Université de Lille

MASTER THESIS

Chaire LOGISTICS CITY

LVMT Laboratoire Ville Mobilité Transport

Université Gustave Eiffel

Mapping Logistics Warehouses and Assessing their Socioeconomic Impacts in France with a Focus on E-commerce Activities

Master Thesis for the University of Lille
Prepared at Laboratoire Ville Mobilité Transport (LVMT) - ENPC ParisTech

École polytechnique universitaire de Lille
Master Speciality: Urban Engineering and Habitat

Thesis presented and defended at Villeneuve d'Ascq, 09th September 2024, by

MOHAMMED YOUNES

Under the supervision of:

Laetitia DABLANC Logistics City Chair Director, Gustave Eiffel University (LVMT)	Academic Supervisor
Marion ALBERTELLI Postdoctoral Researcher, Gustave Eiffel University (LVMT)	Academic Supervisor
Matthieu SCHORUNG Lecturer, Sorbonne University (Médiations)	Academic Supervisor
Ammar ALJER Associate Professor, University of Lille (LGCgE)	Academic Supervisor

Résumé

Cette thèse de Master explore les méthodes spatiales et techniques utilisées pour identifier avec précision les emplacements des entrepôts en France, ainsi que les facteurs influençant la localisation des entrepôts d'e-commerce. Elle se concentre également sur l'étude et la compréhension des relations complexes entre l'urbanisation, les infrastructures et les conditions économiques. De plus, cette étude examine l'impact de différentes variables socio-économiques sur la densité des entrepôts d'e-commerce à travers l'utilisation de méthodes statistiques avancées, telles que la régression linéaire multiple (MLR) et l'analyse en composantes principales (ACP).

Un ensemble de données complet sur les emplacements d'entrepôts, incluant des données collectées manuellement ainsi que des sources gouvernementales telles que les bases de données SIRENE et SITADEL, a été développé. Les données ont été analysées à l'échelle géographique des "*Aires d'Attraction des Villes*" (AAV), qui offre un cadre plus fiable pour l'analyse régionale que les unités administratives plus petites, telles que les communes. L'étude met en évidence l'importance du développement urbain, des infrastructures routières et des caractéristiques socio-économiques dans la répartition spatiale des entrepôts, avec des implications pour l'urbanisme, la logistique et l'e-commerce.

Les résultats montrent que l'urbanisation et les réseaux de transport, en particulier les infrastructures routières, sont des déterminants clés de la localisation des entrepôts. Les facteurs socio-économiques, tels que les revenus et l'emploi, jouent également un rôle important, avec une concentration plus fréquente des entrepôts dans les régions industrielles à faible revenu. Les conclusions de cette étude contribuent à une meilleure compréhension de la logistique de l'e-commerce et offrent des perspectives précieuses pour améliorer l'analyse de la localisation des entrepôts à travers l'intégration de données en temps réel et de techniques économétriques spatiales avancées.

Abstract

This thesis explores the spatial and technical methods used to identify the accurate locations of the warehouses in France, alongside the factors influencing the location selection of e-commerce warehouses. It also focuses on studying and understanding the complex relationships between urbanization, infrastructure, and economic conditions. Moreover, this study investigates the impact of different socioeconomic variables on the density of e-commerce warehouses. Through the use of a combination of advanced statistical methods, such as Multiple Linear Regression (MLR) and Principal Component Analysis (PCA).

A comprehensive dataset of warehouse locations, including both manually collected data and government sources such as the SIRENE and SITADEL databases, was developed. The data were analyzed at the geographic scale of the ”*Aires d'Attraction des Villes*” (AAV), which offers a more reliable framework for regional analysis than smaller administrative units like communes. The study highlights the significance of urban development, road infrastructure, and socioeconomic characteristics in driving the spatial distribution of warehouses, with implications for urban planning, logistics, and e-commerce.

The results show that urbanization and transport networks, particularly road infrastructure, are key determinants of warehouse location selection. Socioeconomic factors, such as income and employment, also play an important role, with warehouses more frequently located in lower-income, industrial regions. The findings contribute to the broader understanding of e-commerce logistics and offer valuable insights for improving warehouse location analysis through the integration of real-time data and advanced spatial econometric techniques.

Contents

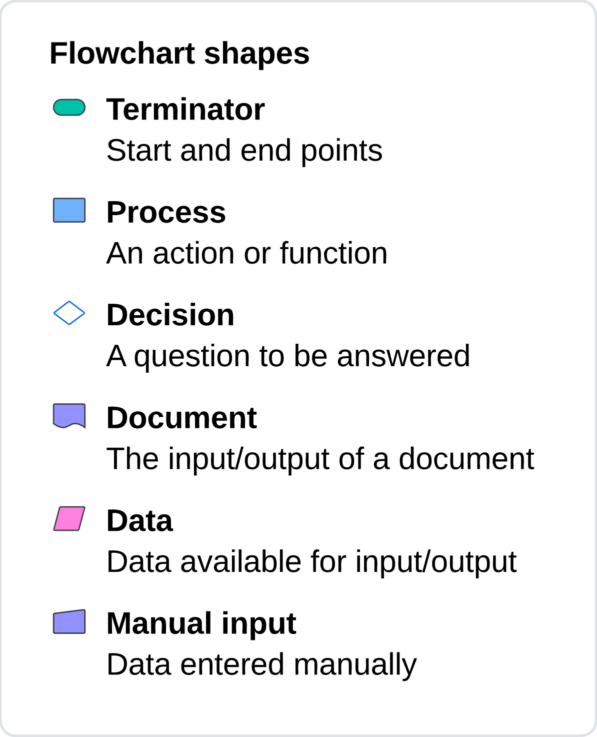
List of Figures

List of Tables

Acronyms and Abbreviations

- WH: Warehouse.
- GWP: Global Warming Potential, *”Pouvoir de Réchauffement Global (PRG)”*.
- DVF: Demande de Valeurs Foncières.
- NAF: Nomenclature d’activités française.
- INSEE: Institut national de la statistique et des études économiques.
- SITADEL: Base des permis de construire et autres autorisations d’urbanisme.
- SIRENE: Système national d’identification et du répertoire des entreprises et de leurs établissements.

Flowchart Shapes Key



Flowchart shapes

⬭ **Terminator**
Start and end points

▭ **Process**
An action or function

◇ **Decision**
A question to be answered

▭ **Document**
The input/output of a document

▱ **Data**
Data available for input/output

▭ **Manual input**
Data entered manually

Chapter 1

Introduction

1.1 Background

The COVID-19 pandemic revealed the importance of logistics in the operation, organization and supply of companies, whether in urban centers or on a regional, national or even international scale. However, warehouse construction projects are increasingly being challenged, as in the case of the Green Dock project in the port of Gennevilliers, where 90,000m^2 of logistics facilities is planned (A. Russo 2024), the project is criticized for the traffic and pollution generated by the warehouse, as well as for the environmental impact caused by the construction of this new building (Collective of Associations 2022). The question of the location and siting of warehouses has thus become a major issue in a context of scarcity of available land in dense areas (R. d. Oliveira et al. 2022), leading to logistics sprawl (Dablanc, Ogilvie, et al. 2014). The lack of development regulation on the periphery has led to an increasingly peri-urban geography for warehouses (Cidell 2010; Bowen 2008).

With the arrival of e-commerce in the late 1990s, the role of warehouses has expanded, especially in the urban and peri-urban areas, which led to significant changes in urban landscapes, putting more stress on the demand for new infrastructure, which eventually affected the transformation of the socioeconomic fabric. The COVID-19 pandemic caused momentous increase in the demand for home delivery products, mainly through e-commerce. This sudden increase in demand accelerated the establishment of several logistics facilities in the urban centers and their periphery areas.

The location of logistic facilities has profound implications for urban areas, causing significant transformation in employment, local economies, and the environment. The spatial distribution of these facilities is driven by various factors, including accessibility, land availability, and proximity to markets, which in turn influence socioeconomic outcomes.

Various studies have extensively explored different aspects of the location selection of logistics facilities and their socioeconomic impacts. However, fewer studies have focused specifically on the location selection of the e-commerce warehouses in the urban and peri-urban areas, particularly in the context of their socioeconomic impacts. A study by Xiao et al. 2021 highlights how the rise of e-commerce has led to a significant restructuring of logistics spaces, especially in rapidly urbanizing cities. This reorganization has been driven by the unique demands of e-commerce, such as the need for faster delivery times and more efficient land use. However, the existing studies, focus primarily on developments on freight activities in the urban areas, often overlook the special case of e-commerce activities in terms of socioeconomic impacts as a whole.

1.2 Structure of the report

The thesis consists of five chapters, each addressing different aspects of the research on e-commerce warehouses and spatial analysis. Chapter 1 now serves as the literature review, providing an in-depth analysis of existing methods, socioeconomic indicators, and spatial factors influencing warehouse location selection. This chapter consolidates previous academic work, theoretical frameworks, and practical approaches relevant to the study. Chapter 2 outlines the methodological approaches, describing the key principles and data collection techniques utilised for the spatial analysis of warehouse locations and introducing different approaches to identify exact warehouse locations. In Chapter 3, the socioeconomic, environmental, and spatial indicators influencing warehouse location selection are investigated through detailed statistical analysis, including regression models and principal component analysis. Chapter 4 focuses on enhancing the visualisation of the chair's data, presenting the results of the internship in terms of developing tools for better data representation and decision-making support. Finally, Chapter 5 presents the conclusions, reflecting on the findings of the research and the personal impact of the internship experience.

1.3 The research environment

The main work of the chair team is focused on various aspects of urban logistics, including different research theme, The Logistics City Chair is structured into three main scientific axes :

1. Urban and suburban logistics real estate;

2. Trends and new consumer practices and their impact on urban logistics and its real estate;

3. Public policies for urban logistics.

This internship is part of the research efforts on warehousing and e-commerce activities (themes 1.1 and 2 of the Chair).

The research topic is based on the chair's work on logistics activities in different metropolitan areas around the world (Dablanc, Schorung, et al. 2023), where the focus was on all warehouses, and on the US and specifically e-commerce warehouses (Schorung et al. 2022). It included "logistics sprawl" studies and some socioeconomic analysis as well as spatial analyses of logistics real estate in the major metropolitan areas in the US, Europe, and some other cities in South America and Japan (Dablanc, Schorung, et al. 2023).

The main theme of this research falls under Theme 1.1, which specializes in applying macro-analyses of the spatial distribution of warehouses.

Laetitia Dablanc initiated the research and was the general supervisor. The main supervisor of the research was Marion Albertelli, in guiding the research's direction and closely monitoring its progress and findings, Matthieu Schorung (Associate researcher at LVMT and an Associate Professor at Sorbonne University in Geography) reviewed the previous methods used by the chair and the methodological work, Abel Kebede REDA (Postdoctoral researcher at LVMT) provided help in the statistical analysis development.

Thanks to Dr. Albertelli, I joined The ElementR Group, within the UMR Géographie-cités, the UMR PRODIG, and the UAR RIATE : the group is a self-training group of researchers from different fields, focusing on organizing seminars and practical workshops on R programming language, demonstrating statistical applications, which helped improving my skills in R. Simultaneously, I completed a series of online courses, concentrating on R, statistics, SQL, Data Governance, and research methodologies.

In May, I attended a research conference titled *"Digitalisation et Décarbonation des Mobilités"*, where various research projects in the field were presented, and round tables were held. I benefited from discussions with several researchers about the methodological part of my research on the data sources and approaches they used to identify warehouses. Their opinions were useful in criticizing and identifying the key issues with common approaches and how I would develop the research.

During the same month, I accompanied the chair's team on a field visit to Metz, where we visited two e-commerce warehouses run by Amazon. This visit provided me with a first-hand look at the inner workings of warehouse operations and the socioeconomic landscape of the area, which I later used as a reference point, a reliable example to verify the validity of data sources and methodological approaches. I gained detailed information from the warehouse managers, which I subsequently used to assess the credibility of any data source or technique I employed.

Since the duration of the internship is too short to create a complete comprehensive approach that is applicable to different countries, after discussion with my supervisors in the chair, we decided to focus the scope of the internship research on France, to concentrate more on the local context of the country in regards to the data sources, spatial data integrity, and the connectivity to the surrounding countries logistics activities.

The main contribution of the internship is methodological, and the provision of a focus on e-commerce warehousing activities in France. In terms of the detailed goals of the internship, they were as follows:

- Identifying potential statistical methods and visualization techniques that could be used in the socioeconomic studies of the warehousing activities.

- Providing suggestions for the development of the website for the chair, enabling easy access and navigation to the different research products of the chair.

- Identifying gaps in methodologies and data used in the studies of logistics infrastructure from the urban scale.

- Introducing different approaches in identifying the locations of the warehouses in France and criticizing each approach in terms of advantages and disadvantages, with a focus on e-commerce warehouses.

- Applying statistical analyses on the collected data using the appropriate approaches.

Chapter 2

Literature Review on Warehousing in France

2.1 Introduction

In the last three decades, the logistics industry has been subject to considerable change, most significantly through the growth of e-commerce and the related demand for efficient and well-located warehousing facilities. Understanding the spatial pattern of logistics and warehousing activities is a crucial input into urban planning, transport infrastructure development, and socio-economic analysis. However, determining the exact locations of logistics facilities remains problematic due to a lack of comprehensive and accurate data. The complexity of spatial patterns, coupled with the issue of data availability, makes understanding the impact of warehousing on both urban and peri-urban environments challenging.

This research aims to contribute to the existing literature by addressing the limitations of prior studies focusing on identifying logistics facilities across Europe, with particular emphasis on France. The study systematically reviews methods for locating logistics facilities, assesses their applicability, and identifies challenges faced by researchers. The correct location of logistics facilities is essential not only for urban planners, policymakers, and logistics managers but also for integrating logistics effectively within contemporary urban life.

The impetus for this research arises from the need to better understand logistics sprawl—the decentralisation of warehousing facilities from urban cores to peri-urban or rural areas—which has significant implications for urban infrastructure and sustainability. Logistics sprawl impacts land use, traffic congestion, and environmental degradation due to longer distances covered by freight vehicles. Therefore, understanding the exact location and distribution of these facilities is vital for developing more sustainable urban planning policies.

In France, the rise of e-commerce over the past two decades has further highlighted these challenges. The growing demand for rapid last-mile deliveries has led to increasingly specialised and dynamic warehouses. However, existing public datasets often fail to capture the nuances of these emerging logistics facilities, particularly smaller distribution centres for last-mile delivery. This research aims to critically evaluate these methods within the French context by reviewing the literature on warehouse locating methods, while discussing the strengths and limitations.

Logistics facilities impact urban areas not only spatially but also socioeconomically and environmentally. Warehouses located in low-income neighbourhoods can provide economic opportunities but also generate negative externalities like pollution and increased traffic. Thus, this in-depth review of methodologies to identify logistics facilities helps improve our understanding of logistics activities and provides insights into their social and environmental dimensions.

Consequently, this study addresses the following key questions: What are the existing methods used to locate logistics facilities in Europe, particularly in France? How effective are these methods in capturing the spatial dynamics of logistics activities? What challenges remain in accurately identifying warehouse locations, and how can these be addressed in future research? The study aims to understand the gap between data availability and real-world logistics challenges, enhancing urban planning processes.

2.2 The Rationale Behind Reviewing Warehouse Location Identification Methods

The identification of logistics facilities, particularly warehouses, has long relied on specific methods that have remained largely unchanged for decades. The methods typically involve the use of the manually and exhaustively collected surveys, and publicly available data sources, such as administrative boundaries and geocoded datasets when available, with a distinguishable use of indirect approximations like barycentres or postal code-based analyses (Dablanc and Rakotonarivo 2010). Despite improvements in the availability and sophistication of data collection methods in the ICT (Information and Communication Technology) fields, as well as the enhanced policies across the governments and logistics companies, the common methods employed in logistics location studies still fall short in accuracy, especially in contexts that require a high level of spatial specificity (Heitz, Launay, et al. 2017). This persistent use of outdated methodologies, while not merely an issue of method choice, highlights a broader methodological gap in the urban planning and logistics city research communities, one that this review seeks to address.

Existing methods for identifying warehouse locations have proven effective to some degree, primarily because they utilise available data sources such as SIRENE, OpenStreetMap (OSM), or other national statistical datasets. However, these approaches are constrained by the quality, granularity, and specificity of the available data (Pajić et al. 2024). The inadequacy of current methods often results from their inability to precisely identify the locations of warehouses, which have unique spatial characteristics and locational drivers compared to traditional logistics facilities (Cidell 2010). Given the increasing complexity of logistics networks and the rise of e-commerce, the shortcomings of these common methods become more pronounced.

The main issue with the prevalent approaches to identifying warehouse locations is that they tend to over-rely on static and administrative data sources, without considering the dynamism of modern logistics. From a philosophical standpoint, the use of static geographies fails to capture the fluid and ever-changing nature of contemporary supply chain dynamics, especially in the complicated socioeconomic context of the urban centres. Furthermore, logistics, by its very nature, is about movement and flow of different aspects, both of which are inherently dynamic. Yet, the static spatial units used in traditional methods do not accurately reflect these temporal and spatial variations (Cresswell 2006). In this context, Henri Lefebvre's Production of Space theory is particularly relevant, as it argues that space is not merely a passive container but is actively produced and reproduced by social relations and economic activities (Lefebvre 1991). The common warehouse identification methods overlook this active production when dealing with specific points over the years, treating space as an inert object rather than as a socially and economically constructed reality and fabric.

