INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data driven spatio-temporal analysis of e-cargo bike network in Lisbon and its expansion: the Yoob case study

Bruno Alexandre Nunes Gil

Master's in **Integrated Business Intelligence Systems**

Supervisor:
PhD José Miguel de Oliveira Monteiro Sales Dias, Associated Professor with
Habilitation,
ISCTE-IUL

Co-supervisor:
Msc Lídia Vitória Pires de Albuquerque, Researcher
NOVA IMS

August, 2022

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Data driven spatio-temporal analysis of e-cargo bike network in Lisbon and its expansion: the Yoob case study

Bruno Alexandre Nunes Gil

Master's in **Integrated Business Intelligence Systems**

Supervisor:
PhD José Miguel de Oliveira Monteiro Sales Dias, Associated Professor with
Habilitation,
ISCTE-IUL

Co-supervisor:
Msc Lídia Vitória Pires de Albuquerque, Researcher
NOVA IMS

August, 2022

“you are smarter than your data. Data do not understand causes and effects; humans do.”

Judea Pearl

Acknowledgments

This work represents for me not only a thesis, it is much more than that.

It is a long-desired achievement. A result of a long path as a student-worker, which has not always been easy, much has been sacrificed, but in the end, the feeling of accomplishment makes everything that has passed worthwhile.

Firstly, I would like to thanks Professor José Miguel Dias for the challenge and all the time invested in advising and guiding me whenever it was necessary.

To Vitória Albuquerque, my co-supervisor, thank you very much, for the availability, motivation and the support that you gave me throughout this journey.

Your knowledge and experience was essential. Thank you!

To my parents I am eternally grateful, they were unwavering during these last month's so that I could dedicate myself in full time to this project.

To my sister for the motivation and strength she gave me at times when I needed it most.

A special thanks to Ana Amaro for the support, encouragement and strength she has gave me over the years in this academic journey and also in everyday life.

Lastly to my friends and work colleagues.

To all sincere "Thank you".

Resumo

A adoção de frotas mais ecológicas e sustentáveis para a distribuição das encomendas na última milha dentro dos grandes centros urbanos tem vindo a crescer. As bicicletas de carga têm sido a alternativa mais comum. A implementação deste tipo de frotas, demonstrou trazer benefícios, mas evidenciou algumas limitações. A rede de infraestruturas, que serve de suporte á logística urbana, teve de se adaptar para poder responder às necessidades deste novo tipo de frotas. A implementação de *micro-hubs* e *nano-hubs* foram a alternativa.

O nosso estudo tem dois objetivos principais. O primeiro objetivo é o de fazer uma caracterização espácio temporal dos comportamentos da frota, através de um estudo de caso onde efetuámos a exploração dos dados da frota de *e-cargo bike* da YOOB (start-up logística de entregas na última milha que atua na área de Lisboa e na periferia). E o segundo consiste em identificar potenciais locais de expansão para a instalação de novos *hubs* no mesmo estudo de caso. Nos processos de trabalho foi seguida a metodologia CRISP-DM e os dados recolhidos foram referentes a um período de 4 meses (Janeiro a Abril de 2022).

Com recurso a técnicas de ciência dos dados e aprendizagem automática, foram identificados cinco tipos de desempenhos da frota da YOOB, com variações em distâncias percorridas, tempos efetuados, volumes transportados e velocidades praticadas. Numa perspetiva de expansão da rede de *e-cargo bike* da YOOB, forma identificados três novos locais na cidade de Lisboa para a instalação potencial de novos *hubs*.

Palavras-chave: bicicletas de carga elétricas, *micro-hub*, *nano-hub*, algoritmo K-Means, entrega na última milha.

Abstract

The adoption of more environmentally friendly and sustainable fleets for last-mile parcel delivery within large urban centers has been on the rise. Cargo bikes have been the most common alternative. The implementation of this type of fleet has proven to bring benefits, but has evidenced some limitations. The infrastructure network, which supports urban logistics, had to adapt to respond to the requirements of this new type of fleet. The implementation of micro-hubs and nano-hubs was the solution.

Our study has two main objectives. The first objective is to perform a spatiotemporal characterization of fleet behavior, by conducting a case study where we explored the data from YOOB (a last mile delivery logistics start-up that operates in the Lisbon area and outskirts) e-cargo bike fleet. And the second is to identify potential expansion locations to the establishment of new hubs.

The work procedures followed the CRIPS-DM methodology and the collected data was based on a 4-month period (January to April 2022).

By adopting data science and machine learning techniques, five types of performances of YOOB fleet were identified, with variations in distances traveled, times, volumes transported and speeds. In the perspective of expanding YOOB's e-cargo bike network, three new locations in Lisbon were signaled for potential new hub installation.

Keywords: *e-cargo bikes, micro-hub, nano-hub, K-Means algorithm, last-mile delivery*

Index

Tables index

Figures index

List of abbreviations

2E-VRP – Two-Echelon Vehicle Routing Problem

AHP - Analytic Hierarchy Process

DBSCAN – Density-based spatial clustering of applications with noise

GRIDBSCAN – GRId Density-Based Spatial Clustering of Applications with Noise

HDBSCAN - High density-based spatial clustering of applications with noise

LCRS - Longest Common Route Subsequence

LRP - Logistics Requirements Planning

PROMETHEE - Preference ranking organization method for enrichment evaluation

SLR – Systematic Literature Review

SP-VRPMPCTW – Split-Delivery Vehicle Routing Problem with Multiple Products, Compartments and Time Windows

ST-HDBSCAN - Spatial-temporal high density-based spatial clustering of applications with noise

ST-TCLUS - Spatial-temporal trajectory clustering

VRP - Vehicle Routing Problem

1. Introduction

1.1. Topic context

Urban logistics and the impact of logistics networks in urban mobility of the large cities are increasingly discussed by policy makers and logistics operators [1].

The impacts caused by the large inflow of vehicles (private and public) in urban centers have degraded the quality of life of residents and workers [2].

For logistics companies operating within urban centers the last mile delivery is one of the most challenging aspects, financially and operationally speaking due to the cities' structure [3].

Several policies and actions have been developed by the European Union (EU) to promote sustainability and improve the quality of life in urban centers [4]. The restrictions to circulation in certain areas are becoming tighter [5], the expansion of bicycle paths and financial support to the adoption of "greener" technologies are becoming more frequent.

Logistics operators and service providers are beginning to introduce more environmentally friendly vehicles into their fleets. E-cargo bikes are one of the most widely implemented vehicles for deliveries within urban centers [6].

However, whether for policy making or optimizing operations, it is necessary to study data generated in the urban settings, to understand the phenomena originated by urban activities to support decision making.

The study of the data generated by urban agents allows us to understand and find patterns and dynamics in the functioning of urban centers.

Spatial-temporal analysis and the application of data science and machine learning algorithms have been widely used in several areas of study, notably in urban transportation, allowing finding relationships and patterns in space and time. [7]

1.2. Motivation and topic relevance

Performing last mile delivery with less impact on urban mobility in a sustainable and ecological way is the main goal of YOOB. This startup is a delivery logistics company operating in Lisbon's urban center, and it is the first of its kind operating in the city, and in Portugal and started its operations in the fall of 2021. At the time of this thesis writing, YOOB had fleet of 10 e-cargo bikes and 2 e-vans supported by 5 logistic hubs.

With the growth of their operations in the city, the need arose to get more insights on the behavior patterns of YOOB's e-cargo bike fleet. With this thesis we aim to provide better strategic decisions for YOOB's future logistic operations and expansion of its network.

1.3. Research questions and objectives

This study aims to analyze and visualize the behavior patterns of e-cargo bike fleet based on anonymized real time data of a logistics company collected from January 2022 to April 2022. It also intends, based on collected data, to evaluate the optimal sites for the new micro or nano hubs locations to expand the delivery area in Lisbon.

The following research questions are addressed by our research:

RQ1: How can we characterize the spatial-temporal traffic of the last mile logistic distribution performed with the electric cargo bike fleet, taking into consideration the open data of the city and the data collected during the performed routes?

RQ2: Based on the fleet behavior and the patterns detected, what are the best possible locations for micro-hubs or nano-hubs expansion?

Table 1.1 - Research questions, objectives and methodology

Research questions	Objectives	Methodology
RQ1: How can we characterize the spatial-temporal traffic of the last mile logistic distribution performed with the electric cargo bike fleet, taking into consideration the open data of the city and the data collected during the performed routes?	Studies of the behavior patterns of the e-cargo bikes fleet.	CRISP-DM and Descriptive Statistics
RQ2: Based on fleet behavior and the patterns detected, what are the best possible locations for micro-hubs or nano-hubs expansion?	Discover the best possible locations to improve last mile delivery service using e-cargo bike.	CRISP-DM and Descriptive Statistics

1.4. Structure and organization of dissertation

This dissertation is organized in four chapters. In chapter 1, we introduce the theme of the dissertation, the topic context, motivation and relevance, research questions and objectives. Chapter 2 presents the methodology applied to our research and using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [8], for systematic literature review, followed by VOSviewer tool to perform bibliometric analysis and visualize our literature survey results. In chapter 3, we apply the CRISP-DM [9] data mining methodology to our case study, following the different phases. Finally, in Chapter 4, we present and discuss our conclusions, limitations, and future work.

2. Literature review

2.1. Methodology

The methodology applied to our literature review was divided into two parts.

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [8] bibliometric research method to identify the most relevant publications for our systematic literature review (SLR). This technique aids to develop the strategy for reporting our SLR, which begins with a large number of records followed by an exclusion phase based on eligibility criteria [8]. The full description of the selection process is documented in the PRISMA flow diagram in Figure 2.1.

The second method applied to analyze the PRISMA results was applied by means of the VOSviewer [10], a bibliometric research tool for network analysis.

This tool builds a bibliometric network based on clusters of authors, keywords, title and abstract text, and allows us to understand the SLR by three methods of analysis: keyword, author and co-authorship and title and abstract text occurrence. Two forms of visualization may be used to illustrate these networks: network visualization and overlay visualization.

In our bibliometric research we were able to determine network aspects including clusters and node centrality, as well as infer SLR paper characteristics. These networks were created using the number of common citations, bibliographic coupling, co-citations, and co-authorship relationships to graphically portray the bibliographic data from our scientific literature.

2.2. Systematic literature review with PRISMA

Preferred Reporting Items for Systematic Studies and Meta-Analyses (PRISMA) [8] is a standard methodology for generating systematic and objective findings from literature reviews. It is an approach that allows other researchers to quickly replicate and verify your results.

PRISMA is used to conduct systematic research reviews in the fields of medicine, social sciences, and exact sciences. It is also a collection of guidelines for writers who want to publish and report on how they arrived at their results during their bibliographic research in a full and clear manner. The PRISMA standards assist writers in describing their research findings, as well as their futures goals.

2.2.1. Keyword identification

The initial phase of PRISMA is the keywords identification, which we accomplished by doing an iterative search for keywords on the articles obtained from the designated sources.

By running the following logical query on academic data repositories, we identified the final list of considered academic papers:

("e-cargo bikes" OR "electric-assist cargo bicycles") OR ("micro-consolidation hubs" OR “hub location”) OR ("Last mile logistic" OR “urban logistic”) OR ("spatial patterns" AND "data mining").

2.2.2. Repositories

Keywords found were applied in Scopus and Web of Science, two well-known and available academic databases.

Web of Science is a research tool and data repository that gives users access to archives of peer-reviewed publications in the fields of science, social science, the arts, and the humanities. It contains 12,000 periodicals and 160,000 conference proceedings and contains articles dating from 1900 to the present.

Scopus is an academic journal abstract and article citation database. It includes 16,500 peer-reviewed journals in the scientific, technological, medical, and social sciences (including arts and humanities) sectors, with over 19,500 titles from more than 5,000 foreign publishers. Subscribers can access it over the internet. Searches on the SciVerse website Scopus includes patent databases as well as scientific searches of web sites via Scirus, another Elsevier offering.

To obtain meaningful and comparable results, the same query parameters as in the previous sub-chapter were used on all two repositories.

2.2.3. Bibliometrics analysis

We end up with a collection of papers for additional quantitative and qualitative analysis as a result from our SLR using PRISMA.

They were all compiled with the help of the Mendeley reference management program [11], which was used to extract metadata and to eliminate duplicate entries from articles.

The author's name, number of publications, publication data, references, and number of citations were all collected from each publication's information.

2.2.4. PRISMA results

Figure 2.1 shows the PRISMA flow diagram, which depicts our SLR approach for additional quantitative and qualitative analysis.

We found the papers in the first stage by conducting a database search using the logical query defined earlier (point 2.3.), which yielded 1564 results (Scopus: 1072; Web of Science: 492).

The criteria selection for papers were being authored in English, published in peer-reviewed publications throughout the last six years, from 2017 to 2022 and limited to Articles and Conference Papers in English in final publications stage. In the following stage we deleted the exact duplicates, there were 52 duplicates articles.

The next step was to read the titles and abstract and exclude the papers that do not fit into our topics.

In the last stage we had 65 articles to read in full, of which only 34 met the requirements to be contemplated in the quantitative synthesis.

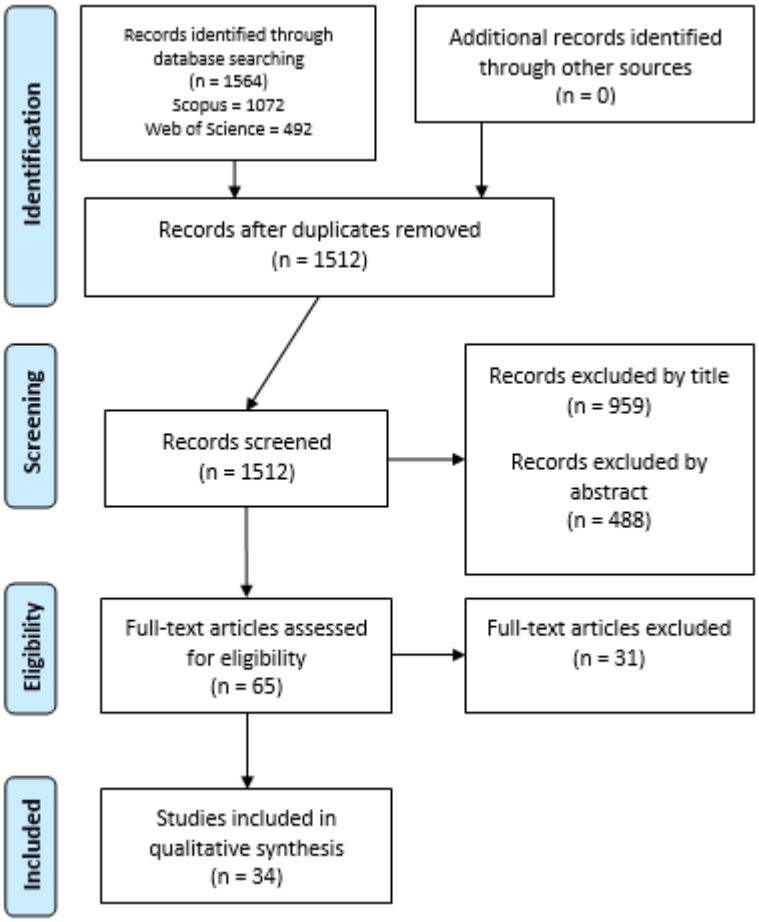

Figure 2.1 - PRISMA flow diagram

A collection of 34 publications was compiled based on the comprehensive literature review indicated in the preceding sub-chapter and submitted to full-text reading and analysis. The topics were closely related to the keywords picked.

In Table 2.1 the identified journals and conferences have a very wide area of coverage.

The source of the articles gathered were Journals and Conferences. Journals were the main source of the selected articles, representing 76% of the total of 34 articles. In the main ranking categories, we have 18 with Q1 ranking, 6 with Q2 ranking and 1 with Q3 ranking. The journal with the largest number of articles is the Sustainability with a total of 5 articles [5], [12]–[15] coming from this source, all the others are represented by only 1 article.

Conferences represent 24%. The most represented source in this group is The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2019 [16], [17], all others are represented by one article.

Table 2.1 - Literature review Journals and Main conferences

Type	Name	No.	Rank	Country	Field	Publisher
Journal	Sustainability	5	Q1	Switzerland	Geography, Planning and Development; Energy Engineering and Power Technology; Environmental Science; Management, Monitoring, Policy and Law; Renewable Energy, Sustainability and the Environment;	MDPI AG
Journal	Cities	1	Q1	United Kingdom	Development; Sociology and Political Science; Tourism, Leisure and Hospitality Management; Urban Studies;	Elsevier Ltd.
Journal	Competition and Regulation in Network Industries	1	Q3	United Kingdom	Business, Management and Accounting; Management Science and Operations Research;	SAGE Publications Inc.
Journal	Mobile Networks and Applications	1	Q2	Netherlands	Computer Networks and Communications; Hardware and Architecture; Information Systems; Software	Springer Netherlands
Journal	European Transport Research Review	1	Q1	Germany	Automotive Engineering; Mechanical Engineering; Transportation	Springer Verlag
Journal	Networks	1	Q1	United States	Computer Networks and Communications; Hardware and Architecture; Information Systems; Software	Wiley-Liss Inc.
Journal	Research in Transportation Business & Management	1	Q1	Netherlands	Business and International Management; Decision Sciences; Economics, Econometrics and Finance; Strategy and Management; Tourism, Leisure and Hospitality Management; Management Science and Operations Research; Transportation	Elsevier BV
Journal	Transport	1	Q2	Lithuania	Automotive Engineering; Mechanical Engineering	Vilnius Gediminas Technical University
Journal	International Journal of Digital Earth	1	Q1	United Kingdom	Computer Science Applications; Earth and Planetary Sciences (miscellaneous); Software	Taylor and Francis Ltd.

Journal	Journal of Transport Geography	1	Q1	United Kingdom	Environmental Science; Geography, Planning and Development; Transportation	Elsevier BV
Journal	Transportation Research Interdisciplinary Perspectives	1	Q2	United Kingdom	Automotive Engineering; Civil and Structural Engineering; Management Science and Operations Research; Transportation	Elsevier Ltd.
Journal	Transportation Research Record: Journal of the Transportation Research Board	1	Q2	United States	Civil and Structural Engineering; Mechanical Engineering	US National Research Council
Journal	Smart Cities	1	NA	Switzerland	Information and communication technology (ICT) in the smart city; Internet of Things (IoT) for smart cities	MDPI AG
Journal	IEEE Transactions on Intelligent Transportation Systems	1	Q1	United States	Automotive Engineering; Computer Science Applications; Mechanical Engineering	Institute of Electrical and Electronics Engineers Inc.
Journal	IET Intelligent Transport Systems	1	Q1	United Kingdom	Law; Environmental Science; Mechanical Engineering; Transportation	Institution of Engineering and Technology
Journal	IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	1	Q1	United States	Computers in Earth Sciences; Atmospheric Science	Institute of Electrical and Electronics Engineers Inc.
Journal	International Journal of Sustainable Transportation	1	Q1	United Kingdom	Automotive Engineering; Civil and Structural Engineering; Environmental Engineering; Geography, Planning and Development; Renewable Energy, Sustainability and the Environment; Transportation	Taylor and Francis Ltd.
Journal	ACM Computing Surveys	1	Q1	United States	Computer Science; Theoretical Computer Science	Association for Computing Machinery (ACM)
Journal	Energies	1	Q2	Switzerland	Control and Optimization; Electrical and Electronic Engineering; Energy Engineering and Power Technology; Energy; Fuel Technology; Renewable Energy, Sustainability and the Environment	MDPI Multidisciplinary Digital Publishing Institute
Journal	Computers & Industrial Engineering	1	Q1	United Kingdom	Computer Science; Engineering	Elsevier Ltd.

Journal	Physica A: Statistical Mechanics and its Applications	1	Q2	Netherlands	Condensed Matter Physics; Statistics and Probability	Elsevier
Journal	International Journal of Management Science and Engineering Management	1	Q1	United Kingdom	Engineering; Information Systems and Management; Management Science and Operations Research; Mechanical Engineering; Strategy and Management	Taylor and Francis Ltd.
Conferences and Proceedings	2019 The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS)	2		Germany	Geography, Planning and Development; Information Systems	
Conferences and Proceedings	2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)	1		United States	Artificial Intelligence; Biomedical Engineering; Computer Vision and Pattern Recognition; Control and Optimization; Human-Computer Interaction; Instrumentation	
Conferences and Proceedings	2017 CARPATHIAN LOGISTICS CONGRESS (CLC)	1		Slovakia	Control and Optimization; Management Science and Operations Research; Modeling and Simulation; Organizational Behavior and Human Resource Management; Strategy and Management	
Conferences and Proceedings	2019 COMPUTATIONAL SCIENCE AND ITS APPLICATIONS (ICCSA)	1		United States	Artificial Intelligence; Computational Mathematics; Computer Networks and Communications; Computer Science Applications; Safety, Risk, Reliability and Quality	
Conferences and Proceedings	2020 23rd EURO Working Group on Transportation Meeting (EWGT)	1		Netherlands	Transportation	
Conferences and Proceedings	2018 4th Conference on Sustainable Urban Mobility, (CSUM)	1		Netherlands	Engineering	
Conferences and Proceedings	2020 5th International Conference on Smart City Applications (SCA)	1		Germany	Geography, Planning and Development; Information Systems	

Table 2.2 shows the list of publications that were collected as well as the authors, methods and applications ranked by number of citations. In the applied methods we can observe a strong emphasis on visualization, in articles focusing on study and detection of pattern [16]–[25]. The main algorithms applied to perform clustering analysis were K-means [16], [17], [26] in travel activity from taxis and bikes and to find the places that gave rise to shorter travel distances, and DBSCAN was implemented to found travel paths made by users of public transportations [18] and to study private car trajectories [21]. In the decision taking for hub location, the two principals algorithms implemented were the Genetic Algorithm [6], [27] and PROMETHEE [26], [28].

Table 2.2 - Literature review, Authors, Methods and applications ranked by number of citations

Title	Author(s)	Application	Method	Citations
Understanding commuting patterns using transit smart card data [18]	Ma X, Liu C, When H, Wang Y, Wu Y	Data mining techniques to analyze patterns to classify commuters' by exploring the travel records from smart card	Visualization, DBSCAN clustering algorithm ISODATA spatial clustering algorithm	190
Understanding the usage of dockless bike sharing in Singapore [19]	Shen Y, Zhang X, Zhao J	Data mining techniques, to identify patterns and impact of dockless bike fleet size on the usage of bikes exploring the GPS records of bike sharing system usage	Visualization, Spatial Autoregressive models: SAR Spatial lag SEM Spatial Error	182
Spatio-temporal data mining: A survey of problems and methods [7]	Atluri, Gowtham Karpatne, Anuj Kumar, Vipin	State of the art of spatio-temporal analysis surveyed, identifying the techniques and methods most used in the process of mining spatio-temporal data	Literature review	127
Spatial–temporal travel pattern mining using massive taxi trajectory data [20]	Zheng L, Xia D, Zhao X, Tan L, Li H, Chen L, Liu W	Application of clustering methods for characterization of urban residents' travel from two aspects: attractive areas and hot paths.	Visualization, GRIDBSCAN algorithm ST-TCLUS(Spatial-temporal trajectory clustering) algorithm	47
City logistics, urban goods distribution and last mile delivery and collection [3]	Cardenas, Ivan Borbon-Galvez, Yari Verlinden, Thomas Van de Voorde, Eddy Vanelslander, Thierry Dewulf, Wouter	Literature review to highlight the current approaches used in the field.	Empirical model	27

Exploring Individual Travel Patterns Across Private Car Trajectory Data [21]	Huang Y, Xiao Z, Wang D, Jiang H, Wu D	Analysis of the individual travel patterns based on a large-scale private car trajectory dataset	Visualization, DBSCAN clustering algorithm	20
Measuring delivery route cost trade-offs between electric-assist cargo bicycles and delivery trucks in dense urban areas [29]	Sheth, Manali Butrina, Polina Goodchild, Anne McCormack, Edward	Mathematical model development to evaluate the cost of delivery made with cargo vans vs. cargo bikes in four different scenarios (changing distance and parcel per stop)	Statistical model	19
Cargo cycles for local delivery in New York City: Performance and impacts [2]	Conway, Alison Cheng, Jialei Kamga, Camille Wan, Dan	Based on real data obtained from GPS-equipped cargo bikes and trucks operating in Manhattan to evaluate three comparable measures of performance: vehicle speed, delay of stopping time relative to travel time, and parking time.	Statistical model	14
Spatio-temporal route mining and visualization for busy waterways [22]	Rong Wen, Wenjing Yan, Zhang A, Nguyen Quoc Chinh, Akcan O	Explore vessels shipping patterns to provide support for decision-making process in optimal shipping route planning and maritime traffic management	Visualization, K-nearest neighbors algorithm	13
AN ASSESSMENT FRAMEWORK TO SUPPORT COLLECTIVE DECISION MAKING ON URBAN FREIGHT TRANSPORT [30]	Golini, Ruggero Guerlain, Cindy Lagorio, Alexandra Pinto, Roberto	Framework development for collecting and classifying information with the goal of enabling the city to be evaluated along the most important dimensions related to UFT.	Conceptual Framework	9
Shortening the Last Mile in Urban Areas: Optimizing a Smart Logistics Concept for E-Grocery Operations [31]	Leyerer, Max Sonneberg, Marc-Oliver Heumann, Maximilian Breitner, Michael H.	Model that contemplates 3 problems (LRP, VRP, VRPWT) and try to minimize costs along each level, always trying to maintain the best possible service	SP-VRPMPCTW model	7
Logistic Modeling of the Last Mile: Case Study Santiago, Chile [12]	Urzúa-Morales, Juan Guillermo Sepulveda-Rojas, Juan Pedro Alfaro, Miguel Fuertes, Guillermo Ternero, Rodrigo Vargas, Manuel	Based on the most widely used LM and CL frameworks, model the behavior of the city's distribution system and enable an optimization of the logistics strategy for the optimization of urban transport processes for freight deliveries.	Multi-criteria Analysis model	7
A tabu search heuristic for the bi-objective star hub location problem [32]	Ghaffarinasab N	Minimizing the total transportation cost and the length of the longest path between the O/D pairs in the hub locations	BOSHLP model Tabu Search algorithm	7

