segmentation points by square approach

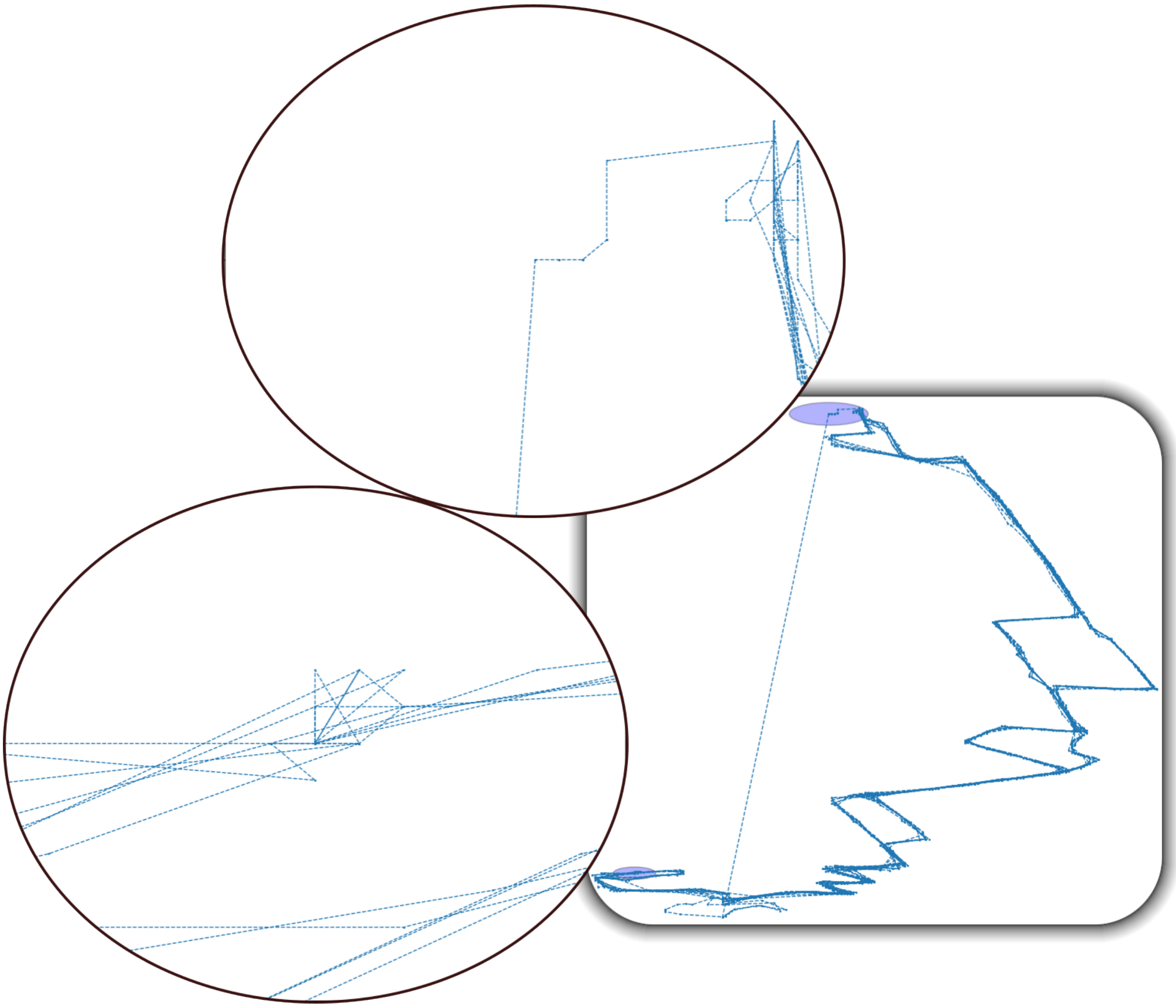


Figure 5 – Segmentation points by circle approach

superimposed which makes them difficult to see, apparently they are few but there are many.

The Figure 5 shows the points that are removed using the approach of eliminating points from rest zones using circles. To choose the center of the circle, the frequency of the GPS geographic points that are repeated in this area was taken into account. Rest zones are areas where it is known that cars wait until they are put back into circulation, these areas are popularly called end of route and are defined by the transport company.

Evaluating the routes created with these two approaches to information segmentation within the rest areas, we reached the following observations: we manually observe that the approach based on circles presents better results, because the circumference can cover a geometric space uniformly, in addition to the fact that it is much easier to define centroids based on frequencies and radius based on scope areas by circumferences.

