

# Planning Delivery Services: Depot Clustering Based on Socio-Economic Indicators and Geospatial Metrics

Iñaki Cejudo<sup>a</sup>, Laura Rabadán<sup>b</sup>, Eider Irigoyen<sup>c</sup> and Harbil Arregui<sup>d</sup>

*Intelligent Systems for Mobility and Logistics, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),  
Mikeletegi 57, Donostia, Spain  
{icejudo, lrabadan, eirigoyen, harregui}@vicomtech.org*

**Keywords:** Depot Clustering, Delivery Logistics, Socio-Economic Indicators, Urban Network, Decision Support Systems.

**Abstract:** People's lifestyles have evolved in recent years, making home deliveries a necessity for various types of services. Moreover, with the growth of big data and Artificial Intelligence, predicting the performance and customer demand of new businesses is a key aspect of logistics and last-mile delivery planning. By using examples and predictions as a foundation for goods delivery services, initial over-sizing costs can be significantly reduced. In this paper, we analyze and compare operational zone similarities for food and parcel delivery services in Spain, considering socio-economic indicators and urban network features. The study leverages motorbike delivery metrics to complement the analysis. The results demonstrate how similar depots can be clustered, providing a foundational performance scenario for decision-making when planning the launch of a new service.

## 1 INTRODUCTION


The rapid evolution of urban lifestyles has significantly transformed our habits in many aspects. One of them is how e-commerce has altered our way of buying things. We now buy from home and expect a fast delivery. As customer habits were evolving, last-mile delivery logistics in cities have adapted too. Furthermore, these habits have been extended to ordering food, where in the last few years there has been a game changer in cities with many riders delivering food.


Opening new commerce now entails planning an efficient last-mile delivery system, and it requires a delicate balance between cost management, operational scalability, and customer satisfaction. Moreover, the diversity of urban environments, shaped by socio-economic and geographical factors, adds layers of complexity to this process. Understanding these dynamics is essential for predicting the performance of a new commerce in terms of demand, and therefore correctly sizing the service needs, optimizing delivery operations, and ensuring their sustainability in an increasingly competitive market.


One of the biggest challenges is to predict the service demand and profile and location of the potential customers. This is essential for successful fleet sizing and demand categorization. Recent studies have highlighted various approaches to tackle these challenges. For instance, Hu et al. (2024) explored how information and communication technology (ICT) impacts the micro-location choices of stores in urban areas, emphasizing the role of digital platforms in optimizing food delivery operations in densely populated districts. Using real data and machine learning methods, the work found the importance of considering location and traffic patterns when designing efficient delivery zones.


Similarly, Ko et al. (2020) proposed a collaboration model for service clustering in last-mile delivery, demonstrating that cooperative approaches can enhance service efficiency and reduce costs. This highlights the relevance of clustering methodologies for planning delivery zones, especially in scenarios with heterogeneous demand patterns.

Another perspective is provided by Ramírez-Villamil et al. (2022), who used clustering techniques to link clients to the satellites and improve last-mile parcel delivery. Their findings reveal that integrating data-driven clustering with logistics algorithms can significantly reduce operational costs and improve delivery times.

<sup>a</sup>  <https://orcid.org/0000-0002-7325-9350>

<sup>b</sup>  <https://orcid.org/0000-0001-5912-046X>

<sup>c</sup>  <https://orcid.org/0000-0001-9486-0906>

<sup>d</sup>  <https://orcid.org/0000-0002-7934-9250>

In another context, Regal et al. (2023) highlighted in their analysis how clustering can capture the diverse socio-economic and functional characteristics of urban regions, easing more tailored and efficient delivery strategies. This work aligns with the need for approaches that integrate socio-economic and geospatial data to address the complexities of urban logistics.

The analysis of Kang (2020) complements the discussion by examining the spatial evolution of warehouse and logistics center locations, emphasizing their tendency to move from urban centers to the periphery despite the significant growth of online shopping and the demand for instant delivery services.