Moving towards “mobile warehouse”: Last-mile logistics during COVID-19 and beyond [33]	Srivatsa Srinivas, S Marathe, Rahul R	A mathematical model was built to test under what conditions it would be feasible to implement and adopt a mobile micro hub.	Analytical model	5
Impact assessment model for the implementation of cargo bike transshipment points in urban districts [15]	Assmann, Tom Lang, Sebastian Müller, Florian Schenk, Michael	Mathematical model development to evaluate the impact on urban quality of life, the implementation of micro hubs, with deliveries made by cargo bicycles.	Statistical model	5
Mobile Access Hub Deployment for Urban Parcel Logistics [5]	Faugère, Louis White, Chelsea Montreuil, Benoit	Mathematical model to evaluate the performance of mobile hub implementations and study the impact on a set of parameters.	Statistical model	5
An OD Flow Clustering Method Based on Vector Constraints: A Case Study for Beijing Taxi Origin-Destination Data [17]	Guo, Xiaogang Xu, Zhijie Zhang, Jianqin Lu, Jian Zhang, Hao	Dynamics of cab travel flows across origin and destination points to find similar patterns of activity to assess how many distinct communities can be detected.	Visualization, ODFCVC method K-means clustering algorithm	5
Hierarchical hub location problem for freight network design [27]	Hwang, Jaemin Lee, Jin Su Kho, Seung-Young Kim, Dong-Kyu	Reformulation of the hub location problem to account for transportation costs, the cost of hub transshipment, and the costs associated with building the link and the hub. And they try to find the optimal locations that minimize these costs.	Genetic Algorithm	5
Integrated sustainable planning of micro-hub network with mixed routing strategy [34]	Huang, Zhihong Huang, Weilai Guo, Fang	By solving the LRPMDP and LRPMDP-DR problems tried to find the locations that minimize the total operating costs	CWIGALNS algorithm	4
Substantiation of loading hub location for electric cargo bikes servicing city areas with restricted traffic [6]	Naumov, Vitalii	Distribution network model and its behavior and iterated the model 30 times to analyze which of the results obtained a lower result in terms of total work of transport.	Floyd–Warshall algorithm, Dijkstra algorithm, Clarke–Wright algorithm, Genetic Algorithm	3
Localization of Relevant Urban Micro-Consolidation Centers for Last-Mile Cargo Bike Delivery Based on Real Demand Data and City Characteristics [28]	Rudolph, Christian Nsamzinshuti, Alexis Bonsu, Samuel Ndiaye, Alassane Ballé Rigo, Nicolas	Application of a multi-criteria analysis approach (based on demand, road type and land use) to find the best location that minimizes travel time and distance	Analytic Hierarchy Process PROMETHEE (Preference Ranking Organization METHod for Enrichment of Evaluations)	2

A Conceptual Framework for Planning Transshipment Facilities for Cargo Bikes in Last Mile Logistics [35]	Assmann, Tom Bobeth, Sebastian Fischer, Evelyn	Literature review to build a theoretical conceptual model for planning UTFs.	Conceptual Framework	2
An exploratory evaluation of urban street networks for last mile distribution [23]	Amaral, Julia Coutinho Cunha, Claudio B	Geographic data collection of cities to perform a characterization of the mobility network and evaluate the difficulties imposed by the network on the distances traveled and the trips made in the last mile deliveries	Visualization, Analytical model	2
A green logistics solution for last-mile deliveries considering e-vans and e-cargo bikes [36]	Caggiani, Leonardo Colovic, Aleksandra Prencipe, Luigi Pio Ottomanelli, Michele	Reformulation of the 2E-EVRPTW-PR model to minimize costs at the two levels by considering travel costs, initial vehicle investment costs, driver salary costs, and micro hub costs.	2E-EVRPTW-PR model	2
A Two-Phase Clustering Approach for Urban Hotspot Detection with Spatiotemporal and Network Constraints [24]	Li, Feng Shi, Wenzhong Zhang, Hua	Application of data mining techniques to detect and characterize the hotspots.	Visualization, ST-HDBSCAN clustering algorithm	2
Defining Urban Freight Micro hubs: A Case Study Analysis [14]	Katsela, Konstantina Güneş, Şeyma Fried, Travis Goodchild, Anne Browne, Michael	Sample of 17 case studies to cross-reference information in order to be able to make achieve an empirical typological definition of the micro hub concept.	Empirical analysis	1
Finding the Best Location for Logistics Hub Based on Actual Parcel Delivery Data [25]	Song, Ha Yoon Han, Insoo	Based on real delivery data, search for hub locations to minimize distances and travel times.	Visualization, Longest Common Route Subsequence (LCRS) algorithm	1
A Model for Solving Optimal Location of Hubs: A Case Study for Recovery of Tailings Dams [26]	Barraza, Rodrigo Sepúlveda, Juan Miguel Venegas, Juan Monardes, Vinka Derpich, Ivan	Data mining techniques of clustering and classification, to find the locations that reduce travel distances.	K-Medoids algorithm K-Means algorithm PROMETHEE algorithm	1
Locating collection and delivery points for goods' last-mile travel: A case study in New Zealand [37]	Kedia, Ashu Kusumastuti, Diana Nicholson, Alan	Based on the location of the demand, the location of the trading facilities and the network connecting the two, application of the LA model to obtain the most reasonable location for the location of the collect and delivery points.	Location Allocation model	0

Evaluating Distribution Costs and CO2-Emissions of a Two-Stage Distribution System with Cargo Bikes: A Case Study in the City of Innsbruck [13]	Büttgen, Anne Turan, Belma Hemmelmayr, Vera	Modeling the 2E-VRP in ARGIS Pro and then using the program's built-in route solver, to determine which hub is best for the location from a list of pre-defined locations.	2E-VRP model	0
LAST MILE LOGISTICS IN THE FRAMEWORK OF SMART CITIES: A TYPOLOGY OF CITY LOGISTICS SCHEMES [38]	Özbekler, T M Karaman Akgül, A.	Literature review to define the various layers of consolidation-distribution schemes and evaluate them from the perspective of the smart city concept.	Empirical model	0
Location of urban micro-consolidation centers to reduce the social cost of last-mile deliveries of cargo: A heuristic approach [39]	Arrieta-Prieto, Mario Ismael, Abdelrahman Rivera-Gonzalez, Carlos Mitchell, John E	Mathematical model (LUMCP) to evaluate the social cost of increasing the number of micro hubs.	LUMCP model	0
BIKEMI BIKE-SHARING SERVICE EXPLORATORY ANALYSIS ON MOBILITY PATTERNS [16]	Toro, J F Carrion, D Brovelli, M A Percoco, M	Application of various data mining techniques to classify bike share users and stations according to their usage pattern	Visualization, K-Means algorithm	0
SEARCHING OPTIMAL HUB LOCATIONS IN POSTAL LOGISTIC NETWORK [40]	Madlenak, R Madlenakova, L Drozdziel, P Rybicka, I	Implementation of two mathematical models, try to find the optimal locations that allow to cover the maximum demand area, with the minimum total costs.	p-median model, UFLP model	0

2.3. Network analysis and visualization with VOSviewer

2.3.1. Keywords analysis

The analysis of the Keywords was performed using a full counting method using the threshold of 2, as the minimum number of occurrences per keyword. As a result, VOSviewer identified a total of 17 keywords which met the threshold and from those, all were selected.

Most of the analyzed words are terms associated with urban logistics. The most frequent words are "city logistic", "data mining", "urban logistics", "human mobility" and "last-mile logistics".

Table 2.3 - Keyword occurrences ranked by total link strength

Keyword	Occurrences	Total link strength
city logistics	5	9
urban logistics	4	9
cargo bike	3	9
human mobility	4	7
public transportation	3	7
urban freight	3	7
trajectory	2	6
clustering algorithms	2	5
urban planning	2	5
last-mile logistics	4	4
patterns	2	4
transportation	2	4
cargo bicycles	2	3
logistics	2	3
data mining	5	2
algorithm	3	2
facility location	2	2

In Figure 2.2 we verify the existence of 3 clusters, the largest is composed of a set of 9 words (identified in the image in red) very focused on urban logistics, while in the data exploration part we can verify the presence of the reference to "clustering algorithms" which in turn is linked with "trajectory" and "patterns"(not visible description), "human mobility" and "public transport".

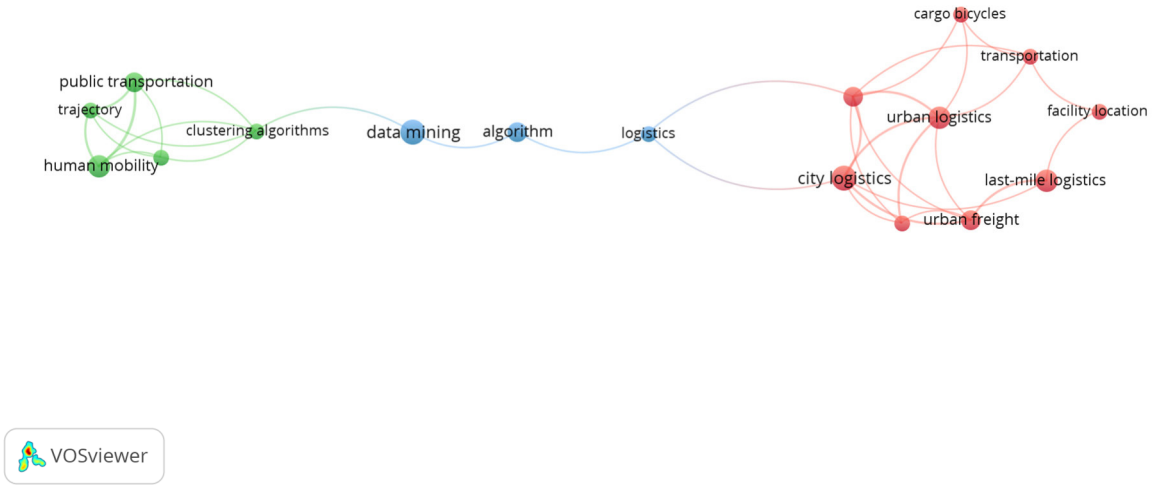

Figure 2.2 - Keyword occurrence network visualization

In Figure 2.3 we can see that "urban freight" and "last mile logistic" have been more frequently mentioned terms in recent literature which may indicate a growing interest in the area.

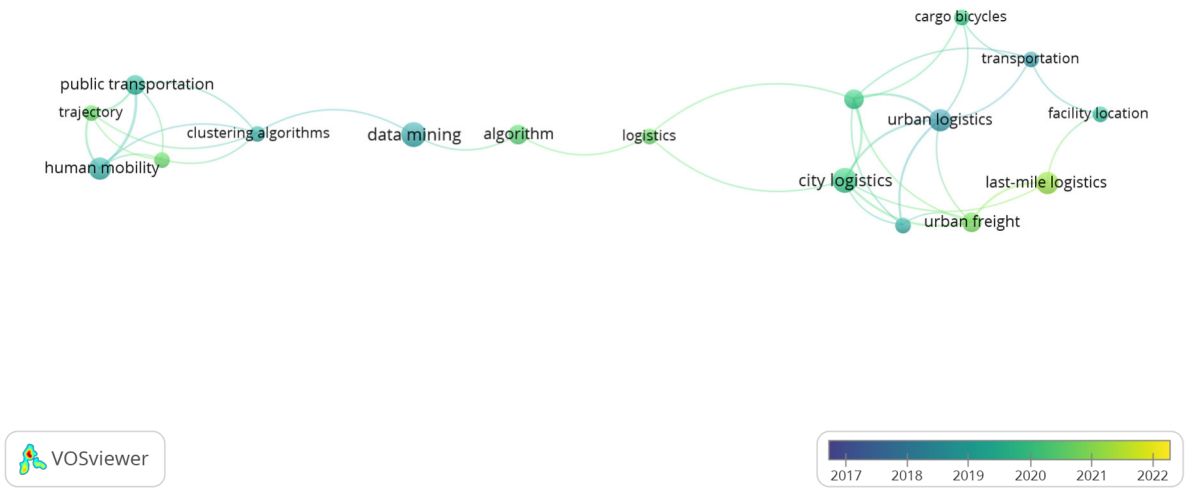

Figure 2.3 - Keyword occurrence network overlay visualization

2.3.2. Title and abstract analysis

The analysis of the title and abstract was performed using a full counting method using the threshold of 3, as the minimum number of occurrences per keyword. As a result, VOSviewer identified a total of 121 words which met the threshold and from those, only 41 keywords were selected.

We found 9 clusters, 113 links and a total link strength of 587. By analyzing Figure 2.4, Figure 2.5 and Table 2.4 we can observe that the first cluster, the largest one is composed of 8 red elements and is composed by terms associated with the delivery of the last mile, we see interesting terms that may represent important characteristics in this type of operation, for example "travel distance" and "demand". But we can also see that in the literature the terminology is not well defined, with several nomenclatures for the same identification (e.g., "e cargo bike" and "ea cargo bike"). In Overlay visualization we can find a mention of electric vehicles in the most recent literature ("e van" and "e cargo bike") confirming the trends that have been seen in the transitions of fleets to more environmentally friendly vehicles.

Table 2.4 - Title and abstract text occurrence keywords occurrence and total link strength

Keyword	Cluster	Occurrences	Total link strength
cdp	1	7	14
demand	1	7	31
cargo cycle	1	6	14
road	1	5	25
stakeholder	1	5	10
loading hub location	1	4	8
travel distance	1	3	14
umcs	1	3	9
sustainability	2	7	42
city logistic	2	6	29
dockless bike	2	5	10
efficiency	2	5	19
urban logistic	2	4	14
microhub	2	3	6
mobile access hub deployment	2	3	18
pattern	3	18	52
fvp	3	4	28
traffic congestion	3	4	25
individual travel pattern	3	3	24

mobility	3	3	25
e van	4	6	69
customer	4	5	29
last mile delivery	4	5	58
e cargo bike	4	4	48
restricted traffic zone	4	3	39
cargo bike	5	7	57
city hub	5	5	50
urban transshipment point	5	4	16
stage distribution system	5	3	36
cluster	6	8	34
optimal location	6	4	2
urban hotspot	6	4	8
od flow	6	3	18
last mile logistic	7	7	52
covid	7	5	35
mobile warehouse	7	4	32
delivery truck	8	8	48
parcel	8	6	61
ea cargo bike	8	3	33
spatio temporal data	9	4	12
spatio temporal data mining	9	3	12

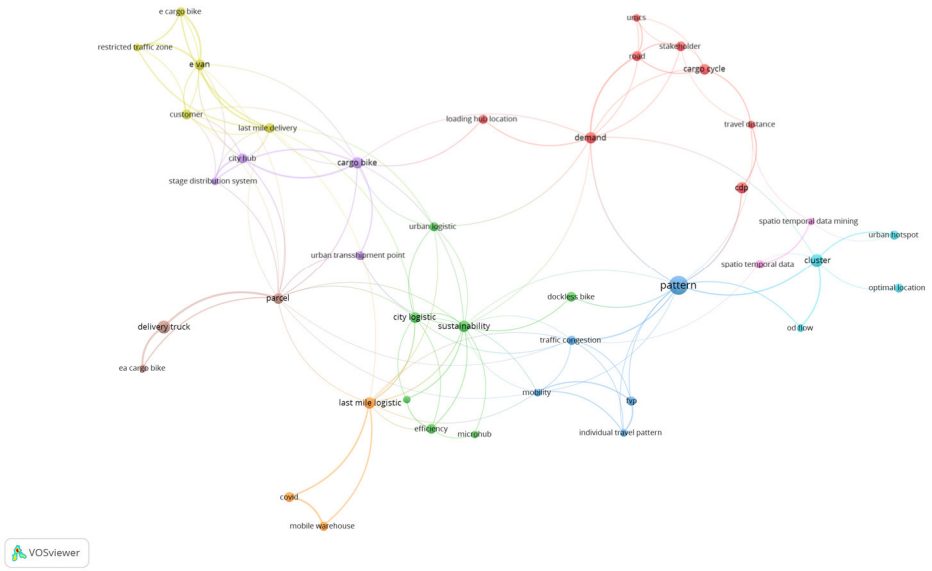

Figure 2.4 - Title and abstract text occurrence network visualization

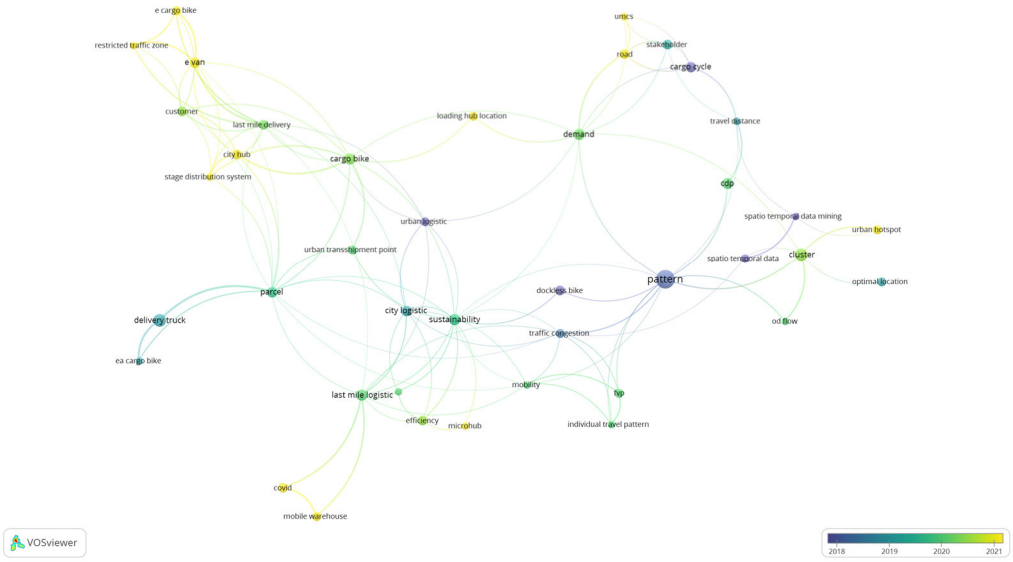

Figure 2.5 - Title and abstract text occurrence network overlay visualization

2.3.3. Author and co-author analysis

In the analysis of authors and co-authors, the selected parameters were full counting, with a maximum number of documents per author of 25 and a minimum of 1. With these settings we obtained 127 authors, which were selected for the visual analysis. We can evaluate that the authors present with a larger number of documents are Assmann, Tom [15], [35] and Goodchild, Anne [14], [29] both with 2, all the others are represented with only 1 document.

In the visual analysis we obtained 32 clusters being the largest cluster composed of 8 elements and composed between the years 2020 and 2022.

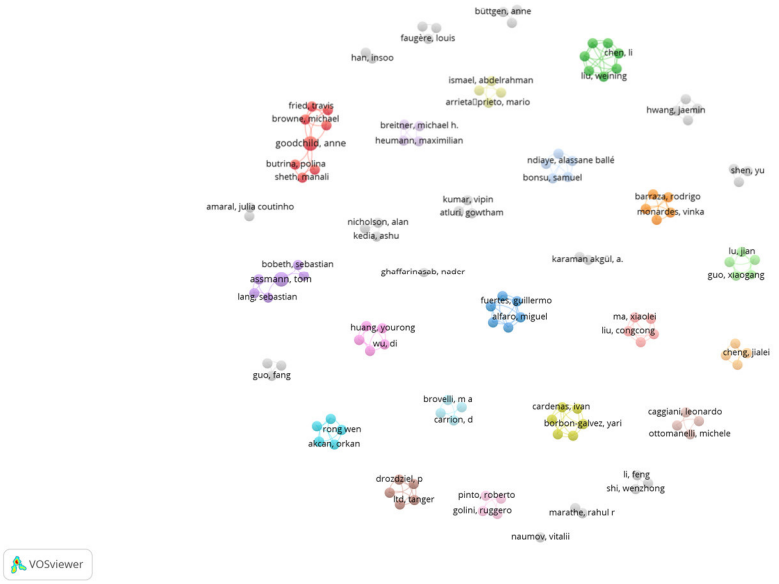

Figure 2.6 - Author and co-author network visualization

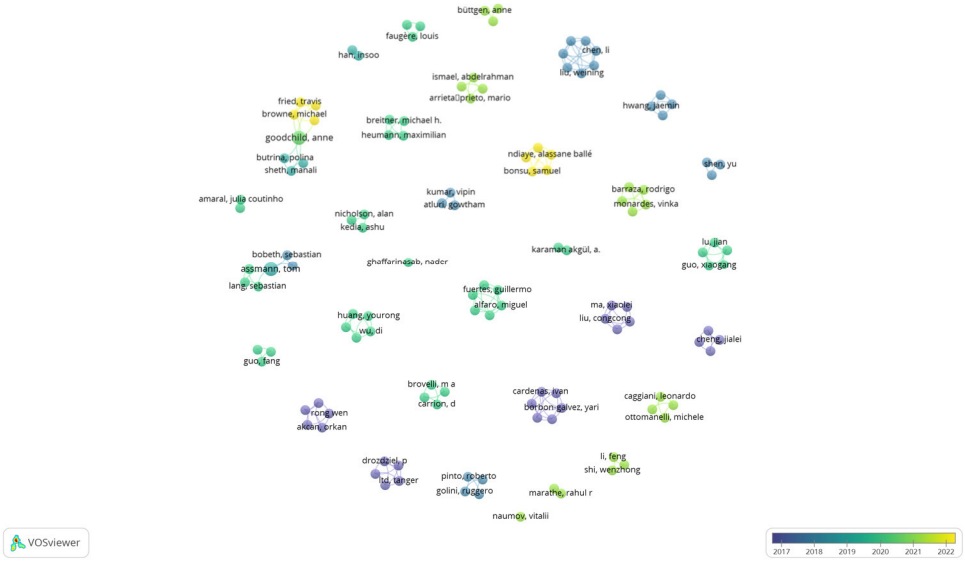

Figure 2.7 - Author and co-author network overlay visualization

2.3.4. Result overview

Within the limits of our bibliographical research, we found that the study of e-cargo bikes is still very limited and focused on the scope on environmental impact and benefit of cargo bike usage. Very few of them analyze the behavior and performance patterns of last mile delivery of e-cargo bikes in urban centers.

The importance of using spatiotemporal analysis in comparison to traditional data mining approaches that consider instances to be "distributed equally and independently", is due to the possibility to find existing links between the various instances of our data in space and time [7]. Ignoring these connections can lead to misinterpretation and results that are difficult to understand [7], [24].

We observe homogeneity in the spatiotemporal analyses evaluated. The clustering technique for pattern detection was the most present in our SLR. Zeng et al. [20] characterized the taxi travel patterns of Chongqing residents from two perspectives, hot spots and hot paths, by applying the GRIDBSCAN and ST-TCLUS (Spatial-temporal trajectory clustering) clustering algorithms. It allowed to conclude that depending on the time of day these areas varied according to their land use. Y. Huang et al. [21] studied the travel patterns of private cars to identify the most frequented sites using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm and Markov chains, allowed them to identify that 59% of car trips exhibit regular spatiotemporal mobility and repeated travel patterns. By applying the ST-HDBSCAN clustering algorithm (combination of ST-DBSCAN and HDBSCAN clustering algorithms) Li et al. [24] made a spatiotemporal characterization of the hotspot characteristics, through the study of "Spatiotemporal Distribution", "Travel Distance Distribution" and "Travel Direction Distribution", concluding that the most frequented areas are the ones where there is a higher density of points of interest. Toro et al. [16] study the mobility patterns of users of Milan's bike sharing systems and using the clustering technique with the K-Means algorithm allowed him to identify which stations have the same usage pattern. In the exploitation of the most frequent paths made in the Singapore Strait Ron, Wen et al. [22] applied the clustering technique with the K-nearest neighbors algorithm, to clustering time series of waterways, allowed them to identify the most congested areas spatially and temporally.

Atluri et al. [7] state, that in exploring problems with spatiotemporal data, finding the similarities or dissimilarities between instances is the key to solving most challenges.

In the studies collected, the evaluation of the performance of cargo bikes is highly focused on comparing with the performance of the cargo vans in the last mile delivery [2], [23], [29], [36].

The cargo bikes presented a greater flexibility and advantage in the routes they made. Most of times the chosen bike route is shorter than the route made by vans [2]. This difference can be up to twice as large on shorter trips [23]. It was also found that cargo bike riders easily break traffic regulations by riding in the opposite direction during short trips [2], Amaral et al. [23] identifies that travel times are not as important for cargo bikes as for motor vehicles, because bicycles can easily "outrun" traffic jams. An interesting observation is made in Conway et al. [2], contrary to what would be expected the speed of cargo bikes on the bike paths is lower than when they go on the roads intended for motor vehicles, with a speed lower by up to 20% on some of the routes. The impact of street topography is mentioned in Amaral et al. [23] and defines a scale between the elevation and the impact on cyclist performance, this scale sets as a reference, below 2% has no effect, between 2% and less than 5% already has a considered impact and above 5% already represents a substantial impact. The speed considered in the studies is not homogeneous, varying between 11.6 Km/h [2] and 24 Km/h [29], a literature review done by Büttgen et al. [13] finds an average speed of this type of vehicles between 8 Km/h and 25 Km/h.

Overall, all studies conclude that cargo bikes represent a more viable and advantageous alternative in last mile delivery, with greater gains in more congested areas [2], but with some constraints. Sheth et al. [29] concluded that the distance and the number of deliveries is the most impacting factors on viability and cannot exceed 3.2 Km and 20 orders per stop. In Amaral et al. [23] the capacity of the vehicle is not considered, but it concludes that beyond 3.0 Km, it is no longer efficient to deliver with this type of vehicle.

The combination of cargo bikes and the implementation of micro hubs has helped the green alternatives to last mile delivery gain momentum [15]. Distribution networks with micro hubs do promote a more organized last mile delivery [14] and benefit from economies of scale [39].