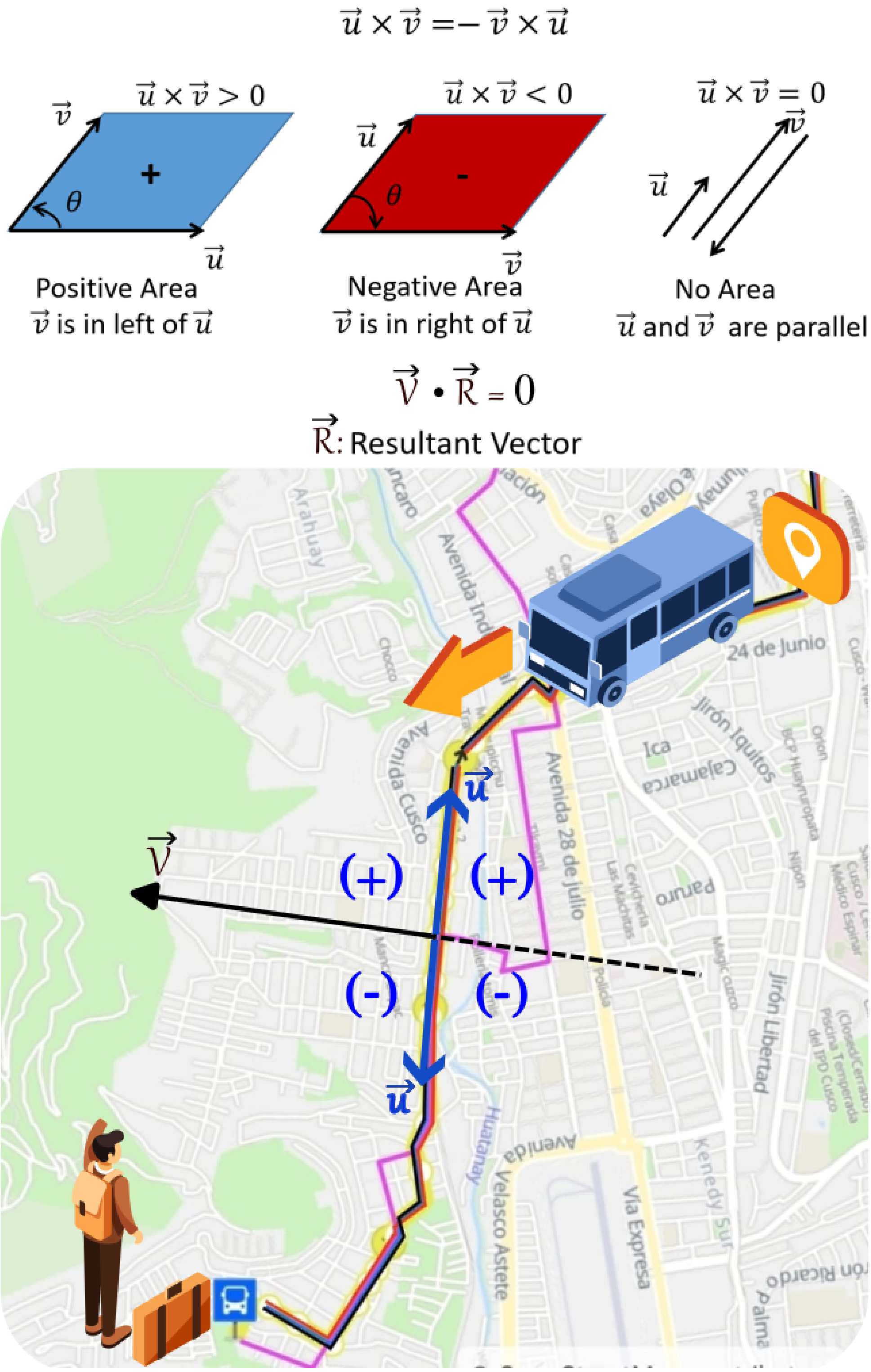


Figure 6 – Trajectory sense detection approach.

4.1.3 Trajectory Sense Detection

Vector algebra is used to detect the direction of a path. The first step is to find the vector resulting from a complete section, seeing a path as a sequence of vectors. With the information of the resulting vector, the cross product is performed between the resulting vector found and the vector perpendicular to both outbound and return sections. The cross product between two vectors results in an orthogonal vector that is incoming or outgoing to the plane that contains both vectors. With the information of the resulting vector, it is possible to classify the directions of the routes. With the information of the third component of the cross product vector, we can classify the transport routes as outbound or return routes, since it is enough to ask if this component is greater than or less than zero.

The Figure 6 illustrates a conceptual approach to detecting the direction of a bus trajectory using GPS data. Following a detailed description:

1. **Vector Analysis**: The upper section of The Figure 6 explains the mathematical principles of vector cross-product and dot-product operations.
   - The Cross Product ($\vec{u} \times \vec{v}$):
     - **Positive Area**: When $\vec{v}$ is to the left of $\vec{u}$. Represented in blue ($\vec{u} \times \vec{v} > 0$).
     - **Negative Area**: When $\vec{v}$ is to the right of $\vec{u}$. Represented in red ($\vec{u} \times \vec{v} < 0$).
     - **No Area**: When $\vec{u}$ and $\vec{v}$ are parallel ($\vec{u} \times \vec{v} = 0$).
   - The Dot Product ($\vec{v} \cdot \vec{R}$): The dot product helps to identify the $\vec{v}$ vector, since it should be orthogonal to $\vec{R}$ which is the resultant vector of a trajectory when the trajectory is treated as a sequence of vectors. If $\vec{v} \cdot \vec{R} = 0$, the vectors are orthogonal.
2. **Bus Route Representation**:
   - A bus trajectory is mapped on a geographic area with GPS data.
   - The vectors $\vec{u}$ and $\vec{v}$:
     - $\vec{u}$ represent a vector of a route produced by a bus. It is alignment with the bus direction.
     - $\vec{v}$ represent the vector defined previously by dot product, this vector is orthogonal to a resultant vector $\vec{R}$. $\vec{R}$ is produced by many more routes belonging to the dataset.
   - Positive and Negative Signs:
     - The cross product between $\vec{u}$ and $\vec{v}$ produce a new vector which has a positive or negative signs, this signs is used to determine de direction of a trajectory.

4.1.4 Trajectory Anomaly Detection

The main idea in this section is the reproduction of the detection of anomalous trajectories based on their shape or morphology, but also taking into account the direction of the trajectories, this direction characteristic is not taken into account in the study of Quispe-Torres *et al.* (2022). Within this case study, the anomaly detection is traduced as detecting when a bus goes off its route, something that can be useful to monitor the work of the drivers of the transport company. The new data set created and cleaned will be used for this reproduction, thus showing new results based on this study, the pipeline is modified and a slight improve is made.