From a technical perspective, the issue also lies in the way spatial data is aggregated and represented. Administrative units like ZIP codes or regions are used for convenience and data availability, including the other data that of indicators that could be used in the analysis of the logistic activities dynamics, but these administrative units lack precision when it comes to the fine-scale localisation of logistics activities (Dablanc 2014). The aggregation of data based on these large spatial units masks micro-level spatial variations that are critical to understanding the nuanced locational factors influencing warehouses. Additionally, there is a critical bias inherent in using these administrative units, as the choice of unit impacts the results and interpretations of spatial analyses. This leads to the well-documented ecological fallacy, wherein assumptions about individual-level behaviour are inferred based on group-level data (Openshaw 1983).

From a critical theoretical perspective, the logistics industry's impact on urban environments is under-addressed due to these methodological shortcomings. David Harvey's concept of spatial fix suggests that logistics infrastructure is often placed in a way that maximises capital accumulation but often at the cost of social and environmental equity (Harvey 2001). Hence, the reliance on traditional, aggregated spatial units tends to obscure the impacts of logistics on urban communities, particularly those that bear the brunt of negative externalities like pollution and traffic. In this context, the lack of precise localisation methods therefore not only represents a technical gap but also a social justice gap, as the true socio-environmental costs of logistics facilities are not fully captured or understood by the stakeholders.

2.3 The Modifiable Areal Unit Problem (MAUP) in the Context of Logistics Facility Identification

One of the major theoretical issues in finding warehouse locations by traditional approaches is MAUP. First coined by Gehlke et al. (1934), and later popularised by Openshaw (1983), it is a kind of statistical error that occurs when all spatial information is aggregated into different areal units, and the result may provide conflicting analytical outcomes for each unit. In the context of warehouse identification, *MAUP* introduces two primary forms of bias: aggregation bias and zoning bias.

Aggregation bias is the result of differential spatial aggregation, yielding different interpretations of the patterns under examination. For instance, a logistics facility located within a wider industrial zone could have markedly different locational attributes depending on the scale of analysis used, which can distort the representation of reality (Kwan 2012).

This issue becomes particularly problematic when these warehouses are located on the outskirts of administrative regions; their influence may be overestimated in one region and, at the same time, be underestimated in another (Rodrigue 2024). This creates significant confusion, especially in urban areas where land uses and economic activities can vary significantly with little distance (Batty 2005). Zoning bias, on the other hand, is due to the arbitrary ways in which spatial divisions are made (Openshaw 1983). In drawing the administrative boundaries, there was no logistical analysis in mind; it was purely influenced by historical reasons, political ones, or administrative reasons. Therefore, warehouses falling into the fringes of such boundaries may be poorly represented within the datasets used, leading to spatial inaccuracies of logistics analysis (Openshaw 1983).

These biases complicate decision-making for urban planners and policymakers who seek to understand and regulate the distribution of logistics activities. A key consequence of MAUP in this context is that it limits our ability to design exact interventions (Jacobs-Crisioni et al. 2014), this limitation aimed at addressing the potential negative externalities created by logistics operations. For instance, when the detailed spatial properties of the warehouses are masked by aggregation and zoning effects, interventions that may be designed to reduce pollution or lower traffic congestion could be poorly directed or not have the desired effects whatsoever.

This therefore calls for the need for more sensitive and contextual approaches that will reduce MAUP impacts and show the actual spatial distribution of logistics facilities. These traditional methods of identifying the locations of warehouses have been the standard because they are based on data that is readily available. However, they are fraught with conceptual and technical limitations regarding dated views of space, reliance on static administrative units, and susceptibility to MAUP-related biases.

This means that in order to enhance our understanding of the spatial aspects of logistics, there is a need to explore innovative techniques that surpass conventional methodologies by embracing more dynamic, accurate and refined spatial data.

2.4 Common Methods Used in Identifying Warehouse Locations

The identification of logistics facilities, including warehouses, has become a focal point for research exploring urban logistics and the spatial organisation of logistics activities. Following the recent literature, methods for the identification of logistics facilities are often different depending on the particular research questions, data available, and scale of the study (Heitz, Launay, et al. 2017; Dablanc 2014). The methodologies employed for locating warehouses are typically embedded in the intersection of urban planning, logistics geography, and data science. The following section outlines the most common methods adopted in identifying warehouse locations across Europe, with a focus on France, by analysing the theoretical and empirical literature in logistics studies.

2.4.1 Barycentre Analysis

Barycentre analysis is one of the most widely used methods in spatial logistics research to determine a central position that best depicts the collective logistics facilities' location. In essence, this methodology approximates the mean geographic centre of logistics operations within a given area (Dablanc 2014). For instance, the barycentre of the locations of warehouses in a municipality or region would give its centroid, computed as the weighted average of the coordinates, where weights usually reflect some characteristics of each facility concerning size or capacity.

When the accurate information on the warehouse locations is not available, this method is often used, and sometimes, it is also used within the available data is accurate, providing a more accurate barycentre calculation. Hence, researchers operate with summarised or aggregated data, such as the number of logistics facilities located within administrative boundaries with the location being simply the centre of the administrative region if neither approximate location is provided (Heitz, P. Launay, et al. 2019). Barycentre analysis serves as a proxy, allowing researchers to conduct more comprehensive regional studies of logistics sprawl or to understand how warehouses are distributed across large areas (Allen et al. 2012). While effective in providing a regional overview, the barycentre approach fails to capture the detailed spatial dynamics of facilities at an individual level, making it more relevant for macro-level analysis than for location-specific studies (Cidell 2010).

2.4.2 Administrative Boundaries and ZIP Code Analysis

Another commonly employed methodology is the use of administrative boundaries, such as ZIP codes or municipalities (communes in the French context), to serve as proxies for warehouse locations (Dablanc and Rakotonarivo 2010). Scholars often depend on existing administrative datasets, correlating logistics operations with geographic units like NUTS regions (Nomenclature of Territorial Units for Statistics) or postal zones. This approach utilises data repositories such as SIRENE, a comprehensive business registry in France that structures enterprises according to their geographical location and activity type (Heitz, Launay, et al. 2017).

This method is particularly prevalent due to the wide availability of administrative data, which allows researchers to identify and quantify the spatial distribution of warehouses across wider geographic units. For instance, studies on logistics sprawl around Paris have applied ZIP code-based methods to investigate the dispersion of warehousing and distribution activities towards the periphery of the city centre (A. Russo 2024). However, using administrative units introduces complexities related to the modifiable areal unit problem (MAUP), which complicates accurate identification of warehouse locations.

2.4.3 Surveys and Field Studies

Surveys and field studies are one of the most common methods for identifying warehouses, particularly in research studies that require primary data collection. This approach is especially useful for collecting detailed information that may not be available from public databases, such as operational characteristics, facility capacity, or the level of technology adoption (Heitz, P. Launay, et al. 2019). In survey methods, researchers typically contact logistics operators, facility managers, or local government officials to request information on location and warehouse characteristics.

(Raimbault et al. 2012) employed surveys to identify logistics hotels in the Paris area, focusing on their spatial patterns and operational features. Fieldwork, on the other hand, involves physical site visits to validate the existence of potential warehouse locations and gather additional data unavailable through desk research (Heitz, P. Launay, et al. 2019). Both surveys and fieldwork are resource-intensive, making them more applicable to detailed, small-scale studies than to broad regional surveys (Allen et al. 2012).

2.4.4 Remote Sensing and GIS

Remote sensing and Geographic Information Systems (GIS) have become increasingly popular in logistics research as geospatial data and satellite imagery have become more accessible. Remote sensing allows researchers to identify warehouses by analysing land cover and land use patterns using high-resolution satellite images or aerial photography (Cidell 2010). GIS tools are then employed to map, analyse, and interpret the spatial distribution of logistics facilities.

In Europe, the integration of CORINE Land Cover data with various geospatial datasets is frequently employed to discern industrial and commercial entities, including warehouses (Institut Paris Région 2024). By cross-referencing land cover information with commercial databases, researchers can identify prospective logistics facilities and assess their spatial relationships with adjacent urban zones. GIS and remote sensing methodologies are particularly effective for large-scale studies that must account for spatial variability, such as investigating the environmental impacts of logistics sprawl (Aljohani et al. 2016).

2.4.5 Business Registers and Official Databases

The use of business registers and official databases is one of the most fundamental methods for identifying warehouse locations. In France, the SIRENE database maintained by INSEE (French National Institute of Statistics and Economic Studies) is widely used. Researchers can filter logistics-related businesses using NAF codes (French activity classification codes) to identify companies involved in transportation, warehousing, and distribution activities (Heitz, Launay, et al. 2017).

By extracting a list of logistics enterprises and cross-referencing them with administrative boundaries such as municipalities, states, and regions, researchers can create a spatial inventory of warehouse locations. This method has been particularly effective in analysing the spatial distribution of logistics activities in urban and peri-urban areas, such as the study of logistics sprawl in the Greater Paris area (Dablanc 2014). However, the accuracy of this method depends heavily on the completeness and currency of the data in business registers.

2.4.6 OpenStreetMap (OSM) and Crowd-Sourced Data

OpenStreetMap (OSM) and similar crowd-sourced platforms provide a valuable alternative for obtaining data on logistics facilities. OSM is an open-access platform that enables contributors to map the world's infrastructure, including roads, buildings, and logistics facilities. By applying tags such as ”warehouse” or ”industrial,” researchers can extract information on warehouses from OSM data (Schorung et al. 2022).

Studies aimed at identifying urban logistics facilities have turned to OSM when official data are unavailable or outdated, as in the case of (Allen et al. 2012). OSM offers detailed information that is often missing from official datasets, especially for smaller or less visible logistics facilities. However, the accuracy of OSM data can vary greatly depending on the contributors, leading to inconsistencies across regions.

2.4.7 Spatial Econometric and Machine Learning Approaches

Recent studies have incorporated spatial econometric models and machine learning algorithms to study different aspects in logistics, Reda et al. 2023 studied the effect of spatial characteristics of establishments on freight (trip) attraction models, mainly using spatial error models (SEM), spatial autoregressive model (SAR), geographically weighted regression (GWR), and multiscale-GWR (MGWR). Such econometric models can be used to predict potential warehouse locations based on logistic facility attributes, land values, accessibility to transportation infrastructure, and demographic data. These models help uncover spatial relationships that may not be evident using conventional mapping methods. The literature specifically focusing on using these methods to identify warehouse locations, however, remains limited.

Machine learning is also gaining significance in logistics research, particularly when working with large datasets (Waller et al. 2013). Clustering algorithms such as K-means can reveal patterns in logistics facility distribution that are not immediately obvious (Zhang et al. 2021). Although these methods are relatively new in logistics research, they offer the potential to address some of the biases associated with conventional approaches. Large datasets combined with machine learning algorithms enable researchers to develop predictive models that could influence logistics strategies and spatial governance (Talwar et al. 2021).

In conclusion, various methods exist for current warehouse locations in France and Europe based on the demand of each research being carried out. Approaches like barycentre analysis and ZIP code-based analysis are indeed popular for regional studies but highly imprecise for detailed locational insight. Surveys, GIS, remote sensing, and business registers are indeed more precise but have their limitations concerning data availability and costs. The fields of machine learning and econometric modelling signify progressive domains within logistics research, tackling several limitations that are characteristic of conventional methodologies, yet necessitating considerable computational capabilities and specialised knowledge. While each approach proves effective within its respective context, it underscores the complexities linked to the dynamic logistics environment and the necessity for more advanced and holistic identification methodologies.

2.5 Practical and Theoretical Issues with Common Methods

The process of identifying the locations of logistics facilities, such as warehouses, presents numerous practical and theoretical challenges that can influence the quality and relevance of research findings. Each of the methods discussed earlier has its own limitations due to methodological constraints and the inherent characteristics of the data employed. This section provides an academic discussion of the practical and theoretical issues associated with these methods, focusing on their advantages and disadvantages to clarify their respective merits and demerits.

2.5.1 Barycentre Analysis and Administrative Boundaries: Constraints of Spatial Resolution

Barycentre analysis, in conjunction with administrative boundaries such as ZIP codes or municipalities, can offer a useful method for estimating the locations of logistics facilities, particularly when detailed locational data is unavailable (Dablanc, Ogilvie, et al. 2014; Heitz, P. Launay, et al. 2019). However, the primary theoretical limitation of these methods lies in their inability to precisely identify facility locations due to their reliance on aggregate data. The modifiable areal unit problem (MAUP) significantly affects the accuracy of spatial analyses that depend on administrative boundaries, as aggregation can obscure the genuine spatial dynamics of logistics facilities (Openshaw 1983).

Utilising either the centroid method or administrative boundaries can result in misrepresenting the spatial relationships of warehouses located near administrative limits (Kwan 2012). Practically, these methods are often employed because of the ease of computation and the accessibility of administrative datasets. However, their lack of granularity poses challenges in evaluating micro-scale impacts, such as environmental and social consequences for local communities (Cidell 2010). This shortcoming is particularly detrimental to studies requiring accurate locational data for assessing logistics-related externalities, such as pollution and noise.

2.5.2 Surveys and Field Studies: Practical Constraints and Generalizability

Surveys and field studies offer a level of precision and detail that aggregated approaches cannot achieve, especially concerning the operational aspects of facilities and real-world verification (Heitz, P. Launay, et al. 2019). Despite their advantages, these methods face significant practical challenges. Surveys demand considerable resources, including time and labour for design and communication with facility operators, making them more suitable for smaller-scale investigations (Allen et al. 2012).

Although fieldwork provides an effective means of verifying warehouse locations, the geographical constraints limit the generalizability of such studies, as they are often confined to specific localities or cities. Furthermore, survey bias and high non-response rates can undermine the reliability of data, thereby reducing the robustness of the findings.

2.5.3 Availability of Data and Resolution Issues: Remote Sensing and GIS

The use of remote sensing and Geographic Information Systems (GIS) for identifying logistics facilities has gained momentum due to improved access to geospatial data and advances in satellite imaging technology. However, these methods are not without practical difficulties. One primary concern is the quality and resolution of satellite imagery. Obtaining high-resolution data can be costly and is not consistently available across all regions, limiting the use of remote sensing in large-scale research projects.

From a theoretical perspective, the analysis of satellite imagery for identifying logistics facilities is often subjective and requires significant expertise, as well as validation with ground data to ensure accuracy. Additionally, ambiguities in land-use categorisation within GIS may impede accurate differentiation between various industrial sites, which could lead to the misidentification of logistics warehouses.

2.5.4 Business Registers and OSM: Data Incompleteness and Quality Issues

Business registers, such as the SIRENE database in France, provide a structured approach to identifying warehouses through the application of NAF codes for classifying logistics activities (Heitz, Launay, et al. 2017). Nonetheless, the incompleteness and outdatedness of business registries pose significant challenges, as logistics facilities may undergo changes that are not promptly reflected in these databases. This is particularly relevant in the fast-evolving e-commerce sector, characterised by high facility turnover rates.

Similarly, OpenStreetMap (OSM) serves as a valuable tool for mapping logistics infrastructure, especially in regions lacking official data. However, inconsistencies and inaccuracies in OSM data, which is often contributed by non-professionals, may result in data quality issues (Allen et al. 2012). Urban areas generally have better coverage compared to rural regions, further contributing to data variability and potentially affecting the reliability of research findings based on OSM.

2.5.5 Econometric and Machine Learning Approaches: Data and Interpretability Challenges

Recent developments in econometrics and machine learning have introduced advanced strategies for in the context of logistics sites, though not directly on the identification of warehouses locations, but it used algorithms for predictive analysis of logistic attributes. Some studies demonstrated the application of machine learning regression techniques to select economic attributes indicative of logistics performance (Jomthanachai et al. 2022), while others explored scalable econometric methods on big data, specifically focusing on logistic regression (Ouattara et al. 2021).

Despite their potential, these methods face both practical and theoretical obstacles. Econometric and machine learning models require substantial amounts of data, often necessitating access to extensive high-quality datasets, which may not always be readily available. Additionally, implementing these models demands sophisticated computational resources and specialised expertise, which may be inaccessible to some researchers or stakeholders.

Theoretically, machine learning models, particularly neural networks, are often criticised for their lack of interpretability (Doshi-Velez et al. 2017). Unlike traditional statistical approaches that yield explicit relationships between variables, machine learning models can function as "black boxes" (Hastie et al. 2009), limiting the ability to understand the factors influencing logistics facility locations, which is a crucial consideration for urban logistics planning and policymaking.

2.5.6 Conclusion

The identification of warehouse locations in logistics research involves a variety of methodologies, each with distinct practical and theoretical strengths and weaknesses. Methods like barycentre analysis and administrative boundary analysis offer broad overviews but suffer from a lack of spatial precision, whereas surveys, field studies, and remote sensing provide more accuracy at the expense of scalability and resource demands. Business registers and OSM are valuable in certain contexts, though they are affected by data quality and completeness issues. Advanced methods, such as machine learning, offer promise but face limitations related to data availability, interpretability, and scalability. Addressing these challenges requires innovative approaches that balance accuracy, cost, and practicality, which will be explored further in the next section on gaps in the literature.

2.6 Academic Attempts to Introduce New Methods for Identifying Logistics Facility Locations

Over the past two decades, researchers have increasingly recognised the limitations of conventional methods for determining the locations of logistics facilities and have sought to introduce innovative techniques that address these deficiencies. The following sections outline some of the newer techniques proposed and advocated in logistics research, particularly in Europe and France, aimed at enhancing the precision, scalability, and applicability of spatial logistics analysis. A significant focus has been placed on advanced integrated spatial analysis, the utilisation of real-time data, and multimodal approaches.

2.6.1 Spatial Econometric Modelling and Multiscale Analysis

One of the more recent academic approaches to improving warehouse identification involves the combination of spatial econometric models with traditional data sources. These models help explain how logistics facilities relate to one another and to socio-economic, environmental, and infrastructural factors in the surrounding area. Spatial econometric modelling offers insights into warehouse dispersion by considering spatial spillover effects, which are often overlooked by conventional methods (Anselin 2003).