The definition of micro hub in the literature is vast, for our paper we will adopt the definition by Katsela et al. [14], which defines it as logistics facilities where commercial transportation providers (or "carriers") consolidate goods near the final delivery point and serve a limited spatial delivery area in a dense urban environment.

Finding and defining a location for micro hubs is an important and complex task [6], [14], the rising costs of urban land, lack of adequate infrastructure, changing demand, changing city characteristics [28] and regulatory requirements [14], do not ease the task of being able to find an optimal solution that minimizes operating costs and impact on communities. The most common characteristics addressed in the literature to study this problem are demand (e.g., residential, commercial, and/or employment density), infrastructure considerations (e.g., pedestrian/bicycle infrastructure provision, road classifications, pedestrian zones, and measures to assess traffic), and land use constraints [12], [28], [30].

When the deliveries are made by cargo bikes, the location of the micro hub should always be the closest to the delivery point [28], [38]. Assman et al. [15] recommends locating them in areas of higher commercial density. This need for proximity comes from the capacity limitation of bikes compared to a delivery van, and multiple trips to the micro hub may be required, so travel time and travel distances should be minimized [14]. According to Assman et al. [15] the maximum distance between the micro hub and the delivery point should not exceed 1000m, in Rudolph et al. [28] a distance between 500 meters and 1200 meters is pointed out as the distance range that allows an economic feasibility for deliveries made by cargo bikes.

In Faugère et al. [5] and Srivatsa Srinivas et al. [33] the implementation of this type of infrastructure in mobile units was evaluated and, in both studies, they conclude that it can be a viable alternative, but under very restricted conditions, Faugère et al. [5] indicates as a condition the requirement to transport a high volume of orders and a very short maximum transit travel time. In Srivatsa Srinivas et al. [33] the need for a strong analytical engine that can accurately predict demand for a given geographic location and the dynamic optimization of the route and parking location of the mobile warehouse is the only way to make this alternative viable.

The study of stationary micro hubs is the most widely covered in the literature, but the methods have proven to be varied among the studies.

When the targets' location points are already known, [6], [13], [31], [37] only an evaluation of the performance of each of the locations was done to find the one that best suited the purpose.

Naumov et al. [6] developed a mathematical model representative of the network and its behavior and by applying Monte Carlos simulation, evaluated which of the five pre-defined locations allowed minimizing the transportation work. In Kedia et al. [37], the Location-Allocation model, was used to find the locations that minimized the distance that had to be traveled. Bütten et al. [13] uses the 2E-VRP model to find an optimal solution that minimizes costs. In Leyerer et al. [31], the SP-VRPMPCTW model is solved, to minimize costs throughout the three stages (LRP, VRP with time window and VRP considering multiple products) that compose model.

When there is no pre-knowledge of such locations, other approaches are needed, and possible solutions can be found based on the knowledge of the demand or the geographical characteristics of the cities. Rudolph et al. [28] uses a multi-criteria method to find the most suitable locations and employs the AHP and PROMETHEE algorithms, defining that the main criteria to use are demand, road type and land use. The optimal locations should minimize travel times and travel distances. Song et al. [25], use the LCRS (Longest Common Route Subsequence) algorithm, complemented with a voting system, to find the paths most traveled and where there is a higher concentration of deliveries. This approach allows them to calculate which locations can minimize the time and distance traveled.

In the literature we found that in the approach to this problem the computational capacity and the time required to explore all possible options, limit the calculation of the optimal points, making one choose to find only the optimal points [25], [27], [32], [34] and the implementation costs of micro hubs and vehicle capacity are often not considered.

We can argue that minimizing distances, travel times, and costs are among the most relevant objectives in locating hubs.

3. Data analysis and modeling

3.1. Data mining with CRISP-DM

The CRISP-DM [9] (CRoss Industry Standard Process for Data Mining) methodology will be applied in our research. This methodology attempts to reduce the cost and increase reliability, repeatability, manageability, and speed of big data mining operations. According to this methodology the life cycle of data mining projects is divided into six parts, as shown in Figure 3.1.

Figure 3.1 - CRISP-DM methodology flowchart

Business understanding is the first phase of CRISP-DM and focuses on gaining business understanding of the project objectives and requirements, transforming that information into a data mining issue definition and a preliminary project plan to meet the goals. The second phase is data understanding phase, which begins with data collection, finding data quality issues, getting early insights into the data, or uncovering subsets to create hypotheses about hidden information. Data understanding and business understanding are inextricably linked. A basic grasp of the available data is required for the creation of the data mining challenge and the project strategy. The third phase is data preparation and encompasses all operations that will result in the final dataset or datasets. The fourth phase is modelling, and various modeling approaches can be chosen (such as data science and machine learning models) and employed during this phase, and their parameters can be calibrated to ideal levels. For the same data mining issue type, there are usually many possible modelling approaches. The fifth phase is evaluation and aims to evaluate the models that were built in the previous phase, using established performance evaluation criteria. Here we can evaluate both the modelling approach and the extent to which the business requirements have been met. The main goal is to see whether there is any significant business issue that has not been adequately addressed. A choice on how to use the data mining results should be made at the end of this step. The sixth phase is the deployment. At this point the information gathered must be structured and presented in a way that the stakeholders can understand, can be provided as a simple report, a publicly or privately deployed tool, or more complex explanation to make possible the repeatable data mining process. [41]

3.2. Business understanding

The data explored was provided by the e-cargo bike urban logistics startup YOOB [42]. As mentioned, the purpose of the study is two-fold: the first, to provide a spatiotemporal characterization of the behaviors of YOOB e-cargo bike fleet in the parcel collection and delivery processes in Lisbon as well in its outskirts; the second, to propose locations for the new hubs and adjustments to the existing network in order to strengthen and expand the fleet operations. The company has two types of hubs, the micro-hub with an area of 36 m^2, a relatively smaller option compared to the values found in SLR, which range between 92 m^2 to 920 m^2 [43]. The functional definition is in line with that found at SLR, with various services being done at the micro hub has, namely, consolidation of goods, storage and recharging of e-cargo bikes. The nano-hubs, which is an innovative concept of YOOB, emerged from the adaptation of the pick-up/drop-off concept to last mile delivery logistics, characterized by having relatively small areas ranging between 3 m^2 and 120 m^2, exclusively dedicated as a temporary transition point were the goods remain no longer than 48 hours. The type of associated physical infrastructure varies depending on where it is implemented, given it only requires temporary storage capacity for goods.

3.3. Data understanding

The data provided comes from YOOB's database, and the period covered ranges from January 1st to April 30th 2022, with 9 175 records and 34 variables. The data does not provide the routes (trajectories) taken by the fleet. The geographic information on the route is characterized by the latitude and longitude of origin and destination. There are some variables that are generated based on a mobile device used by the employees during their entire operation. Table 3.1 describes the database variables, their origin, and the decision option regarding their use or not in this thesis.

To clarify the terminology used throughout the thesis, each record in the data received represents a “story”, which is geographically composed of two points, one for pickup and the other for delivery. Within each story there are two “sub-stories”, where each “sub-story” refers to a geographical location (pickup or delivery) and is always associated to a “route”, where the “routes” can be composed of one or more “stories”.

Table 3.1 – Database schema provided by YOOB

Column Name	Description	Origin	Format	Excluded	Reason
_id	YOOB Database id		object	Yes	YOOB Database id
uid	Unique id of record		object	No	
type	Type of service (pickup or delivery)		object	No	

reference	Reference code of the parcel		object	Yes	Not relevant for our analysis
date	Date when route was completed		datetime64[ns]	No	
Parcels	Number of Parcels	Mobile Device	int64	No	
Weight	Weight of the parcel(s)	Mobile Device	float64	Yes	Many missing values and non-standard filling. Out of scope.
driverUID	Unique id of driver		object	No	
pickupUID	Unique ID associated to YOOB known location		object	No	
pickupName	Name associated to the pickupUID		object	No	
pickupLongitude	Longitude coordinate where the pickup was made	Mobile device	float64	No	
pickupLatitude	Latitude coordenate where the pickup was made	Mobile device	float64	No	
dropoffLongitude	Longitude coordinate where the delivery/dropoff was made	Mobile device	float64	No	
dropoffLatitude	Latitude coordinate where the delivery/dropoff was made	Mobile device	float64	No	
State	State of the record (number code)		int64	No	
stateText	Text describing the state variable code		object	No	
failedReason	Reason of the failure delivery	Mobile device	object	Yes	Not relevant for our analysis
Assignee	Driver assigned to the route		object	No	
assignedRoute			object	Yes	Duplicated info (assignedRouteUID)
assignedRouteUID	Unique ID of Route		object	No	

failedStates	Datetime of the last failed delivery state	Mobile device	datetime64[ns]	Yes	
failedStatesCount	Counter with the number of delivery failures		int64	No	
history.cancelled	Datetime of the failed state	Mobile device	datetime64[ns]	No	
history.failed	Datetime of the failed state	Mobile device	datetime64[ns]	No	
history.pickupFailed	Datetime when the pickup failed	Mobile device	datetime64[ns]	No	
history.created	Datetime when the history is created for the first time	Mobile device	datetime64[ns]	No	
history.sorted	Datetime when parcel is sorted	Mobile device	datetime64[ns]	Yes	Many missing values (99%)
history.loaded	Datetime when parcel is loaded	Mobile device	datetime64[ns]	Yes	Many missing values (93%)
history.pickupEnRoute	Datetime when the pickup is in movement	Mobile device	datetime64[ns]	No	
history.pickupArrived	Datetime when the pickup arrived to the place	Mobile device	datetime64[ns]	No	
history.pickupCompleted	Datetime when the pickup is declared as completed	Mobile device	datetime64[ns]	No	
history.enRoute	Datetime when the delivery/dropoff is in movement	Mobile device	datetime64[ns]	No	
history.arrived	Datetime when the delivery/dropoff arrived to the place	Mobile device	datetime64[ns]	No	
history.completed	Datetime when the full history is declared as completed	Mobile device	datetime64[ns]	No	

3.4. Data preparation

For the process of data preparation, modeling and visualization of the data, we adopted the python programming language (v3.10.4) [44], compiled with Visual Studio Code (v1.69.1) [45] on Jupyter

Notebooks extension [46]. The packages used to handle geographical data were: Descartes [47], geog [48], geopy [49], geopandas [50], rasterio [51] and osmnx [52]. For visualization: contextily [53], geoplot [54], folium [55], mapclassify [56], matplotlib [57], plotly [58], matplotlib_scalebar [59] and seaborn [60]. To statistical and data processing, numpy [61], pandas [62], pyproj [63], scipy [64], shapely [65], sklearn [66] and kneed [67].

The first data preparation procedure was the individual evaluation of all variables. The actions taken in this step were as follows:

- Dropped variables ['_id'], [‘reference’], [‘Weight’], ['failedReason'], ['assignedRoute'], ['failedStates'], [‘history.sorted’] and ['history.loaded'].
- In the variable [‘Parcels’] we have detected an outlier value with 620 parcels. We fixed to 62, and the records with 0 values (14%) were corrected to 1(mode value) in the first briefing with the company we were informed, that sometimes the employees did not assign any parcel to the travel wrongly.
- The variable [‘assignedRouteUID’] had 3.3% of null values and for those records we gave them a unique identifying composed by “RTD-” plus the value of the variable [‘uid’]. We took this option to not loose records and to always be able to identify them if needed.
- Dropped all records that were not filled in the variable stateText with “complete” or “failed” (7,12%)

The result was a dataset with 8 516 (92,8%) records each one with 26 variables. Some of the variables had “NULL” values as they are not required to be fill.

The second procedure was to classify the stories as “complete” or “incomplete”, based on the filled timestamps, resulting in the variable [‘Classf_Route’].

To considered a story as “complete” the follow requirements must be met:

- If the type of the story is a delivery, then the variable [‘history.enRoute’] must be filled together with one of the variables [‘history.completed’], [‘history.failed’] or [‘history.arrived’].
- If the type of the story is a pickup, then the variable [‘history.pickupEnRoute’] must be filled with one of the variables [‘history.pickupCompleted’], [‘history.completed’], [‘historyPickup.failed’] or [‘history.arrived’].

This procedure has resulted in the reduction of our dataset by 1.58% (135 records).

The third procedure was to duplicate the dataset in order to have each row referring to a sub-story.

To avoid removing information, we created another variable, named ['ct_type'], which assumed the value of “pickup” if the sub-story represented the pickup of the story or “delivery” if the sub-story represented the dropoff/delivery of the story.

The values on the variables [‘history.pickupEnRoute’], [‘history.pickupArrived’] and [‘history.pickupCompleted’] were merged to the columns [‘history.enRoute’], [‘history.arrived’] and [‘history.completed’].

The values on the variables [‘pickupLongitude’], [‘pickupLatitude’], [‘dropoffLongitude’] and [‘dropoffLatitude’] were merged into the variables [‘Latitude’] and [‘Longitude’].

After this duplication our dataset was characterized by 16 762 records. Although, after validation of the filled timestamps we found more incomplete records and had to remove 148 records (0.89% of the total records). The result of this final step was a dataset with 16 614 records and 23 variables

To enrich our dataset, we added extra features:

- ['Elevation_point'] - Elevation of the sub-story geographic location, was obtained by consulting a DEM (Digital Elevation Map) [68], previously converted from Coordinate Reference System 3 035 to 4 326. Then the points were mapped and the respective elevations were obtained.
- ['order route']: Number indicating the order in which the location is visited within the route sequence. In order to find this sequence, we had to recreate the routes, grouping them by the variable ['assignedRouteUID'] and then sorting ascending by the time of the variable ['history.arrived']. Each route has an associated day that corresponds to the day on which the route was completed, within the route we only took in count sub-stories that were within that day. In the briefing with YOOB it was referred that the routes were only valid for one day. However, we detected some cases where a route was performed but the delivery failed, and the route was carried over to the next day with the addition of more deliveries. In these cases, we chose to keep the sub stories that were within the day on which the route was completed. After this step it was not possible to reconstruct 7 routes. The sub-stories associated with these routes and the sub-stories that were not within the route day were removed representing 786 records a 4.7% in the 16614 total records.
- [‘time_enRoute_sec]: Time period in seconds between the [‘history.enRoute’] and [‘history. arrived’].
- [‘time_points_sec’]: Time period in seconds between two consecutive points on the same route, measured by the difference of the two values of [‘history.arrived’].

The first dataset with sub-story granularity resulted in 15 828 records and 27 variables.

Table 3.2 - Schema dataframe one - sub-story granularity

Column Name
uid
type
date
Parcels
driverUID
pickupUID
pickupName
Longitude
Latitude
state
stateText
Assignee
assignedRouteUID
failedStatesCount
history.cancelled
history.failed
history.pickupFailed
history.created
history.enRoute
history.arrived
history.completed
ct_type
Classf_Route
order_route
time_enRoute_sec
time_points_sec
Elevation_point

In the monthly and weekday distribution of sub-histories in Figure 3.2 and Figure 3.3, we observed a higher concentration of records in the month of March, but during the days of the week it is more evenly distributed, except in the weekends, where there were very few records, mostly due to the fact that the company operates the most from Monday to Friday.

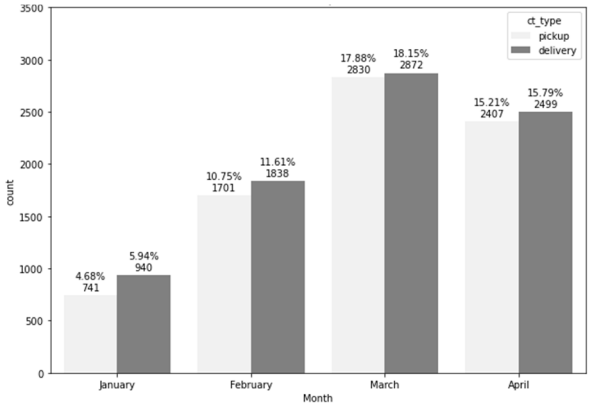

Figure 3.2 - Sub-story month distribution

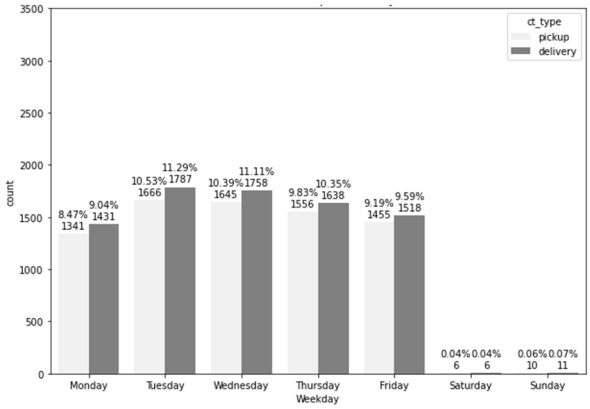

Figure 3.3 - Sub-story day week distribution

In Figure 3.4 we observe the hours of highest activity. The pickups have a prominent incidence during the period of 9 am, while in deliveries we have a distribution with two peak incidence periods, the first during the morning at 11 am, and the second in the afternoon at 3 pm.

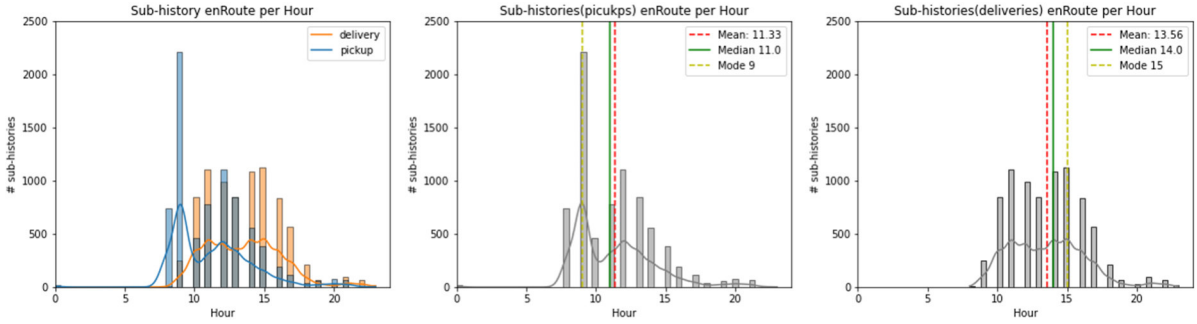

Figure 3.4 - Sub-stories per hour

In the volume of parcels transported we can infer from Figure 3.5 that at the end of the day in the periods after 3 pm the sub-stories tend to have more parcels associated. This trend reflects the redistribution of the goods by the hubs, for the distribution in the next day.

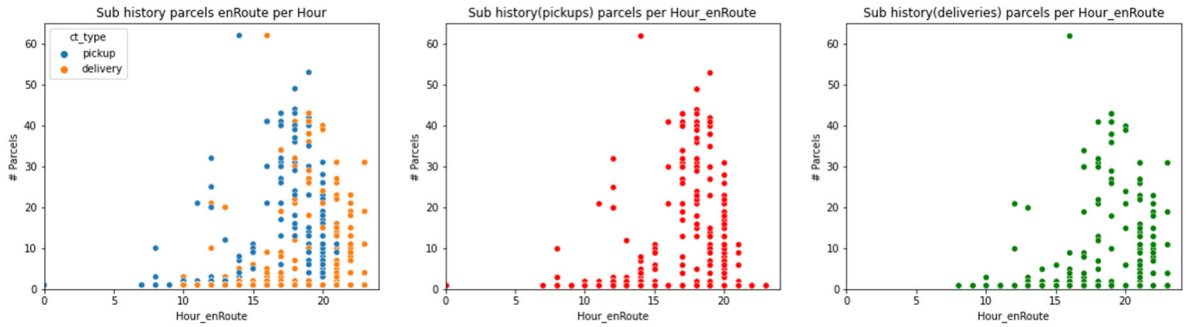

Figure 3.5 - Parcels per sub-story

To perform the spatial analysis two geographic data frames were generated with geopandas [50]. In the first the granularity was the route level, and second the granularity was the sub-story level. We performed a loop based on the first dataset, which went through all the unique values of the variable ['assignedRouteUID'] to filter all the sub-stories of the route and generate the two geographic data frames represented in Table 3.3 and Table 3.4. To be considered valid, a route must have two or more associated sub-stories. Routes that do not meet this requirement were removed.

Table 3.3 – Schema geodataframe with routes granularity

Column Name	Description	Format
rotaUID	unique id for route identification based on the variable ['assignedRout'UID']	object
dia	Day when the route was completed	datetime64[ns]
dia_semana	Weekday of the route	int64
semana_iso	Calendar number of the Week based on ISO 8601	int64
mes	Number of the month of the route	datetime64[ns]
InicioH	Hour of the first sub-story based on the variable ['history.enRoute'] with a movement assigned	int64
InicioM	Minutes of the first sub-story based on the variable ['history.enRoute'] with a movement assigned	int64
FimH	Hour of the last sub-story based on the variable ['history.arrived']	int64
FimM	Minutes of the last sub-story based on the variable ['history.arrived']	int64
geometry_ini	Geographical location where the route starts	geometry
uid_do_ini	Unique identification of the start location of the route	int64
geometry_fim	Geographical location where the route ends	geometry
line3d	Sequence list with the geographic route sequency of all sub-stories formed by [latitude,longitude,elevation]	object

direcao_ini_seq	Sequence list with cardinal direction of all sub-stories relative to the initial location	object
direcao_prev_seq	Sequence list with cardinal direction of the route relative to previous visited location	object
distancia_ini_seq	Sequence list with distances in meters of all sub-stories relative to the initial location	object
distancia_prev_seq	Sequence list with distances in meters of all sub-stories relative to the previous location	object
dif_elev_seq	Sequence list with the elevation difference in meters relative to the previous location of all sub-stories	object
distancia_total	Sum of the list ['distancia_prev_seq'] in meters	float64
distancia_media_entre_pontos	Mean value of the list ['distancia_prev_seq'] in meters	float64
distancia_maxima_do_ini	Max distance value on the list ['distancia_ini_seq'] in meters	float64
Parcels_pickup	Sum of all parcels in the pickup sub-stories	int64
Parcels_delivery	Sum of all parcels in the delivery sub-stories	int64
Pontos_totais	Count of sub-stories	int64
Pontos_com_deslocacao	Count of sub-stories with distance traveled detected	int64
Pontos_pickup	Count of pickup sub-stories	int64
Pontos_delivery	Count of delivery sub-stories	int64
pickups_extra	Count of pickup sub-stories that occure after the route starts	int64
entregas_falhadas	Sum of ['failedStatesCount']	int64
media_enRoute_sec	Mean measured in seconds of all sub-stories['time_enRoute_sec'] with distance traveled detected	float64
media_entrepontos_sec	Mean measured in seconds of all sub-stories ['time_points_sec'] with distance traveled detected	float64
tavg	Average temperature on the day of the route (Celsius degrees)	float64
tmin	Minimum temperature on the day of the route (Celsius degrees)	float64
tmax	Maximum temperature on the day of the route (Celsius degrees)	float64
prcp	Precipitation on the day of the route (mm)	float64
total_enRoute_sec	Sum of all sub-stories ['time_enRoute_sec'] in seconds	float64
total_entrepontos_sec	Sum of all sub-stories ['time_points_sec'] in seconds	float64

Table 3.4 -Schema geodataframe with sub-story granularity

Column Name	Description	Format
date_day	Day of the sub-story	datetime64[ns]
Parcels	Number of Parcels	int64
Longitude	Longitude coordinate	float64
Latitude	Latitude coordinate	float64

assignedRouteUID	Unique identification of the route which the sub-story belongs	object
failedStatesCount	Number of failed attempts	int64
ct_type	Type of the sub-story	object
order_route	Number indicating the order in which the local is visited within the route sequence.	int64
time_enRoute_sec	Represents the difference of time in seconds between the ['history.enRoute'] and ['history. arrived'] in seconds	float64
time_points_sec	Time period between two consecutive locals on the same route, measured by the difference of the two values of ['history.arrived'] in seconds	float64
Classf_hist	Sub-story classification	object
Elevation_point	Elevation of the geographical local in meters	float64
linear_distance_last_to_point	Euclidean distance measured in meters	float64
Hour_arrived	Hour of the sub-story based on the variable ['history.arrived']	int64
Minute_arrived	Minutes of the sub-story based on the variable ['history. arrived']	int64
Hour_enRoute	Hour of the sub-story based on the variable ['history.enRoute']	int64
Minute_enRoute	Minutes of the sub-story based on the variable ['history.enRoute']	int64
Weekday	Weekday of the sub-story	int64
Month	Number of the month of the sub-story	int64
geometry	Coordinates (latitude, longitude)	geometry
uid_do_ini	Unique identification of the sub-story start location	int64
dist_to_prev	Euclidean distance calculated with great circle and elevation difference, distance to previous visited local measured in meters	float64
dif_elev_prev_point	Difference of elevation between previous visited location	float64
orient_Prev_point	Cardinal direction of the sub-story relative to previous visited location	object
dist_INI	Euclidean distance calculated with great circle and elevation difference, distance to initial start location of the route measured in meters	float64
orient_INI_point	Cardinal direction of the sub-story relative to the initial start location of the route	object
tavg	Average temperature on the day of the sub-story (Celsius degrees)	float64
tmin	Minimum temperature on the day of the sub-story (Celsius degrees)	float64
tmax	Maximum temperature on the day of the sub-story (Celsius degrees)	float64
prcp	Precipitation on the day of the sub-story (mm)	float64

With this procedure we were able to reconstruct 664 routes, representing 95% of the total routes in the original dataset (699 routes), at sub-story level. We have removed 20 records (<0.002%), ending with 15 808 records.

To visualize where the highest density of sub-stories were located, a heatmap was created with folium [55] with the built-in OpenstreetMap [69] tileset feature. In Figure 3.6 the spatial distribution of the sub-stories is very wide, but there were two outstanding higher density zones. Therefore, it was necessary to analyze and characterize in more detail the type of density areas. Separating the plot by the variable ['ct_type'], we observed the difference between the density of the pickup sub-stories, Figure 3.7 , and delivery sub-stories in Figure 3.8.

The pickup density was restricted to few locations, while the deliveries were widely spread with only two main detected hot spots.