Regarding machine learning algorithms, Sarkar (2024), and Wangwattanakool and Laesanklang (2024) explored customer segmentation and delivery zone partitioning, both using advanced clustering algorithms. Their research demonstrates how clustering can be leveraged not only to understand customer behavior but also to establish delivery zones. K-means algorithm is shown to be effective for these tasks. Besides, Dupas et al. (2024) propose a K-means clustering approach to allocate customers to depots and optimize vehicle routing, evaluating both the operational efficiency and the impact of last-mile depot locations.

Lastly, Zheng et al. (2023) applied fuzzy clustering analysis to optimize logistics distribution based on customer demand attributes.

Overall, these studies have demonstrated how clustering urban areas and clients is a meaningful method for a first-base idea for business sizing and logistics planning. Most of them focus on clustering customer demand based on geographical coordinates or on client profiles.

This paper proposes a clustering approach for food and parcel delivery services in Spain, that incorporates socio-economic indicators and urban network features, based on the amount of demand from specific small areas within a city. By analyzing the clustering differences and similarities in operation zones concerning the group of indicators used, our study provides insights into depot performance and decision-making in last-mile delivery planning.

## 2 DATA DRIVEN APPROACH

Building depot clustering algorithms requires knowing specific metrics of stores performance. Using delivery electric and combustion motorbike data from food and parcel delivery services, in a prior work the process to link that data to specific depots and extract performance metrics and statistics of each service was made (Arregui et al., 2024). The result was the de-

tection of many food and parcel delivery services in Spain, the motorbikes that operate within each service, geolocated trips, and delivery data of each motorbike. The data used in this work is from 2024 and overall, we have detected 854 parcel and 579 food delivery services.

Additionally, this data is enhanced with open data population indicators and geospatial information of the urban areas.

### 2.1 Depot's Performance Metrics

After the corresponding data cleaning and processing, the obtained daily performance metrics of each depot are the following:

- Number of bikes
- Time of use of each bike
- Distance covered by each bike
- Number of deliveries per km
- Consumption per km
- Average delivery radius
- Maximum delivery radius
- Average time per trip
- Number of deliveries per trip
- Total number of deliveries

With this data, we are capable of knowing the delivery demand in every area of a city and the fleet metrics, therefore we are able to use this information as a benchmark for future service planning. The depot clustering analysis increases the usability of this information, and it relies on socio-economic and network features.

### 2.2 Socio-Economic Data

In Spain, the National Statistics Institute (Instituto Nacional de Estadística, 2024) offers insights into many social, demographic, and economic indicators with a high granularity. These indicators show, for instance, inhabitants, genre, origin, educational level, working status, marital status, and housing. The data is updated to the year 2022. The full list of indicators is depicted in Table 1.

### 2.3 Network Features

The heterogeneity of urban areas can be captured using Open Street Map (OSM) data enhanced with elevation data. We can obtain interesting geospatial indicators to cluster the services based on metrics such as

Table 1: Socio-economic indicators published by the Statistics National Institute (INE) of Spain.

Code	Indicator	Code	Indicator
t1_1	Total people	t17_3	Percentage of widowed people
t2_1	Percentage of women	t17_4	Percentage of people with unknown marital status
t2_2	Percentage of men	t17_5	Percentage of people legally separated or divorced
t3_1	Average age	t18_1	Total dwellings
t4_1	Percentage of people under 16	t19_1	Primary dwellings
t4_2	Percentage of people aged 16 (inclusive) to 64 (inclusive)	t19_2	Non-Primary dwellings
t4_3	Percentage of people over 64	t20_1	Owner-occupied dwellings
t5_1	Percentage of foreigners	t20_2	Rented dwellings
t6_1	Percentage of people born abroad	t20_3	Dwellings under other tenure types
t7_1	Percentage of people pursuing higher education over population 16+	t21_1	Total households
t8_1	Percentage of people pursuing university education over population 16+	t22_1	Single-person households
t9_1	Percentage of people with higher education over population 16+	t22_2	Two-person households
t10_1	Percentage of unemployed people over active population	t22_3	Three-person households
t11_1	Percentage of employed people over population 16+	t22_4	Four-person households
t12_1	Percentage of active population over population 16+	t22_5	Five-or-more-person households
t13_1	Percentage of disability pensioners over population 16+	r1	Average net income per person
t14_1	Percentage of retirement pensioners over population 16+	r2	Average net income per household
t15_1	Percentage of people in other inactive situations over population 16+	r3	Average income per unit of consumption
t16_1	Percentage of students over population 16+	r4	Median income per unit of consumption
t17_1	Percentage of single people	r5	Average gross income per person
t17_2	Percentage of married people	r6	Average gross income per household