Spatial models have demonstrated considerable potential in real-world applications, especially in explaining the relationship between warehouse locations and regional socio-economic factors such as income levels, land prices, and employment rates. This applies to spatial lag models to highlight the spatial dynamics of neighbouring areas, offering detailed insights into regional logistics ecosystems. Furthermore, the scalability of these models makes them suitable for comparing urban and peri-urban logistics clusters, thus proving valuable for regional and local analysis.

2.6.2 Machine Learning Techniques in Remote Sensing

The field of remote sensing has seen significant advancements, particularly through the integration of machine learning methodologies. Recent research has explored the use of convolutional neural networks (CNNs) and other machine learning algorithms to automate the detection of logistics facilities using satellite imagery (Li et al. 2018). These approaches allow for the identification of warehouses based on their distinct spatial characteristics, such as size, shape, and proximity to transportation networks.

By combining high-resolution satellite imagery with machine learning, researchers can improve the resolution at which logistics facilities can be detected, particularly in densely populated urban areas where industrial, residential, and commercial land uses are closely intertwined. In 2022, The National Institute of Geographic and Forest Information (IGN) in France has started employing the use of machine learning methodologies in studying and surveying the land use using deep learning models that are made open source for researchers to use, these models depends on the imageries collected using remote sensing (Actuia 2022). The power of deep learning methods used in this project makes it possible to automatically identify different types of land use on the ground, such as buildings, green cover, and roads. This advancement overcomes many of the limitations and misclassifications associated with traditional remote sensing techniques. (IGN 2022).

2.6.3 Digital Twin and Scenario Modelling in Urban Logistics

The concept of a digital twin has recently been introduced to logistics research as a means of enhancing the understanding of logistics networks, particularly in urban areas. A digital twin is a virtual representation of the logistics framework, powered by real-time information to model, simulate, and forecast logistics operations (M. Batty 2018).

Digital twins allow researchers to test various scenarios concerning warehouse locations and their subsequent effects on road traffic, environmental conditions, and socio-economic factors. For instance, digital twins can simulate the impact of constructing new warehouses on urban ecosystems, enabling stakeholders to foresee potential consequences such as increased traffic congestion or environmental degradation before any physical changes are made. This methodology has proven useful in decision-making processes for selecting optimal warehouse locations.

Furthermore, digital twin models provide an interactive platform for engagement among policymakers, researchers, and community stakeholders to explore prospective logistics infrastructure projects, considering their impacts across multiple dimensions.

2.6.4 Participatory GIS and Crowdsourcing

Participatory Geographic Information Systems (PGIS) and crowdsourcing have emerged as innovative methods for augmenting logistics research, particularly in data-scarce regions. PGIS leverages community input to gather spatial data on logistics infrastructures, serving as a bottom-up alternative to traditional data collection methods. In metropolitan areas where logistics facilities are not highly visible, researchers have utilised PGIS to involve local stakeholders—such as truck operators and warehouse employees—in collecting spatial and operational information (Rodrigue 2024).

Crowdsourcing platforms like OpenStreetMap have also played a significant role in gathering geospatial data relevant to logistics facilities. Crowdsourcing is especially beneficial in areas where official datasets are outdated or incomplete (Alamri 2024). Although concerns remain regarding data reliability and accuracy, crowdsourcing provides a flexible and cost-effective method for improving spatial knowledge of logistics networks, particularly at the micro-scale. Nevertheless, research in this area remains limited.

2.6.5 Conclusion

The growing recognition of the limitations in conventional methods for identifying logistics facility locations has prompted the development of several innovative approaches. These include spatial econometric models, machine learning-enhanced remote sensing, digital twins, and participatory GIS. These new methodologies offer enhanced spatial and temporal precision, allowing for a better understanding of logistics activities and their integration into broader urban planning frameworks. The next section will further explore gaps in the existing literature on these methods and highlight potential areas for future research.

2.7 The Role of Socioeconomic Indicators and Spatial Factors in Warehouse Location Determination

The literature on the socioeconomic indicators that influence the location selection for the e-commerce warehouses highlights a complex interplay between economic, spatial, and environmental factors, all of which significantly shape the dynamics of modern logistics (M. Hesse et al. 2004; Rao et al. 2015). The increasing role of e-commerce in urban and sub-urban areas has driven a nuanced examination of these factors (Bowen 2008; Sakai et al. 2015), each contributing to an understanding of the location selection trends of these warehouses (Markus Hesse 2008).

One of the most fundamental indicators explored in the literature is proximity to urban centres (Sheffi 2012; Cidell 2010). The proximity of warehouses to high-density urban areas is often seen as essential for optimizing last-mile deliveries, which is a key demand in e-commerce (Allen et al. 2012; Woudsma et al. 2008). Studies have consistently emphasized that the siting of logistics facilities within a reasonable radius of urban centres not only ensures shorter delivery times but also enhances the competitiveness of e-commerce players (M. Hesse et al. 2004; Sakai et al. 2015). However, this also comes at a cost, as urban spaces are limited and expensive, which forces logistics operators to navigate the balance between operational feasibility and spatial constraints (Dablanc, Heitz, et al. 2020; Cidell 2012).

The availability and cost of land are other central factors in warehouse site selection decisions (Cidell 2010). The high cost of urban land has driven a trend known as ”logistics sprawl,” where warehouses relocate to more affordable peri-urban or suburban areas (Dablanc and Rakotonarivo 2010; Heitz 2017). This move allows logistics operators to acquire larger facilities, often necessary for the scale of e-commerce warehousing, but it simultaneously results in increased transportation distances, thus impacting overall logistics efficiency (M. Hesse et al. 2004). This tension between land affordability and transportation costs is a recurring theme in logistics literature, particularly in France and across Europe, where urban policies heavily influence land use and availability (Dablanc 2007).

Transportation infrastructure, particularly proximity to major highways, airports, and ports, has been widely recognized as a critical indicator of warehouse location. Warehouses located near transportation nodes benefit from reduced logistics costs, improved efficiency, and better access to both inbound and outbound flows. This strategic positioning not only serves the operational needs of e-commerce but also aligns with broader urban planning goals of creating multimodal and interconnected logistics networks (M. Hesse et al. 2004).

Another significant, yet often controversial, indicator is the socioeconomic profile of neighbourhoods. Literature shows that warehouses are frequently established in lower-income areas, where land values are lower, and resistance from communities tends to be less pronounced (L. K. d. Oliveira et al. 2022). These neighbourhoods, however, bear the external costs, such as increased traffic, noise, and pollution, leading to concerns about environmental justice and equitable spatial planning. Such findings highlight the dual impact of logistics facilities—while they may offer local economic opportunities and job creation, they also bring externalities that disproportionately affect vulnerable populations.

The environmental implications of warehouse location selection are also increasingly important in the literature, particularly in the context of urban sustainability (Rodrigue 2024; F. Russo et al. 2011). The expansion of logistics facilities into sub-urban areas has raised concerns about urban sprawl, emissions, and land artificialization (Heitz, Dablanc, et al. 2017). The regulatory frameworks around these environmental concerns significantly influence location selection decisions, with stricter regulations often leading to the adoption of innovative logistics models (Dablanc and Rakotonarivo 2010), such as multi-story warehouses and urban consolidation centres, which aim to minimize environmental impacts while maximizing space efficiency.

Finally, the literature points to the growing role of technology and automation in influencing warehouse location selection. Technological advancements, such as inventory automation and routing optimization, are transforming the logistics landscape (Boysen et al. 2019). By integrating spatial econometric models and real-time data analytics, logistics operators can better navigate complex urban environments and make informed location selection decisions that balance multiple socioeconomic and operational factors simultaneously. This data-driven approach is a significant departure from traditional location theories, which often relied on static data and generalized assumptions.

In conclusion, the socioeconomic indicators influencing e-commerce warehouse location selection are diverse and complex, each interacting with others in shaping the logistics landscape. These indicators—urban proximity, land cost, transportation infrastructure, socioeconomic context, environmental impact, and technological advancement—collectively provide a framework for understanding the dynamics of logistics facility siting. While some indicators, like transportation infrastructure and land availability, are consistently highlighted across studies, others, such as socioeconomic impacts and technological integration, represent evolving areas of focus that underscore the changing nature of logistics in the e-commerce era. Moving forward, a more integrated approach that considers these multifaceted indicators holistically is crucial for sustainable and efficient logistics planning in urban and peri-urban areas.

Chapter 3

Methodological Approaches to Locating Warehouses

3.1 Introduction

As presented in the literature review, the primary challenge with studying logistic activities usually lies in acquiring detailed data on the logistic facilities and their characteristics. Research on the logistic facilities often uses the administrative area centre (e.g. ZIP code area) as a focal point when analysing logistics activities in a spatial context, as seen in Dablanc (2014). In certain cases, the barycentre of warehouse locations is calculated and used in the analysis, as demonstrated in the works of Dablanc, Ogilvie, et al. (2014) and Heitz, P. Launay, et al. (2019).

Obtaining accurate warehouse location data is often a significant challenge with the insufficient information in a standard framework. Precise geospatial data is rarely publicly available and typically comes from private, paid databases. Moreover, publicly accessible datasets are not only limited in scope but are often inaccurate or outdated, as I noticed when working with resources like the *MWPVL* website which is used mainly by several studies. These issues complicate spatial analysis, particularly in logistics research where the lack of detailed and reliable data is a persistent problem, especially for e-commerce warehouses. Public datasets fail to provide the necessary level of detail for in-depth analysis, highlighting the pressing need for an open-source, up-to-date database that can keep pace with the rapidly changing logistics landscape.

3.2 Limitations of Data Collection

Throughout my work, different difficulties were faced due to technical limitations and sometimes due to hardware limitations for processing big datasets using the PC, and sometimes the use of an open-source software such as RStudio and QGIS proved limited functioning.

For the technical limitations of the data, they are as follows:

- **Encoding Issues:** Different encoding systems (e.g., UTF-8, Windows-1256) caused issues when importing and cleaning data, particularly with French names of the areas and the establishments containing special characters (e.g., words like "*Rhône*" becomes "*RhÃ ´ne*" if not imported properly), for both the CSV files and the spatial data.

- **Projection Conflicts:** Spatial data with different projections (IGN-based, WGS 1984) led to conflicts and errors that required careful handling for each dataset and for each processing.

- **CSV Delimiters:** Variations in CSV delimiters (comma or semicolon) needed close attention during data import to ensure consistency. In R, this could cause issues because of the use of different functions for each seperator. Also, the use of datasets from different French and European or international sources means using different encoding systems that needs more attention. In the cases of large datasets, it often required the use of data exploration tools like FME (Feature Manipulation Engine).

- **Large Datasets:** The constant issues when dealing with exceptionally large datasets such as SIRENE and SITADEL in R when the datasets are converted into data frames for manipulation purposes. Using SQL is one solution, but this also means the need to forward the data after handling it in SQL. Using PostgreSQL and R togather may solve some issues but sometimes when not all the datasets are stored locally, this can pose new issues of connection and transferring. Also, when the data needs to go through advanced spatial analysis in R in one dataset, this can cause the RStudio to crash, or it may take days of processing. Further handling techniques could be utilized such as applying parallel programming techniques in R or segmentation actions, depending on the nature of the data and the available resources.

In this chapter, we are going to explore different methodologies experimented during the internship to identify the locations of warehouses. This includes combining public data with different data obtaining methods such as databases, GIS, and web scraping. The goal is an attempt to find a more accurate way of collecting data that can help us better understand the locating of the warehouses to improve research results.

3.3 Using French Classification of Activities (NAF) with SIRENE Database

3.3.1 Concept

The most basic method is used in identifying warehouse locations is using logistics-related NAF codes, selected from the relevant literature to identify which of the NAF codes indicates a commercial or industrial activity that is related to the logistics. The chosen NAF codes for this approach are based on the literature; mainly a study by Heitz, Launay, et al. (2017) which also worked on proposing a new data collection methodology on logistics facilities in the french context. Only those codes related to the storage are used in this methodology. Later, these NAF codes were used to identify potential warehouses using SIRENE database which provides geocoded locations for the establishments. This approach aims to categorize and explore the potential use of these codes for pinpointing warehouse locations. Therefore, the chosen NAF codes are listed below:

- **52.10A**: Cold warehousing and storage.

- **52.10B**: Non-cold warehousing and storage.

3.3.2 Methodology

Figure 3.1 shows the steps taken in this methodology.

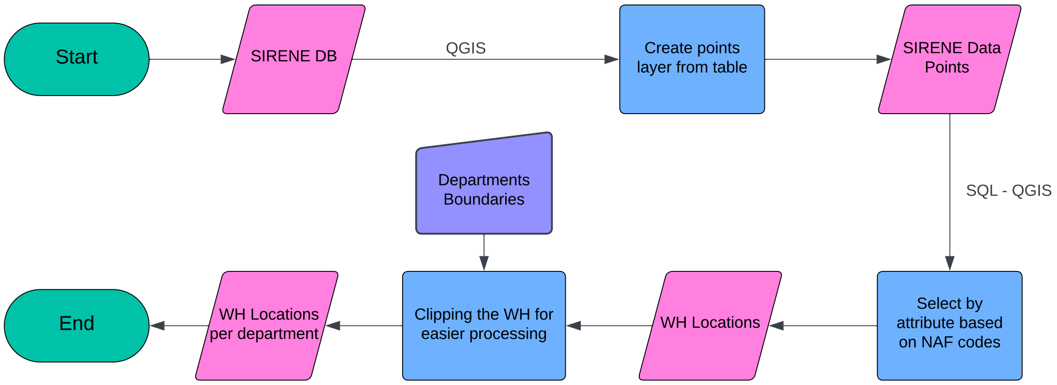

Start
SIRENE DB
QGIS
Create points layer from table
SIRENE Data Points
SQL - QGIS
Departments Boundaries
Select by attribute based on NAF codes
WH Locations
Clipping the WH for easier processing
WH Locations per department
End

Figure 3.1: First Approach Methodology

Data Extraction

- NAF codes specific to logistics and storage activities were chosen to filter relevant establishments.

- SIRENE database was queried using **SQL** to retrieve the establishments' records with the selected NAF codes.

Goespatial Mapping

- **Conversion of Data into Spatial Points:** The extracted data, which included addresses and geographical coordinates, was imported into a Geographic Information System (GIS) tool to generate a point layer for visual analysis.

- **Overlay with Department Boundaries:** To understand the spatial distribution of these warehouses, the point data was overlaid with administrative boundaries of the French departments.

- **Cross-referencing with External Sources:** To compare the results for inital evaluation, I added the last land-use data produced by Grand Région (GR-SIG) on the logistic zones bigger than 25 hectares.

Figure 3.2 shows a map of the analysis results in Île-de-France area. After that, to evaluate the results of the analysis, the summary statistics for the intersection between the selected NAF codes and the GR logistic zones are calculated and it is shown in the Table 3.1.

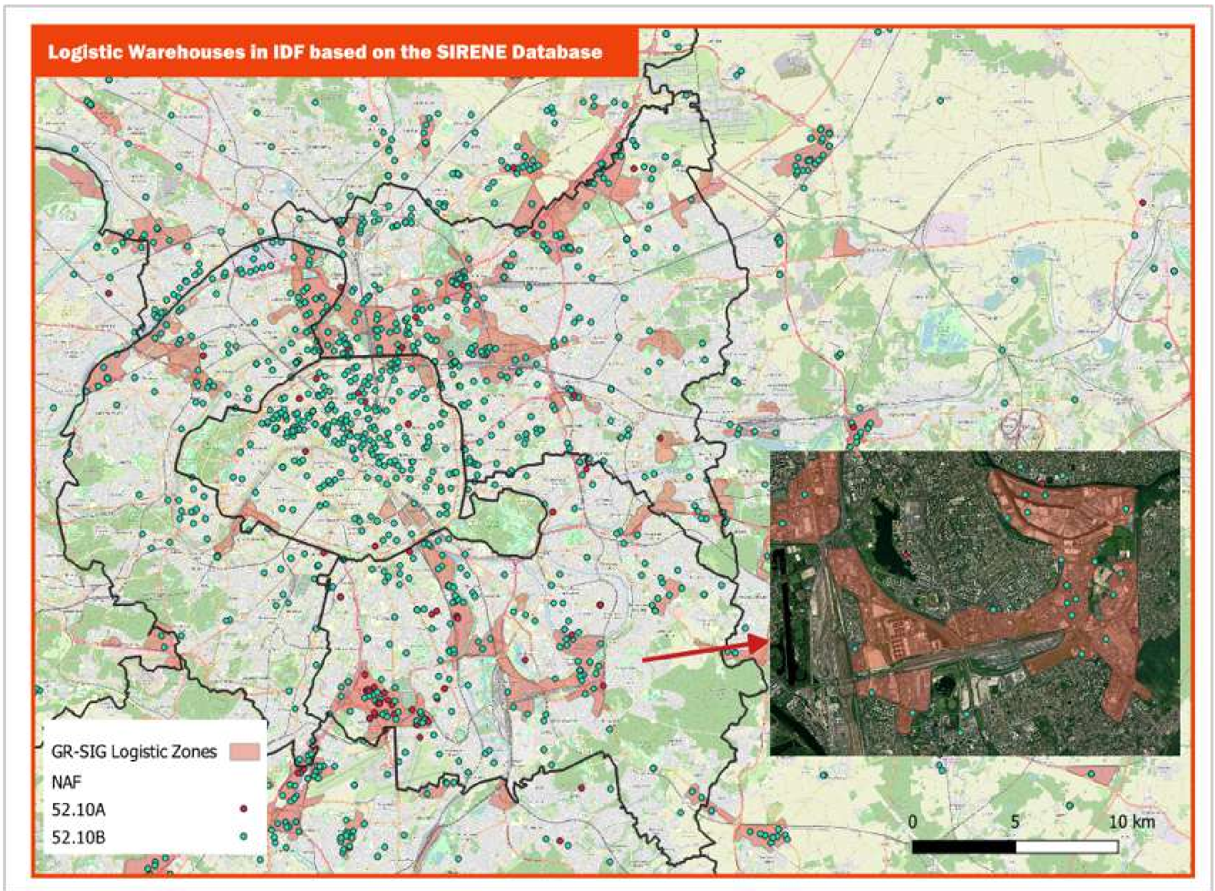


Figure 3.2: This map displays the location of the resulted warehouses in Île-de-France, identified using NAF codes 52.10A and 52.10B. Map by Mohammed Younes, based on geocoding of SIRENE database data from April 2024. The logistic zones are as of 2018, with the basemap sourced from OpenStreetMap, visualized via QGIS

Statistic	Value
Count	945
Mean	5.88889
Median	2
St dev (pop)	17.6279
Minimum	0
Maximum	412
Minority	27
Majority	0
Variety	51

Table 3.1: Descriptive Statistics.