The Figure 7 illustrates the pipeline for trajectory-based anomaly detection used in Quispe-Torres *et al.* (2022), which consists of three main stages: **Data**, **Modeling**, and **Detection**.

1. **Data**:

    - This stage involves obtaining and processing trajectory data, for our case we use Try-Bus dataset.
    - The input consists of raw GPS trajectories, visualized as maps and trajectory plots.
    - The main objective is to extract and prepare trajectories for subsequent analysis.

2. **Modeling**:

    - In this stage, trajectories are transformed into a usable format through *Trajectory Embedding* and *Sub-trajectories Generation*.
    - The approach generate sub-trajectories that preserve the morphology of the original trajectory.
    - Techniques such as **Polynomial Interpolation**, **Discrete Fourier Transform**, and **Wavelet Discretization** are applied to effectively model and represent the trajectories. For our case study we perform Wavelet descriptor and all results are performed with this descriptor.

3. **Detection**:

    - This stage focuses on identifying anomalies in the trajectories.
    - It starts with *Trajectory Clustering*, which clusters similar trajectories to establish typical patterns.
    - Subsequently, *Anomaly Detection* is performed by detecting clusters with low number of elements, identifying those that deviate significantly from majority pattern, this anomaly detection method is denominated as anomaly detection based on distances.

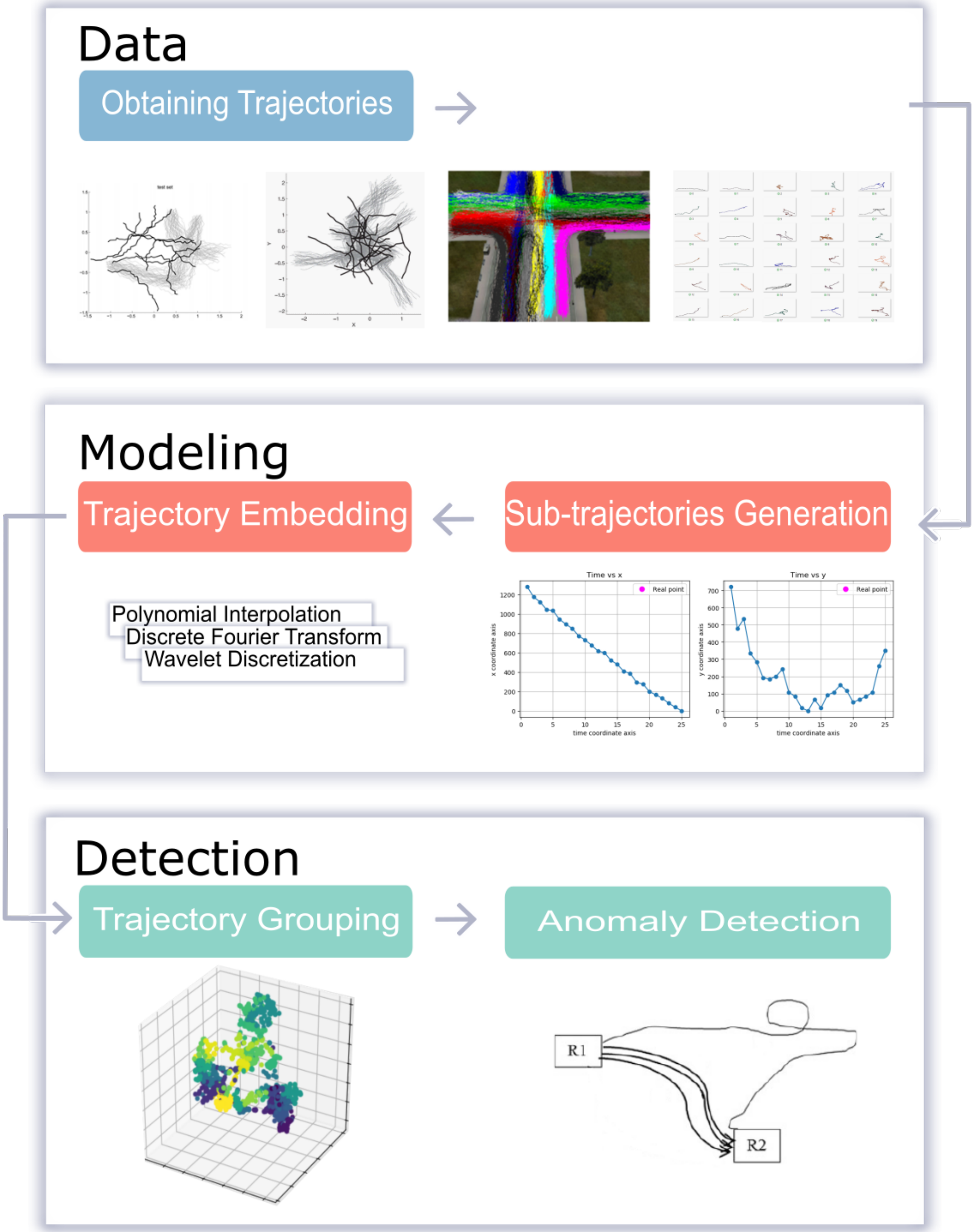


Figure 7 – Trajectory Anomaly Detection pipeline. Image extracted from Quispe-Torres *et al.* (2022).

The pipeline was reproduced with our data (Try-Bus) and using only the Wavelet descriptor since according to the study of Quispe-Torres *et al.* (2022), Wavelet descriptor obtain the best results.