area, elevation, road speed, slope, etc. After a process to extract the metrics, these are depicted in Table 2.

### 3 METHODOLOGY

#### 3.1 Weighting the Variables

Socio-economic and network features are obtained at the census section level. A census section is the smallest administrative unit used for statistical purposes in Spain. It is defined by the INE and typically corresponds to a neighborhood or a similar small geographic area within a municipality.

For each depot, this data is aggregated from all the individual census sections where deliveries are made. However, the number of deliveries in each census section can vary a lot. For instance, we can have a big census section with just a few deliveries and a small one with many deliveries. This particular case makes the smallest census section's indicators and metrics more meaningful for that service than the ones of the bigger area. Therefore, it is necessary to weigh every

socio-economic and geospatial indicator according to the number of deliveries made in each area.

Clustering analysis is carried out separately for parcel and food delivery. This separation attends to the performance, demand, and delivery differences between these two service types. For each of them, socio-economic and network indicators from all depots are analyzed. Although some clustering methods, such as hierarchical clustering or density-based methods could be used, because of the regular distribution of data, the ease of finding an optimal number of clusters, and results interpretability, K-means algorithm, a widely used unsupervised machine learning algorithm was chosen. It divides the data by a pre-defined number of clusters, where each data point belongs to the cluster with the nearest mean, minimizing the variance within clusters. It has proven to be a good method for customer segmentation among other applications.

Once the clustering is made, a classification dataset is created with the 854 parcel delivery services and their corresponding socio-economic and network features. The clustering group is added to the dataset as the target variable. The same is done with the 579

Table 2: Network features.

Code	Description
surface_m2	Average area of the census sections
way_distance_meters	Average road distance within the census sections
num_of_nodes	Average number of nodes in the census sections
avg_speed	Average maximum speed of roads in the census sections
max_max_speed	Average of the maximum speed of roads in the census sections
min_max_speed	Average of the minimum speed of roads in the census sections
speed_percentil_10	10th percentile of the average maximum speeds in the census sections
speed_percentil_90	90th percentile of the average maximum speeds in the census sections
avg_elev	Average of the average elevation in the census sections
max_elev	Average of the maximum elevations in the census sections
min_elev	Average of the minimum elevations in the census sections
elev_percentil_10	10th percentile of the average elevations in the census sections
elev_percentil_90	90th percentile of the average elevations in the census sections
avg_slope	Average of the average slope in the census sections

food delivery services. These datasets are used to create classification models with random forest machine learning algorithms, and these models can serve as a tool for decision-making.

### 3.2 Parcel Delivery Clustering

K-means algorithm needs the optimal number of clusters to be predefined. For choosing the best number in each iteration, we use 3 different methods: K-means Inertia, GMM (Gaussian Mixture Model) BIC, and GMM AIC methods. The optimal number depends on the data used, therefore we have different numbers of optimal clusters when using socio-economic indicators or network features. These numbers are:

- For socio-economic indicators: 6 clusters
- For network features: 4 clusters

The generated clusters distribution is depicted in Figure 1.