The table shows that on average, each logistics zone has around six warehouses, and the maximum number of 412 warehouses intersected out of the 945 original results suggests that around half of the resulted warehouse locations from this approach are not in the logistics zones. Moreover, the median number of warehouses per logistics zone is just two, which means that half of the zones include a relatively small number of the resulted warehouses. This suggests that a few zones have a lot more than the others. Finally, the standard deviation of 17.63 indicates that there is a vast range in the number of warehouses from one logistics zone to another. This suggests that some zones are absolute hotspots for warehouses, with many more than the others. All these indications suggest that this approach is not accurate in terms of locating the warehouses using the NAF codes.

3.4 Using OpenStreetMap (OSM)

3.4.1 Concept

OSM services could be useful to extract potential logistic warehouse locations based on the keywords used in the attribute table of the OSM data, specifically, the buildings layer contains two main fields that may be helpful to the analysis: `name`, and `type`. One of the main studies (Schorung et al. 2022) produced by the chair's team introduced similar approach focusing on the USA, focusing on the buildings types and their area, unlike this approach which is using the names of the buildings too. The proposed approach is basically to select the buildings with keywords related to the warehousing activities, in either of the buildings layer fields, in both English and French. Figure 3.3 shows the conceptual idea of using the keywords in extracting this information using SQL, either in GIS or any database management system.

3.4.2 Methodology

Data Extraction

For cost and time efficiency due to the vast size of OSM data, the query was implemented only in Île-de-France region. The data was downloaded using Geofabrik Download Server. Following that, the building layer were queried usign SQL in QGIS to select and extract the buildings to a new shapefile. The query resulted in 885 buildings in Île-de-France region that are potentially warehouses. Table 3.2 shows the keywords used in the query, and Figure 3.3 shows a flowchart illustrating the methodology used in this approach.

Type	Name
’warehouse’	’logistique’
’warehouses’	’logistics’
’entrepôt’	’logistics’
	’entrepôt’
	’entrepôts’

Table 3.2: The keywords used in the SQL query.

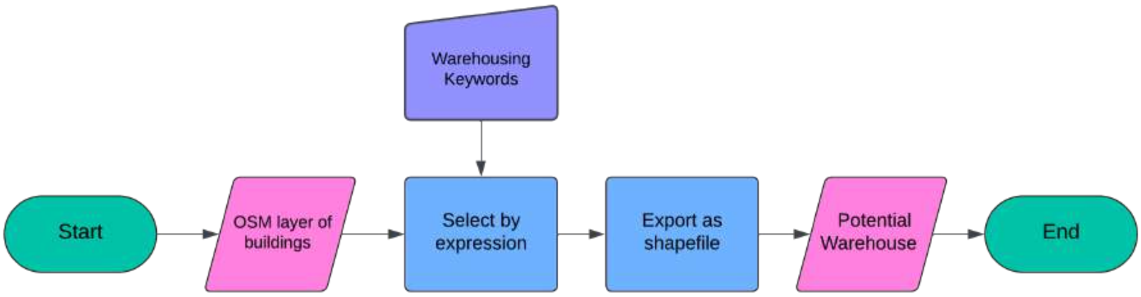


Figure 3.3: Flowchart diagram shows the extraction process using OpenStreetMaps data.

3.4.3 Results

The advantages of using this approach are that OSM database offers rich dataset with extensive coverage of different areas, and its usually utilising national data sources for the buildings such as IGN. Also, it allows for flexible querying using the keywords approach.

The main issue with these results that the definition of the warehouse categories in OSM is quiet general and is not necessarily limited to storage or fulfillment facilities, and it could also be small warehouses attached to retail stores.

Figure 3.4 shows a facility labeled with a type value of `warehouse` in OSM data, but it is actually a bus station (*Centre Bus Rive Gauche*), showing the ambiguity. On the other hand, Figure 3.5 identifies a Decathlon warehouse that the NAF codes-based approach failed to capture. Moreover, some warehouses from the buildings layer were not identified using this approach, while other warehouses were not even present in the buildings layer. This could be attributed to outdated data, leading to incomplete or inaccurate results.

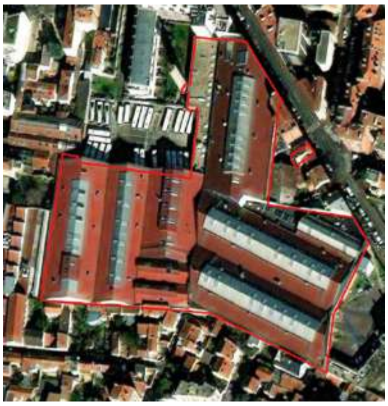


Figure 3.4: Centre Bus Rive Gauche Bus Station.

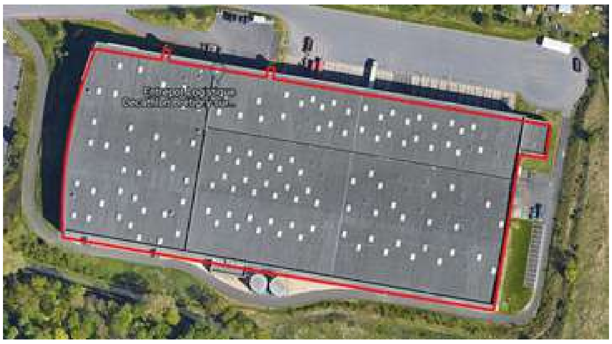


Figure 3.5: Decathlon Warehouse south of Paris.

This approach may be more useful if combined with other approaches that solve the false positives that could be produced by this approach, in order to refine the results and achieve a more accurate identification of the warehouses. For a moment, it could be helpful to study the pattern of the false positives produced, and then remove them using other data sources to refine the results.

3.5 Using FEVAD List of E-commerce Companies

3.5.1 Concept

One main organization providing information on e-commerce, one major part of the logistics activity is *La Fédération du e-commerce et de la vente à distance* (FEVAD). It produces regular reports and studies on the e-commerce activities. This approach proposes to identify the warehouses' locations, but focusing on e-commerce warehouses, by scrapping FEVAD website to get a list of the e-commerce companies and then extract potential warehouses with similar names from SIRENE dataset and SITADEL database to get the areas of these establishments.[1] The locations of the warehouses are also based on SIRENE dataset geolocations.

3.5.2 Methodology

Data Extraction

After studying the structure of FEVAD webpage in HTML, R is used to scrape the FEVAD website into a tabular form, and then saving them as a CSV list of around 560 e-commerce company names and websites.

The initial results showed some establishments with establishments associated with hospitals and schools, so, using a set of keywords associated with the schools and hospitals, they are filtered out, using **R**. Table 3.3 shows the keywords used in the filtering process.

Category	Keywords
Hosptials	'hôpital', 'hopital', 'clinique'
Schools	'school', 'école', 'ecole', 'lycée', 'college'

Table 3.3: The keywords used in the filtering process.

Subsequently, using R, partial filtering was applied to the SIRENE dataset, utilising the e-commerce companies list from the scraping results. This filtering process provided the coordinates of each identified establishment. Following this, the area information for these establishments was extracted using their SIRET numbers from both the SIRENE and SITADEL datasets. Establishments with areas smaller than 5000m^2 were then excluded, based on the commonly accepted definition in the literature that warehouses are at least 5000m^2 in size. The final data, consisting of 152 potential e-commerce warehouses, was exported as a CSV file. Figure 3.6 illustrates the methodology used in this proposed approach.

[1] Database of Building Permits and Other Urban Planning Authorisations (SITADEL).

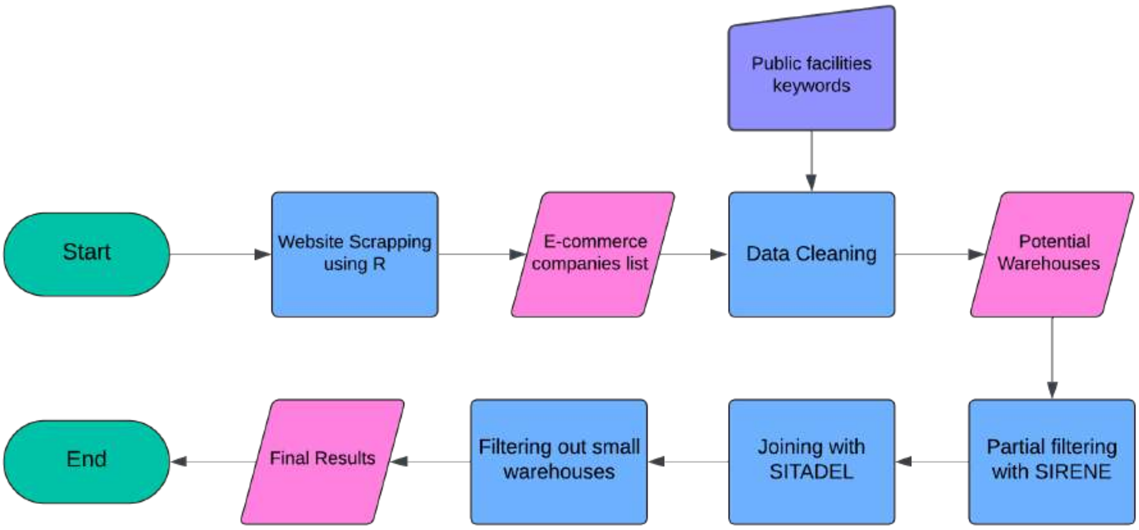

Public facilities keywords
Start
Website Scrapping using R
E-commerce companies list
Data Cleaning
Potential Warehouses
End
Final Results
Filtering out small warehouses
Joining with SITADEL
Partial filtering with SIRENE

Figure 3.6: Flowchart diagram shows the methodology using FEVAD list of e-commerce.

3.5.2.1 Results

Figure 3.7 shows a map illustrating the distribution of potential e-commerce warehouses across France.

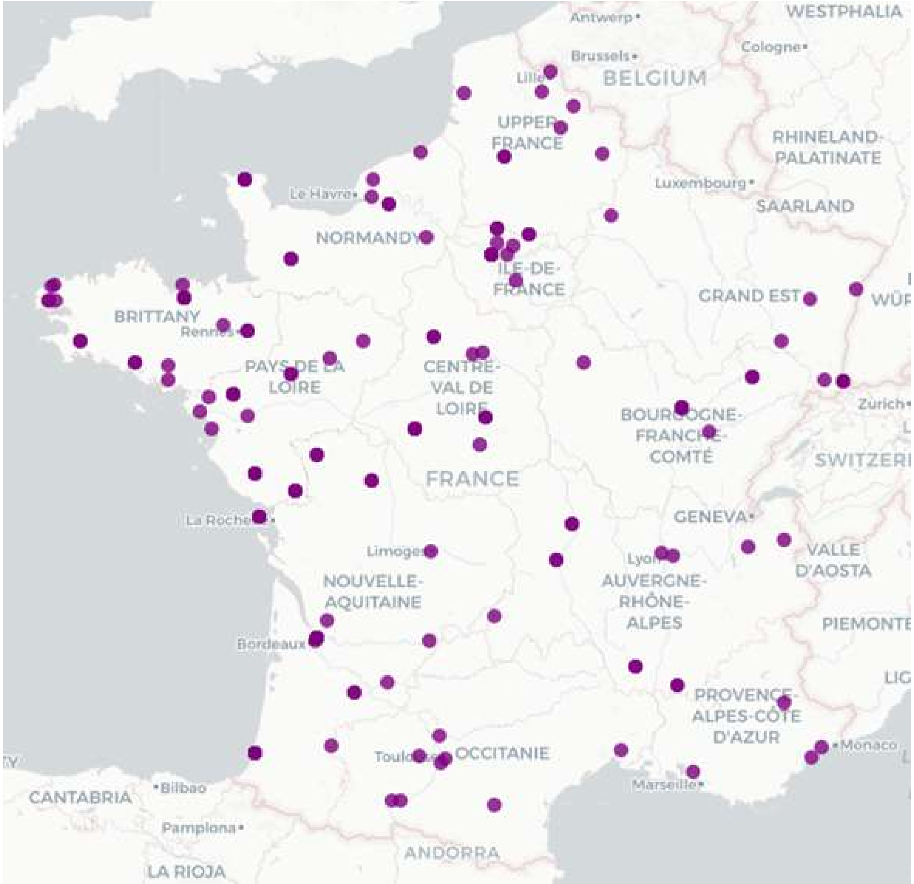


Figure 3.7: Geographical distribution of potential e-commerce warehouses across France, identified using the FEVAD list. Map by Mohammed Younes, with the basemap sourced from CARTO, visualized via Leaflet in R.

The advantage of using this approach is providing a targeted method to identify the e-commerce warehouses based on an authentic source of e-commerce activities in France such as FEVAD. Furthermore, the use of both SITADEL and SIRENE datasets in the process improve the accuracy and obtain detailed information. The main limitation of this approach being relying on the accuracy and completeness of the scrapped data from the website, and the names needs further filtering and cleaning which may result of losing the actual names of the companies. Also, the cleaning and filtering processes, especially the partial filtering, may select data that are not related to the e-commerce establishments, being general words as I saw when I encountered the hospitals and schools' data after the initial cleaning and extraction. Hence, more consideration should be taken to include other kinds of potential matching.

Finally, one significant systematic error encountered is in the area attribute, which inaccurately corresponds to the land plot or parcel size rather than the actual built-up area of the building. Such data will result in inaccurate results if the area attribute is used in any analysis. Moreover, the filtering may exclude smaller but significant warehouses that are small because they are small e-commerce companies or they are close to the dense urban areas, or simply because the data provided in SITADEL dataset are not always accurate.

3.6 Using Web-sourced E-commerce Logistics Companies List

3.6.1 Concept

Many e-commerce companies do not own warehouses, and depend on other companies warehouses or rent spaces in logistic hubs, which makes it had to identify them using SITADEL database. Therefore, in this approach, we are going to introduce an approach that identifies the e-commerce warehouses using a list of the logistic services companies that are involved in e-commerce activities. Thus involves using the data of e-commerce companies obtained directly from the website of french Supply Chain Magazine.[2]

3.6.2 Methodology

After extracting and cleaning the logistics companies list into an Excel sheet, it was cleaned and structured into a tabular format to prepare it for further analysis. Following this, the structured data was imported into a PostgreSQL database in CSV format for further processing and filtering. Subsequently, the company names were cross-referenced with the SITADEL dataset to obtain their siret numbers and area info.Then SIREN dataset was used to cross-reference company names and extract location details. Then, using SQL within *PostgreSQL*, the data went through cleaning and filtering process to remove irrelevant entries and establishments smaller than 4800 m^2 to avoid filtering out potential warehouses that are not exactly 5000m^2. Then the cleaned data was exported as a CSV file and imported into RStudio for mapping using Leaflet package.

Figure 3.8 shows a flowchart illustrating the methodology used in this approach.

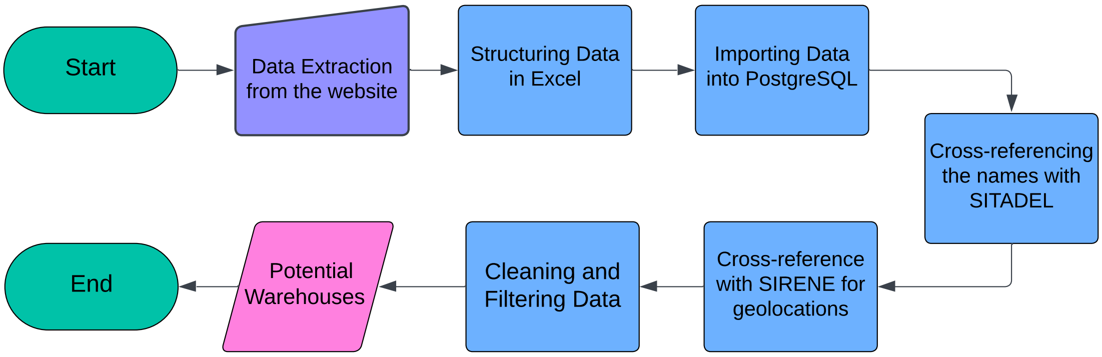

Start
Data Extraction from the website
Structuring Data in Excel
Importing Data into PostgreSQL
Cross-referencing the names with SITADEL
Cross-reference with SIRENE for geolocations
Cleaning and Filtering Data
Potential Warehouses
End

Figure 3.8: Flowchart diagram shows the extraction process using logistics companies working in e-commerce activities

3.6.3 Results

In this approach, using the names of logistics companies proved leading to more accurate identification of warehouse owners, as these companies are typically listed in governmental databases such as SIREN and SITADEL. And in this approach, I also found out that SQL-based cleaning and filtering allowed for efficient and fast data cleaning and filtering compared to using R when used with big datasets such as SIREN.