4.1.5 Performance Analysis

For the generation of performance reports by time and distance we create three diagrams, those showing the performance of the buses.

Once the data cleaning and segmentation process is completed, the analysis of the bus routes begins. As an initial analysis, the times and distances traveled are measured. For this first task, the data previously explained in 4.1.1 is used. This data corresponds to a fleet of buses from a public transport company. This data is within an established schedule and is also limited to one month.

The Figure 8 shows the times used by each of the buses in the entire fleet being studied. Once the data has been cleaned and the times used by a bus during an entire day have been extracted, these daily extracted times are added up throughout the month. This gives a monthly report of the performance in terms of time.

Similarly, Figure 9 shows the distances traveled in an entire month by an entire fleet of buses from a company. With this type of graph it is possible to observe which buses traveled the greatest distances in a given month. The extraction of distance information is done in a similar way to how time information was extracted. The time information was extracted by adding the differences in the GPS coordinates of two consecutive points that belong to a trajectory. To find these distances, the python library called *vincenty* (Pietrzak, 2024) is used. This library implements Vincenty's solution to the inverse geodetic problem. In addition, this library can return the distance between two GPS points in both miles and meters.

Figure 10 shows the performance behavior of the entire bus fleet based on time vs distance, efficiency can be interpreted as the greatest amount of distance traveled in the shortest time. Taking this definition into account, we can say that the point that is furthest to the upper left of this graph will have the highest performance, since it will

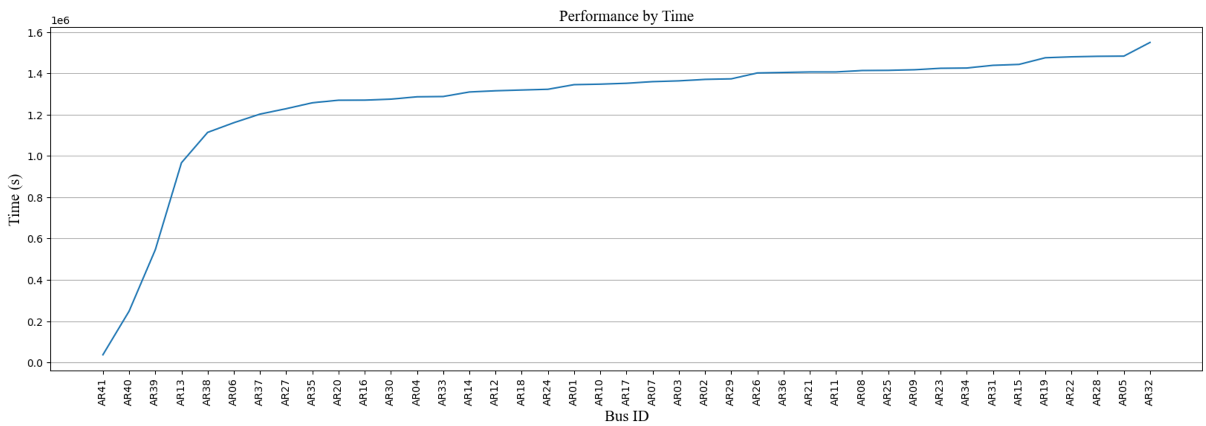


Figure 8 – A fleet performance analysis by time.

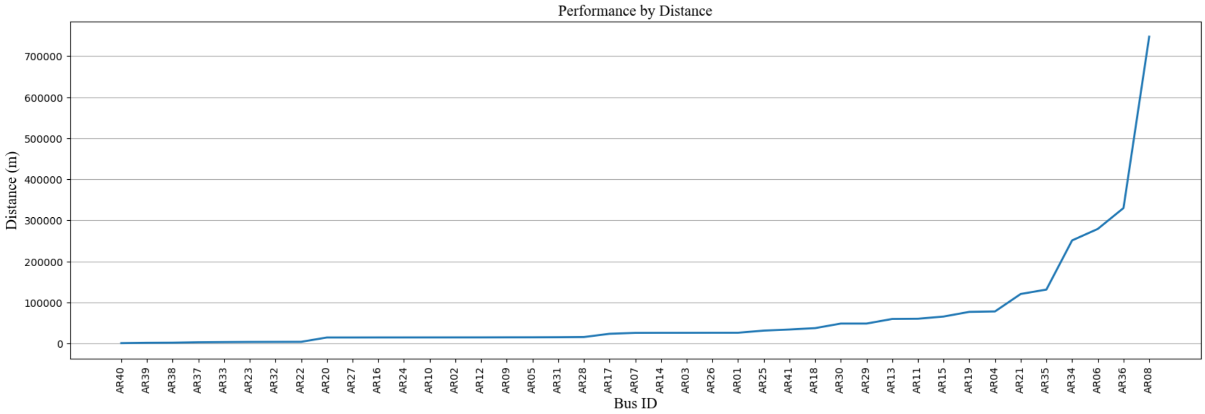


Figure 9 – A fleet performance analysis by distance.