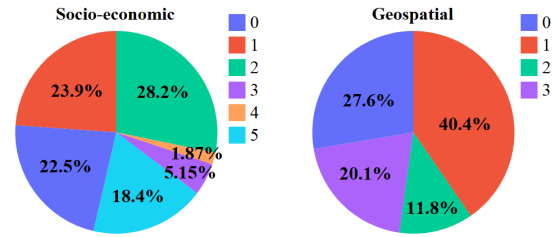


Figure 1: Parcel socio-economic vs network features cluster distribution.

An overview of the indicator's impact on forming each cluster is shown in Figure 2. We can observe for example the variables that do not have almost an impact or the ones that have an impact in more than one cluster. Focusing on geospatial clustering, elevation metrics are meaningful to form two of the clusters.

With all the depots linked to a cluster, we have created random forest classification models to predict new depots. The models show an accuracy of 90% for the socio-economic analysis and 93% for the geospatial analysis.

To better understand the model performance and have more information for future decision-making, we have looked into the model explainability through a feature importance method. SHAP (SHapley Additive exPlanations) values allow understanding a machine learning model prediction by assigning each feature a contribution to the output. It shows which indicators have a bigger impact on each class prediction, Figure 3. In the case of socioeconomic indicators, apart from the salary variables, t6\_1, t20\_2, and t4\_3 are the ones with a higher importance. These are related to age, origin, and housing. Although there are other indicators like t12\_1 and t11\_1, related to employment, that have a considerable impact in some specific clusters. For the network features, we appreciate that speed\_percentil\_10 is taken into account for two clusters followed by avg\_elev in four clusters.

### 3.3 Food Delivery Clustering

Food delivery services work on a different basis than parcel ones. For instance, in every food delivery trip, the rider usually serves a few customers and then returns to the depot. Therefore, we have different performance metrics and a separate clustering study. The optimal number of cluster groups, using the same methods as in parcel analysis, are:

- For socio-economic indicators: 5 clusters
- For network features: 6 clusters

The generated clusters and their distribution are depicted in Figure 4. Unlike for socio-economic clus-

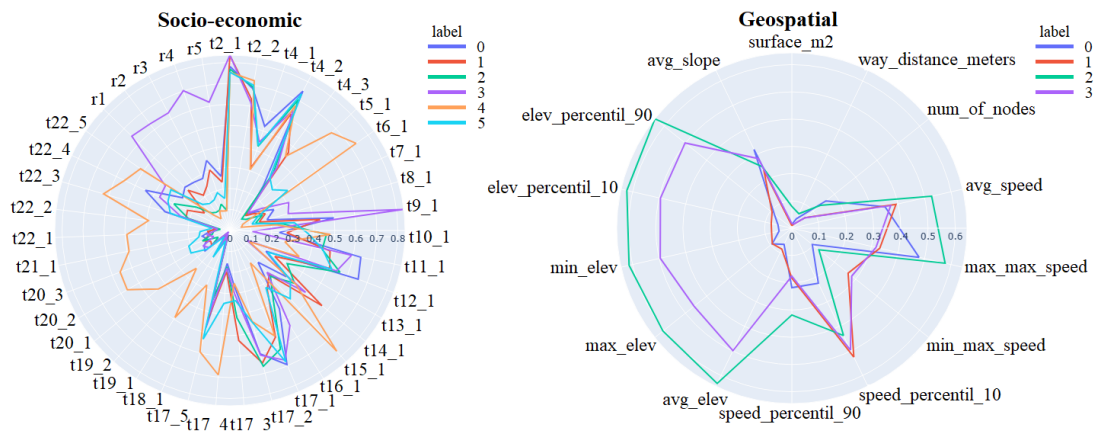


Figure 2: Parcel socio-economic vs network variables in clustering.