[2] See the website: https://urlis.net/989nxm8m.

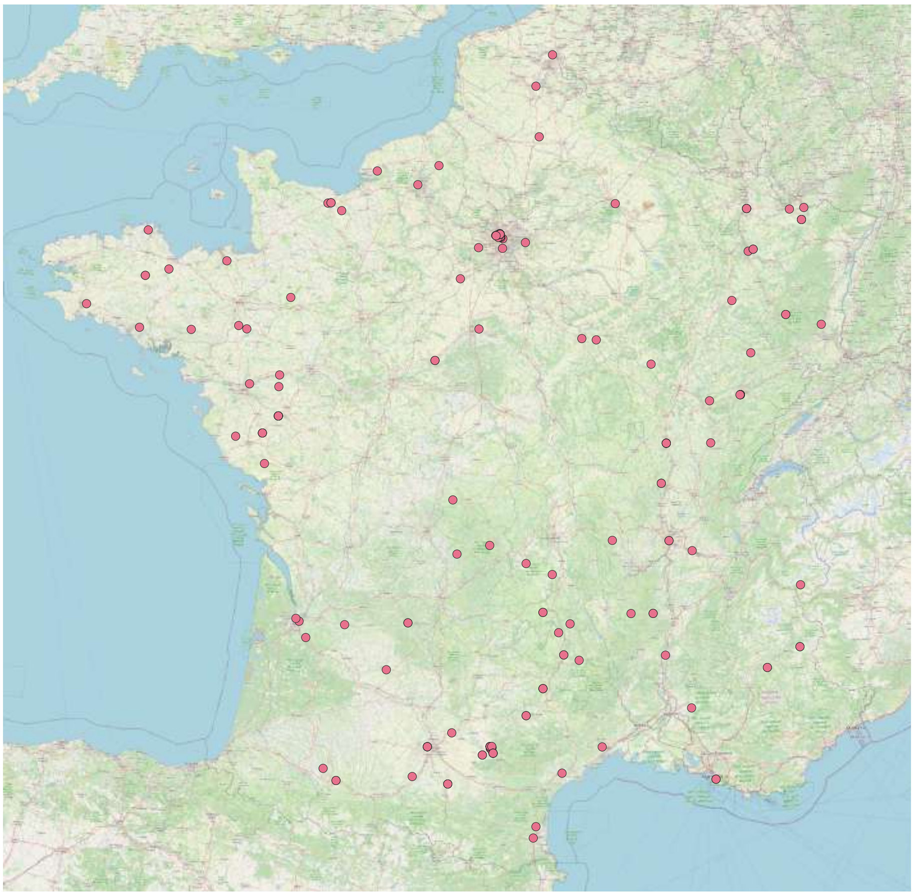


Figure 3.9: The resulted establishments based on the logistics companies specialised in e-commerce.

However, as in the previous approach, there are several limitations to this approach. Inaccurate location data from the SIRENE dataset led to inaccurate, and sometimes wrong locations. Systematic errors in the area data provided by the SITADEL dataset can affect the accuracy of the analysis, therefore this needs further investigation. Additionally, the data used from the website is not up to date, as the list provided was produced in 2022.

When comparing the mapped results of both approaches in figure 3.9 and figure 3.7, we can notice some obvious differences in the distribution of the points across France, the second one, figure 3.7 shows more distinguished clustering compared to the first one (figure 3.7), and the latter is also more consistent with the main logistics routes in France in most of the results, but it also reveals some new patterns in the logistic activities that may not be covered comprehensively in the logistics studies, such as potential new locations where logistic companies are expanding even though they are not on the main logistic routes. Additionally, the two figure may show some political changes in terms of distribution of the warehouses in some regions, but this assumption needs further studies and evaluation that takes into account the essence difference between the two methods used in this context, and also compare them to the other methods.

Finally, the process of matching names across datasets can also be challenging due to inconsistencies and variations in naming conventions. This approach had more better results from the rest of the previous approaches so far, but we still need for a better approach to decrease the processing time and provide better results.

3.7 Manual Enumeration of E-commerce Warehouses

3.7.1 Concept

Before getting into the manual collection and verification of the warehouse locations, it is crucial to acknowledge that despite the different semi-automated approaches presented so far, challenges remained in accurately identifying warehouses locations, let alone the e-commerce warehouses. The reason behind this is mostly associated with the lack of accurate integrated and up-to-date data, and the integrity of the different data sources. Moreover, the lack of a simple automated verification tool, keeping us from leveraging high quality data sources together to dynamically produce more accurate data on the warehouse locations.

The most effective method for identifying e-commerce warehouses involves manually collecting and verifying locations based on a list of the top 20 e-commerce companies, published by FEVAD. This approach is basically depending on the names of these companies to look up their warehouses on Google Maps and OpenStreetMap, and then trying to make sure of their activity combines multiple data sources, including the SIRENE database, IGN Orthophotographs, Google Earth, and Google Maps, to ensure accuracy and reliability.

3.7.2 Methodology

This approach involved spending long hours manually looking up e-commerce companies' names online, mainly Google search engine and Google Maps, then collecting and verifying as much information as possible.

To streamline the manual identification process of e-commerce warehouses, several key steps were taken to ensure accuracy and reliability. The process began with cross-referencing, where the top 20 e-commerce companies, as listed by FEVAD, were used to look up potential warehouse locations on Google Maps and, at times, OpenStreetMap. Following this, visual verification was conducted using IGN Orthophotographs, which provided updated aerial images to confirm the precise location of the warehouses.

Table 3.4: The manually collected dataset contents description.

Value name	**Description**	**Example of the values**
company	Company name	Amazon
code	The warehouse code if exists	ORY8
address	The address of the warehouse, preferably in the way written in the SIRENE catalogue for easier matching in any future processes with BAN database of French locations	ZAE DE GIDY 150 RUE DES VERGERS 45520 GIDY
area	The area of the warehouse based on a direct measurement using Google Earth or QGIS, or based on info from the company website.	15000
opened	The opening year, based on SIRENE or the company website.	01/02/2017
closed	The closure date, if closed.	-
details	The type of the warehouse and the nature of its operations, and any useful details such as the name of logistics operator if known.	Sorting Centre connected to ORY1 DC, Managed by AMAZON FRANCE TRANSPORT SAS
latitude	The latitude of the warehouse	47.96549218500225
longitude	The longitude of the warehouse	1.8443414642401705
status	The operating status, active or inactive.	Active
NAF code	The NAF code of the warehouse according to SIRENE catalogue, if found.	52.29B
siret	The SIRET code of the warehouse according to SIRENE catalogue, if found.	82324437100030
siren	The NAF code of the company operating the warehouse according to SIRENE catalogue, if found.	823 244 371

LogisticsCoApproach	Is this warehouse found based on the logistics companies approach?	no
link	Any additional information on the warehouse	https://www.loiret.gouv.fr/file/AMAZON.pdf

Once a potential warehouse was found, its address was copied into the SIRENE catalogue to check for the establishment's registration, regardless of the operating company. Multiple variations of the address were often tested to locate the correct information. Subsequently, manual validation was performed, involving a thorough check of Google reviews, news articles, and essays to ensure the warehouse was still active and engaged in e-commerce activities. Finally, after all validations, the confirmed warehouse data was manually entered into an Excel sheet, organized in a structured table to support further analysis. Table 3.4 shows the description of the table.

Data Used

- **FEVAD List**: The top 20 most visited e-commerce sites and applications in France in 2023, (Besnard 2023).
- **SIRENE Online Catalogue**: Used for cross-referencing additional information on the warehouses.[3]
- **IGN Orthophotograph**: Used as an updated base-map for visual verification of warehouse locations.
- **Google Maps and OSM**: The main sources in verifying and pinpointing exact warehouse locations.
- **Amazon Warehouses Map**: A map presented by Amazon team during the field visit to Metz, showing the distribution of Amazon Warehouses across France and their types (Distribution Centre, Sorting Centre, Delivery Station).
- **Google Earth**: Provided a quick tool to measure the area of the warehouse manually.

3.7.3 Results

The manual process of locating warehouses resulted in about 130 warehouses locations in Metropolitan France area, which are verified and accurate enough for any potential analysis. Figure 3.10 shows the mannually collected locations of the e-commerce warehouses.

This approach proved to have higher accuracy and reliability due to manual verification when compared to the other approaches. It also provided an ability to cross-reference multiple data sources for more comprehensive results. The dependency of visual verification such as aerial photographs helps to ensure that the identified locations are indeed warehouses. However, the nature of this approach is also a time-consuming and labour-intensive process, and it took a lot of efforts and time to look-up each warehouse, and sometimes some warehouses were not clear if they are active or inactive. Also, when SIRENE catalogue is used directly to look up the e-commerce company in the cases where the company itself manages a chain of warehouses, it does not include all the warehouses, like in the case of Amazon ETZ2 warehouse, which is not managed by AMAZON FRANCE TRANSPORT SAS like most of Amazon's warehouses, leaving it out of the results and posing the need to look up all the potential options.

[3] Annuaire des Entreprises: le moteur de recherche officiel. Available at: https://annuaire-entreprises.data.gouv.fr/. Accessed: September 2023.

It also has a potential for human error during manual verification, and it presents a limited scalability due to the manual nature of the method, this being said, the current results are about 100 located warehouses that took a lot of time. But it provides a basis to the data analysis with accurate information.

However, this approach proved to offer the most accurate information on e-commerce warehouses locations that no other approach or data source provided. Therefore, it would be used later in the research for the further statistical analysis to evaluate the efficiency of the collected data in any potential statistical analysis in the e-commerce warehouses context.

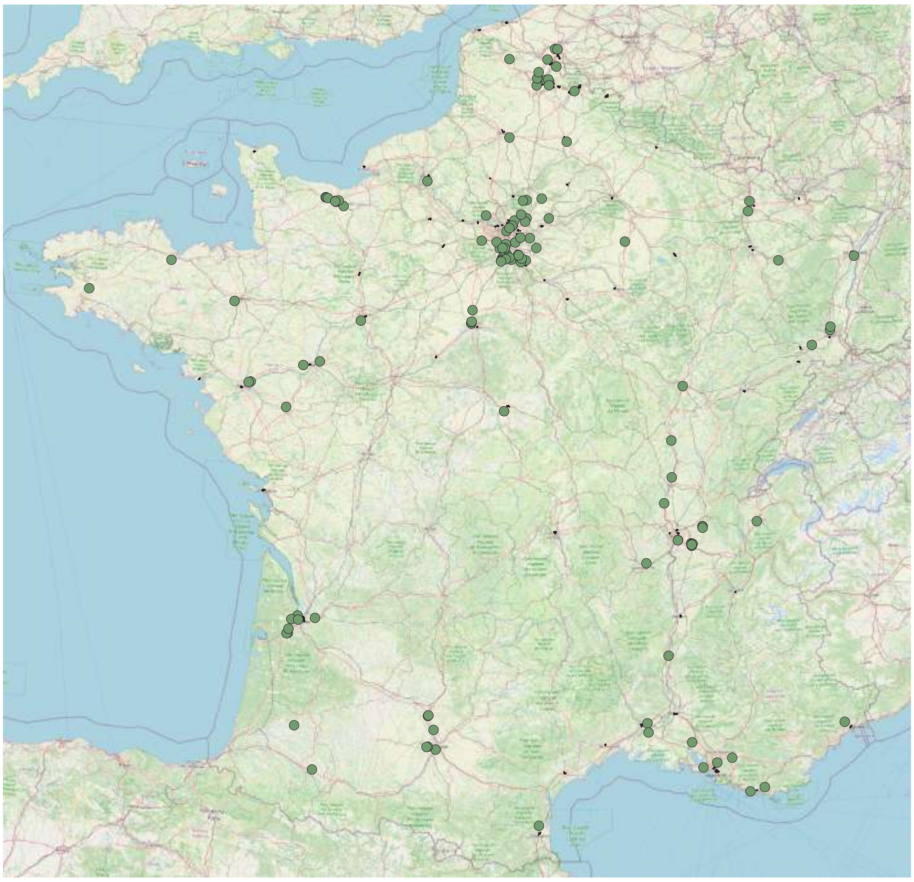


Figure 3.10: Geographical distribution of the manually identified e-commerce warehouses across France.

3.8 Conclusion

In this chapter, various methodological approaches were explored to accurately identify and map the locations of warehouses, with a particular focus on e-commerce warehouses. The first approach, using NAF codes within the SIRENE database, provided a broad but somewhat inaccurate identification of warehouses, largely due to the generalization of the codes and incomplete data. Despite offering comprehensive coverage, the method revealed its limitations in distinguishing between logistics offices and actual warehouse locations.

The SIRENE database offers extensive coverage of businesses throughout France, which makes it a valuable resource for identifying logistic facilities. Its accessibility for research purposes, along with monthly-updated detailed information on business activities, locations, and classifications, enhances its utility in logistics analysis. However, the broad classification of the NAF codes presents challenges, as it groups a wide range of logistics-related establishments, which may not always include warehouses. This can lead to the misclassification of entities like transport hubs or offices as warehouses. Additionally, some businesses may not regularly update their registration in the SIRENE database, resulting in outdated or incomplete data that affects the accuracy of the analysis.

The second approach utilizing OpenStreetMap data highlighted the flexibility and richness of open-source data. However, its reliance on keyword searches in building layers produced ambiguous results, leading to potential misclassifications and false positives. This limitation pointed to the need for refined filtering techniques or the combination of datasets for more precise outcomes.

The third method, leveraging the FEVAD list of e-commerce companies, proved to be more focused on e-commerce-specific logistics activities. By scraping and cross-referencing company names with SIRENE and SITADEL datasets, this approach significantly improved the accuracy of identifying potential warehouse locations. However, challenges such as inaccurate area data and difficulties in matching company names persisted, necessitating further refinement.

The fourth approach, which involved using web-sourced data from logistics companies, demonstrated promising results, especially when combined with SQL-based filtering in PostgreSQL. This method provided efficient data cleaning and processing, but inconsistencies in location data from the SIRENE dataset and outdated information remained a concern.

Finally, the manual enumeration approach emerged as the most accurate method, allowing for direct verification of warehouse locations using a combination of Google Maps, IGN Orthophotographs, and other data sources. While labor-intensive, this approach yielded the most reliable results, with verified and precise locations of e-commerce warehouses.

The manually identified e-commerce warehouses 3.10 show a marked concentration around Île-de-France, especially Paris, and in Northern France. This suggests a focus on identifying highly visible or larger-scale warehouses, potentially through the concentration of the ecommerce companeis from the data sources around the main cities. The logistics companies' warehouse locations 3.9 reveal a more distributed pattern, expanding significantly across the southern parts of France, suggesting that logistics networks extend beyond the major urban centres into peripheral regions. The FEVAD list results as shown in 3.7 also shows a wide distribution, but with notable clusters in regions like Brittany and Provence-Alpes-Côte d'Azur, which may reflect lists provided by industry associations that highlight potential rather than confirmed warehouses.

Another remark on the warehouses distribution of the e-commerce warehouses in 3.10 shows the potential influence of the proximity to the ports, airports, and borders with the main neighbouring countries especially on the eastern borders of the country. On the other hand, as mentioned before, the warehouses are also concentrated around the main urban centers in France, other cases may involve newly emerging economic active areas.

Another observation regarding the distribution of e-commerce warehouses in 3.10 highlights the possible impact of proximity to ports, airports, and the country's borders, particularly along the eastern border with key neighbouring nations. Additionally, as previously noted, many warehouses are concentrated near France's major urban centres, while other locations may reflect regions experiencing recent economic growth.

The differences in warehouse coverage across the three maps highlight how different methods and data sources yield varying results. Manually identifying warehouses might be more precise for large and top e-commerce facilities, but it may miss smaller or more rural locations. Conversely, using logistics companies' data or industry lists (such as from FEVAD) captures a broader range of establishments, including smaller or less well-known locations, though they may include speculative or inactive sites.

Moreover, in all maps, Île-de-France is a major hub for e-commerce warehouses, which is consistent with the region's role as a logistics and distribution centre. However, certain regions, like Brittany, show greater prominence in the FEVAD map 3.7 compared to the manually identified and logistics company data. This suggests that potential warehouses in these areas may be more speculative or emerging, not yet fully captured by businesses records.

In summary, the chapter demonstrates that no single method is flawless, but by integrating various datasets and refining techniques, significant strides can be made in accurately identifying warehouse locations. Future work should focus on improving data integration, reducing processing time, and addressing the limitations encountered in each approach, particularly with regard to incomplete or outdated data. Further work could include integrating two or more of these approaches to benefit from their advantages and reduce the effect of their disadvantages.

The next chapter will focus on using the data collected during the last approach which proved to be the most accurate one. This dataset of the e-commerce warehouses in France will undergo statistical analysis to test its usability in any potential advanced analysis that aims to provide answers to more questions using more progressive methods. This can be an indication also for other researchers on the capacity of using such detailed locations of the warehouses for advanced spatial statistical analysis such as Geographically Weighted Regression (GWR) which focuses on the exact location of the warehouse and takes into account more advanced parameters that are beyon the scope of this internship.

Chapter 4

Statistical Analysis of Socioeconomic Impacts

4.1 Introduction

Research into a new methodology for collecting the location of e-commerce warehouses in France has resulted in a usable database of 133 warehouses. The aim now is to assess the views of a logistics provider, Amazon, interviewed in May 2024, regarding the criteria for choosing the location of e-commerce warehouses. During the meeting, the importance of the employment area in the choice of location was emphasized. The job is difficult because it is manual and repetitive, which makes it hard for companies to recruit: Amazon told us that employment areas with high unemployment not only enabled them to recruit, but also to meet the objectives of local authorities to create jobs in their areas. In the last part of the internship, and following the methodological research in Chapter 2, we arrive to the statistical analysis work. The goal of this chapter is to provide a diligent examination of the operability of the data collected on warehouse locations, specifically, on the warehouses locations collected manually during the fifth approach in section 3.7. Although this part took some time from the work in the last month, it was due to the iteration in experimenting different approaches in the statistical analysis, more specifically on using different methods in the data wrangling and data modeling which is not my background. Therefore this needed more continuous studying of the advanced statistical analysis using R and more discussions and meetings with my supervisor to discuss each attempt in the analysis and understand its implications on later steps.