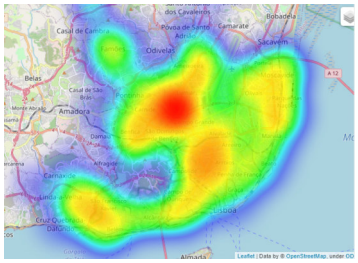

Figure 3.6 - Heatmap of all sub-stories

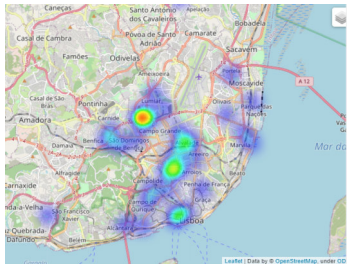

Figure 3.7 - Heatmap of pickup sub-stories

Figure 3.8 - Heatmap of delivery sub-stories

To be more accurate, we created a choropleth map with all Lisbon boroughs. Before performing this visualization, we had to count the number of sub-stories that were in each borough. With the geopandas [50] feature "contains", we found if the location of our sub-story was inside the geometric polygon of each borough. A temporary geodataframe was created containing the name of the borough, the borough geometry (polygon) and the number of sub-stories that polygon contains.

In the choropleth map of Figure 3.9, we plot with all sub-stories on the boroughs with the highest activity were Lumiar and Avenidas Novas with 28,77% and 15,70% respectively (see Figure 3.10). Analyzing by type of sub-story we can observe in Figure 3.11 that the deliveries were in higher number in Lumiar, Alvalade, São Domingos de Benfica and Avenidas Novas, and the pickups in Figure 3.12 were higher in Lumiar, Marvila,Avenidas Novas and Alcântara. These four main pickups areas are located where the YOOB has its hubs: one micro-hub in Lumiar and nano-hubs in Marvila, Avenidas Novas and Alcântara.

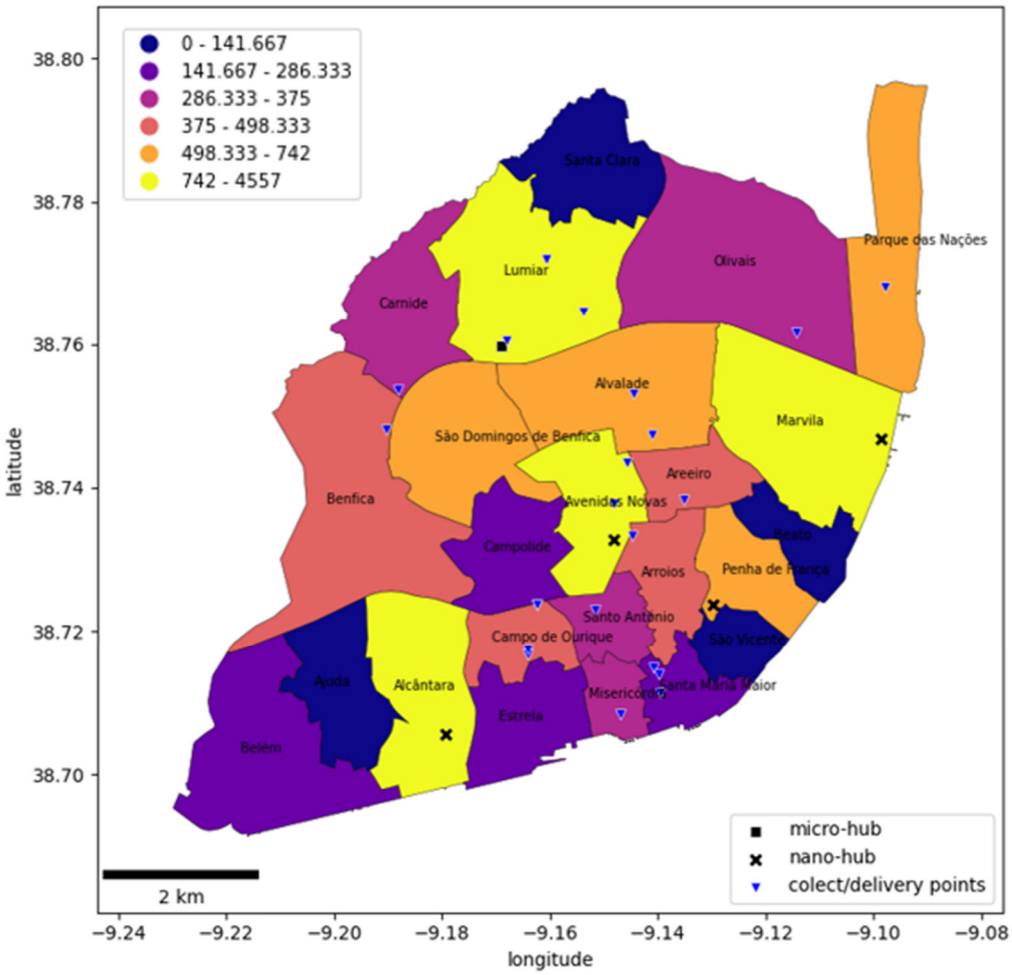

Figure 3.9 - Sub-stories numbers per borough

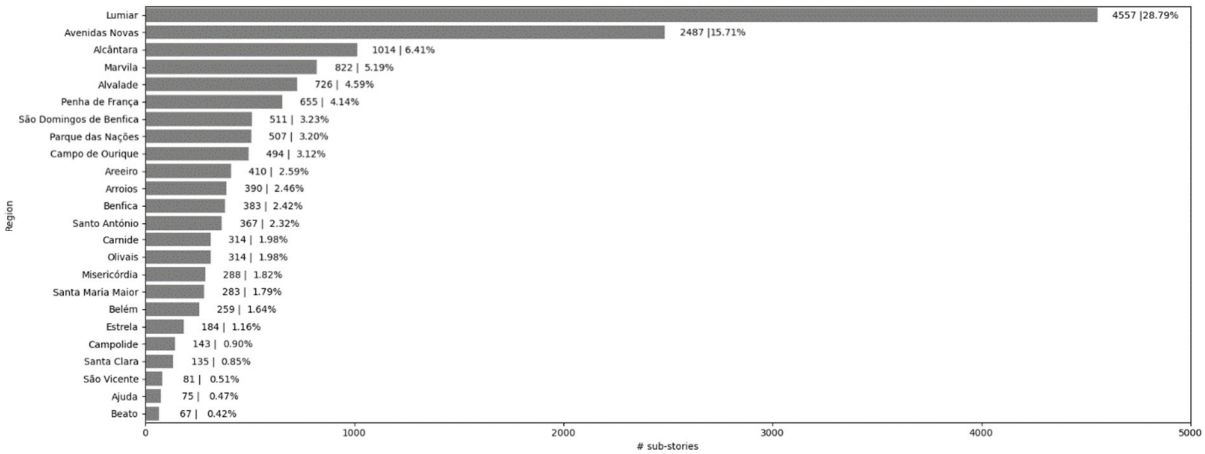

Figure 3.10 - Sub-stories number per borough and percentage

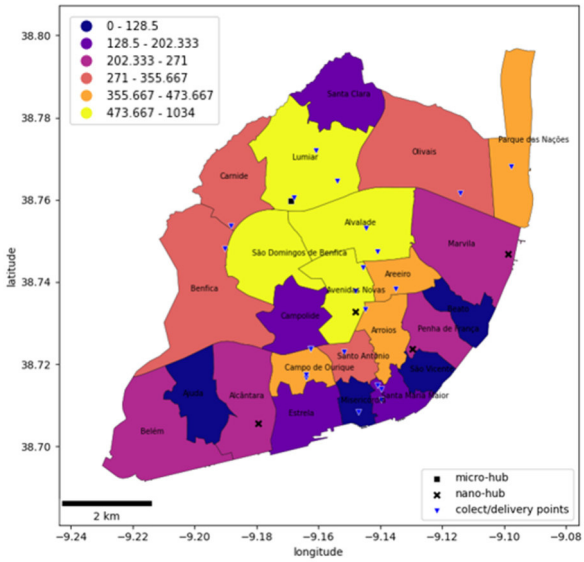

Figure 3.11 - Sub-stories deliveries per borough

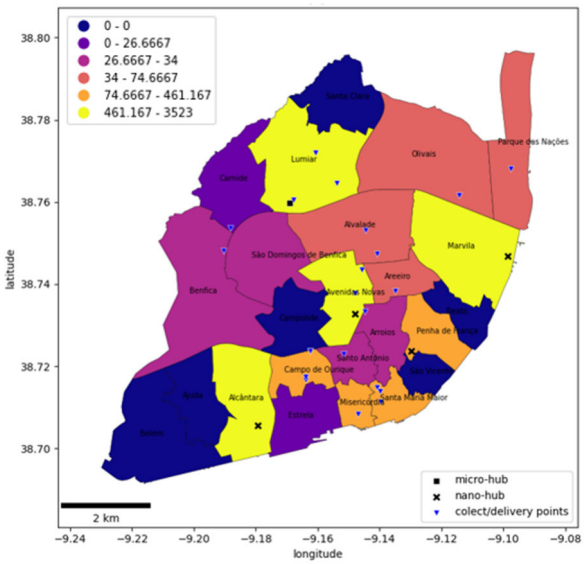

Figure 3.12 - Sub-stories pickups per borough

Analyzing the routes in a monthly, weekly and daily distribution in Figure 3.13, respectively, in Figure 3.14 and Figure 3.15, most of the routes were performed in February and March, with an average of 166 routes per month, 39 per week and 7 per day of the week, except on Fridays where the average is higher due to a specific client. On 18th of February (week 7) and on March 1st YOOB performed a stress test to its system and changed the normal routine, by having the riders performing more routes than usual.

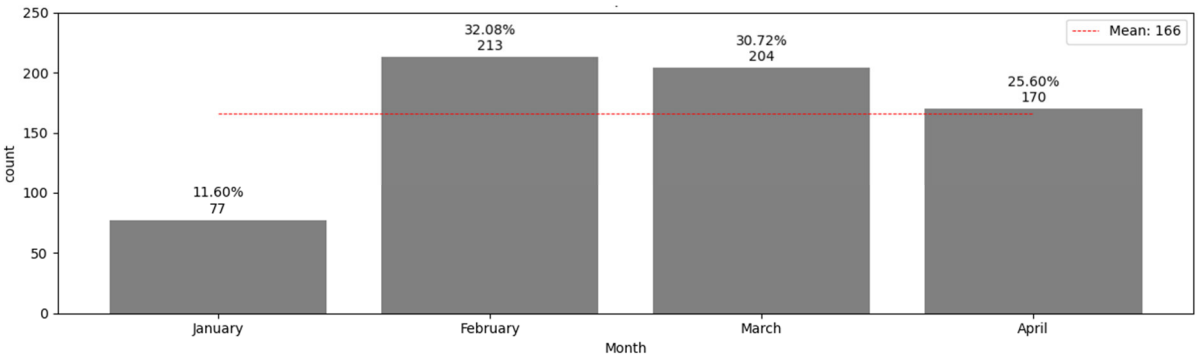

Figure 3.13 - Routes number per month

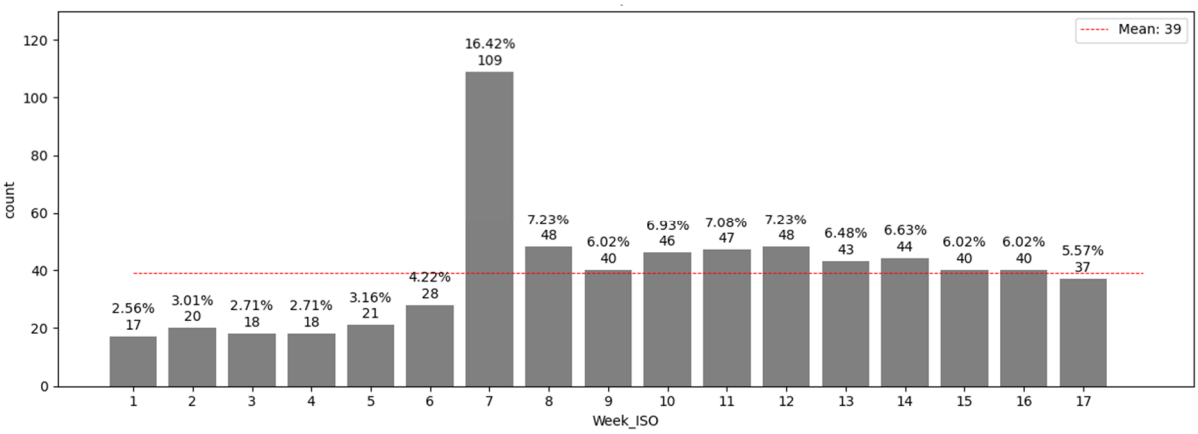

Figure 3.14 - Routes number per week

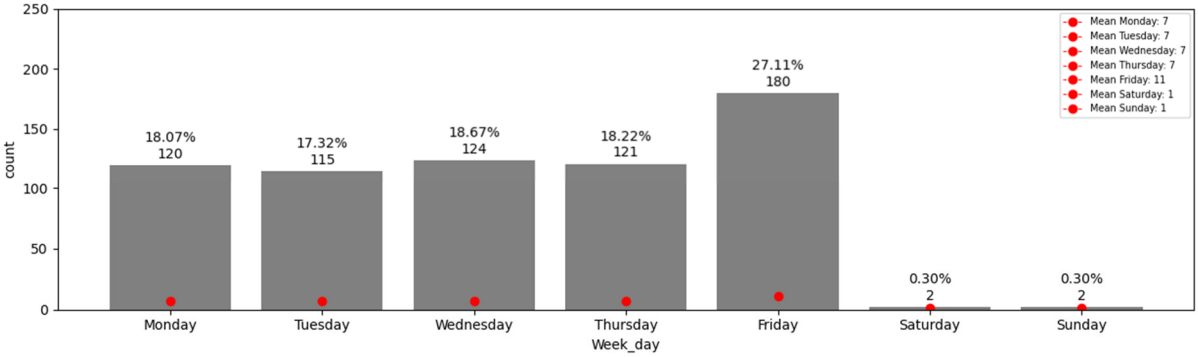

Figure 3.15 - Routes number per day of the week

The average euclidean distance per week made by the entire fleet was 541 Km (Figure 3.16), with an average of 105 Km per day (Monday to Friday) represented in Figure 3.17.

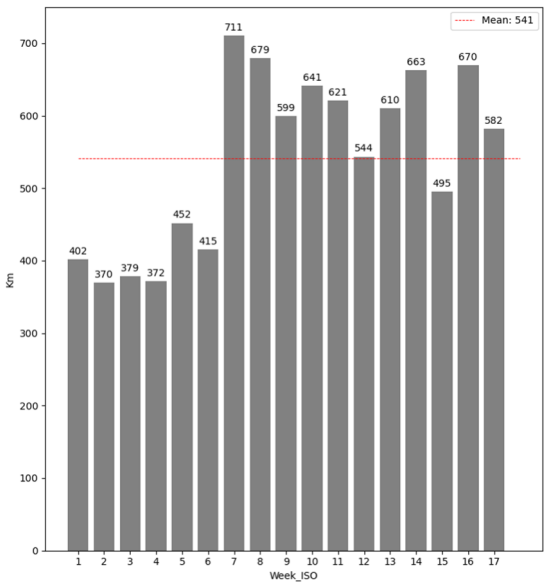

Figure 3.16 - Euclidean distance per week

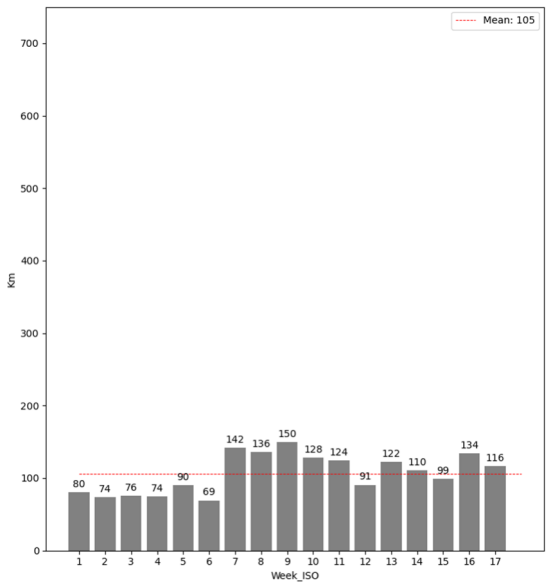

Figure 3.17 - Euclidean distance per day in the week

To begin exploring the characteristic of the routes we begin in terms of distances, three measures of distance were taken. The first was the Euclidean distance of the route, where on average a route has 13 860 meters of distance, but with 50% of the routes below 11 940 meters (see Figure 3.18). The second measure was the distance between two consecutive points inside the route, where the average is 1 705 meters. In this last measure, we found also, two main ranges were most of the values fit, the first being between 0 and 1 000 meters with 42.17% of the distances being inside of this range, and the second one >1 000 meters to 2 000 meters with 32.98% of the distances (see Figure 3.19). The third measure was the maximum distance at which the route travels from its starting location. On average the routes go as far as 4 641 meters from their initial location, but three principal ranges are mostly represented, the first one between >3 000 meters to 4 000 meters with 20.48% of all the distances, the second one between >2 000 meters and 3 000 meters, with 16.87%, and lastly between >6 000 meters and 7 000 meters, with 16.72% (see Figure 3.20).

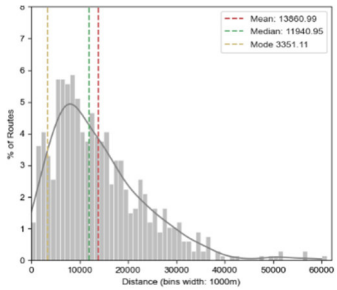

Figure 3.18 - Total distance distribution

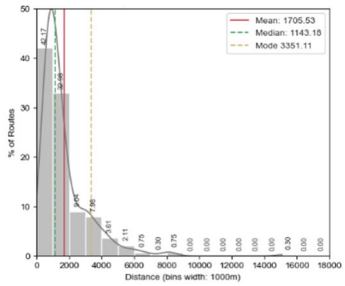

Figure 3.19 - Average distance between location distribution

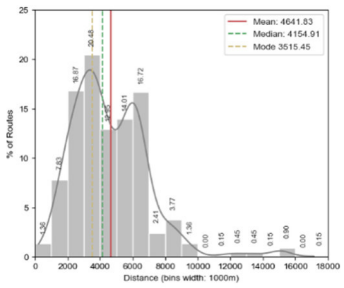

Figure 3.20 - Maximum distance from initial location distribution

On average the routes visit approximately 15 (14.74) different locations, Figure 3.21.

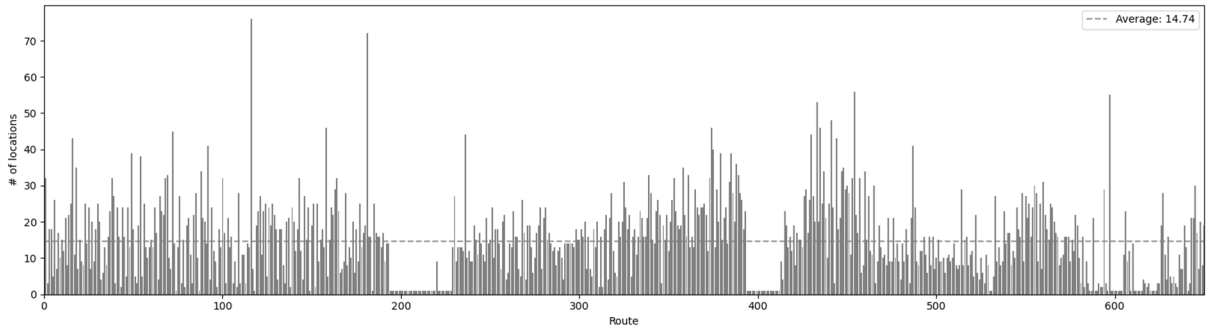

Figure 3.21 - Number of visited location on the route

When evaluating the number of parcels transported on the routes, we got an average of 15 pickups parcels and 15 delivery parcels per route (see Figure 3.22 and Figure 3.23).

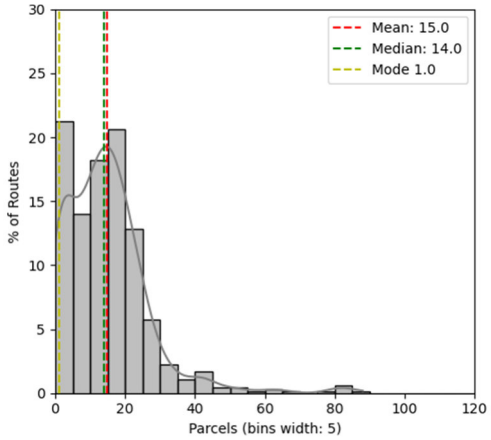

Figure 3.22 - Distribution of parcels pickups

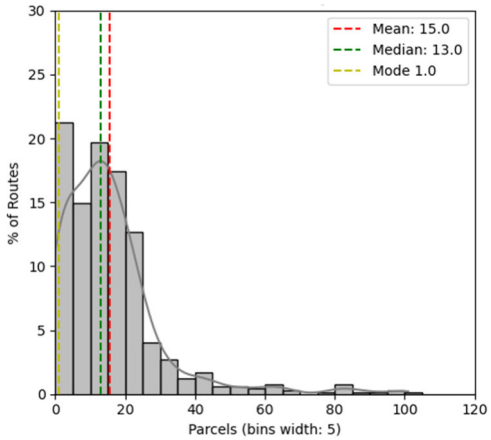

Figure 3.23 - Distribution of parcels deliveries

After a route starts, approximately 15% of the routes had one extra service assigned, while approximately 10% of the routes had two extra services assigned, shown in Figure 3.24. In this metric we counted the number of pickups made after the bike starts moving. Some sensitivity is needed in interpreting this value as YOOB redistributes goods at the end of the day, causing extra pickups. As such our interpretation was limited to a range of 0 to 2. The number of failed deliveries is near zero, as can be seen in Figure 3.25.

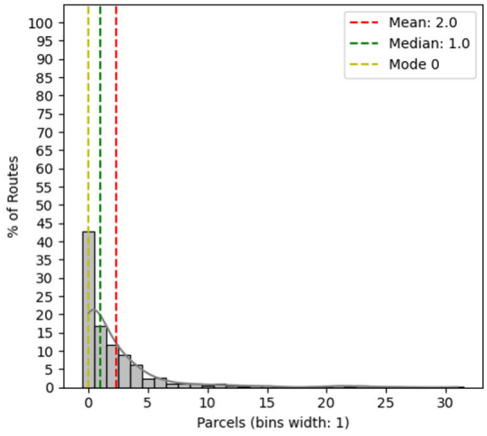

Figure 3.24 - Extra pickups

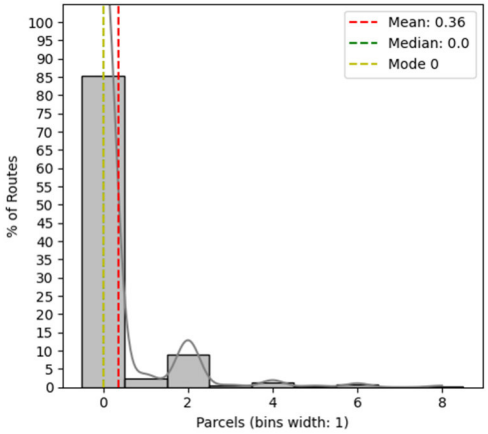

Figure 3.25 - Failed deliveries

The last characteristic of the route explored was time. We measured the time en route, and the time between two consecutive locations in the same route. These metrics gave us the register time of movement in the route, and the time spent from one location to the other.

The average time en route spent was 8.5 minutes (8 minutes and 30 seconds), with 50% of the routes being under 6.5 minutes (6 minutes and 30 seconds), Figure 3.26.

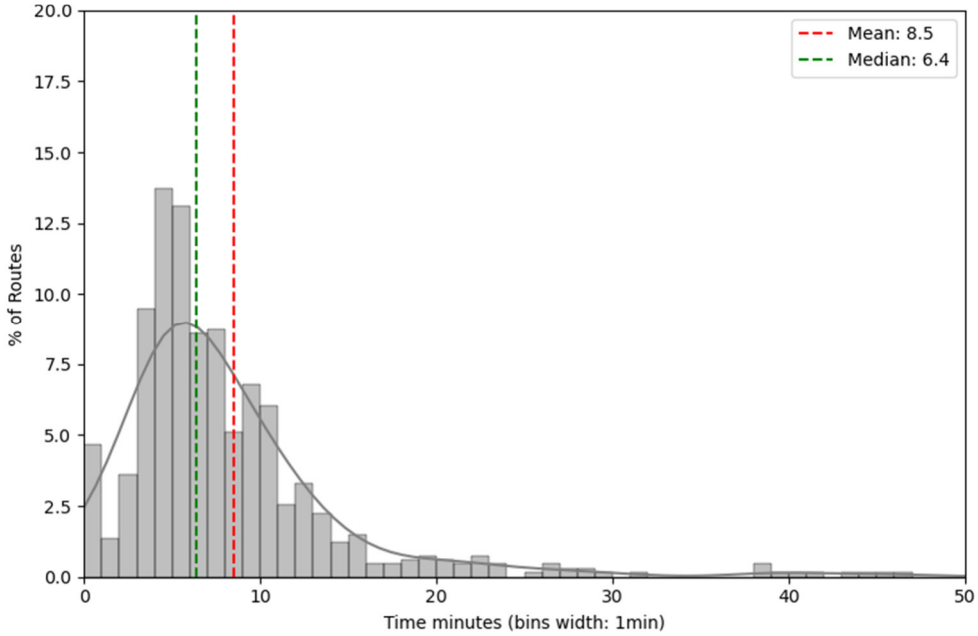

Figure 3.26 - Average time en route distribution

The average time spent between two consecutive locations was 34.6 minutes (34 minutes and 36 seconds), 50% of the routes had an average time below 16.5 minutes (16 minutes and 30 seconds), Figure 3.27. The median value was the most plausible one to characterize the routes as the average was impacted by the largest values. In the briefing with YOOB it was highlighted that there in some clients, the waiting time for the parcels was extended to one hour or more.