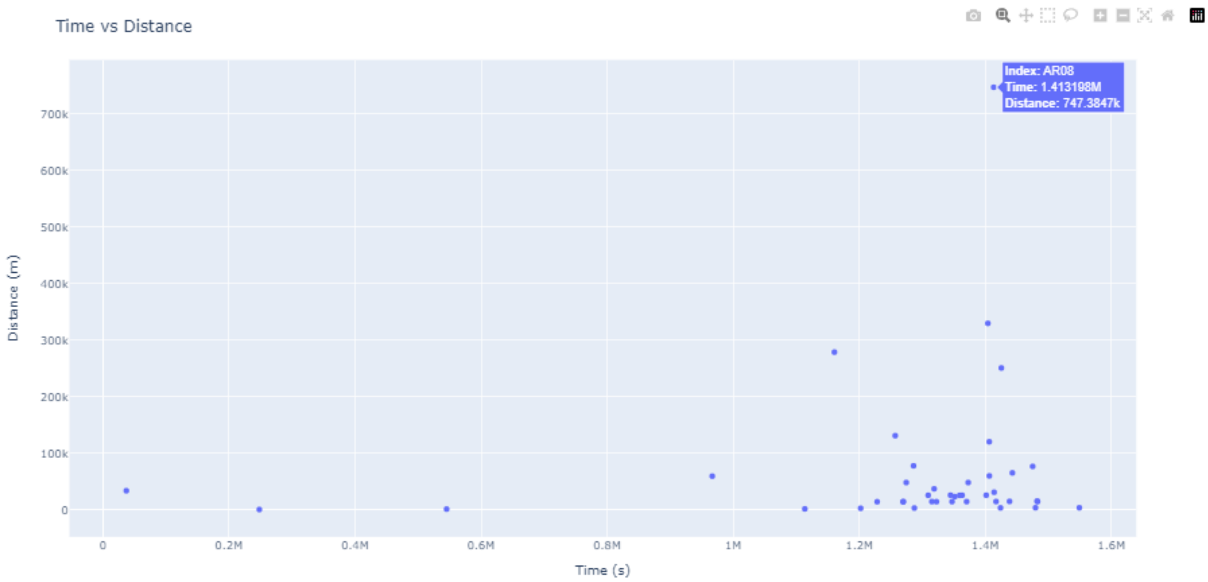


Figure 10 – Performance based on time vs distance, each point correspond to the performance of one bus during a month, all buses are from a fleet of a company

correspond to a shorter time over a greater distance. The creation of this graph was useful to detect optimal performance within the bus fleet, the graph was created interactively with the *plotly* (Plotly Technologies Inc., 2015) library.

5 EXPERIMENTS

This chapter presents experiments designed to validate our data exploration process using the introduced dataset. We independently analyze and discuss the resulting data. Manual validation is performed on our findings, and detected trajectory anomalies are corroborated with existing information.

The experiments were performed on 13th Gen Intel(R) Core(TM) i5-13400F 2.50 GHz with 32GB RAM, Windows 11 Pro 64 bits version 23H2 employing Python and C++ languages, our framework utilizes the following Python libraries: Pandas and *Numpy* (Walt; Colbert; Varoquaux, 2011), *Scikit-learn*(Pedregosa *et al.*, 2011); to draw the graphs, we used *matplotlib* (Hunter, 2007), *seaborn*(Waskom, 2015) and *plotly*(Plotly Technologies Inc., 2015).

5.1 Validation process

Both the approach and the research design for our study are drawn from the work of Hernández-Sampieri and Mendoza (2020) which is where we draw from to describe the scope of the work.

From a holistic point of view, the present work has a qualitative approach, because in our research we define a course (problem statement), but it is not a straight line. It acts as the traffic and navigation application (it repositions or recalculates the best route according to the circumstances to arrive at the place we want). The qualitative approach also studies phenomena systematically. However, instead of starting with a theory and then “turning” to the empirical world to confirm whether it is supported by the data and results, the researcher begins the process by examining the facts themselves and reviewing previous studies, both simultaneously, in order to generate a theory that is consistent with what he or she is observing happening.

Qualitative research usually produces questions before, during or after data collection and analysis. The inquiry action moves dynamically between the facts and their interpretation, and results in a rather “circular” process in which the sequence is not always the same; it may vary in each study.

The route is discovered or constructed according to the context and the events that occur as the study develops, in addition to generating reports and conclusions, in order to establish guidelines and propose a quantitative study. According to Hernández-Sampieri and Mendoza (2020), there are scopes within the research, and these should not be considered as “types”, since rather than being a classification, they constitute a continuum of “chance” that a study may have.

According to Hernández-Sampieri and Mendoza (2020), the same research can include different scopes, likewise, it is possible for a research to begin as exploratory and then become correlational and even explanatory. Taking this into account, it is confirmed that our research is of the exploratory and qualitative type, because the goal of exploratory research is to prepare the ground; that is, it precedes research with descriptive, correlational or explanatory scopes.

Descriptive studies seek to specify the properties, characteristics and profiles of objects or any other phenomenon under analysis. That is to say, they only seek to measure or collect information independently or jointly on the concepts or variables to which they refer; their objective is not to indicate how these are related. Descriptive studies are useful to accurately show the angles or dimensions of a phenomenon, event, community, context or situation (Hernández-Sampieri; Mendoza, 2020), this type of descriptive research is what we aspire to, proposing that this study will serve for a second part later on. In this sense, our results are flexible and do not have a ground-truth.