4.1.1 Purpose and Objectives

The objective of the Statistical Analysis chapter is to apply a range of statistical methods to explore the spatial and socioeconomic factors that influence the location selection and distribution of e-commerce warehouses in metropolitan France based on the available collected data. This chapter aims to address key research questions by utilizing data-driven approaches, including ***Regression Studies*** and ***Principal Component Analysis*** (**PCA**), to identify the most significant determinants of e-commerce warehousing activity across different regions. These statistical techniques aims to help in quantifying the relationships between different variables in order to provide a clearer understanding of the spatial patterns of e-commerce warehouse development.

4.1.2 The Rationale for the Statistical Analysis

The available studies on the warehouses are often focused on its functional activities and do not take into account the location unless it is concerning the delivery time, which is usually limited to specific warehouses that has enough detailed data to conduct a study on them. The research on Logistic City often lack or overlook the detailed locations of the warehouses due to their unavailability or the more complicated techniques used if the researchers wanted to account for each warehouse location, therefore, research tends to use areal spatial analysis to facilitate the analysis, which holds major issues on the detailed scale or when accounting for large scale variables that could benefit from detailed classifications, due to the MAUP bias issue that we discussed in section 3.1.

Moreover, the collected data of e-commerce warehouses are distributed based on various factors such as population density, proximity to transport infrastructure, land use, and economic activity, all of which interact in complex ways. So using and applying statistical methods could move the analysis beyond descriptive insights to empirically test hypotheses and quantify the relationships between these factors and e-commerce warehouse locations, which is not always covered in the literature of e-commerce warehousing more than general descriptive results or limited quantitative indications.

4.1.3 The Rationale for the geographic scope

In conducting spatial analysis for warehouse location determinants, one of the key challenges lies in the availability and granularity of data. While governmental sources such as INSEE provide substantial datasets, detailed information at the commune level is often restricted due to privacy concerns, particularly when it comes to socioeconomic data. This limits the accessibility of highly localized data, which makes it difficult to perform detailed analysis at such a small geographic scale. In response to these constraints, the **Aires d'attraction des villes (AAV)** has become a more viable and widely available alternative. The AAV offers aggregated data that captures the urban influence over surrounding areas, which provides a more comprehensive view of regional dynamics while ensuring privacy protections.

Additionally, focusing on smaller administrative units like communes can exacerbate the modifiable areal unit problem (MAUP), where the scale and boundaries of analysis can distort spatial patterns. The commune level, with its very small and inconsistent scales, risks introducing unnecessary noise into the analysis. In contrast, the AAV provides a more stable geographic unit for analysis, offering a balanced scale that mitigates the MAUP issue while still providing sufficient detail for robust statistical analysis. Therefore, the decision to use AAV-level data not only addresses data availability issues but also enhances the analytical robustness of this study.

When considering whether to include all AAV areas, even those without warehouses, or only the regions where warehouses are present, it is essential to weigh the statistical implications against the research objectives. Including all AAV areas ensures that the analysis covers a comprehensive geographic scope, allowing for a fuller understanding of why warehouses exist in some regions and not in others. In many cases, the presence of numerous regions without warehouses might suggest the need for advanced statistical techniques, such as zero-inflated models, to handle zero counts effectively. By including all regions, the analysis becomes more generalizable, as it reflects the diverse conditions across both urbanized, industrial zones and less-developed rural regions, providing valuable insights into broader spatial patterns.

On the other hand, focusing exclusively on areas with existing warehouses offers a more targeted, simplified analysis, particularly when this study aims to understand the characteristics of these warehouse-hosting regions. By excluding areas with zero warehouse counts, researchers can concentrate on the factors influencing warehouse placement without the added complexity of managing zeros in the data. However, this approach can introduce statistical risks, as regions without warehouses often have distinct characteristics (e.g., rural vs. urban or per-urban) that could skew the analysis if omitted. While excluding zero-warehouse areas can streamline the modeling process, it limits the generalizability of the findings and may result in biased interpretations. Ultimately, the choice between these two approaches depends on the research goal: if the objective is to examine broad trends and potential future warehouse locations, a comprehensive analysis including all AAV areas is preferable. However, if the focus is solely on understanding areas with existing warehouses, the narrower scope might suffice, though it comes with trade-offs regarding statistical accuracy and scope. Figure 4.1 shows the boundaries and distribution of the AAV areas in France.

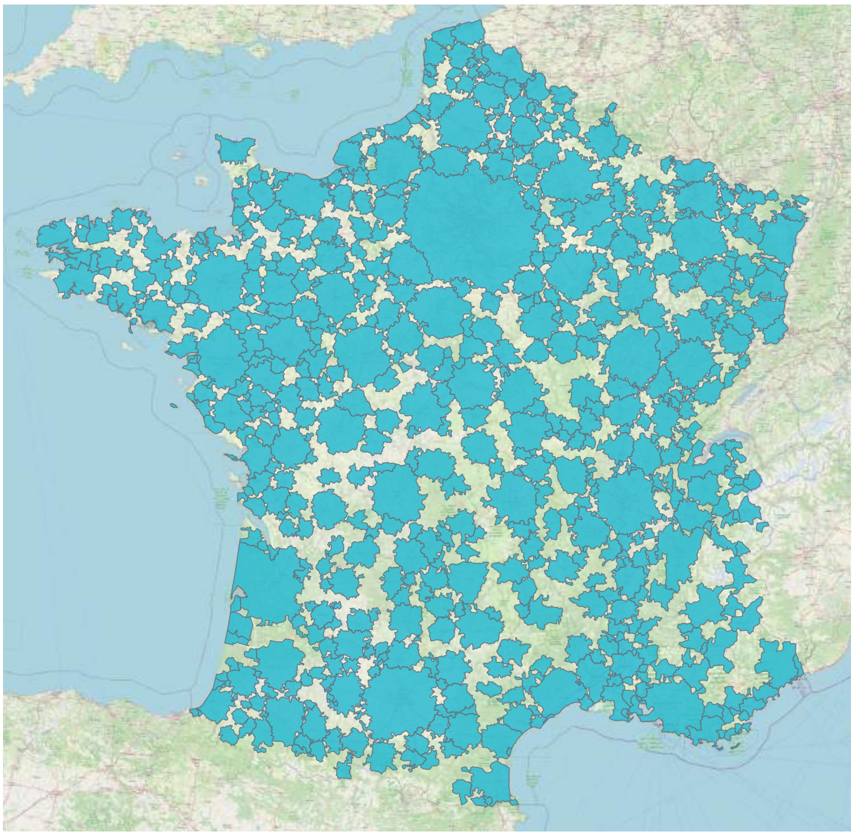


Figure 4.1: The geographical distribution of the Aires d'attraction des villes (AAV) areas.

4.2 Data Preparation and Exploratory Data Analysis (EDA)

4.2.1 Exploratory Data Analysis

4.2.1.1 Data Preparation

The data preparation phase is a critical step before the statistical analysis. For this part, different socioeconomic indicators were collected based on the literature and the previous studies of the Chair, to identify the used variables and how were they used in the analysis, to justify their use in the statistical analysis. Mainly, a wide group of the indicators were collected to allow more experimentation on potential use of them when correlated with e-commerce warehouses, but most of them were based on the literature. Also, some of them were based on the discussions with my supervisors in the laboratory who provided insightful notes on potential indicators, and finally, some of the indicators were chosen to test them based on the field visit for Amazon warehouses in Metz, where we were told about some of the factors behind building their warehouses in specific locations.

Other variables can be explained in details for their choice. For example, employment levels are a direct measure of economic health, and regions with stable or increasing employment are likely to attract more businesses (Moraga 2023). Employment stability over time reflects the region's ability to sustain industries and workforce, which is crucial for long-term economic stability. A stable employment market can provide a reliable source of labour, reducing the risks associated with labour shortages and turnover, making it an attractive location for logistics and distribution centres. Population density, as suggested by Glaeser et al. (2009), can be a proxy for the economic attractiveness of a region. Higher densities are often associated with more dynamic economic environments, conducive to stable growth. The role of population density in regional development is well-documented, with denser populations tend to support higher levels of economic activity, which in turn can lead to greater stability (Carlino et al. 1987). Changes in population density can indicate shifts in regional attractiveness or economic vitality, which are essential components of economic stability. Furthermore, the relationship between population density and economic vibrancy is complex and multifaceted. On one hand, higher population densities can lead to increased economic activity, innovation, and entrepreneurship, which can contribute to economic stability (Florida 2002). On the other hand, high population densities can also lead to increased competition for resources, infrastructure, and labor, which can negatively impact economic stability (Bettencourt 2013). The concept of Aire d'attraction d'une ville (AAV) is particularly relevant in the French context, as it refers to the geographic area that comprises a city and its surrounding territories, which are economically and socially integrated (INSEE 2020). The AAV is a useful framework for analysing the relationships between cities and their surrounding territories, and for understanding the dynamics of economic development and stability in these regions.

Later, these indicators underwent thorough classical data cleaning and collected into one data set on different geographical scales based on their availability. The dataset was cleaned by identifying and handling missing values, for variables with missing entries, different techniques such as imputation, or in cases where missing values were too numerous, careful exclusion of the data points was employed, but since most of the data are governmentally collected, the quality of the data were better and barely had to do any exclusion. Additionally, variables with inconsistent formats were standardized to ensure compatibility.

Once the dataset was fully cleaned and filtered, it was ready for exploratory data analysis (EDA), which allowed for initial insights and guided the selection of appropriate statistical techniques later. Table 4.1 shows the main variables used in this analysis.

Table 4.1: Collected Variables Description

Category	**Variable**	**Unit**	**Definition**	**Source**
Economic	dv3f_d5	euro/capita	Median income per adult in AAV area	Cerema
Economic	tx_pauv	%	Share of people below 60% of median income	INSEE
Economic	p_emplt	person	Number of employed people	INSEE
Economic	unemp15_64	%	Unemployment rate (ages 15-64)	INSEE
Economic	buisness_creation	%	Rate of business creation per AAV	INSEE
Economic	emplt_dens_area	resident/km^2	Employment density per km^2	INSEE, Manual
Economic	emplt_09_20	%	Change in employment (2009-2020)	INSEE, Manual
Demographic	pop_dens	resident/km^2	Population density in AAV	INSEE
Demographic	pop	resident	Total population of AAV (2021)	INSEE
Demographic	pct_change_09_20	%	Population change (2009-2020)	INSEE, Manual
Housing	evol_logt	%	Annual growth in housing units	INSEE
Housing	evol_logt_vacant	%	Annual change in vacant housing	INSEE
Infrastructure	ZFU_Dist	km	Avg. distance to nearest ZFU	INSEE, Manual
Infrastructure	AirportDist	km	Distance to nearest airport	OSM, Manual
Infrastructure	RoadDist	km	Distance to nearest regional road	IGN, Manual
Environmental	prg_tertiary	tCO_2/capita	GWP of tertiary sector per capita	Citepa
Environmental	prg_road	tCO_2/capita	GWP from road transport per capita	Citepa
Environmental	evol_sau	%	Change in agricultural area (2010-2020)	Agreste
Environmental	surf_artif	%	Share of artificial surfaces (2009-2021)	Soil Obs.
Geometric	AAV area	km^2	Area of AAV in km^2	Manual
Warehouse Metric	wh_count	warehouse	Number of e-commerce warehouses	Manual
Warehouse Metric	wh_dens_pop	warehouse/1M res	Warehouse density per million people	Manual
Warehouse Metric	wh_dens_area	warehouse/1000 km^2	Warehouse density per 1000 km^2	Manual

Table 4.1: The Collected Variables Description

4.2.1.2 Correlation Analysis

As part of the EDA, a correlation matrix was calculated. It provides an overview of the relationships between key variables influencing warehouse density. In the heatmap in Figure 4.2, strong correlations are displayed in shades of red, indicating a positive correlation, while shades of blue indicate negative correlations. Some key correlations that may affect the regression analysis include:

- `Surf_artif` (Share of Artificial Surface) shows a strong positive correlation with `wh_dens_area` (Warehouse Density) (0.42), indicating that regions with more artificial surfaces tend to have a higher density of warehouses. This variable is expected to play a significant role in the regression model.

- GWP from the Roads (`Prg_routier`) also shows a strong positive correlation with `wh_dens_area` (0.37), suggesting that areas with higher road emissions are likely to have higher warehouse density.

- Median Income (`dv3f_d5`) exhibits a strong negative correlation with `surf_artif` (-0.45) and `wh_dens_area` (-0.42), which could imply that warehouses are often located in lower-income areas.

- Employment Change (`emplt_09_20`) and Population Density (`pop_dens_k`) show a moderate positive correlation with `wh_dens_area` (0.22 and 0.31, respectively), indicating that regions with growing employment and higher population densities may also see more warehouses.

These correlations highlight important relationships between variables that will be crucial in understanding the factors influencing warehouse location selection in the regression analysis. Additionally, the presence of multicollinearity between some variables (e.g., between `pop_dens_k` and `surf_artif`) might need to be addressed to avoid biased estimates.

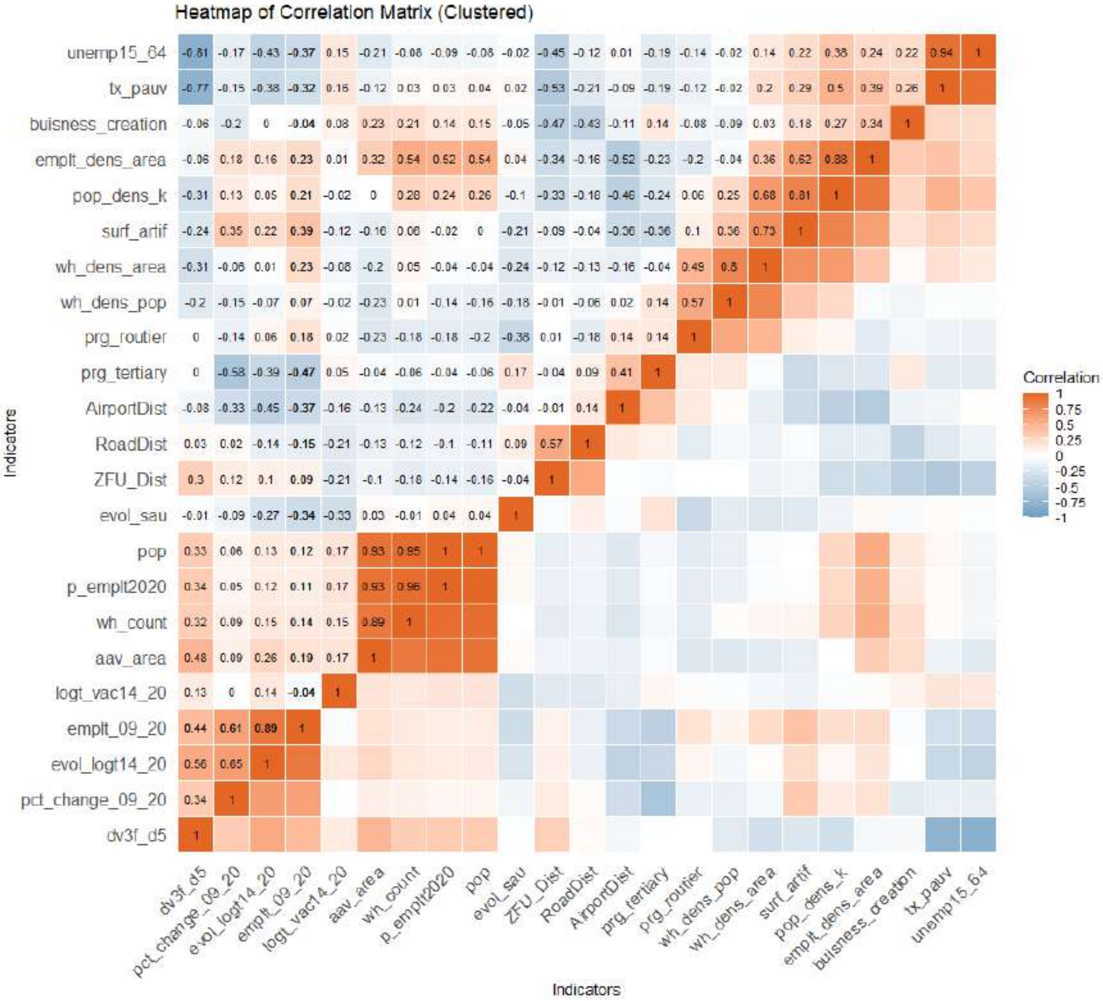


Figure 4.2: The correlation matrix of socioeconomic indicators.

After reviewing the correlation results, there was variability in the strength of the correlation between the indicators and the different warehouse metrics (warehouse count, warehouse density/km^2, warehouse density/100K capita). Consequently, I evaluated the correlations for each warehouse metric individually to determine which of metric most effectively explains the relationship with the indicators. Table 4.2 shows the correlation between each variable with each warehouse metric.

The findings indicate that the warehouse density metric (Warehouse Density/km^2) is the most efficient warehouse metric, therefore, it would be used for the subsequent analysis. The correlation results identified the following indicators with strong correlations:

1. Share of artificial surfaces: 0.727, meaning that regions with a higher proportion of artificial surfaces, such as urban areas, tend to have a higher warehouse density per km^2.