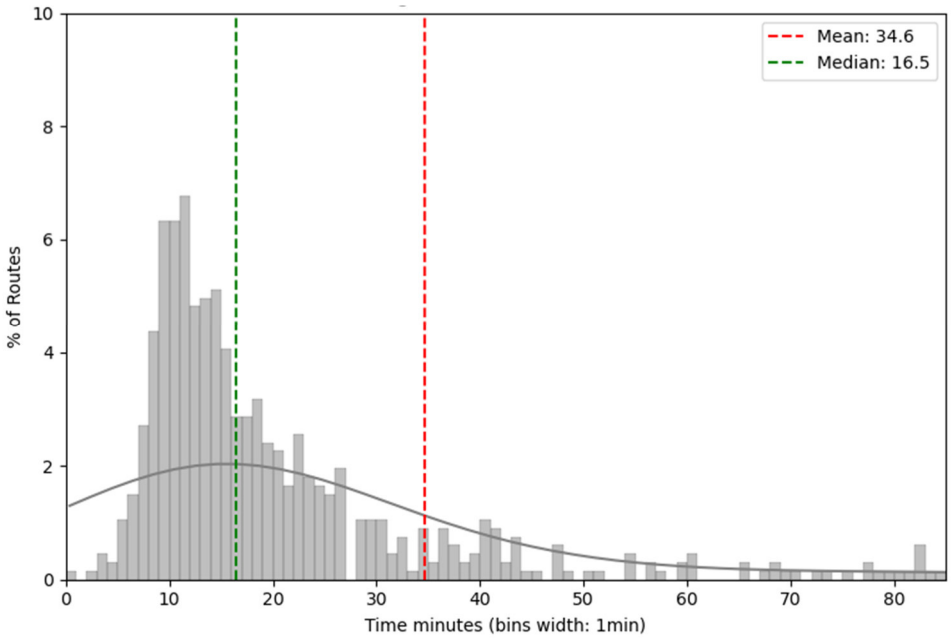

Figure 3.27 - Average time spent between locations distribution

All the initial locations of the routes were identified, and 42 unique locations were detected. To differentiate known locations from unknown locations we have based on the variable ['pickupUID'] from the initial data frame, which is only filled if the location is identified in the YOOB system. Most of the routes start from the micro-hub of Telheiras (49.77%), followed by the four other established nano-hubs, as depicted in Figure 3.28 and Figure 3.29.

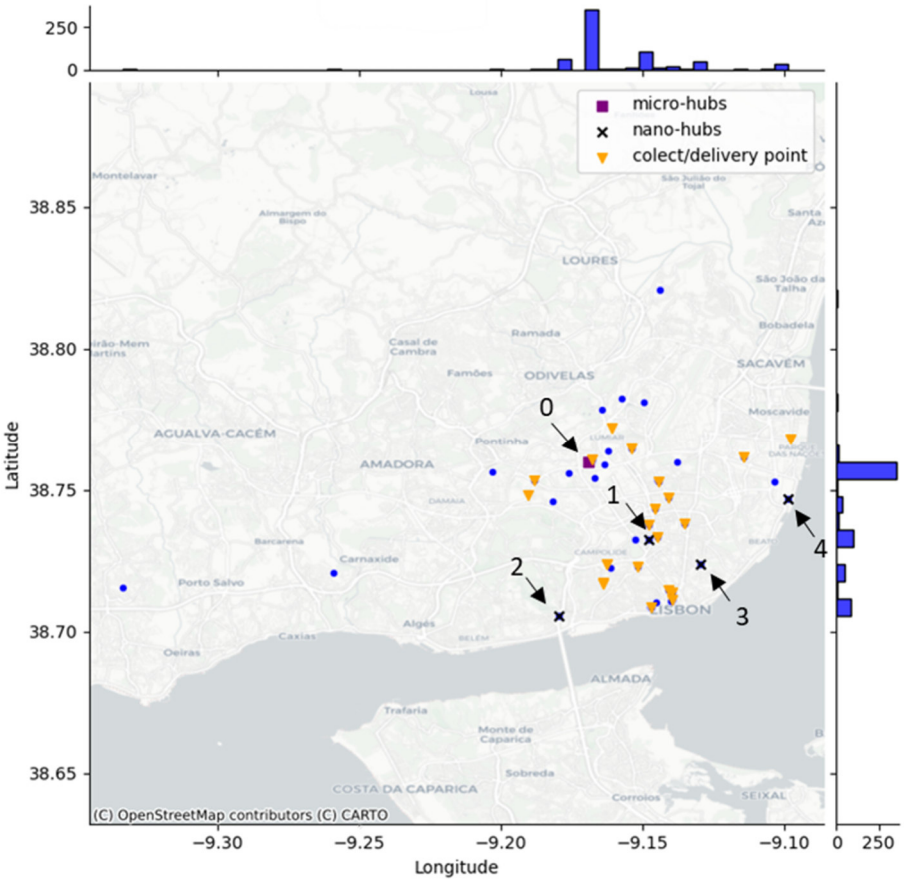

Figure 3.28 - Start locations distribution

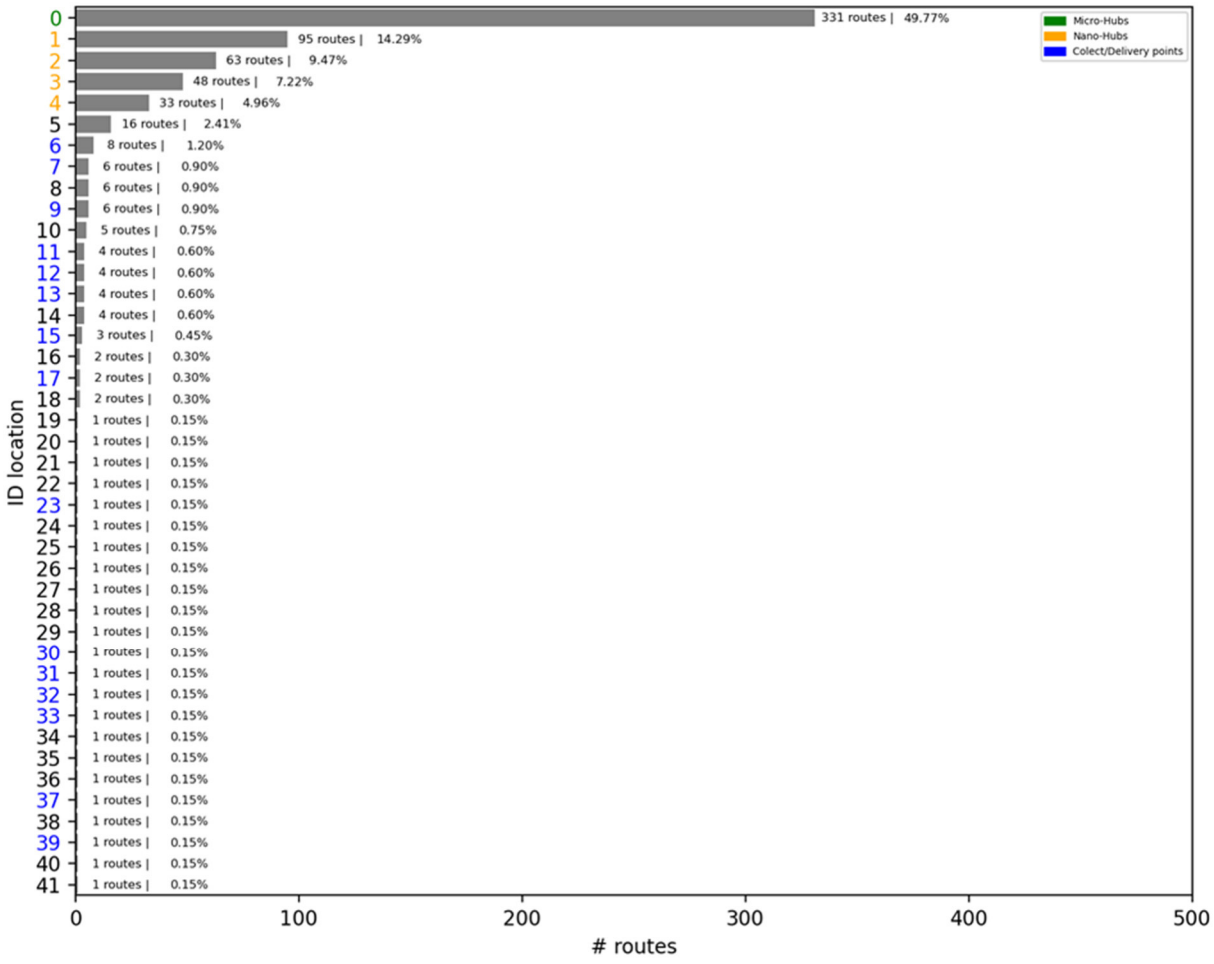

Figure 3.29 - Number of routes per start locations

In our first routes visualization, we grouped them by total distance, creating three groups: in the first, the small-size routes considered were under quartile 25 which gave us a range of distances from 0 to 6 785 meters shown in Figure 3.30. The average-size routes considered ranged from 6 785 meters to 16 965 meters, in Figure 3.31 and the long-size routes considered a distance above 16 965 meters, as depicted in Figure 3.32.

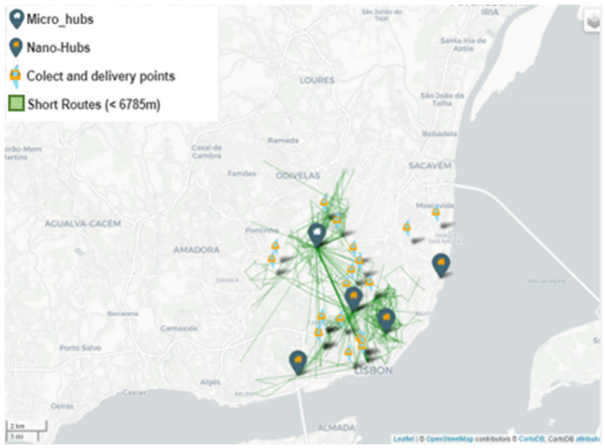

Figure 3.30 – Small-size routes

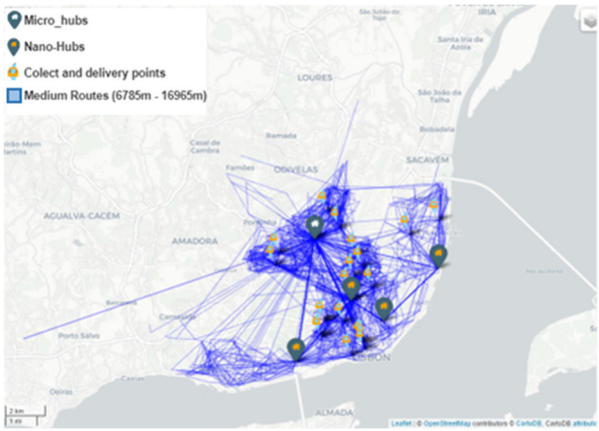

Figure 3.31 - Average-size routes

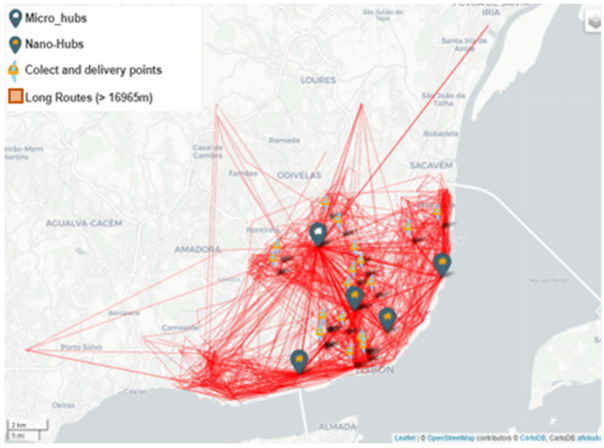

Figure 3.32 - Long-size routes

To analyze which origin-destination pairs were most performed, a data frame was created with the existing pair combinations and the respective count of how many times the pair was repeated, into an origin-destiny matrix. To identify the known from the unknown locations, we have differentiated the locations with a color code: green refers to micro-hub, orange refers to the nano-hubs, blue refers to the collect/delivery points the black are the unknown locations. The minimum frequency analyzed was 5 and the result is shown in Figure 3.33.

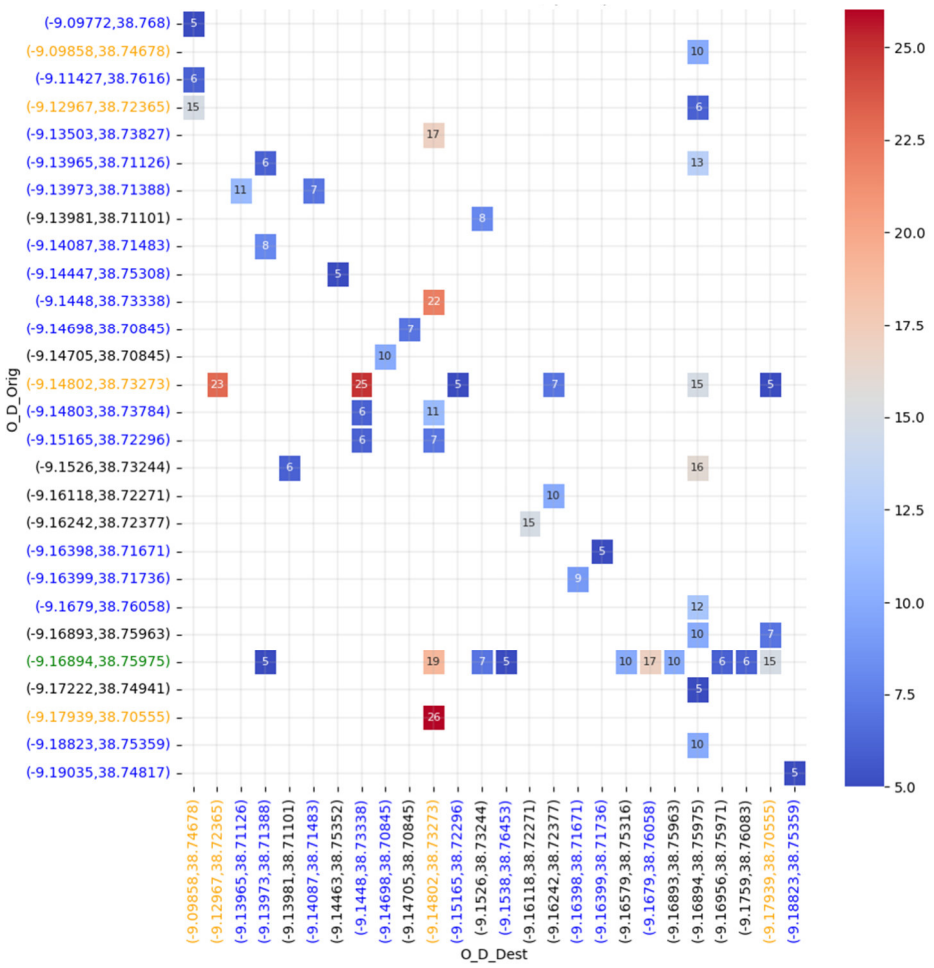

Figure 3.33 - Origin-destiny matrix (frequency >=5)

The most travelled pairs were between known locations. In Figure 3.34 the pairs are linked with a color coding according to their frequency. If the frequency is less than 10 the pair has a blue connection. If the frequency is equal or more than 10 but less than 18, the connection is yellow and if the frequency is equal to 18 or more the color of the connection is red. The top four travelled pairs were, by order, between nano-hub 2 to 1, the second from nano-hub 1 to 3, the third from micro-hub to nano-hub 1 (in the visualization it is hide behind the yellow connection that links nano-hub 1 to micro-hub) and the fourth, between the collect/delivered point located in Saldanha, shown in the zoomed part of the Figure 3.34.

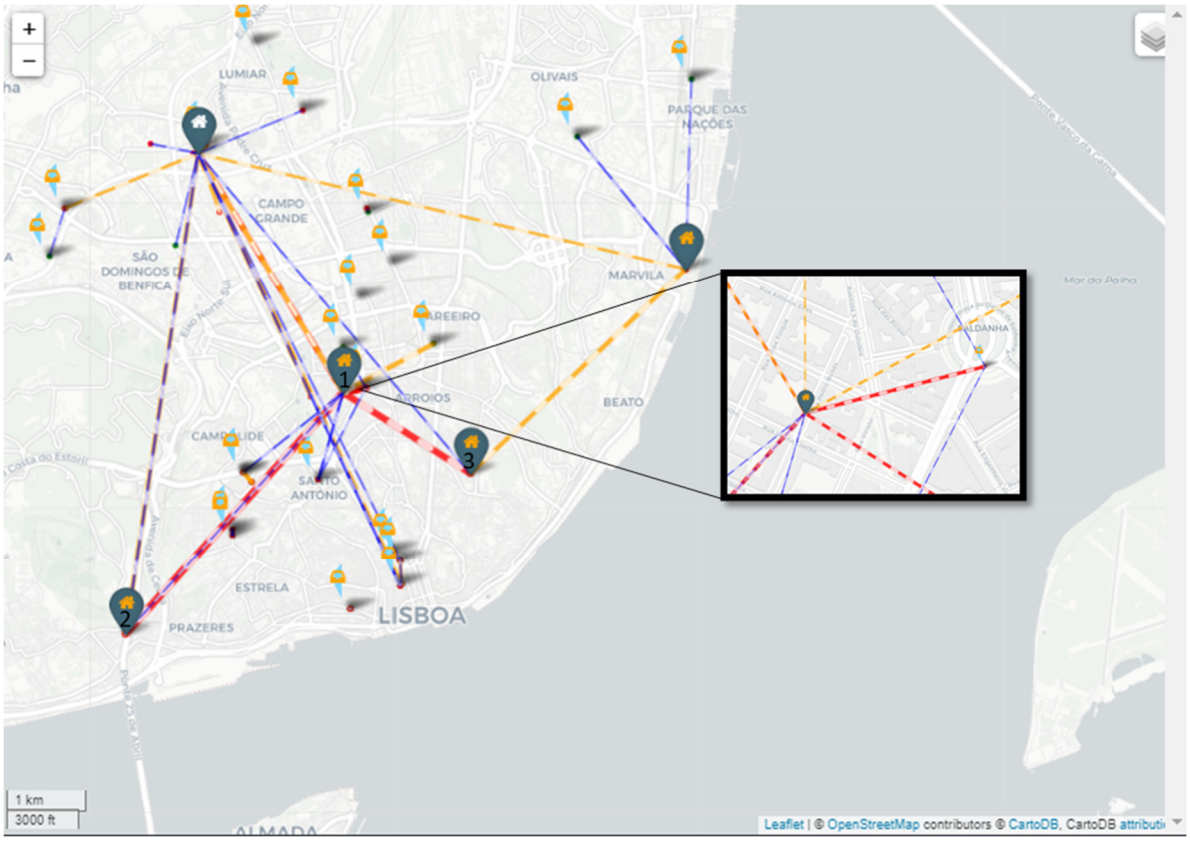

Figure 3.34 - Most traveled pairs

Three unknown origin locations, seen in Figure 3.35, and two destiny unknown locations, depicted in Figure 3.36, were detected. Due to privacy concerns we do not identify those in detail in this text.

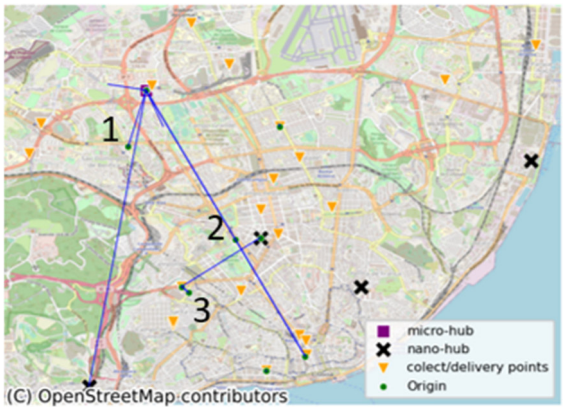

Figure 3.35 - Origin unknown locations

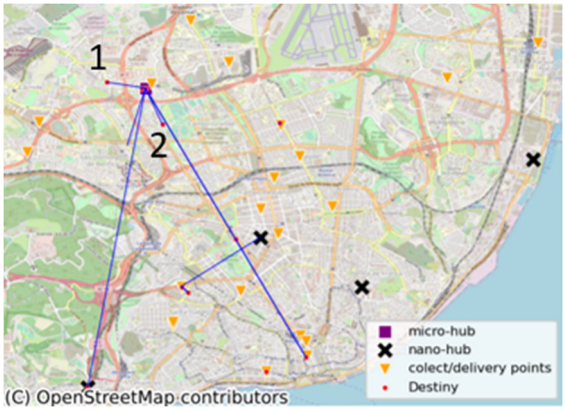

Figure 3.36 - Destiny unknown locations

To analyze the average operation radius of the micro-hub and nano-hubs, we performed a loop in our Python code for each micro-hub and nano-hub, filtering the routes with start location in that place and then calculated the average of the distances from the starting point (micro-hub or nano-hub). Based in the computed average distance (corresponding to a circle radius) we calculated the respective covered area. Consulting Figure 3.37, Figure 3.38 and Figure 3.39, we see that the average radius of the micro-hub is 2.00 Km, covering an area of 15.80 Km^2. Regarding the nano-hubs we have a minimum range (nano-hub 3), with a radius of 0.72 Km and a covered area of 2.07 Km^2, and with maximum range (nano-hub 2) with a radius of 2.05 Km and a covered area of 12.39 Km^2. The covered area of all five hubs is 41.13 Km^2, representing 38% of the Lisbon city area.

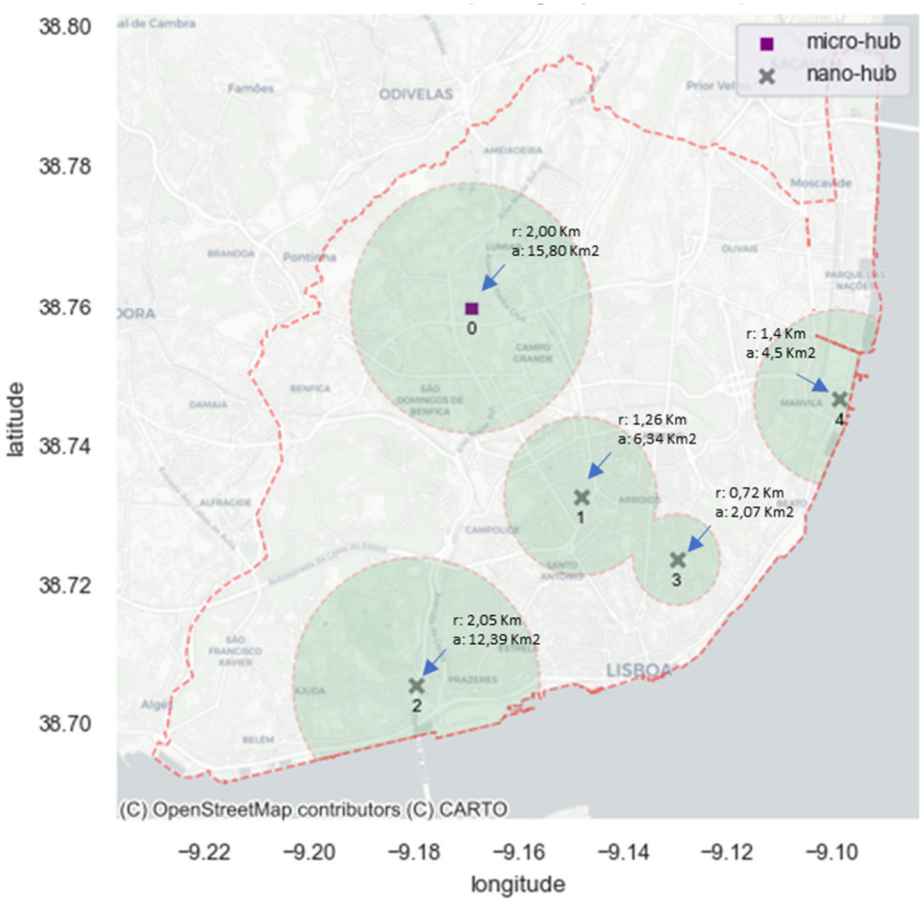

Figure 3.37 – YOOB hub locations average radius and covered area

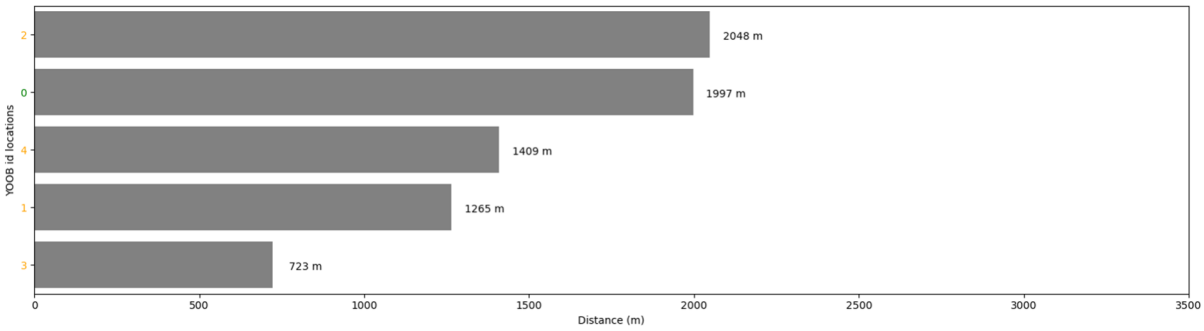

Figure 3.38 – YOOB hub locations average radius

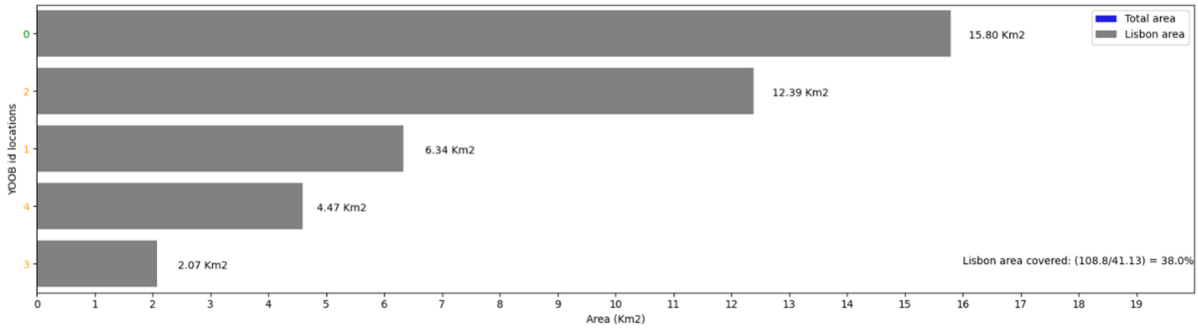

Figure 3.39 – YOOB hub locations average covered area

The collect/delivery locations extend the operation area of YOOB. Representing in the Figure 3.40 these locations, we see that the covered area extends to 73% of the Lisbon city area. We found that two of these locations have a wider range. The collect/delivery location nº 33 has a radius of 2.61 Km covering 21.70 Km^2 of the area of Lisbon and extends 5.30 Km^2 outside Lisbon area covering a total of 27.00 Km^2. This location was used as a temporary nano-hub to perform distribution for a specific client. The location nº 17, located in downtown Lisbon, is the one with the largest radius, with 3.22 Km and a covered area of 24.97 Km^2 and is regularly used as a temporary nano-hub for nearby distributions.