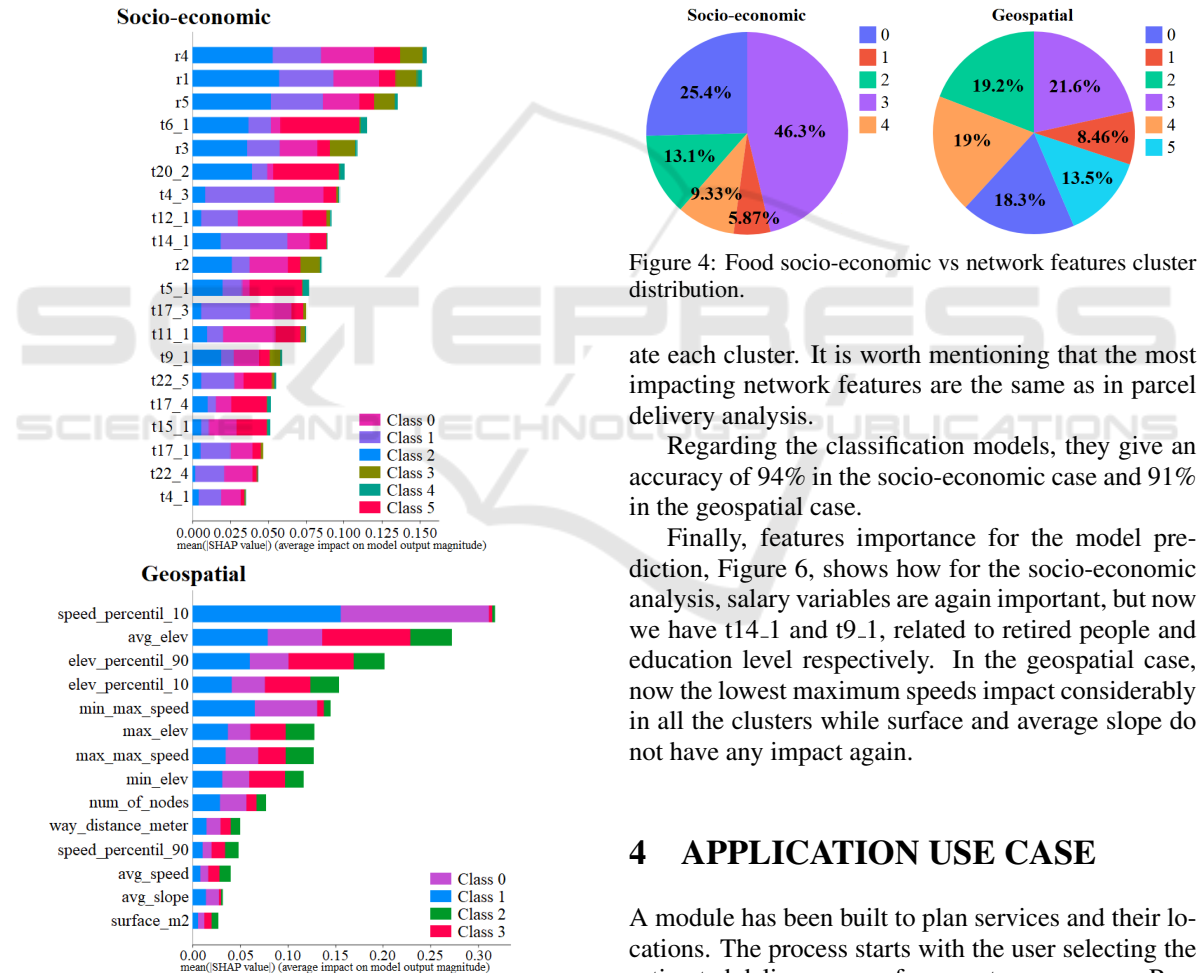


Figure 3: Parcel socio-economic vs network features importance.

ters, we see a balanced distribution for those with the network features.

Figure 5 shows the impact of the indicators to cre-

Figure 4: Food socio-economic vs network features cluster distribution.

ate each cluster. It is worth mentioning that the most impacting network features are the same as in parcel delivery analysis.

Regarding the classification models, they give an accuracy of 94% in the socio-economic case and 91% in the geospatial case.

Finally, features importance for the model prediction, Figure 6, shows how for the socio-economic analysis, salary variables are again important, but now we have t14\_1 and t9\_1, related to retired people and education level respectively. In the geospatial case, now the lowest maximum speeds impact considerably in all the clusters while surface and average slope do not have any impact again.