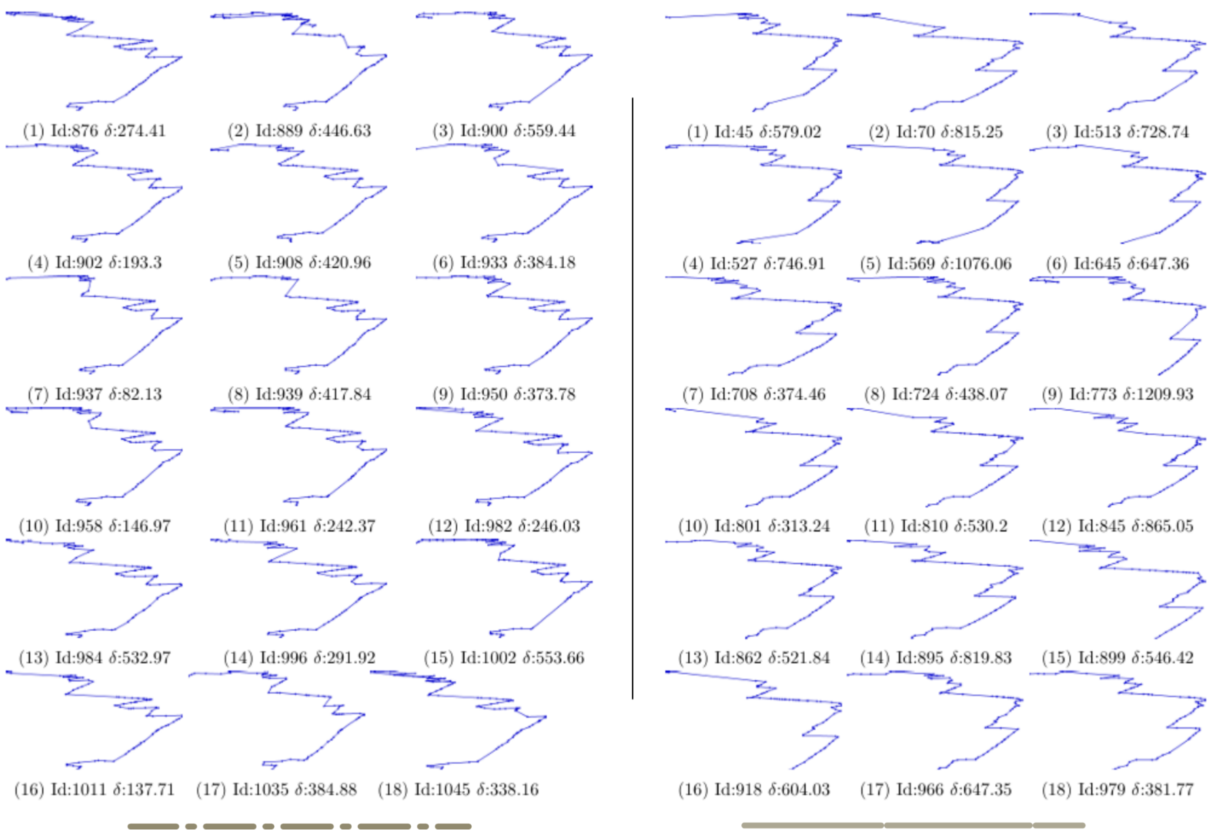


Figure 11 – Two groups of no anomaly trajectories. These groups was defined by clustering using the feature vectors extracted from trajectories.

5.2 Results

In this section, we describe the qualitative results obtained with the anomaly detection algorithm, based on the experiments conducted.

Despite the fact that the trajectories have a variable number of points, and some of them contain a considerable amount of data, the method successfully detects anomalous trajectories. These results could be improved by further cleaning the data. One of the advantages of this dataset is that the variation in shapes is not extensive, making it easier to cluster the trajectories by shape.

The Figure 11 shows the results of trajectory clustering using the algorithm proposed by Quispe-Torres *et al.* (2022), which classifies trajectories based on their shape. In this case, two sets of trajectories are presented, identified as “non-anomalous” because they follow regular or expected patterns corresponding to routes defined by a transportation company. Each subfigure contains multiple trajectories represented by blue lines, where the trajectories within the same group share similar characteristics in terms of shape and path. Additionally, each trajectory is identified with a unique ID (e.g., “Id:876”) followed by the degree of dispersion of each cluster, calculated based on a centroid.

On the left and right sides of the image, two different groups are observed. Although both categories show regular trajectories, the patterns may vary slightly between groups. This clustering demonstrates that the algorithm can organize trajectories according to their shape, using extracted structural features. Each figure is accompanied by a legend that details the trajectory IDs, providing greater clarity. Overall, this image highlights the