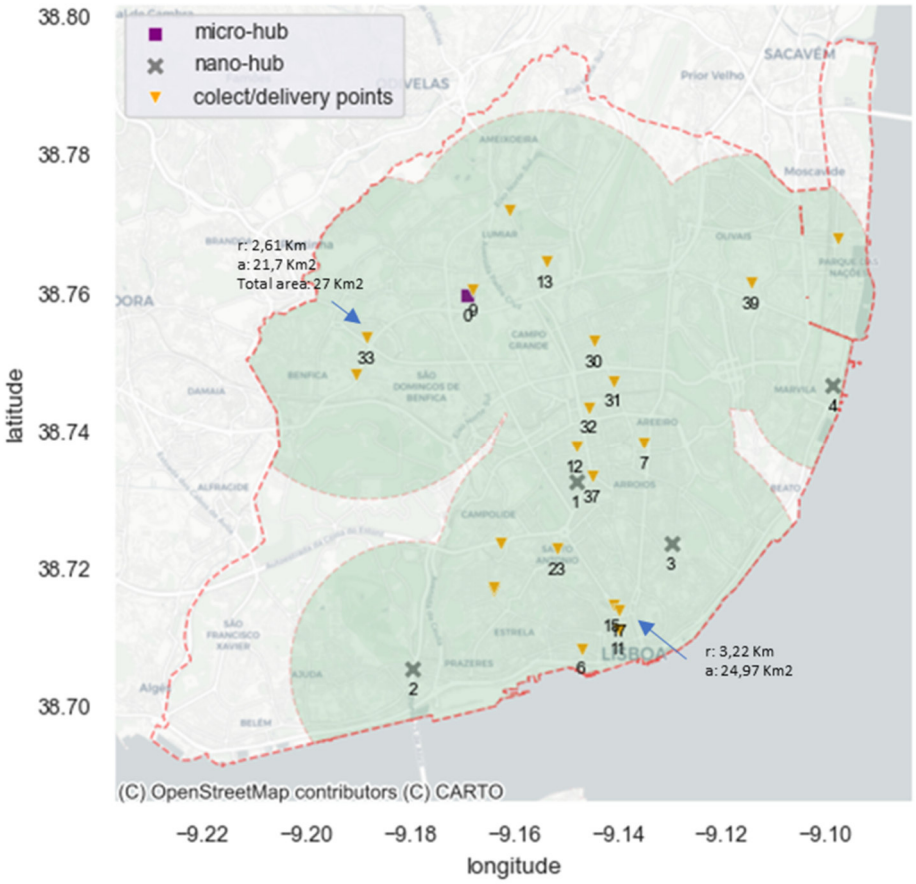

Figure 3.40 - YOOB hub locations plus collect/delivery locations average radius and covered area

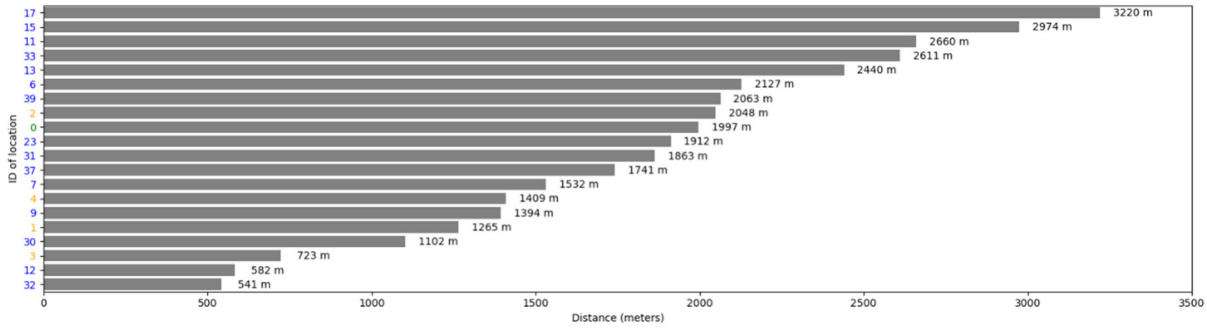

Figure 3.41 – YOOB hub locations plus collect/delivery locations average radius

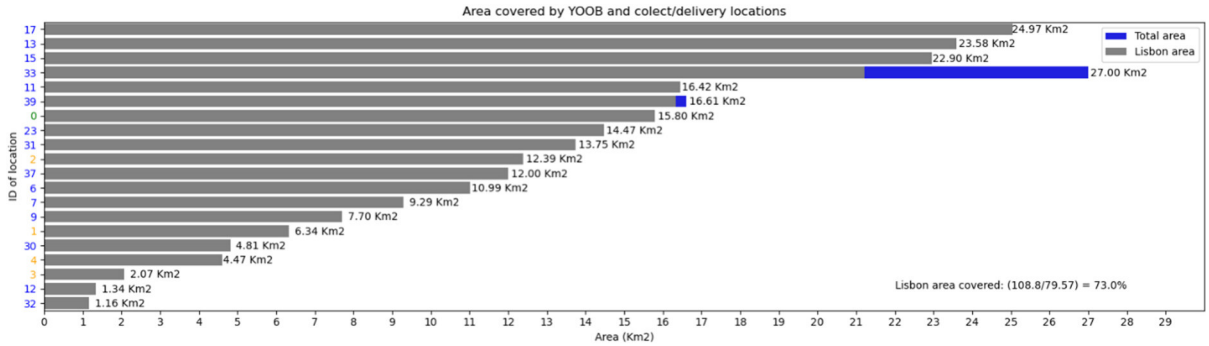

Figure 3.42 - YOOB hub locations plus collect/delivery locations average covered area

We analyzed the impact of opening new locations (nano-hub and collect/delivery) on the distances traveled. With the opening of new locations, the total distances were reduced, see Figure 3.43 and Figure 3.44. The opening of the first three nano-hubs had a greater contribution to the current average distance's values. Traveled distances were not penalized by the addition of more collection/delivery locations. And the number of locations that the routes visits is on an increasing trend, while the distances between those same locations have an inversed trend, pointing to a greater proximity between the deliveries, Figure 3.45.

Figure 3.43 - Average route distance vs expansion

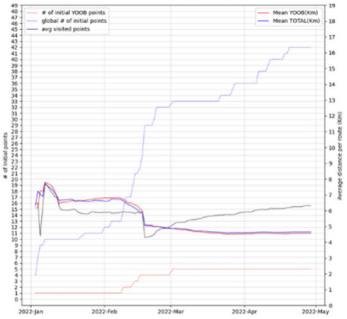

Figure 3.44 - Average maximum distance from initial location vs expansion

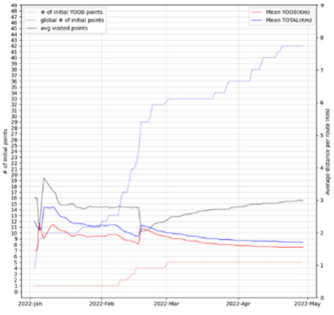

Figure 3.45 - Average distance between locations vs expansion

3.5. Data modelling

In the modeling phase, we developed three models, allowing us to further answer our research questions. We applied machine learning techniques, namely using K-means, for clustering the data and to perform center of gravity analysis. K-means was selected given that some attempts were made to implement the DBSCAN algorithm in model one and two as shown in annex A and B. When adjusting the eps value of the algorithm, we observed that for a low number of clusters, there is an excessive concentration of data in a single cluster and as more clusters are created, the percentage of data classified as noise increases. This effect does not allow us to make an objective and focused analysis. The parameterization of the density function of the algorithm as a function of the structure of our data was the biggest challenge regarding our DBSCAN implementation. For the reasons mentioned above we chose not to further pursue with the DBSCAN technique in our study.

We developed and evaluated three data models. In the first model we created clusters that allowed us to identify the behavior of routes in certain geographical areas, in the second model we clustered the routes and evaluated their characteristics, providing answers to our first research question. In the last model we performed a gravity center analysis, with the goal to explore new locations for the implementation of new hubs, answering our second research question.

To build the models we used the sklearn [66] library, for pre-processing the data we use MinMaxScaler [70] and LabelEnconder [71] and to perform cluster and the center of gravity analysis we use the K-Means algorithm [72]. To evaluate the optimal K value in the two first models we used the Knee Elbow method with knee library [67] and Davies-Bouldin index [73].

3.5.1. Model one – Clustering the sub-stories with K-Means

As mentioned, in the first model we created clusters that allowed us to identify the behavior of routes in certain geographical areas. The feature selection was made from the geodataframe with sub-story granularity. To the variables [‘assignedRouteUID’] and [‘ct_type’] the label encoder method was applied to allow the visualization of results in the correlation matrix. The result can be seen in Figure 3.46.

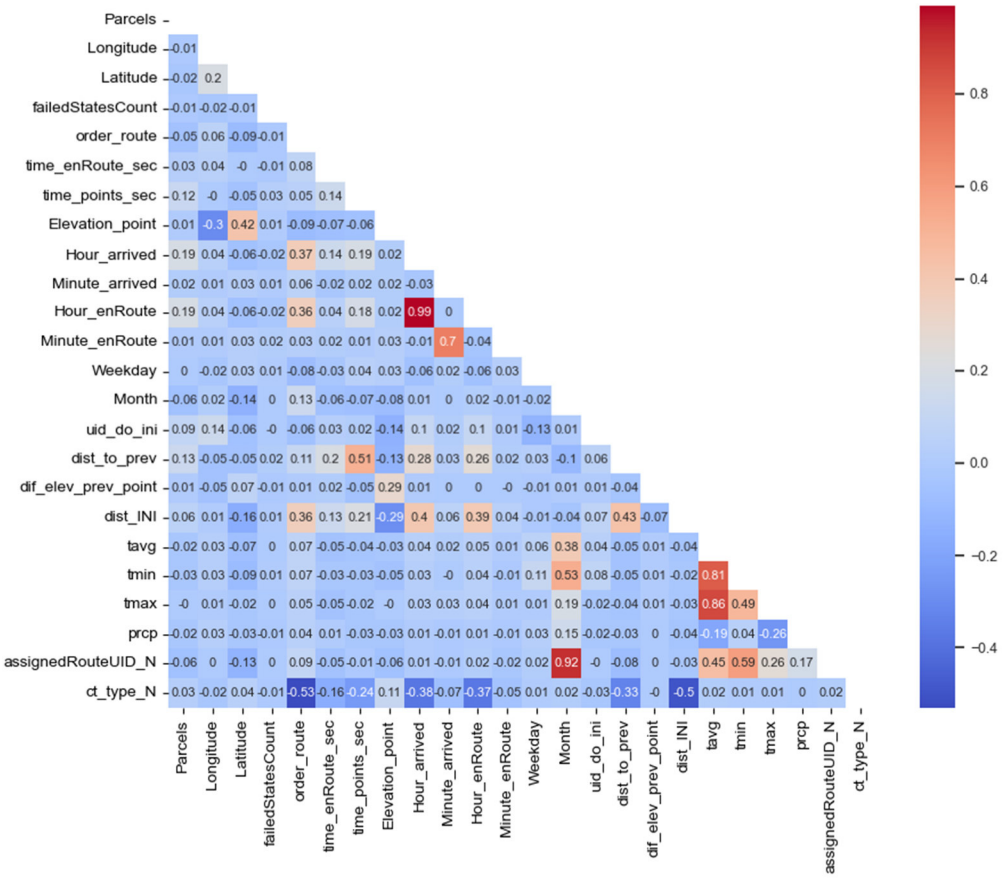

Figure 3.46 - Correlation matrix for model one

The features selected to our model were [‘Latitude’], [‘Longitude’], [‘Elevation_point’], [‘time_points_sec’] and [‘distance_to_prev’]. Before running the clustering model in our data, we had to scale the data as it had different measurement units. For that process we applied MinMaxScaler. Evaluating the Knee Elbow method and the Davies-Bouldin index through a range from 1 to 30 clusters,

the optimal value for K was 5 in knee elbow method (Figure 3.47) and in Davies-Bouldin (Figure 3.48) the optimal value was 4. After testing the model with both values, the knee elbow value gave us more information (confirmed with YOOB briefings) that would be hidden if we chose the value of the Davies-Bouldin index. For our final implementation we adopted the K value indicated by the knee elbow method.

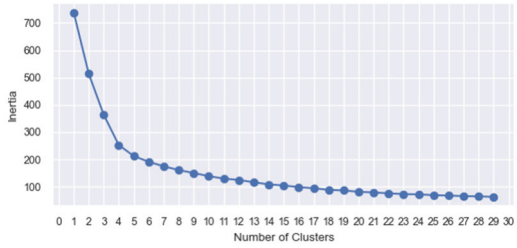

Figure 3.47 - Knee elbow method

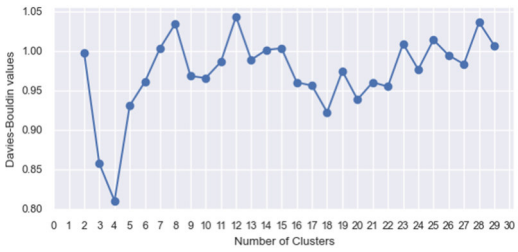

Figure 3.48 – Davies-Bouldin index

We applied the K-Means algorithm with a K value of 5 to our data, and the output is shown in Figure 3.49. Four main clusters (C0 to C3) outstand in the visualization, and a fifth cluster (C4) with dissipated grey dots among the four other main clusters.

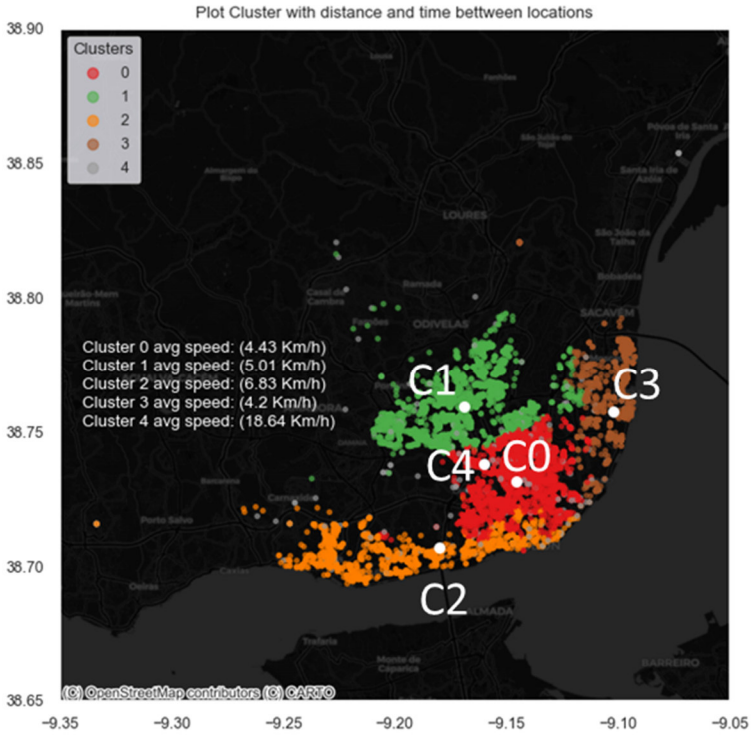

Figure 3.49 – K-Means clustering results

3.5.2. Model two – Clustering the routes with K-Means

In the second model the selected features were based on the geodataframe with granularity of the route. The correlation matrix is shown in Figure 3.50.

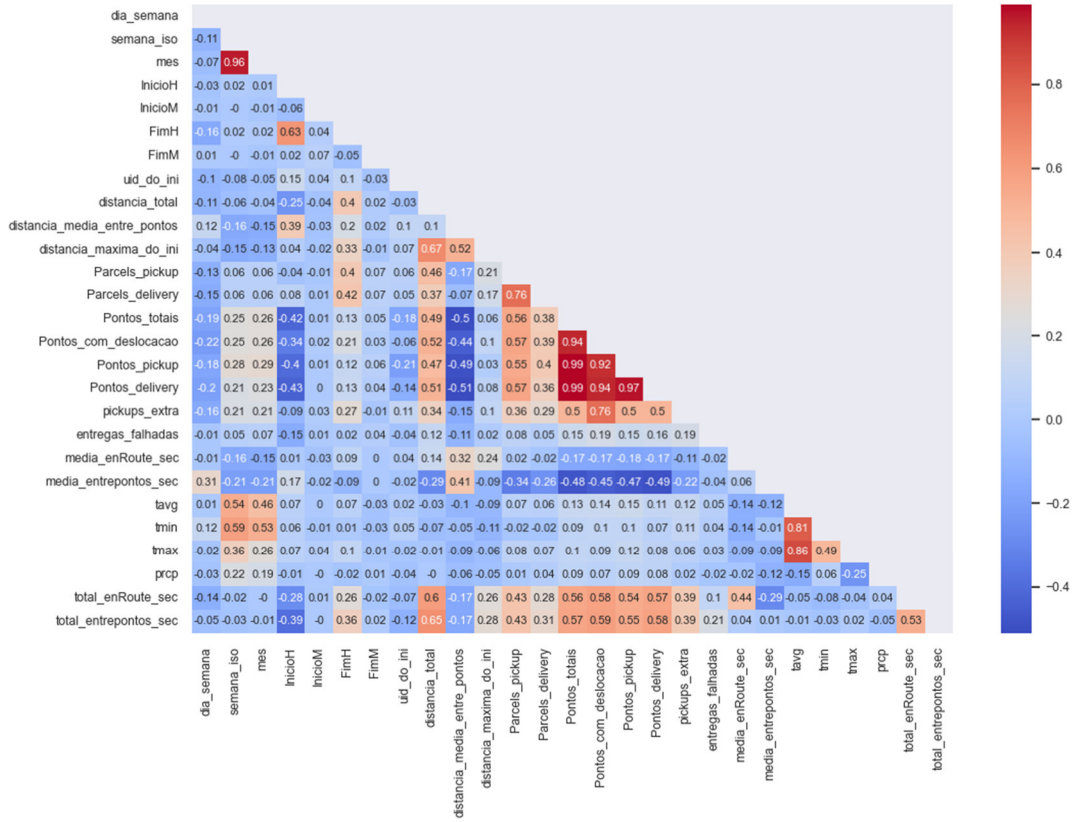

Figure 3.50 - Correlation matrix for model two

The selected features were [‘distancia_total’] and [‘distancia_maxima_do_ini’], and the features were scaled with MinMaxScaler. Evaluating the Knee Elbow value and the Davies-Bouldin value in a range from 1 to 30 clusters, the optimal value for K in Knee Elbow method was 5 (Figure 3.51). The optimal value in the Davies-Bouldin index is far bigger (Figure 3.52). For our K value we chose the value of the Knee Elbow for better and cleaner interpretation.

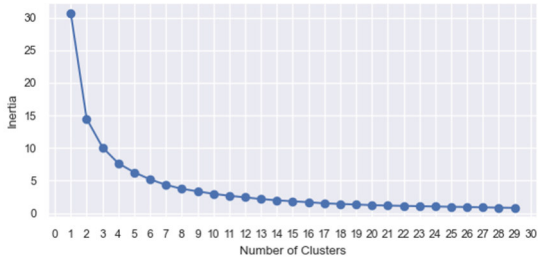

Figure 3.51 - Knee Elbow Method

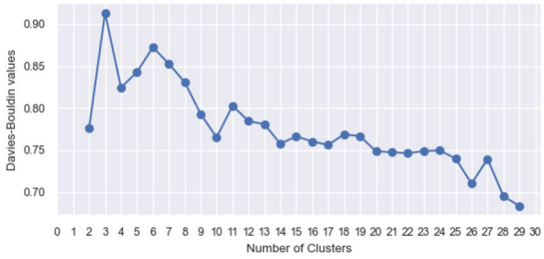

Figure 3.52 – Davies-Bouldin index

Applying the data to our model with the K-Means algorithm with a K value of 5, the output results in five clusters, shown in Figure 3.53, Figure 3.54, Figure 3.55, Figure 3.56 and Figure 3.57.

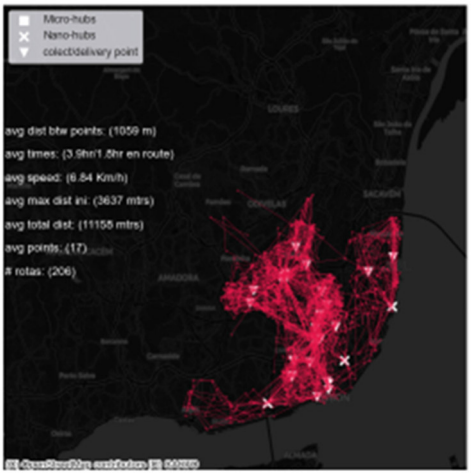

Figure 3.53 - Routes cluster 0

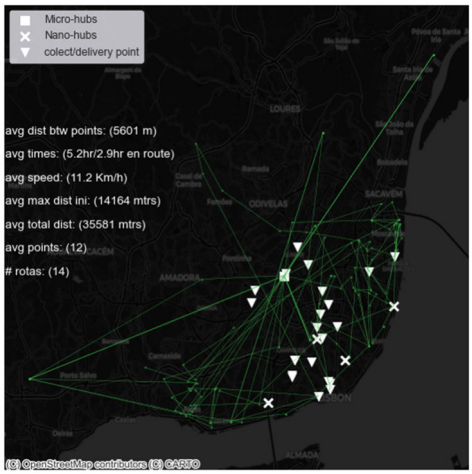

Figure 3.54 - Routes cluster 1

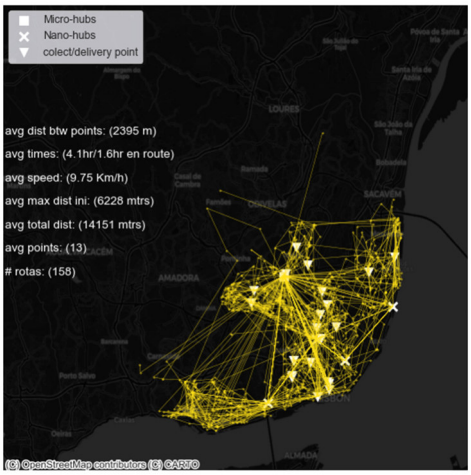

Figure 3.55 - Routes cluster 2

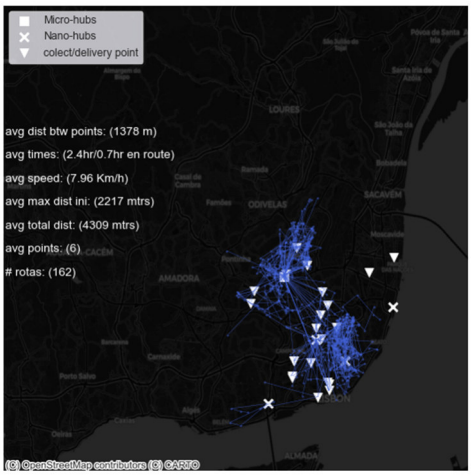

Figure 3.56 - Routes cluster 3

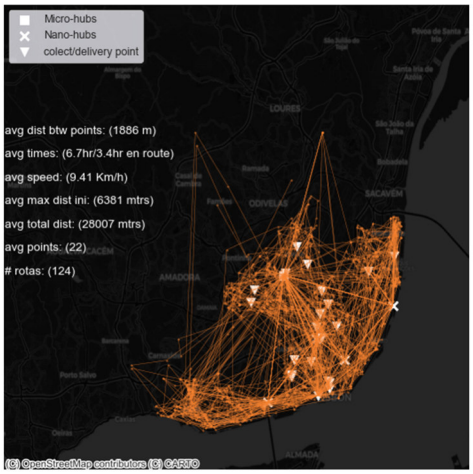

Figure 3.57 - Routes cluster 4

In Figure 3.58, the number of routes per cluster is shown. Each cluster was analyzed by total distance in Figure 3.59. Figure 3.60, depicts the average maximum distance from the initial location.

Figure 3.61 shows the average number of stop locations and Figure 3.62, the average speed in each cluster, this value is the result of calculating the mean of the total distance and divide it by the mean of the total time en route. Figure 3.63 visualizes the total time en route and Figure 3.64, the total time between locations. The last metric, depicted in Figure 3.65, was the operation time and this value was calculated by subtracting the average total en route time from the total time spent between two locations and divided the result by twice the number of locations visited, representing the operation time spent at each location.