#### 4 APPLICATION USE CASE

A module has been built to plan services and their locations. The process starts with the user selecting the estimated delivery area of a new store on a map. Possible depot locations can also be selected, for which accessibility and centrality road network metrics are obtained. These metrics help decide the best possible location for the depot.

The socio-economic and network features of the



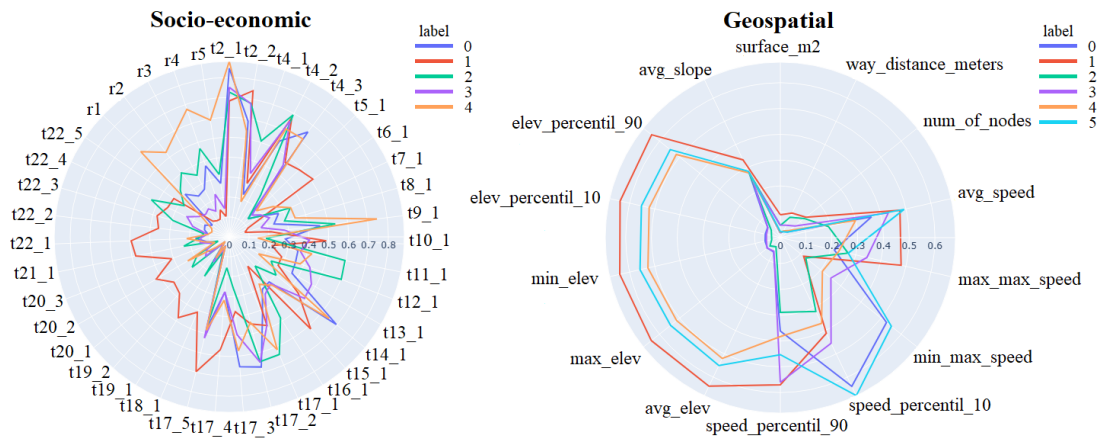


Figure 5: Food socio-economic vs network variables in clustering.

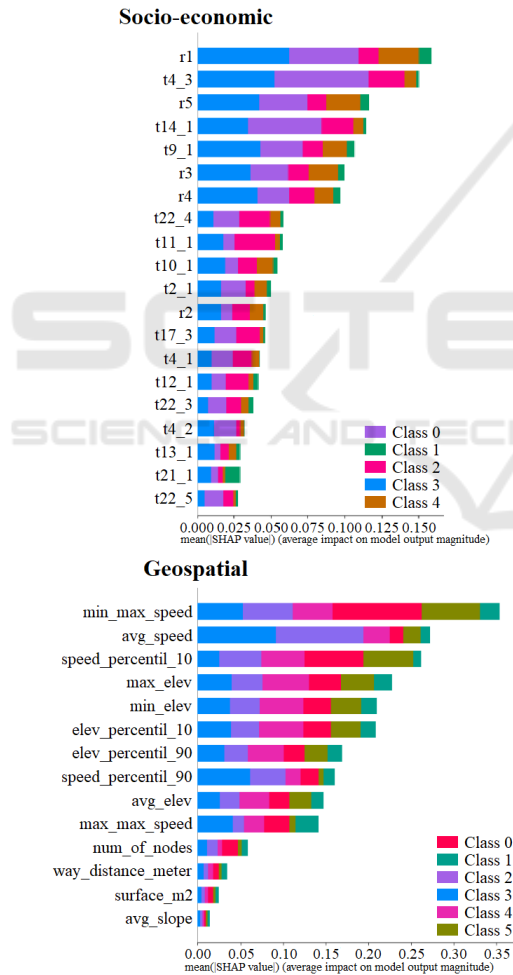


Figure 6: Food socio-economic vs network features importance.

census sections within the designed delivery area are aggregated and serve as input for the classification model. The model classifies the new depot within a

group, and every depot's performance metrics of that cluster are shown. Then, filtering can be made to keep only the most similar services according to some specific metrics, such as depots with a similar delivery area extension. Additionally, the average demand and fleet performance indicators of the filtered group are shown.