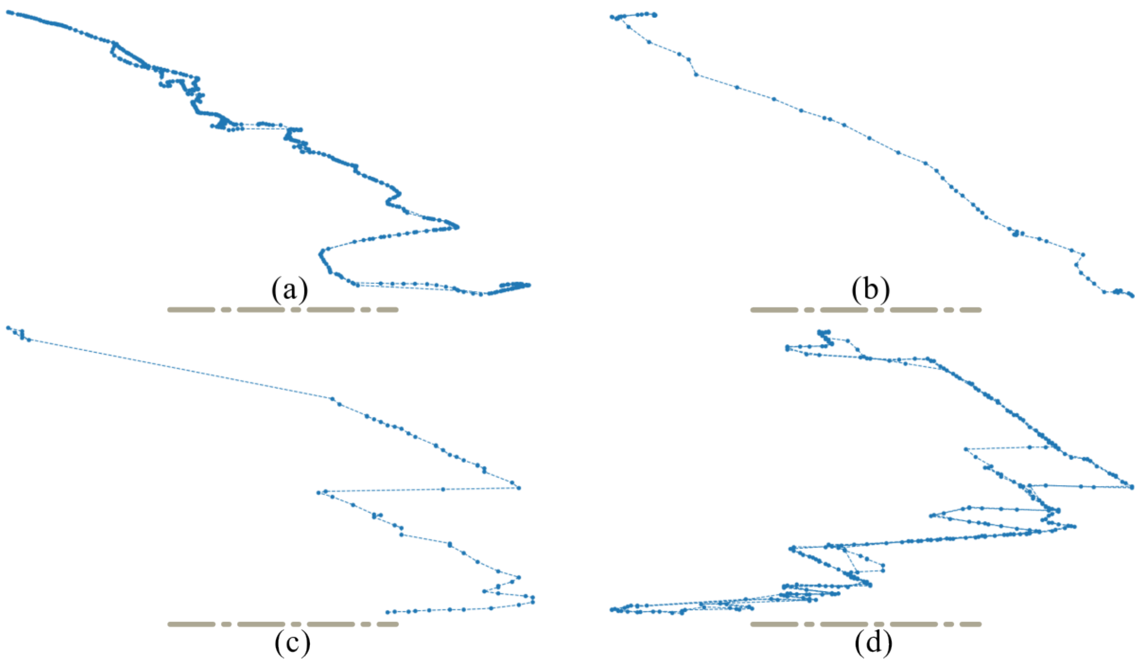


Figure 12 – Anomaly Trajectories founds with Quispe-Torres *et al.* (2022) approach.

effectiveness of the algorithm in identifying and organizing regular trajectories, which is crucial for the anomaly detection process.

Figure 12 illustrates the results of an anomaly detection algorithm to bus trajectories, highlighting irregular patterns based on their shapes. It showcases four trajectories, visualized as sequences of blue points connected by lines, which represent movement paths identified as anomalies. These trajectories deviate significantly from the expected patterns, underscoring their irregularity when compared to standard routes. The algorithm flagged these paths as anomalous due to their pronounced divergence in shape from the regular trajectories defined in the dataset.

Finding the origin of each trajectory we have the following:

(12.a) **Date-of-the-route-section:** 2019-04-14 **Sunday**
Start-and-End-of-section: [05:19:32 – 21:49:17]
Trajectory-name: 2019-04-14_AR13_0

(12.b) **Date-of-the-route-section:** 2019-04-07 **Sunday**
Start-and-End-of-section: [05:08:23 – 21:56:14]
Trajectory-name: 2019-04-07_AR02_0

(12.c) **Date-of-the-route-section:** 2019-04-12 **Friday**
Start-and-End-of-section: [17:48:17 – 18:24:17]
Trajectory-name: 2019-04-12_AR06_13

(12.d) **Date-of-the-route-section:** 2019-04-27 **Saturday**
Start-and-End-of-section: [05:18:07 – 10:31:07]
Trajectory-name: 2019-04-27_AR04_0

The anomalies depicted in Figures 12.a and 12.b occurred on Sundays. On these days, drivers often take alternate routes due to the informal nature of operations, as Sunday is a day of rest and buses may be used for other activities or purposes outside their regular schedules.

Figure 12.c represents a route segment spanning approximately 36 minutes. This highlights the need for adjustments in the data cleaning process to ensure better segmentation. Nonetheless, it also demonstrates the robustness of the anomaly detection method, which effectively identifies irregular patterns even in incomplete data.

Figures 12.c and 12.d reveal a lack of proper segmentation in the dataset, particularly in dividing a full day's route into sub-routes. Despite this, the anomaly detection method proves to be resilient, effectively identifying anomalies regardless of segmentation inconsistencies.

Additionally, the trajectories in Figures 12.c and 12.d differ significantly in the number of points. This further illustrates the robustness of the anomaly detection method, as it remains unaffected by variations in the number of points within a trajectory.

5.2.1 ***Observation***

One limitation observed during the experimentation phase was the prolonged processing time required to obtain results for datasets containing more than 3,000 feature vectors. This issue arises from the algorithmic complexity of the Affinity Propagation clustering method, which restricts its ability to efficiently handle large numbers of trajectories. For datasets exceeding this threshold, the algorithm's runtime becomes excessively long, posing a challenge for its application in scenarios involving extensive data.

6 CONCLUSIONS

This study demonstrates the feasibility and utility of an automated system for detecting anomalies in public transport networks by months based on the analysis of GPS-generated data. The results highlight the importance of implementing automated analysis tools to improve efficiency and monitoring in urban transport systems.

6.1 Applicability in other cities

The findings of this study are highly replicable in other urban contexts, as the anomaly detection approach is not dependent on specific geographic configurations. By utilizing contextual data and processing it automatically, the model can adapt to the conditions and characteristics of any city, provided sufficient and high-quality data is available.

6.2 Contribution to public transport analysis

The ability to identify mobility patterns and detect deviations enables the optimization of routes and the reduction of inefficiencies in transport systems. Moreover, this approach can be used by transport managers to monitor vehicle performance in real-time and ensure a reliable service for users.

6.3 Reflections on challenges encountered

Several challenges were identified during the development of this work:

- **Processing large volumes of data**: The dataset used was extensive, complicating its analysis in its entirety. To overcome this challenge, it was necessary to divide the information into more manageable fragments ("chunk"), optimizing the analysis process.
- **Complexity of traffic analysis**: Despite the progress made, detecting specific patterns related to delay times and their impact on transport efficiency requires further investigation. This aspect could be strengthened by integrating predictive models and advanced machine learning techniques.

In summary, this study not only provides practical solutions for the analysis of public transport networks in Cusco but also lays the groundwork for future applications in other cities with similar characteristics. Implementing this system can represent a significant step toward the modernization of public transport systems and the overall improvement of urban mobility.

7 FUTURE WORKS

The methodologies employed in this study demonstrate significant potential for improving urban transportation management, not only in Cusco-Peru but also in other cities with comparable challenges. This research establishes a foundation for integrating predictive analytics and machine learning models to enhance public transport network management, making systems more efficient and responsive to urban mobility demands.