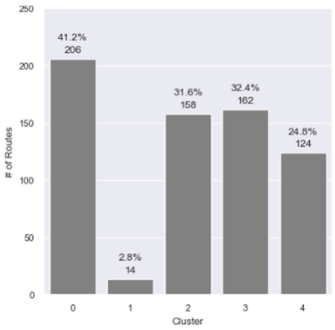

Figure 3.58 - Routes per cluster

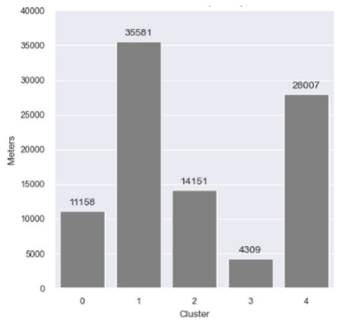

Figure 3.59 - Average total distance per cluster

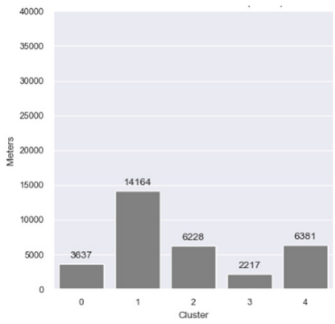

Figure 3.60 – Average maximum distance from initial location per cluster

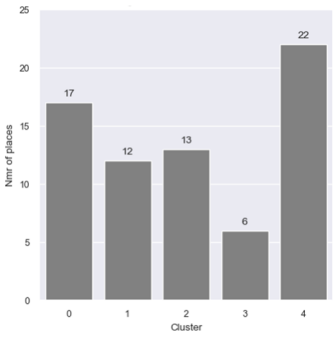

Figure 3.61 – Average visited locations per cluster

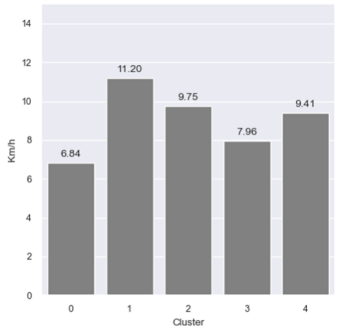

Figure 3.62 - Average speed per cluster

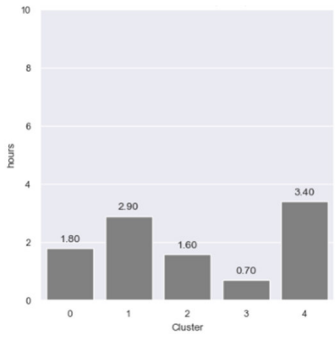

Figure 3.63 - Average total time en route per cluster

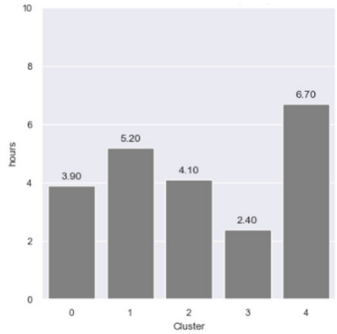

Figure 3.64- Average total time between locations per cluster

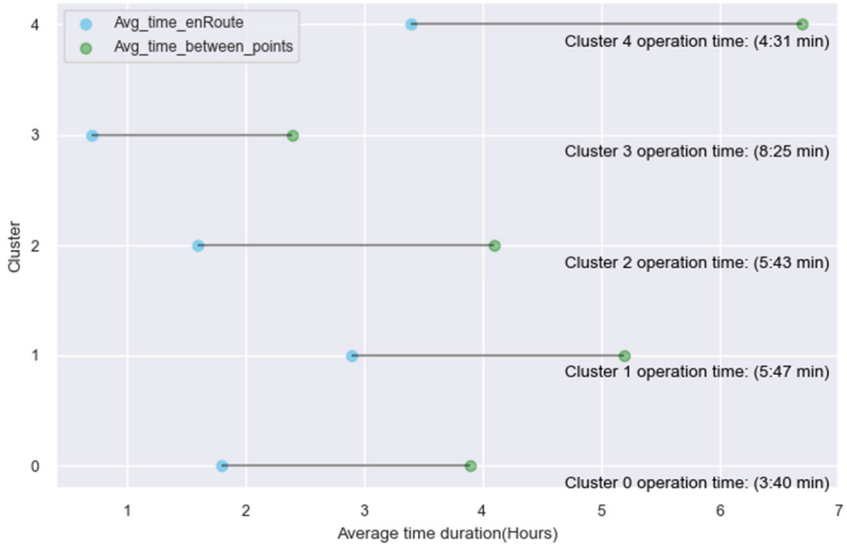

Figure 3.65 - Average operation time per cluster

3.5.3. Model three – K-Means center gravity analysis

In the third model, we analyzed the centers of gravity of the sub-stories of our data. This model analysis was requested by YOOB in one of the meetings held. Although in our initial SLR there was no direct references to this specific topic, by doing some additional research in the literature, we found that Wen et al. [74] and Cai et al. [75], both approached this problem by applying K-means techniques with a weighted featured to find the best hub locations. In our approach we adopted a similar approach with a weighted K-Means algorithm.

The parameters needed for this algorithm were the K value. Our model applied the number of locations intended to simulate, and a new variable was considered in the weighting of the cluster. In our model the number of parcels was considered, as the effort needed to carry out the delivery. As most of the time the pickup parcels were in the hubs or at the collect/delivery locations, we added a penalty value in the delivery parcels, considering these last ones three times bigger in effort than the pickup ones. This would force the algorithm to locate the centroids of the cluster in places where distance and effort would be reduced. The data applied in this model was based on the variables [‘latitude’], [‘longitude’] and [‘Parcels’] from the geodataframe with sub-story granularity. A new variable was created designated [‘Calc_ajusto_de_custo_se_houver’], to include the penalty value. By looping through all the sub-stories the new variable was created based on the condition: if the sub-story was a pickup, the value of the variable maintains equal to the [‘Parcels’] variable; if the sub-story was a delivery, the value of the variable is equal to the value of the [‘Parcels’] multiplied by three. We simulated the center of gravity for 8 hubs (a business figure that was transmitted to the author by YOOB, in the framework of the company hub expansion policy) and the result is shown in Figure 3.66.

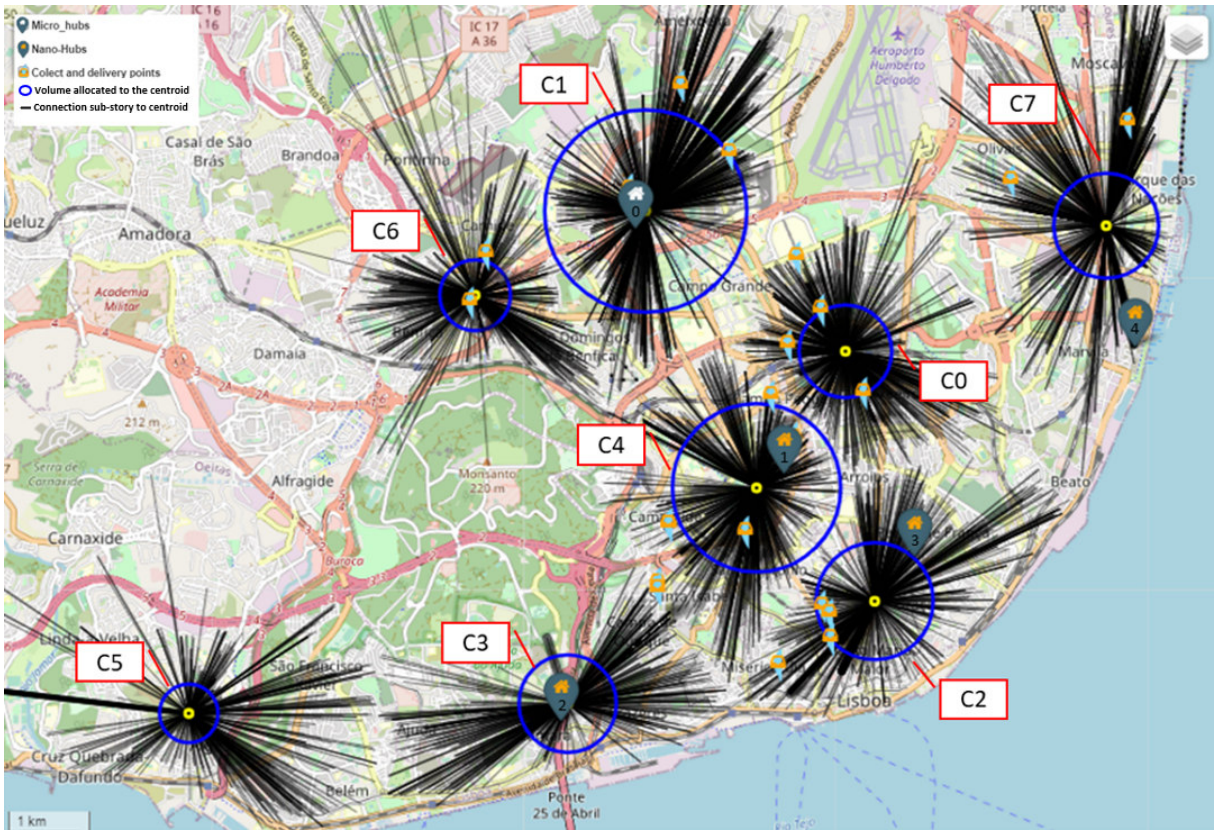

Figure 3.66 - Center of gravity for eight hubs, using K-Means

The found hubs coordinates are presented in Table 3.5, and the volume associated for each location is depicted in Figure 3.67.

Table 3.5 - Centroids coordinates of the found 8 hubs

Location	Latitude	Longitude	Address
C0	38.746051	-9.139322	Av. Frei Miguel Contreiras
C1	38.761480	-9.167363	Praceta Prof. Gonçalves Ferreira
C2	38.718589	-9.135055	Rua do Benformoso
C3	38.707362	-9.178456	Rua Prof. Vieira Natividade
C4	38.731047	-9.151716	Av. António Augusto de Aguiar
C5	38.706249	-9.231764	Rua Sofia Carvalho
C6	38.752185	-9.191358	Largo Revista Militar
C7	38.759893	-9.102499	Rua Centieira

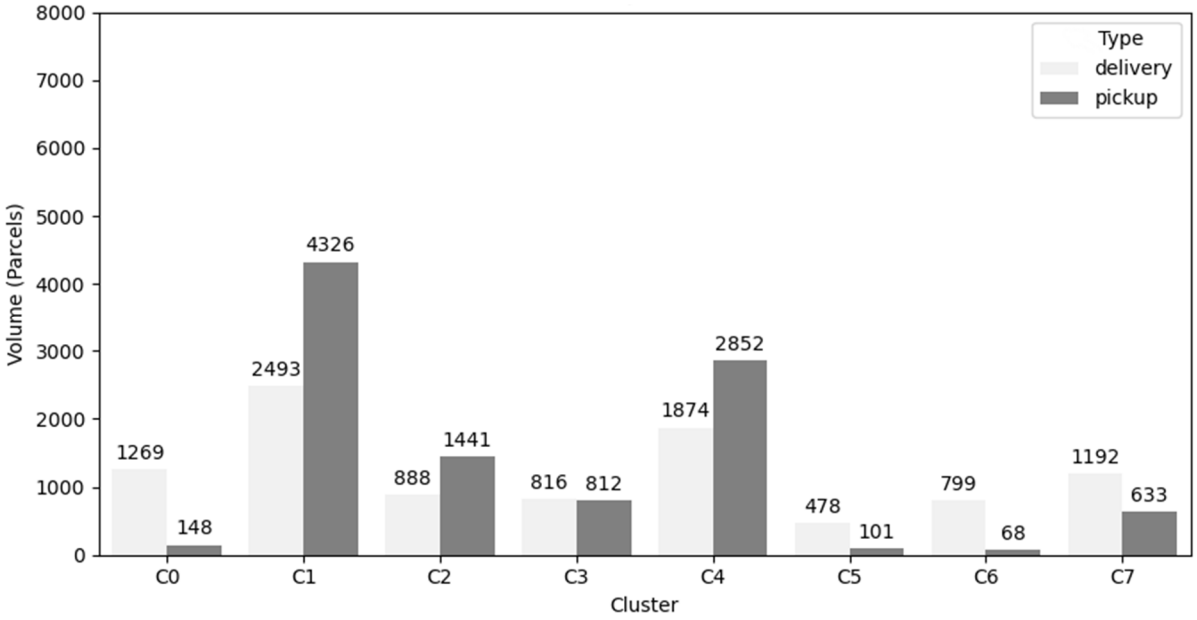

Figure 3.67 - Volume parcels per proposed new cluster centroid

3.6. Evaluation

3.6.1. Model evaluation

The results obtained with the implementation of our models allowed us to gather valid information on the various topics covered in the scope of the respective thesis.

In the first model we were able to observe the e-cargo bikes' performance according to the geographical area. In Figure 3.49, we can see the four well defined clusters and a more disperse cluster (C4) where the e-cargo bikes have a higher average speed of 18.64 Km/h, indicating that these are acceleration areas. In the other clusters the average speed is significantly lower.

The zones with the second highest average speed were the ones in cluster C2 where e-cargo bikes achieved average speeds of 6.83 Km/h, followed by the zones covered by cluster C1 with average speeds of 5.01 Km/h. The areas covered by clusters C0 and C3 have a more homogeneous performance, however in the areas covered by cluster C3 the e-cargo bikes tend to be slower, with average speeds of 4.20 Km/h vs 4.43 Km/h of the speeds practiced in the C0 areas.

In our second model we obtained 5 types of routes. The most common routes were the ones grouped in cluster 0, representing 41.2% of all routes. Characteristics associated to this cluster were routes of a total average distance of 11 158 meters, to a maximum of 3 637 meters in average from their starting location, visiting an average of 17 different locations, at an average speed of 6.84 Km/h. The total observed time en route was 1 hour and 48 minutes with a total route duration of 3 hours and 54 minutes. The average operational time associated with each location visited was 3 minutes and 40 seconds, the shortest time observed of all the clusters.

The routes in cluster 1 had the smallest representation with only 2.8 % of all routes. They are characterized by having the highest average total distance of all with 35 581 meters, with a maximum of 14 164 meters in average from its initial place of departure, visiting an average of 12 different locations, at an average speed of 11.20 Km/h, making them the fastest routes of all clusters. The total time en route was 2 hours and 54 minutes with a total route duration of 5 hours and 12 minutes. The operational time associated with each location visited is 5 minutes and 47 seconds.

In cluster 2 we have 31.6% of the routes, with an average distance of 14 151 meters, moving a maximum of 6 228 meters in average from their initial starting location, visiting an average of 13 different locations, at an average speed of 9.75 Km/h. The total time en route is 1 hour and 36 minutes with a total route duration of 4 hours and 6 minutes. The operational time associated with each site visited is 5 minutes and 43 seconds.

Embedded in cluster 3 were 32.4% of the total number of routes, characterized by being the routes with the shortest distances, with an average total distance of 4 309 meters, distancing themselves from the start location 2 217 meters, visiting on average 6 different locations, at an average speed of 7.06 Km/h. These routes had an average time en route of 42 minutes and a total route duration of 2 hours and 24 minutes. An interesting statistic fact is that these routes present the highest operation time, with 8 minutes and 25 seconds. This may be associated with the high waiting time for customers according to YOOB feedback.

The last cluster identified, cluster 4, representing 24.8% of the routes, characterized by having a total distance of 28 007 meters, being 6 381 meters away from its starting location. These routes, presented the highest number of visited locations, with an average of 22 different locations visited per route at a speed of 9.41 Km/h. Also, presented the highest times, with a total time en route of 3 hours and 24 minutes and a total route duration of 6 hours and 42 minutes. The operational time for these routes is 4 minutes and 31 seconds.

In the third model we obtained a proposal for the expansion of the network of YOOB hubs, composed of a total of eight, with an additional three of the five ones existing at the time of the thesis writing. In Figure 3.66 we observed that the current micro-hub 0 and nano-hub 2 are located very close to the locations suggested by our model. For nano-hub 1 the deviation is relatively short, approximately 400 meters, for nano-hub 3 the deviation is approximately 700 meters and for nano-hub 4 the deviation is larger, with a value of approximately 1.5 Km (Euclidean distances measured in Google Maps [76]). The three new proposed locations by the model, are in the Alvalade borough, identified as "C0", Benfica borough, identified as "C6" and the third one in Algés identified in the figure as "C5". In Figure 3.67 it is possible to have an overview of the volume allocated with each location. This indicator can be related to the type of hub (micro or nano) more suitable for the area. By analyzing

this indicator in more detail, we could understand better if the existing nano hub 1 would benefit to be transformed (and upgraded) into a micro hub.

3.6.2. End-user evaluation

The end-user evaluation verifies that the findings are consistent with the proposed research objectives and the accuracy of the business requirements.

While working on this thesis, three briefing meetings and presentations were organized with the YOOB partners. This ensure that YOOB partners observed and supported the development of this thesis using the company's knowledge and experience in the field.

In the end of the study a questionnaire was sent to the two YOOB partners, with the questions and answers indicated in Table 3.6.

Table 3.6 - Method assessment questionnaire

Criteria	Objective statement	Evaluator #1	Evaluator #2
Utility	It can help business decisions regarding the behavior of the fleet and hub expansion	FA	FA
Understandability	Provides understandable results	FA	FA
Accessibility	Can be used without training	LA	LA
Level of detail	Provides knowledge from the mobility of the fleet and detailed location for expansion.	FA	FA
Consistency	Gives consistent results.	LA	LA
Robustness	Has enough detail to be used in other cases of e-cargo bikes and hub expansion	FA	FA

The development of the questionnaire follows the standards defined by the ISO/IEC TS 33061 [77], primarily used to assess software development processes. Four levels of the NLPF were employed for evaluation:

- Not Achieved (NA) - [0-15%]
- Partially Achieved (PA) -]15-50%]
- Largely Achieved (LA) -]50-85%]
- Fully Achieved (FA) -]85-100%]

In the evaluation from the company, made by the two YOOB partners and co-supervisors of this thesis, we obtained a rating of FA, in the criteria of usefulness, understanding, level of detail and robustness, and LA rating in the criteria of accessibility and consistency. Overall, this indicates that the work done represents an added value for the company, providing useful, detailed and clear information, with appropriate to support decision making, in the context of the e-cargo bike fleet as well as for the expansion of new hubs.

As this is the first thesis in the field of data science applied to e-cargo bikes in urban settings in a Portuguese city as far as the authors are aware, the YOOB evaluators consider that this study can be replicated to other case studies with potential for improvement, and implementation readiness.

Moreover, the outcomes are aligned with the objectives and requirements proposed for this study.

3.7. Deployment

The models created were not applied in a real production environment. Observations and results obtained were compiled in the thesis writing and summarized in a report presented in power point format to the YOOB partners. All software development was done in Python on a personal computer equipped with Windows 10 (64bits) operating system, Intel$^{(R)}$ Core$^{(TM)}$ i7-11370H 3.30GHz, with 40Gb of memory ram. We adopted the python programming language (v3.10.4) [44], compiled with Visual Studio Code (v1.69.1) [45] on Jupyter Notebooks extension [46]. The packages used are detailed in section 3.4 Data Preparation. The reproducibility of the whole process can be accomplished by running the Jupyter Notebooks and auxiliary files provided along with the thesis. All the developed software material and data sets is available for use by the YOOB company and for further academic research purposes.

4. Conclusions

4.1. Discussion

In the work developed throughout the thesis, we analyzed data allowing us to answer our research questions:

RQ1: “How can we characterize the spatial-temporal traffic of the last mile logistic distribution performed with the electric cargo bike fleet, taking into consideration the open data of the city and the data collected during the performed routes?”

RQ2: “Based on the fleet behavior and the patterns detected, what are the best possible locations for micro-hubs or nano-hubs expansion?”

The results of the second model (Clustering the routes with K-Means), served as a proxy to answer the first research question, allow us to characterize the behavior of the e-cargo bike fleet through the traveled distance, time, speed and number of visited locations.

Five types of performances were found in our cluster analysis. The most common performance is the one observed in cluster 0, accounting to 41.2% of the total trips (see Figure 3.58). This cluster features an average speed of 6.84 Km/h and is the lowest speed of the five clusters, corresponding to a total average traveled distance of 11.16 Km. YOOB´s e-cargo bikes travel, on average, at a maximum distance of 3.64 Km, from their starting location. On average, the total duration of cluster 0 trips is 3 hours and 54 minutes, and the e-cargo bikes are only in motion for an average period of 1 hour and 48 minutes. Seventeen different locations are visited on average, and 3 minutes and 40 seconds is the shortest operating time per location visited, during trips of cluster 0.

The second largest type of performance is observed in cluster 3, which includes 32.4% of the total trips (see Figure 3.58). It is characterized by an average total distance traveled of 4.31 Km, at an average speed of 7.96 Km/h. In cluster 3, e-cargo bikes travel a maximum distance of 2.22 Km in average, from their starting location. These trips have the shortest and closest travel distances. On average they have a total duration of 2 hours and 24 minutes, and bikes are only in motion for an average of 42 minutes. With an average of six different locations, cluster 3 has the fewest number of locations visited from all five performances, but has the longest operation time per location visited, requiring an average of 8 minutes and 25 seconds.

A third most predominant type of performance is the one observed in cluster 2, with 31.6% of total trips (see Figure 3.58). On average the total distance traveled is 14.15 Km at an average speed of 9.75 Km/h. The e-cargo bikes travel on average at a maximum distance of 6.23 Km from the starting location. The total travel time is 4 hours and 06 minutes, on average, with the e-cargo bikes being in motion for an average time of 1 hour and 36 minutes. Thirteen different locations are visited, on average, and bikers spend an average of 5 minutes and 43 seconds for each location.

The fourth most observed performance type is the one of cluster 4, with 24.8% of the total trips (see Figure 3.58). It is characterized by a total traveled distance, on average of 28.01 Km, at an average speed of 9.41 Km/h. The e-cargo bikes travel on average at a maximum distance of 6 381 meters from their starting location, with a total duration of the route, on average, of 6 hours and 42 minutes. Bikes are in motion for an average period of 3 hours and 24 minutes. These are the trips with the longest travel time and with the largest number of places visited, with an average of twenty-two different places. At each location visited bikers spend an average of 4 minutes and 31 seconds in operation time.

The least observed type of performance is the one corresponding to cluster 1 (see Figure 3.58), with only 2.8% of the total trips. These are the longest trips with the wider range, but also the fastest ones, with an average of total distance traveled of 35.58 Km, at an average speed of 11.20 Km/h. In this cluster, the e-cargo bikes travel at a maximum distance of 14.16 Km, from their starting location. On average, the total travel time of trip is 5 hours and 12 minutes, with the e- bikes being in motion for an average period of 2 hours and 54 minutes. An average of twelve different locations are visited, and bikers spend an average of 5 minutes and 47 seconds in each location.

Overall, the average of total traveled distance ranges between 4.31 Km and 35.5 Km, distancing from their start location, on average, between 2.2 Km and 14.10 Km. 63% of the routes were very close or even lower than the values for which Sheth et al. [29], which considered cargo bikes to have an efficient performance under 3.20 Km. The average number of different locations visited per route ranges between 6 and 22. The average observed speed varies between 6.84 Km/h and 11.20 Km/h, a value close to the study by Bütten et al. [13], where these authors looked at several cargo bike projects, and calculated average speeds between 8.00 Km/h and 25.00 Km/h. The temporal characteristics revealed an average time in movement per route from 42 minutes up to 3 hours and 24 minutes and the average total route duration times, ranged between 2 hours and 24 minutes and 6 hours and 42 minutes. Required transaction time within each route ranged from 3 minutes and 40 seconds to 8 minutes and 25 seconds. This higher time may be due to the particularities of certain customers requiring more waiting time. Excluding this last observation, the time metric ranges between 3 minutes and 40 seconds and 5 minutes and 43 seconds. This set of characteristics gave us an overview of the needs of each route and the respective performance of the e-cargo bikes in their operation conditions.

As for the second question, the third model (K-Means center gravity analysis), was used as our basis for analysis. The choice of new hubs locations, in the context of an expansion of the e-cargo bikes network, is a complex process due to the high number of constraints that are to be considered in the site search [6], [25], [27], [32], [34].

In the search for new locations the factors considered for the cost function of our model, were the distance and the cost associated with each visited location. Then for evaluation of the hub type, the volume associated with each hub of this new structure was analyzed.

When simulating an expansion of three more hubs beyond the five that are currently part of YOOB's network, our model suggests that the implementation of these new hubs should be located in the boroughs of Alvalade, Benfica and Algés. When presenting the results of this model, the YOOB partners considered that these three new proposed locations are valid options that required further analysis in terms of economic viability. Regarding the 3 remaining computed locations, in the case of C2, the choice of the current location of the hub (nr 1), which is within the radius of this cluster, was due to the geographical characteristics of the area, which is on top of a hill, causing the trips to have a downward direction, facilitating the effort required by the biker. In the case of C7, the divergence between the location of the hub (nr 4) and the location proposed by the model raises additional challenges of further changes of location due to the high price of real estate in the area where the centroid calculated by our model is located. In the remaining clusters no particular observations were mentioned by the YOOB partners, other than agreement. By analyzing the volume of parcels associated with each hub in Figure 3.67, we can discuss what type of hub is the most adequate for micro-hub or nano-hub requirements. In our study all three new locations, are more suitable for nano-hubs. In the already existing nano-hub located in the Saldanha, we observed that due to the high associated volume of parcels it could shift to a micro-hub, and this observation was positively validated by YOOB partners.

Deeper research should be conducted to explore further details, including the economic viability of this hub expansion proposal.

4.2. Research limitations

This study is innovative as the first study in the data science field with YOOB data operating in Lisbon, as far as the authors are aware.

The most significant limitations of our study are related to the dimension, granularity, and structure of the data. The information on the routes was limited to the visited geographical points, lacking information about the order of each visited location, and lacked complete information about the route trajectory (its 3D coordinates) taken from pickup to delivery. Having trajectory data would allow a deeper and more rigorous analysis of the e-cargo bike fleet route patterns, namely the real trajectories in each which route was performed and the actual distances traveled. We collected data in the period from January to April of 2022, corresponding to the first four months of the company's registered activity (YOOB started operations in Lisbon in the fall of 2021). After data pre-processing, we came up with a dataset comprising 15 828 records and 27 variables, which was considered sufficient for our analysis, but that nevertheless can be limited for long-term trend analysis.