With this application, when a company designs a new service delivery area, it will be classified into a group of services with similar socio-economic and geographic conditions. These similar services show their fleet characteristics, performance, and delivery metrics. Therefore, this information can be used as a benchmark for decision-making regarding the sizing needs of the new service.

## 5 CONCLUSIONS

In this work, we have followed a methodology to cluster food and parcel delivery services from delivery motorbikes data, based on socio-economic indicators and geospatial metrics of the census sections where deliveries are made. The clustering results show patterns to classify these services based on how the inhabitants are or where they live. Additionally, the classification models show high accuracy and serve as a tool to obtain insights into the most meaningful variables and similarities of services already tested and working at the moment. Although the study has been done in the context of Spain, the same methodology can be followed in other places where service performance metrics and socio-economic or urban characteristics data could be obtained.

Opening a commerce poses several challenges and uncertainty regarding the social scenario and delivery fleet needs. Depot's performance metrics such as number of motorbikes, average delivery radius, con-

sumption per km, or total number of deliveries are crucial indicators to ease this uncertainty.

This study can be used as a benchmark for store owners to plan and size new stores, their location, and delivery logistics.

mization: Fuzzy clustering analysis of e-commerce customers' demands. *Computers in Industry*, 151:103960.

## REFERENCES

- Arregui, H., Cejudo, I., Arandia, I., Mujika, A., Eider, I., Laura, R., and Estibaliz, L. (2024). Last-mile delivery through electric motorbikes: Modelling considerations for parcel vs. food delivery. In *10th Conference of Transport Research Arena, TRA 2024*. In press.
- Dupas, R., Hsu, T., and Taniguchi, E. (2024). A clustering-routing approach for assigning customers to depots in last mile delivery. *Transportation Research Procedia*, 79:13–20.
- Hu, X., Zhang, G., Shi, Y., and Yu, P. (2024). How information and communications technology affects the micro-location choices of stores on on-demand food delivery platforms: Evidence from xinjiekou's central business district in nanjing. *ISPRS International Journal of Geo-Information*, 13(2).
- Instituto Nacional de Estadística (2024). Instituto Nacional de Estadística. <https://www.ine.es/>. Retrieved February 5, 2024.
- Kang, S. (2020). Relative logistics sprawl: Measuring changes in the relative distribution from warehouses to logistics businesses and the general population. *Journal of Transport Geography*, 83:102636.
- Ko, S. Y., Sari, R. P., Makhmudov, M., and Ko, C. S. (2020). Collaboration model for service clustering in last-mile delivery. *Sustainability*, 12(14).
- Ramírez-Villamil, A., Montoya-Torres, J. R., Jaegler, A., Cuevas-Torres, J. M., Cortés-Murcia, D. L., and Guerrero, W. J. (2022). Integrating clustering methodologies and routing optimization algorithms for last-mile parcel delivery. In de Armas, J., Ramalhinho, H., and Voß, S., editors, *Computational Logistics*, pages 275–287, Cham. Springer International Publishing.
- Regal, A., Gonzalez-Feliu, J., and Rodriguez, M. (2023). A spatio-functional logistics profile clustering analysis method for metropolitan areas. *Transportation Research Part E: Logistics and Transportation Review*, 179:103312.
- Sarkar, M., P. A. R. . C. F. R. (2024). Optimizing marketing strategies with rfm method and k-means clustering-based ai customer segmentation analysis. *Journal of Business and Management Studies*, 6(2):54–60.
- Wangwattanakool, J. and Laesanklang, W. (2024). Delivery zones partitioning considering workload balance using clustering algorithm. In *14th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, SIMULTECH 2024*, pages 378–385. Science and Technology Publications, Lda.
- Zheng, K., Huo, X., Jasimuddin, S., Zhang, J. Z., and Battaia, O. (2023). Logistics distribution opti-