Future research should focus on strengthening anomaly detection mechanisms in public transportation systems. An important step will involve accurately identifying bus stops, which can improve both anomaly detection precision and route analysis. Additionally, comparing the current anomaly detection approach with methods such as isolation forest will help assess its effectiveness and uncover potential advantages. Adaptive Affinity Propagation could also be employed to refine clustering techniques, improving the system's ability to adapt to diverse operational conditions.

To further enhance detection capabilities, integrating K-Nearest Neighbors (KNN) for automatic anomaly threshold determination is proposed. This approach would streamline the process, minimizing the need for manual adjustments. Furthermore, creating a reliable ground truth dataset for various routes using the TryBus dataset will provide a standardized benchmark to validate and improve these methods, ensuring scalability and accuracy across different scenarios.

Beyond anomaly detection, further exploration of robust traffic analysis methods is essential. This includes developing tools to predict anomalies using time-series data and investigating how delay-time analysis can optimize route planning. Incorporating predictive models to anticipate demand variations and dynamically adjust transport services can offer a transformative approach to urban mobility, enhancing efficiency and passenger satisfaction.

REFERENCES

BORGES, J. *et al.* **Towards spatiotemporal integration of bus transit with data-driven approaches**. 2024. Available at: https://arxiv.org/abs/2402.17866.

DIGHE, B.; NIKAM, A.; MARKAD, K. Intelligent traffic management systems: A comprehensive review. **International Journal of Creative Research Thoughts**, v. 12, n. 4, p. 912–917, 2024. ISSN: 2320-2882. Available at: https://ijcrt.org/papers/IJCRT2404906.pdf.

DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. 2010. Official Journal of the European Union, 6 August 2010, pp. 1–13. Available at: http://data.europa.eu/eli/dir/2010/40/oj [Accessed 20 Dec. 2023].

GONZÁLEZ, F.; ANAPOLSKY, S. Identificando la desigualdad en los patrones de movilidad en transporte público. IADB: Inter-American Development Bank, 2022.

GORDON, R. Intelligent transportation systems. **Cham: Springer**, Springer, 2016.

GUO, Y. *et al.* The impact of urban transportation infrastructure on air quality. **Sustainability**, MDPI, v. 12, n. 14, p. 5626, 2020.

HERNÁNDEZ-SAMPIERI, R.; MENDOZA, C. **Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta**. [*S.l.: s.n.*]: Mcgraw-hill México, 2020.

Hunter, J. D. **Matplotlib Plotting Library**. 2007. Available at: https://matplotlib.org/.

KARLAFTIS, M. G.; TSAMBOULAS, D. Efficiency measurement in public transport: are findings specification sensitive? **Transportation Research Part A: Policy and Practice**, Elsevier, v. 46, n. 2, p. 392–402, 2012.

LIESHOUT, R. N. van; DALMEIJER, K. **A Unified Approach to Evaluation and Routing in Public Transport Systems**. 2024. Available at: https://arxiv.org/abs/2207.09969.

MAMANI, E. A. Sistema de transporte inteligente (sti), para el control y monitoreo del servicio urbano en la ciudad del cusco. Universidad Nacional de San Antonio Abad del Cusco, 2018.

MAZLOUMI, E.; CURRIE, G.; ROSE, G. Using gps data to gain insight into public transport travel time variability. **Journal of transportation engineering**, American Society of Civil Engineers, v. 136, n. 7, p. 623–631, 2010.

ÖZGÜN, K.; GÜNAY, M.; BAŞARAN, B. D. Determination of peak times in public transportation. *In*: IEEE. **2021 Innovations in Intelligent Systems and Applications Conference (ASYU)**. [*S.l.: s.n.*], 2021. p. 1–6.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011. Available at: https://doi.org/10.48550/arXiv.1201.0490.

PIETRZAK, M. **vincenty: Calculate geodesic distances with Vincenty's formula**. PyPI, Python Software Foundation, 2024. Available at: https://pypi.org/project/vincenty/.

Plotly Technologies Inc. **Plotly Interactive Graphing Library**. 2015. Available at: https://plotly.com/.

PUEBLA, J. G. *et al.* Cómo aplicar big data en la planificación del transporte: El uso de datos de gps en el análisis de la movilidad urbana. Banco Interamericano de Desarrollo (BID), 2020.

QUISPE-TORRES, G. F. *et al.* Feature-based trajectory anomaly detection. **CLEI electronic journal**, v. 25, n. 2, 2022.

VERMA, S. K. *et al.* Management of intelligent transportation systems and advanced technology. *In*: **Intelligent transportation system and advanced technology**. [*S.l.: s.n.*]: Springer, 2024. p. 159–175.

VUKIĆ, L. *et al.* Model of determining the optimal, green transport route among alternatives: data envelopment analysis settings. **Journal of marine science and engineering**, MDPI, v. 8, n. 10, p. 735, 2020.

WALT, S. v. d.; COLBERT, S. C.; VAROQUAUX, G. The numpy array: a structure for efficient numerical computation. **Computing in Science & Engineering**, IEEE, v. 13, n. 2, p. 22–30, 2011. Doi:10.1109/MCSE.2011.37. Available at: https://ieeexplore.ieee.org/abstract/document/5725236.

WANG, Y. *et al.* A feature-based method for traffic anomaly detection. *In*: **Proceedings of the 2Nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics**. [*S.l.: s.n.*], 2016. p. 1–8.

Waskom, M. L. **Seaborn Statistical Data Visualization**. 2015. Available at: https://seaborn.pydata.org/.

WU, X. *et al.* **Joint Design of Conventional Public Transport Network and Mobility on Demand**. 2024. Available at: https://arxiv.org/abs/2407.01700.