2. Population density: 0.677, meaning that high population density areas tend to have higher warehouses density per km^2, meaning that warehouses are more concentrated in densely populated regions.

Table 4.2: Correlation with Warehouses Metrics

Variable	**WH Count**	**WH Density/100K capita**	**WH Density/km^2**
dv3f_d5	0.323	−0.202	−0.310
tx_pauv	0.029	−0.024	0.201
p_emplt2020	0.956	−0.142	−0.039
unemp15_64	−0.079	−0.025	0.138
emplt_09_20	0.137	0.075	0.234
pop_dens_k	0.283	0.246	0.677
pop	0.954	−0.158	−0.041
evol_logt14_20	0.153	−0.066	0.014
logt_vac14_20	0.152	−0.018	−0.085
ZFU_Dist	−0.175	−0.007	−0.119
AirportDist	−0.242	0.017	−0.164
RoadDist	−0.118	−0.064	−0.126
prg_tertiary	−0.059	0.142	−0.037
prg_routier	−0.181	0.574	0.490
evol_sau	−0.010	−0.181	−0.244
surf_artif	0.064	0.362	0.727
aav_area	0.895	−0.229	−0.196
emplt_09_20.1	0.137	0.075	0.234
pct_change_09_20	0.088	−0.150	−0.060
buisness_creation	0.215	−0.090	0.030
emplt_dens_area	0.543	−0.041	0.364

3. GWP of the road sector: 0.490, suggesting that regions with higher greenhouse gas emissions from the road traffic are associated with higher warehouse density, possibly indicating areas with significant logistics and transport activities.

4. Median income: -0.310, meaning that areas with higher median income tend to have lower warehouse density per km^2, suggesting that warehouses are less concentrated in wealthier regions.

5. Employment density: 0.364, meaning that areas with higher employment density tend to have more warehouses per km^2.

6. Employment change from 2009 to 2020: 0.234, meaning that Regions that experienced employment growth between 2009 and 2020 also show a moderate increase in warehouse density.

7. Poverty rate: 0.201, meaning that higher poverty rates are moderately associated with higher warehouse density, potentially indicating that warehouses are often located in less wealthy areas.

8. Agricultural land use changes: -0.244, meaning that regions with significant agricultural land use changes tend to have lower warehouse density, indicating that warehousing is less prevalent in areas undergoing agricultural development.

As a reminder, one of the objectives was to test the link between the characteristics of the employment area and the location of e-commerce warehouses. These variables will be so used in the modeling for further analysis because of their strong correlation. Other variables from the original set is left out because of the visible correlation between them and other variable that could also explain the same theme.

4.3 Regression Analysis

4.3.1 Introduction to Multiple Linear Regression

Linear regression is a widely used statistical method that models the relationship between a dependent variable and one or more independent variables. In its simplest form, known as simple linear regression, the method assesses the effect of a single independent variable on the dependent variable. When multiple independent variables are involved, it is called multiple linear regression. The goal is to estimate how changes in the independent variables correspond to changes in the dependent variable by fitting a linear equation to the observed data (Craig et al. 2016).

In the context of analyzing warehouse density, linear regression is particularly useful because it allows us to quantify the influence of various factors—such as population density, infrastructure, employment rates, and income levels—on the spatial distribution of warehouses. By incorporating these independent variables, the model can help determine which factors are most significantly associated with higher warehouse densities. For example, variables like proximity to airports may have a strong positive relationship with warehouse locations, while socioeconomic factors like income levels may show a different kind of influence.

The use of linear regression provides a clear, interpretable model that can guide decision-making in logistics and urban planning. It allows for hypothesis testing and generates insights into the nature and strength of the relationships between the studied variables. In this research, linear regression helps in understanding how different spatial, economic, and infrastructural variables contribute to the location and clustering of warehouses across various regions.

4.3.2 Regression Model

The regression studies took different treatment and attempts to get to the right formula of the model, this subsection presents the summary of the work on the regression studies to understand the socioeconomic impact on the e-commerce warehouses locating.

The final regression model represents the culmination of several iterative attempts to refine the relationships between warehouse density and key socioeconomic and spatial factors. Previous models explored transformations and interactions among variables, focusing on the role of population density, artificial surfaces, GWP of road infrastructure, and other predictors of warehouse location. Initial models revealed multicollinearity issues, particularly between population density and other predictors, which prompted the introduction of logarithmic transformations to improve model stability and interpretability. Additionally, the high variance inflation factor (VIF) values in earlier models indicated the need for non-linear transformations and careful selection of independent variables, therefore, the indicators that showed multicollinearity were transformed, such as the log transformation of population density, and the scaling of the income data.

We used a zero intercept (no constant term) in the regression model to ensure that the relationship between the independent variables and warehouse density is measured relative to the origin, assuming that when all predictors are zero (as a baseline), the outcome should also be zero. This approach is typically used when it is reasonable to assume that the dependent variable (warehouse density) would be zero when all the independent variables (e.g., population density) are zero.

Model Specification:

The regression model is expressed as follows:

$$
\begin{aligned}
\text{wh\_dens\_area} = {} & \beta_1 \cdot \text{surf\_artif} + \beta_2 \cdot \log(\text{pop\_dens\_k}) \\
& + \beta_3 \cdot \text{prg\_routier} + \beta_4 \cdot \text{tx\_pauv} \\
& + \beta_5 \cdot \text{emplt\_09\_20} + \beta_6 \cdot \text{income10k} \\
& + \beta_7 \cdot \text{evol\_sau} + \beta_8 \cdot \text{AirportDist} + \epsilon
\end{aligned}
\tag{4.1}
$$

Where:

- `wh_dens_area` represents warehouse density (per 1000 km^2),
- `surf_artif` is the percentage of artificial surfaces in the area,
- `log_pop_dens_k` is the log-transformed population density (residents/km^2),
- `prg_routier` is the GWP of roads (tCO_2/capita),
- `tx_pauv` is the poverty rate,
- `emplt_09_20` is the employment growth from 2009 to 2020,
- `income10k` is the scaled income,
- `evol_sau` is the evolution of agricultural land use,
- `AirportDist` is the distance to the nearest airport,
- ϵ is the error term.

The coefficients $\beta_1, \beta_2, \ldots, \beta_8$ represent the estimates for each predictor.

Results

The regression model results were as follows:

Residuals

```
    Min      1Q  Median      3Q     Max
-4.2341 -1.1409  0.0319  1.2955  4.1943
```

Coefficients

```
    Min      1Q  Median      3Q     Max
-4.234  -1.141  0.032  1.296  4.194
```

Coefficients

```
    Min      1Q  Median      3Q     Max
-4.234  -1.141  0.032  1.296  4.194
```

Coefficients

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Variable	Estimate	Std. Error	t value	p-value
`surf_artif`	4.300	1.112	3.866	0.000 ***
`log_pop_dens_k`	1.258	0.863	1.457	0.154
`prg_routier`	1.926	0.402	4.787	0.000 ***
`tx_pauv`	-0.198	0.127	-1.558	0.128
`emplt_09_20`	-0.034	0.067	-0.506	0.616
`income10k`	-4.643	1.507	-3.081	0.004 **
`evol_sau`	0.058	0.111	0.519	0.607
`AirportDist`	0.005	0.009	0.621	0.538

Residual Standard Error and R-squared

Residual standard error: 2.049 on 35 degrees of freedom
Multiple R-squared: 0.8115, Adjusted R-squared: 0.7685
F-statistic: 18.84 on 8 and 35 DF, p-value: 1.418e-10

The regression results show that:

- **surf_artif (4.30,** $p = 0.0005$**):** Change in the artificial surfaces was a significant predictor of warehouse density, reinforcing the idea that urbanization and land development are strongly associated with warehouse presence.

- **log_pop_dens_k (1.26,** $p = 0.154$**):** While population density shows a positive relationship with warehouse density, it is not statistically significant, suggesting that its effect may be more context-dependent or weaker when controlling for other factors.

- **prg_routier (1.93,** $p < 0.001$**):** Global warming potential from road transport continue to have a highly significant and positive relationship with warehouse density, showing the critical role of the availiability of road infrastructure in logistics and warehouse location.

- **income10k (-4.64,** $p = 0.004$**):** Interestingly, higher income (scaled down) has a significant negative effect on warehouse density, indicating that areas with lower incomes may be more likely to host warehouses.

- **tx_pauv (-0.20,** $p = 0.128$**), emplt_09_20 (-0.03,** $p = 0.616$**), and AirportDist (0.005,** $p = 0.538$**):** These variables do not show significant effects, suggesting that poverty rates, employment growth, and proximity to airports do not have a strong direct influence on warehouse density in this model.

Model Performance

The R-squared value of 0.8115 indicates that approximately 81% of the variance in warehouse density is explained by the model, while the adjusted R-squared of 0.7685 accounts for the number of predictors, indicating strong model performance.

The F-statistic of 18.84 and p-value of 1.418×10^{-10} suggest that the overall model is statistically significant, meaning that the included variables collectively explain a significant amount of the variation in warehouse density.

Conclusion

This final model effectively captures the significant predictors of warehouse density, particularly, the role of change in the artificial surfaces, and the GWP of the roads. The transformations applied, such as the log transformation of population density, improved model interpretability and reduced multicollinearity, as evidenced by lower VIF values. Despite some non-significant predictors, the model provides a robust framework for understanding the spatial distribution of warehouses, particularly in urbanized and lower-income areas with strong road networks. Although no strong correlation could be determined between employment, income and warehouse location, the model did show a link suggesting that employment partly determines the location choices of e-commerce warehouses by logistics providers.

To address potential issues of variable interdependence and to further verify the relationships uncovered in the regression model, Principal Component Analysis (PCA) is employed next. It helps reducing the dimensionality of the data and capturing the core variance across independent variables. Therefore, PCA serves as a complementary tool to confirm the robustness of the regression results and ensure that the significant factors identified are not unreasonably influenced by multicollinearity or other confounding factors.

4.4 Principal Component Analysis (PCA)

4.4.1 Introduction to PCA

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of datasets while retaining the most important variance in the data. In this study, PCA is employed as a complementary method to the regression analysis to verify the results and address any potential multicollinearity issues present among the independent variables. By transforming the original variables into uncorrelated principal components, PCA helps to simplify the complex relationships between variables, providing a clearer understanding of the key factors that drive warehouse density (Johnson et al. 2014).

PCA allows us to focus on the core dimensions that capture the majority of the variance in the dataset, thus reducing redundancy and ensuring that the significant relationships identified in the regression model are not influenced by overlapping or highly correlated predictors. This method will help confirm the robustness of the regression results and provide additional insights into the underlying spatial and socioeconomic patterns that influence warehouse location (Johnson et al. 2014).

4.4.2 Key Findings

From the cumulative variance plot:

- The first 2 PCs explain about 63% of the total variance.
- The first 4 PCs explain about 85% of the total variance.
- The first 6 PCs explain about 94% of the total variance.

- PC1 is strongly influenced by pop_dens_k (0.444), and surf_artif (0.405).
- PC2 has strong negative loadings for emplt_09_20 (-0.500), dv3f_d5 (-0.445), and income10k (-0.445).

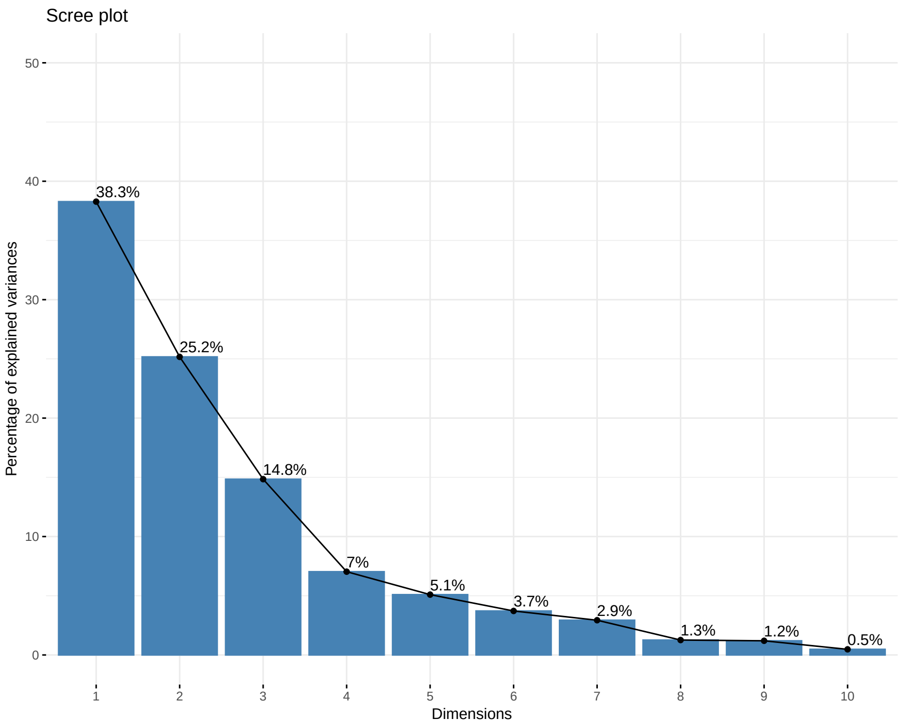


Figure 4.3: Scree plot showing the variances explained by each principal component.

- PC3 is heavily influenced by prg_routier (0.629) and negatively by evol_sau (-0.471).

This scree plot visually represents the variance captured by each principal component (PC) in the dataset. The sharp decline after the first two components indicates that these components explain most of the variance, while subsequent components contribute less significantly. The ”elbow” of the plot, typically where the slope starts to flatten (around the 3rd to the 4th component), suggests that the point where adding more components offers diminishing returns in explaining variance, meaning we can focus on the first few components for analysis.

Figure 4.4 shows the results presenting each variable contributions:

4.4.3 Discussion

The results of the Principal Component Analysis (PCA) reveal several key components that explain the majority of the variance in the dataset, which simplifies the complex relationships between the original variables. For example, PC1 shows strong positive loadings for artificial surfaces (`surf_artif`), population density (`log_pop_dens_k`), and poverty rate (`tx_pauv`), indicating that this component captures the urbanization and socioeconomic characteristics of the regions. This aligns well with the regression results, where urbanization and population density were significant predictors of warehouse density. PC2, which has strong loadings for employment growth (`emplt_09_20`) and GWP of road infrastructure (`prg_routier`), highlights the role of infrastructure and economic activity in warehouse location selection. These findings further support the regression model’s identification of `prg_routier` as a key factor influencing warehouse density.

The PCA biplot helps in visualizing the relationships between the variables based on their contributions to the principal components (PCs). Variables that are clustered together (closely positioned on the plot) share a strong relationship, indicating that they move together or exhibit similar patterns in the data. Figure 4.5 shows several distinct clusters:

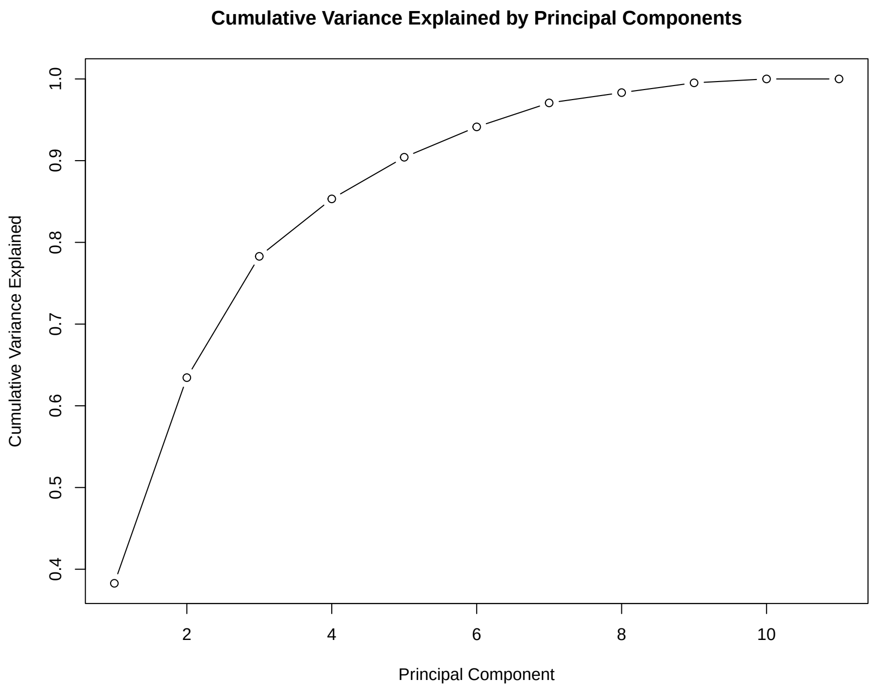


Figure 4.4: Cumulative Variance Explained by Principal Components.