The proposed hub locations can be considered the best possible locations with limitations, as many factors were not considered, such as street elevations or socio-economic factors.

4.3. Future work

The following suggestions are made for upcoming research work:

- Expand the number of observations analyzed to detect long-term trends and produce more insightful results, given that YOOB has the possibility to collect stories and route data on a regular basis.
- Study the shortest and flattest path
- Perform more detailed cluster analysis, with an increased number of clusters when analyzing route typologies.
- Include centrality metrics and more data sources such as real estate prices, other economic data, city data (traffic congestion, mobility data from cell operators), weather data, pollution data and detailed terrain information, to further study the expansion of the YOOB network with new hub locations.
- With enriched route data collection, including time stamped trajectory data adopt a data fusion approach with the mentioned data sources, to predict the demand of the use of the YOOB e-cargo bike network service.

References

[1] "The Future of the Last-Mile Ecosystem Transition Roadmaps for Public-and Private-Sector Players," 2020. [Online]. Available: www.weforum.org

[2] A. Conway, J. Cheng, C. Kamga, and D. Wan, "Cargo cycles for local delivery in New York City: Performance and impacts," *Research in Transportation Business & Management*, vol. 24, pp. 90–100, Sep. 2017, doi: 10.1016/j.rtbm.2017.07.001.

[3] I. Cardenas, Y. Borbon-Galvez, T. Verlinden, E. van de Voorde, T. Vanelslander, and W. Dewulf, "City logistics, urban goods distribution and last mile delivery and collection," *Competition and Regulation in Network Industries*, vol. 18, no. 1–2, pp. 22–43, Mar. 2017, doi: 10.1177/1783591717736505.

[4] European Environment Agency, *Urban Sustainability in Europe - Learning from nexus analysis*, no. 07. 2021.

[5] L. Faugère, C. White, and B. Montreuil, "Mobile Access Hub Deployment for Urban Parcel Logistics," *Sustainability*, vol. 12, no. 17, p. 7213, Sep. 2020, doi: 10.3390/su12177213.

[6] V. Naumov, "Substantiation of loading hub location for electric cargo bikes servicing city areas with restricted traffic," *Energies (Basel)*, vol. 14, no. 4, Feb. 2021, doi: 10.3390/en14040839.

[7] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *ACM Comput Surv*, vol. 51, no. 4, 2018, doi: 10.1145/3161602.

[8] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews", doi: 10.1136/bmj.n71.

[9] "CRISP-DM: A Framework For Data Mining & Analysis." https://thinkinsights.net/digital/crisp-dm/ (accessed May 20, 2022).

[10] "https://www.vosviewer.com/ (accessed April 13, 2022)."

[11] "https://www.mendeley.com (accessed April 23, 2022)."

[12] J. G. Urzúa-Morales, J. P. Sepulveda-Rojas, M. Alfaro, G. Fuertes, R. Ternero, and M. Vargas, "Logistic Modeling of the Last Mile: Case Study Santiago, Chile," *Sustainability*, vol. 12, no. 2, p. 648, Jan. 2020, doi: 10.3390/su12020648.

[13] A. Büttgen, B. Turan, and V. Hemmelmayr, "Evaluating Distribution Costs and CO2-Emissions of a Two-Stage Distribution System with Cargo Bikes: A Case Study in the City of Innsbruck," *Sustainability*, vol. 13, no. 24, p. 13974, Dec. 2021, doi: 10.3390/su132413974.

[14] K. Katsela, Ş. Güneş, T. Fried, A. Goodchild, and M. Browne, "Defining Urban Freight Microhubs: A Case Study Analysis," *Sustainability*, vol. 14, no. 1, p. 532, Jan. 2022, doi: 10.3390/su14010532.

[15] T. Assmann, S. Lang, F. Müller, and M. Schenk, “Impact assessment model for the implementation of cargo bike transshipment points in urban districts,” *Sustainability (Switzerland)*, vol. 12, no. 10, May 2020, doi: 10.3390/SU12104082.

[16] J. F. Toro, D. Carrion, M. A. Brovelli, and M. Percoco, “BIKEMI BIKE-SHARING SERVICE EXPLORATORY ANALYSIS ON MOBILITY PATTERNS,” in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Aug. 2020, vol. XLIII-B4-2, no. B4, pp. 197–203. doi: 10.5194/isprs-archives-XLIII-B4-2020-197-2020.

[17] X. Guo, Z. Xu, J. Zhang, J. Lu, and H. Zhang, “An OD Flow Clustering Method Based on Vector Constraints: A Case Study for Beijing Taxi Origin-Destination Data,” *ISPRS Int J Geoinf*, vol. 9, no. 2, p. 128, Feb. 2020, doi: 10.3390/ijgi9020128.

[18] X. Ma, C. Liu, H. Wen, Y. Wang, and Y.-J. Wu, “Understanding commuting patterns using transit smart card data,” *J Transp Geogr*, vol. 58, pp. 135–145, Jan. 2017, doi: 10.1016/j.jtrangeo.2016.12.001.

[19] Y. Shen, X. Zhang, and J. Zhao, “Understanding the usage of dockless bike sharing in Singapore,” *Int J Sustain Transp*, vol. 12, no. 9, pp. 686–700, Oct. 2018, doi: 10.1080/15568318.2018.1429696.

[20] L. Zheng *et al.*, “Spatial–temporal travel pattern mining using massive taxi trajectory data,” *Physica A: Statistical Mechanics and its Applications*, vol. 501, pp. 24–41, Jul. 2018, doi: 10.1016/j.physa.2018.02.064.

[21] Y. Huang, Z. Xiao, D. Wang, H. Jiang, and D. Wu, “Exploring Individual Travel Patterns Across Private Car Trajectory Data,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 12, pp. 5036–5050, Dec. 2020, doi: 10.1109/TITS.2019.2948188.

[22] Rong Wen, Wenjing Yan, A. N. Zhang, Nguyen Quoc Chinh, and O. Akcan, “Spatio-temporal route mining and visualization for busy waterways,” in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2017, pp. 000849–000854. doi: 10.1109/SMC.2016.7844346.

[23] J. C. Amaral and C. B. Cunha, “An exploratory evaluation of urban street networks for last mile distribution,” *Cities*, vol. 107, p. 102916, Dec. 2020, doi: 10.1016/j.cities.2020.102916.

[24] F. Li, W. Shi, and H. Zhang, “A Two-Phase Clustering Approach for Urban Hotspot Detection With Spatiotemporal and Network Constraints,” *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 14, pp. 3695–3705, 2021, doi: 10.1109/JSTARS.2021.3068308.

[25] H. Y. Song and I. Han, “Finding the Best Location for Logistics Hub Based on Actual Parcel Delivery Data,” in *COMPUTATIONAL SCIENCE AND ITS APPLICATIONS - ICCSA 2019, PT I: 19TH INTERNATIONAL CONFERENCE, SAINT PETERSBURG, RUSSIA, JULY 1-4, 2019, PROCEEDINGS, PT I*, 2019, vol. 11619, no. 19th International Conference on Computational Science and Its Applications (ICCSA), pp. 603–615. doi: 10.1007/978-3-030-24289-3_45.

[26] R. Barraza, J. M. Sepúlveda, J. Venegas, V. Monardes, and I. Derpich, “A Model for Solving Optimal Location of Hubs: A Case Study for Recovery of Tailings Dams,” in *INTELLIGENT*

METHODS IN COMPUTING, COMMUNICATIONS AND CONTROL, vol. 1243, no. 8th International Conference on Computers Communications and Control (ICCCC), I. Dzitac, F. G. Filip, M. J. Manolescu, S. Dzitac, J. Kacprzyk, and H. Oros, Eds. Univ Santiago, Santiago, Chile, 2021, pp. 304–312. doi: 10.1007/978-3-030-53651-0_26.

[27] J. Hwang, J. S. Lee, S. Kho, and D. Kim, "Hierarchical hub location problem for freight network design," *IET Intelligent Transport Systems*, vol. 12, no. 9, pp. 1062–1070, Nov. 2018, doi: 10.1049/iet-its.2018.5289.

[28] C. Rudolph, A. Nsamzinshuti, S. Bonsu, A. B. Ndiaye, and N. Rigo, "Localization of Relevant Urban Micro-Consolidation Centers for Last-Mile Cargo Bike Delivery Based on Real Demand Data and City Characteristics," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2676, no. 1, pp. 365–375, Jan. 2022, doi: 10.1177/03611981211036351.

[29] M. Sheth, P. Butrina, A. Goodchild, and E. McCormack, "Measuring delivery route cost trade-offs between electric-assist cargo bicycles and delivery trucks in dense urban areas," *European Transport Research Review*, vol. 11, no. 1, p. 11, Dec. 2019, doi: 10.1186/s12544-019-0349-5.

[30] R. Golini, C. Guerlain, A. Lagorio, and R. Pinto, "AN ASSESSMENT FRAMEWORK TO SUPPORT COLLECTIVE DECISION MAKING ON URBAN FREIGHT TRANSPORT," *Transport*, vol. 33, no. 4, pp. 890–901, Dec. 2018, doi: 10.3846/transport.2018.6591.

[31] M. Leyerer, M.-O. Sonneberg, M. Heumann, and M. H. Breitner, "Shortening the Last Mile in Urban Areas: Optimizing a Smart Logistics Concept for E-Grocery Operations," *Smart Cities*, vol. 3, no. 3, pp. 585–603, Jul. 2020, doi: 10.3390/smartcities3030031.

[32] N. Ghaffarinasab, "A tabu search heuristic for the bi-objective star hub location problem," *International Journal of Management Science and Engineering Management*, vol. 15, no. 3, pp. 213–225, Jul. 2020, doi: 10.1080/17509653.2019.1709992.

[33] S. Srivatsa Srinivas and R. R. Marathe, "Moving towards 'mobile warehouse': Last-mile logistics during COVID-19 and beyond," *Transp Res Interdiscip Perspect*, vol. 10, p. 100339, Jun. 2021, doi: 10.1016/j.trip.2021.100339.

[34] Z. Huang, W. Huang, and F. Guo, "Integrated sustainable planning of micro-hub network with mixed routing strategy," *Comput Ind Eng*, vol. 149, p. 106872, Nov. 2020, doi: 10.1016/j.cie.2020.106872.

[35] T. Assmann, S. Bobeth, and E. Fischer, "A conceptual framework for planning transhipment facilities for cargo bikes in last mile logistics," in *4th Conference on Sustainable Urban Mobility, CSUM 2018*, 2018. doi: 10.1007/978-3-030-02305-8_69.

[36] L. Caggiani, A. Colovic, L. P. Prencipe, and M. Ottomanelli, "A green logistics solution for last-mile deliveries considering e-vans and e-cargo bikes," in *Transportation Research Procedia*, 2021, vol. 52, pp. 75–82. doi: 10.1016/j.trpro.2021.01.010.

[37] A. Kedia, D. Kusumastuti, and A. Nicholson, “Locating collection and delivery points for goods’ last-mile travel: A case study in New Zealand,” in *Transportation Research Procedia*, 2020, vol. 46, pp. 85–92. doi: 10.1016/j.trpro.2020.03.167.

[38] T. M. Özbekler and A. Karaman Akgül, “LAST MILE LOGISTICS IN THE FRAMEWORK OF SMART CITIES: A TYPOLOGY OF CITY LOGISTICS SCHEMES,” in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Nov. 2020, vol. XLIV-4/W3-, no. 4/W3, pp. 335–337. doi: 10.5194/isprs-archives-XLIV-4-W3-2020-335-2020.

[39] M. Arrieta-Prieto, A. Ismael, C. Rivera-Gonzalez, and J. E. Mitchell, “Location of urban micro-consolidation centers to reduce the social cost of last-mile deliveries of cargo: A heuristic approach,” *Networks*, p. net.22076, Aug. 2021, doi: 10.1002/net.22076.

[40] R. Madlenak, L. Madlenakova, P. Drozdziel, I. Rybicka, and T. Ltd, “SEARCHING OPTIMAL HUB LOCATIONS IN POSTAL LOGISTIC NETWORK,” in *CARPATHIAN LOGISTICS CONGRESS (CLC’ 2017)*, 2017, no. Carpathian Logistics Congress (CLC), pp. 216–221.

[41] R. Wirth and J. Hipp, “CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39,” *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, no. 24959, pp. 29–39, 2000, [Online]. Available: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining

[42] “https://yoob.pt/.” https://yoob.pt/ (accessed Mar. 17, 2022).

[43] S. Clarke and J. Leonardi, “Data Report Agile Gnewt Cargo: parcels deliveries with electric vehicles in Central London Multi-carrier central London micro-consolidation and final delivery via low carbon vehicles.” [Online]. Available: www.london.gov.uk

[44] “www.Python.org.” https://www.python.org/ (accessed Jul. 13, 2022).

[45] “Visual Studio Code.” https://code.visualstudio.com/ (accessed Jul. 13, 2022).

[46] “Project Jupyter.” https://jupyter.org/ (accessed Jul. 13, 2022).

[47] “descartes.” http://descartes.sourceforge.net/ (accessed Jul. 13, 2022).

[48] “jwass/geog: Quick and easy geographical functions in Python.” https://github.com/jwass/geog (accessed Jul. 13, 2022).

[49] “Welcome to GeoPy’s documentation! — GeoPy 2.2.0 documentation.” https://geopy.readthedocs.io/en/stable/ (accessed Jul. 13, 2022).

[50] “GeoPandas 0.11.0 — GeoPandas 0.11.0+0.g1977b50.dirty documentation.” https://geopandas.org/en/stable/ (accessed Jul. 13, 2022).

[51] “Rasterio: access to geospatial raster data — rasterio documentation.” https://rasterio.readthedocs.io/en/latest/ (accessed Jul. 13, 2022).

[52] “OSMnx 1.2.1 — OSMnx 1.2.1 documentation.” https://osmnx.readthedocs.io/en/stable/ (accessed Jul. 13, 2022).

[53] “contextily: context geo tiles in Python — contextily 1.1.0 documentation.” https://contextily.readthedocs.io/en/latest/index.html (accessed Jul. 13, 2022).

[54] “geoplot: geospatial data visualization — geoplot 0.5.0 documentation.” https://residentmario.github.io/geoplot/ (accessed Jul. 13, 2022).

[55] “python-visualization/folium: Python Data. Leaflet.js Maps.” https://github.com/python-visualization/folium (accessed Jul. 13, 2022).

[56] “pysal/mapclassify: Classification schemes for choropleth mapping.” https://github.com/pysal/mapclassify (accessed Jul. 13, 2022).

[57] “Matplotlib — Visualization with Python.” https://matplotlib.org/ (accessed Jul. 13, 2022).

[58] “Plotly: Low-Code Data App Development.” https://plotly.com/ (accessed Jul. 13, 2022).

[59] “ppinard/matplotlib-scalebar: Provides a new artist for matplotlib to display a scale bar, aka micron bar.” https://github.com/ppinard/matplotlib-scalebar (accessed Jul. 13, 2022).

[60] M. Waskom, “seaborn: statistical data visualization,” *J Open Source Softw*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[61] “NumPy.” https://numpy.org/ (accessed Jul. 13, 2022).

[62] “pandas - Python Data Analysis Library.” https://pandas.pydata.org/ (accessed Jul. 13, 2022).

[63] “pyproj4/pyproj: Python interface to PROJ (cartographic projections and coordinate transformations library).” https://github.com/pyproj4/pyproj (accessed Jul. 13, 2022).

[64] “SciPy.” https://scipy.org/ (accessed Jul. 13, 2022).

[65] “The Shapely User Manual — Shapely 1.8.2 documentation.” https://shapely.readthedocs.io/en/stable/manual.html (accessed Jul. 13, 2022).

[66] “scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation.” https://scikit-learn.org/stable/ (accessed Jul. 13, 2022).

[67] “Welcome to kneed’s documentation! — kneed 0.6.0 documentation.” https://kneed.readthedocs.io/en/stable/ (accessed Jul. 13, 2022).

[68] “EU-DEM v1.1 — Copernicus Land Monitoring Service.” https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1?tab=metadata (accessed Jul. 14, 2022).

[69] “OpenStreetMap.” https://www.openstreetmap.org/#map=13/38.7588/-9.1358 (accessed Jul. 18, 2022).

[70] “sklearn.preprocessing.MinMaxScaler — scikit-learn 1.1.1 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (accessed Jul. 18, 2022).

[71] “sklearn.preprocessing.LabelEncoder — scikit-learn 1.1.1 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html (accessed Jul. 18, 2022).

[72] “sklearn.cluster.KMeans — scikit-learn 1.1.1 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (accessed Jul. 18, 2022).

[73] “sklearn.metrics.davies_bouldin_score — scikit-learn 1.1.1 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html (accessed Jul. 18, 2022).

[74] R. Wen, W. Yan, and A. N. Zhang, “Weighted clustering of spatial pattern for optimal logistics hub deployment,” in *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 2016, pp. 3792–3797. doi: 10.1109/BigData.2016.7841050.

[75] C. Cai, Y. Luo, Y. Cui, and F. Chen, “Solving multiple distribution center location allocation problem using k-means algorithm and center of gravity method take jinjiang district of chengdu as an example,” in *IOP Conference Series: Earth and Environmental Science*, Oct. 2020, vol. 587, no. 1. doi: 10.1088/1755-1315/587/1/012120.

[76] “Google Maps.” https://www.google.pt/maps/@38.9954378,-9.1411938,10z?hl=pt-PT (accessed Jul. 19, 2022).

[77] “ISO - ISO/IEC TS 33061:2021 - Information technology — Process assessment — Process assessment model for software life cycle processes.” https://www.iso.org/standard/80362.html (accessed Jul. 28, 2022).

Annexes and appendix

Annex A – DBSCAN Model one

# clusters	4	6	6	6	6	5	4	5	6	7	8	10	12	10	9	12	18	25	25	30	31	33	42	42	57	56	62	57	53	48
% Noise	3%	3%	3%	3%	4%	4%	4%	4%	5%	5%	6%	6%	6%	7%	8%	9%	9%	10%	12%	13%	15%	16%	18%	21%	24%	27%	30%	34%	38%	42%
eps value	0,07	0,068	0,066	0,064	0,062	0,06	0,058	0,056	0,054	0,052	0,05	0,048	0,046	0,044	0,042	0,04	0,038	0,036	0,034	0,032	0,03	0,028	0,026	0,024	0,022	0,02	0,018	0,016	0,014	0,012
cluster	# of sub-stories per cluster																													
-1 (noise)	441	465	484	514	561	612	656	688	732	799	870	930	1007	1104	1225	1360	1483	1639	1865	2067	2331	2560	2904	3314	3781	4320	4790	5409	6008	6589
0	15306	15266	15246	15221	15173	15132	15098	15046	14984	14917	14783	14695	14601	14523	13887	13302	11608	11480	11272	11086	10896	10744	5269	5138	3835	3816	3790	3705	3629	3586
1	30	30	30	30	30	30	30	30	27	27	65	12	11	64	62	60	59	48	45	20	16	16	75	72	166	137	108	80	75	58
2	13	13	13	13	13	13	13	13	13	13	26	64	64	24	23	23	22	22	22	26	26	23	15	13	349	150	65	53	39	28
3	18	14	13	11	10	10	11	20	21	21	13	25	25	11	542	525	501	27	26	390	380	375	22	22	67	160	102	32	29	28
4		15	14	11	11	11		11	20	11	21	13	13	13	10	156	156	442	434	120	111	106	3849	3598	13	62	57	29	48	47
5		5	8	8	10				11	11	11	21	20	20	13	13	1001	148	124	45	41	30	371	321	22	11	21	90	2004	1664
6										9	11	18	18	16	20	19	24	966	48	43	892	874	102	96	3196	21	2961	51	124	116
7											8	10	13	14	10	10	13	41	932	897	30	29	27	23	311	3080	174	19	52	49
8												11	7	9	16	52	566	25	37	30	31	26	529	487	93	181	86	2471	463	463
9												9	10	10		264	19	6	23	34	31	24	873	845	15	86	464	169	125	109
10													10			16	11	13	34	32	27	24	28	28	473	470	137	59	13	89
11													9			8	49	566	23	12	24	13	26	14	464	217	37	463	110	11
12																	247	18	13	27	13	565	27	26	843	42	90	213	88	96
13																	16	11	565	27	42	20	54	27	28	134	834	123	16	87
14																	6	10	18	13	565	18	13	36	22	842	28	90	819	13
15																	10	15	11	565	18	18	13	50	25	28	71	819	18	818
16																	9	31	14	18	17	10	565	13	35	20	20	25	24	18
17																	8	25	29	24	10	10	17	13	39	25	43	136	302	103
18																		221	186	8	13	10	17	104	35	19	15	55	131	124
19																		16	15	10	7	9	18	565	20	90	79	42	29	96
20																		5	23	13	16	148	9	17	13	29	27	15	31	116
21																		7	20	18	156	13	501	18	13	34	34	50	34	77
22																		9	9	181	17	10	21	42	99	38	36	24	42	13
23																		8	10	15	13	16	112	492	14	17	19	13	11	20
24																		9	10	24	13	10	10	15	563	13	29	13	86	13
25																				19	15	14	25	110	26	98	14	86	31	70
26																				14	10	8	16	25	17	563	21	562	13	41
27																				13	9	13	31	24	15	21	21	9	85	10
28																				8	13	13	12	12	21	17	13	15	9	560
29																				9	15	12	59	50	39	13	91	18	561	28
30																					10	12	25	16	214	33	563	213	30	204
31																						10	47	11	8	213	9	32	15	21
32																						25	10	37	14	9	16	13	205	20
33																							14	10	7	16	31	22	19	10
34																							14	11	49	27	15	8	18	21
35																							11	24	17	10	213	14	10	15
36																							12	15	257	153	10	132	106	82
37																							10	22	143	65	143	30	29	29
38																							10	7	14	15	45	10	84	82
39																							10	10	10	103	10	18	15	12
40																							10	10	104	36	101	13	10	11
41																							25	25	20	133	11	96	17	12
42																									49	18	10	12	12	82
43																									20	22	123	10	10	10
44																									27	10	27	10	10	10
45																									37	26	14	110	10	12
46																									7	28	10	10	88	10
47																									29	16	12	10	10	25
48																									11	11	11	12	6	
49																									19	14	10	10	10	
50																									21	16	13	9	7	
51																									14	18	19	11	13	
52																									10	12	9	12	25	
53																									10	15	12	25		
54																									10	10	12	11		
55																									10	25	11	11		
56																									25		6	6		
57																											10			
58																											10			
59																											25			
60																											9			
61																											11			

Annex B – DBSCAN Model two

# clusters	5	5	5	5	6	5	4	4	5	4	5	5	6	6	12	12	12	14	15	19	23	29	28	26	24	28	33	30	29	22
% Noise	6%	6%	6%	7%	7%	8%	9%	9%	10%	11%	11%	13%	14%	16%	17%	18%	19%	21%	23%	25%	28%	29%	34%	40%	46%	52%	58%	65%	70%	78%
eps value	**0,035**	**0,034**	**0,033**	**0,032**	**0,031**	**0,03**	**0,029**	**0,028**	**0,027**	**0,026**	**0,025**	**0,024**	**0,023**	**0,022**	**0,021**	**0,02**	**0,019**	**0,018**	**0,017**	**0,016**	**0,015**	**0,014**	**0,013**	**0,012**	**0,011**	**0,01**	**0,009**	**0,008**	**0,007**	**0,006**
cluster	# of routes per cluster																													
-1 (noise)	39	40	42	47	48	55	59	61	65	74	74	83	93	103	110	118	129	139	156	169	186	194	228	264	303	342	383	430	463	516
0	600	600	599	597	593	593	592	590	578	574	558	554	502	497	427	422	351	330	321	316	292	64	64	61	61	33	33	21	20	17
1	5	5	5	5	5	5	5	5	8	8	8	15	26	18	18	18	17	17	17	17	11	11	127	122	108	34	5	10	10	10
2	5	5	5	4	4	4	4	4	5	4	16	4	20	15	6	6	6	6	6	6	14	206	14	12	12	11	10	9	8	8
3	11	10	9	7	7	4	4	4	4	4	4	4	4	20	11	11	68	67	58	35	20	14	6	6	50	40	11	6	6	6
4	4	4	4	4	4	3			4		4	4	15	7	31	31	9	8	8	19	4	7	56	53	12	15	7	6	6	6
5					3								4	4	14	11	30	19	19	6	19	5	19	12	4	12	6	6	6	11
6															18	18	11	9	4	17	6	19	6	6	6	6	7	18	14	5
7															4	4	17	17	8	8	12	6	6	7	6	10	8	6	6	4
8															7	7	4	11	11	11	16	10	5	5	7	29	20	15	6	4
9															4	4	7	15	16	10	11	8	13	12	9	6	12	6	4	4
10															3	3	4	4	15	9	10	16	16	14	12	7	7	4	4	5
11															11	11	11	7	4	4	9	18	18	17	11	9	16	7	8	6
12																		4	7	6	5	9	9	9	8	5	13	4	6	8
13																		11	4	7	6	5	5	5	4	9	9	8	9	14
14																			10	6	4	5	4	4	6	10	5	9	9	5
15																				4	6	4	4	4	4	6	8	11	4	7
16																				4	6	6	6	6	6	7	4	4	6	4
17																				6	6	6	6	6	4	4	8	6	6	5
18																				4	4	6	11	6	6	28	4	6	4	5
19																					5	4	4	4	6	5	6	7	5	4
20																					4	5	8	7	5	4	7	5	4	4
21																					4	9	4	4	4	4	4	4	8	6
22																					4	4	4	4	4	4	28	13	14	
23																						4	3	5	6	4	4	8	5	
24																						4	5	4		6	4	15	7	
25																						2	4	5		4	4	4	4	
26																						5	4			4	5	4	4	
27																						4	5			6	4	4	4	
28																						4					4	4	4	
29																											4	4		
30																											4			
31																											4			
32																											6			