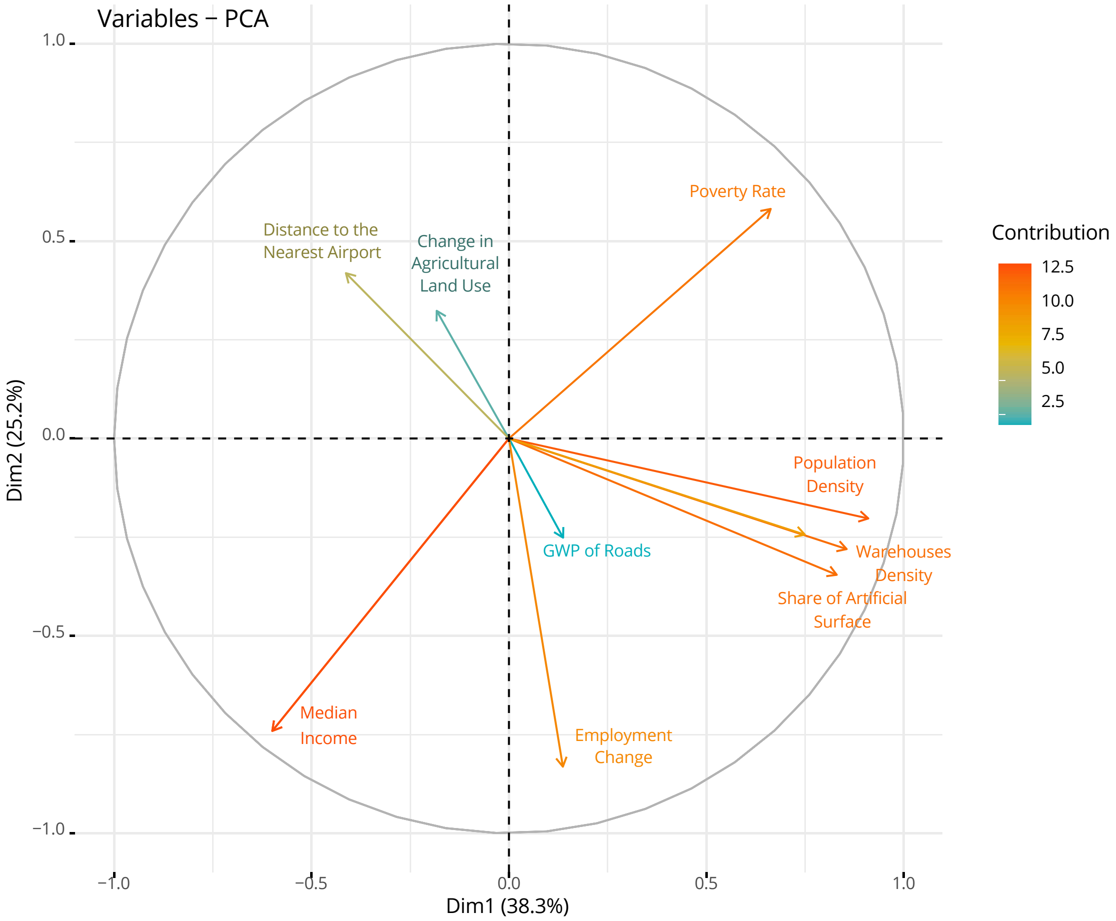


Figure 4.5: The Clustering of the variables by Principal Components.

- Cluster of Urban Variables: Variables like population density, warehouse density, and artificial surfaces are closely aligned, showing that they contribute similarly to the first principal component (Dim1). This suggests that urbanization factors, such as densely populated areas and the extent of developed land, are strongly correlated with the presence of warehouses. These variables were also identified as significant in the regression model, supporting the idea that urbanized areas with higher population density and land development are key drivers of warehouse location selection.

- Cluster of Socioeconomic Variables (Bottom Left): Median income and poverty rate are negatively correlated with the urbanization variables, as seen in their opposite directions on the plot. This inverse relationship makes sense in reality—areas with lower income levels often attract more warehouses, as indicated by the regression model's findings. These socioeconomic variables highlight a disparity between more affluent areas and those that may serve as logistical hubs due to lower land costs or proximity to transportation networks.

- Infrastructure and Environmental Variables (Top Left): Road infrastructure and employment growth, which cluster together, suggest that regions with better road infrastructure and higher employment growth are important for warehouse location selection. Additionally,Distance to the nearest airport and utilized agricultural area variables are less tightly clustered with the main urbanization variables, showing that they have less influence on warehouse locations compared to factors like the GWP of road networks or even urban density.

This analysis showed that the first four components cumulatively explain a significant portion of the dataset's variance, as shown in the scree plot and the cumulative variance plot. As a result, the ten variables were dimensionally reduced to four components, which can be seen in table 4.3. The table shows the correlation of each variable with the component.

Variables	**PC1**	**PC2**	**PC3**	**PC4**
Artificilized Surface	0.40	-0.21	-0.01	0.08
Population Density	0.44	-0.12	-0.13	0.12
GWP of Roads	0.07	-0.15	0.63	0.31
Poverty Rate	0.32	0.35	-0.09	-0.24
Change in Employed People	0.07	-0.50	0.02	-0.24
Evolution of SAU	-0.09	0.19	-0.47	0.76
Dist. to nearest Airport	-0.20	0.25	0.43	0.12
Median Income	-0.29	-0.45	-0.12	0.09

Table 4.3: Principal Component Analysis Loadings for the Top 4 Components

4.4.3.1 Verification of Regression Results

The PCA results confirm the regression findings by visually reinforcing the relationships between key variables. The clustering of urbanization, socioeconomic, and infrastructure variables in the PCA biplot aligns well with the variables that were statistically significant in the regression model. The PCA effectively verifies that factors like urban development and road infrastructure are critical to understanding warehouse density, while socioeconomic factors such as income play an important, yet inverse, role. This consistency between the two analyses strengthens the overall conclusions about the drivers of warehouse location selection.

Overall, the PCA complements the regression analysis by confirming that the most significant predictors—urbanization, infrastructure, and socioeconomic factors—are consistently important in explaining warehouse density. The principal components provide a clearer representation of these factors by reducing the complexity of the data, ensuring that the regression results are not affected by multicollinearity or overlapping variables. In this way, PCA validates the robustness of the regression model while offering additional insights into the spatial and socioeconomic dimensions of warehouse location selection.

4.5 Conclusion

4.5.1 Summary of Statistical Insights

The key findings from both the Multiple Linear Regression analysis and Principal Component Analysis (PCA) highlight several significant variables influencing warehouse location selection. In the regression analysis, change in the artificial surfaces (`surf_artif`) and GWP of road (`prg_routier`) emerged as the most significant predictors, indicating that urbanization and the development of transport networks are crucial factors in determining where warehouses are located. Additionally, income (`income10k`) showed a significant negative relationship with warehouse density, suggesting that warehouses are more likely to be found in lower-income regions.

The PCA confirmed these results by identifying principal components that captured the same patterns. PC1, for example, was heavily influenced by variables related to urbanization and population density, while PC2 reflected the impact of infrastructure and economic activity, aligning with the significance of `prg_routier` from the regression. Overall, both analyses point to the conclusion that warehouse location is largely driven by urbanization, infrastructure availability, and socioeconomic conditions. These complementary findings reinforce the importance of considering these variables in logistics and spatial planning decisions.

4.5.2 Implications for Policy and Practice

The findings from this analysis have significant implications for urban planning, logistics, and e-commerce practices. The clear relationship between urbanization, road infrastructure, and warehouse density suggests that policymakers need to prioritize infrastructure investments in regions where logistical hubs are growing or anticipated to grow. Urban planners can use these insights to better manage land use, ensuring that warehousing facilities are efficiently integrated into urban and peri-urban areas without overwhelming local communities or ecosystems. Additionally, the negative correlation between income and warehouse density raises equity concerns, signaling a need for policies that balance economic opportunities in lower-income areas with potential environmental and social costs.

In logistics, companies can leverage this analysis to optimize warehouse placement, ensuring proximity to key infrastructure while considering regional socioeconomic factors. For future improvements in warehousing analysis, more refined data collection—potentially through public-private partnerships to share real-time logistics data—can enhance decision-making and planning processes.

4.5.3 Future Research Directions

Future research could explore incorporating real-time data and more advanced spatial econometric models to further refine the understanding of warehouse location selection.

Moreover, real-time data, such as traffic patterns, land use changes, and e-commerce activity, could significantly improve predictive modeling. By leveraging dynamic data sources, researchers can better capture the evolving nature of logistics infrastructure and its interaction with urban and regional growth.

This approach could provide insights into how quickly warehouses emerge in response to shifts in consumer demand or transportation developments. Furthermore, integrating real-time data with spatial econometric models, such as Spatial Lag Models (SLM) and Spatial Error Models (SEM), would allow for a deeper examination of spatial dependencies. (Reda et al. 2023) For example, the presence of a warehouse in one area might affect the likelihood of other warehouses being established in neighboring regions, indicating a spatial spillover effect. This could also help better capture the influence of regional characteristics and how proximity to infrastructure, such as roads or ports, affects the concentration of warehouses.

Additionally, methods like Geographically Weighted Regression (GWR) offer promising avenues for future research. GWR allows relationships between variables, such as population density or road emissions, to vary across space, which could be useful in understanding how different regions react to logistics demands. For instance, while road infrastructure might be a key driver of warehouse density in some regions, other areas might be more influenced by socioeconomic factors such as income or employment rates. Using GWR would allow researchers to pinpoint these regional differences, offering a more nuanced understanding of e-commerce warehousing dynamics. Moreover, incorporating spatial lag into studies of socioeconomic impacts could shed light on how the benefits of warehousing spread to adjacent areas, providing policymakers with more robust tools for fostering balanced regional development. This would ensure that logistics planning and urban development strategies are informed by comprehensive, real-time, and spatially sensitive analyses.

Chapter 5

Conclusions and Perspectives

5.1 Contributions

This research on the e-commerce warehouses location selection and their potential socioeconomic impacts in metropolitan France provides several significant contributions to the academic and practical understanding of logistics activities. First, the methodological approaches developed, particularly through manual and semi-automated techniques, have resulted in a usable database of 133 e-commerce warehouses. This work addresses key gaps in the literature by offering different potential frameworks for identifying and mapping warehouse locations, using a combination of manual verification and spatial extraction methods.

Additionally, the findings from the regression analysis and Principal Component Analysis (PCA) offer new insights into how urbanization, employment rates, income levels, and transportation infrastructure influence warehouse distribution patterns. The identification of key indicators—such as population density, road GHG emissions, and artificial surfaces—further enhances our understanding of warehouse location selection drivers, supporting both academic discourse and urban planning initiatives. Finally, the research provides a foundation for further investigations into the socioeconomic implications of logistics activities, particularly in urban and peri-urban regions, while also demonstrating the utility of detailed, geolocated warehouse data for spatial econometric analysis.

5.2 Perspectives

To extend this study, some research works could be developed for expanding upon this work. Incorporating real-time data sources, such as transportation flows and labor market dynamics, as well as the political layer, could provide a more dynamic understanding of warehouse location selection, allowing for continuous updates to the database and more responsive analyses. Additionally, exploring advanced spatial econometric models such as *Geographically Weighted Regression (GWR)* and spatial lag models could offer deeper insights into the spatial dependence and heterogeneity in warehouse distribution patterns. These techniques would be especially valuable for identifying localized factors that influence warehouse siting decisions and for refining policy recommendations aimed at balancing logistics needs with socioeconomic goals.

Expanding the geographic scope beyond metropolitan France and integrating cross-border logistics dynamics could also provide a broader perspective on the European logistics network. Moreover, there is potential to integrate environmental sustainability metrics into future studies, examining the environmental impacts of warehousing activities and exploring the role of green logistics practices in e-commerce supply chains.

Bibliography

Actuia (2022). *IGN Relies on AI and Deep Learning to Enrich Land Use Data.* URL: https://www.actuia.com/english/ign-relies-on-ai-and-deep-learning-to-enrich-land-use-data/.

Alamri, S. (2024). "The Geospatial Crowd: Emerging Trends and Challenges in Crowdsourced Spatial Analytics". In: *ISPRS International Journal of Geo-Information* 13.6. ISSN: 2220-9964. DOI: 10.3390/ijgi13060168. URL: https://www.mdpi.com/2220-9964/13/6/168.

Aljohani, K. et al. (2016). "Impacts of logistics sprawl on the urban environment and logistics: Taxonomy and review of literature". In: *Journal of Transport Geography* 57, pp. 255–263. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2016.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0966692316300692.

Allen, J. et al. (Sept. 2012). "Investigating relationships between road freight transport, facility location, logistics management and urban form". In: *Journal of Transport Geography.* Special Section on Theoretical Perspectives on Climate Change Mitigation in Transport 24, pp. 45–57. ISSN: 0966-6923. DOI: 10.1016/j.jtrangeo.2012.06.010. URL: https://www.sciencedirect.com/science/article/pii/S0966692312001615 (visited on 05/14/2024).

Anselin, L. (2003). "Spatial externalities, spatial multipliers, and spatial econometrics". In: *International Regional Science Review* 26.2, pp. 153–166.

Batty, M (2005). *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals.* Cambridge, MA: The MIT Press.

— (2018). "Digital Twins". In: *Environment and Planning B: Urban Analytics and City Science* 45.5, pp. 817–820. DOI: 10.1177/2399808318796416. URL: https://doi.org/10.1177/2399808318796416.

Besnard, G. (2023). *Baromètre de l'audience du e-commerce: 1er trimestre 2023 - Fevad, la Fédération du e-commerce et de la vente à distance.* https://www.fevad.com/1er-trimestre-2023-barometre-de-laudience-du-e-commerce/. Fevad, La Fédération Du E-commerce Et De La Vente À Distance.

Bettencourt, L. M. A. (2013). "The Origins of Scaling in Cities". In: *Science* 340.6139, pp. 1438–1441. DOI: 10.1126/science.1235823.

Bowen, J. T. (2008). "Moving places: the geography of warehousing in the US". In: *Journal of Transport Geography* 16.6. Growing Public Transport Patronage, pp. 379–387. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2008.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0966692308000185.

Boysen, N. et al. (2019). "Warehousing in the e-commerce era: A survey". In: *European Journal of Operational Research* 277.2, pp. 396–411. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2018.08.023. URL: https://www.sciencedirect.com/science/article/pii/S0377221718307185.

Carlino, G. A. et al. (1987). "The Determinants of County Growth". In: *Journal of Regional Science* 27.1, pp. 39–54. DOI: 10.1111/j.1467-9787.1987.tb01142.x.

Cidell, J. (2010). “Concentration and decentralization: The new geography of freight distribution in US metropolitan areas”. In: *Journal of Transport Geography* 18.3. Tourism and climate change, pp. 363–371. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2009.06.017. URL: https://www.sciencedirect.com/science/article/pii/S0966692309001008.

— (2012). “Flows and Pauses in the Urban Logistics Landscape: The Municipal Regulation of Shipping Container Mobilities”. In: *Mobilities* 7.2, pp. 233–245. DOI: 10.1080/17450101.2012.654995.

Collective of Associations (Sept. 20, 2022). *Avis critique sur le projet Green Dock : Expertise par les associations.* fr. Technical Report. Ville de Gennevilliers. URL: https://www.ville-gennevilliers.fr/fileadmin/www.ville-gennevilliers.fr/MEDIA/actualites/urbanisme/green-dock/avis-critique-green-dock_collectif-associations.pdf (visited on 11/06/2024).

Craig, B. A. et al. (2016). *Introduction to the Practice of Statistics.* 8th. W.H. Freeman and Company.

Cresswell, T. (2006). *On the Move: Mobility in the Modern Western World.* Routledge.

Dablanc, L. (2007). “Goods transport in large European cities: Difficult to organize, difficult to modernize”. In: *Transportation Research Part A: Policy and Practice* 41.3, pp. 280–285. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2006.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0965856406000590.

— (Jan. 2014). “Logistics Sprawl and Urban Freight Planning Issues in a Major Gateway City: The Case of Los Angeles”. In: *Urban Studies.* EcoProduction, pp 49–69. DOI: 10.1007/978-3-642-31788-0_4. URL: https://hal.science/hal-00943491.

Dablanc, L., A. Heitz, et al. (2020). “The evolution of logistics facilities: A spatial analysis of warehousing in Paris”. In: *Urban Studies* 57.10, pp. 2105–2122.

Dablanc, L., S. Ogilvie, et al. (2014). “Logistics Sprawl: Differential Warehousing Development Patterns in Los Angeles, California, and Seattle, Washington”. In: *Transportation Research Record* 2410.1, pp. 105–112. DOI: 10.3141/2410-12. eprint: https://doi.org/10.3141/2410-12. URL: https://doi.org/10.3141/2410-12.

Dablanc, L. and D. Rakotonarivo (2010). “The impacts of logistics sprawl: How does the location of parcel transport terminals affect the urban environment?” In: *Procedia - Social and Behavioral Sciences* 2.3, pp. 6087–6096.

Dablanc, L., M. Schorung, et al. (2023). *Locational patterns of warehouses in 78 cities around the world, a comparative meta-analysis.* Tech. rep. Report for Logistics City Chair, University Gustave Eiffel. URL: https://hal.science/hal-04487271.

Doshi-Velez, F. et al. (2017). “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv: Machine Learning.* URL: https://api.semanticscholar.org/CorpusID:11319376.

Florida, R. (2002). *The Rise of the Creative Class: And How It’s Transforming Work, Leisure, Community and Everyday Life.* New York: Basic Books.

Gehlke, C. E. et al. (1934). “Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material”. In: *Journal of the American Statistical Association* 29.185, pp. 169–170.

Glaeser, E. L. et al. (2009). “The Economics of Place-Making Policies”. In: *Brookings Papers on Economic Activity* 2009.1, pp. 155–239. DOI: 10.1353/eca.0.0041.

Harvey, D. (2001). *Spaces of Capital: Towards a Critical Geography.* Routledge.

Hastie, T. et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd. New York: Springer. ISBN: 978-0-387-84857-0.

Heitz, A. (2017). “Logistics spatial patterns in Paris: Understanding urban freight through warehouse locations”. In: *Urban Studies* 54.6, pp. 1232–1250.

Heitz, A., L. Dablanc, et al. (2017). “Logistics sprawl in monocentric and polycentric metro- politan areas: the cases of Paris, France, and the Rand- stad, the Netherlands”. In: *Region: the journal of ERSA* 4.1, pp. 93–107. DOI: 10.18335/region.v4i1.158. URL: https://hal.science/hal-02614568.

Heitz, A., P Launay, et al. (2017). “Rethinking Data Collection on Logistics Facilities: New Approach for Determining the Number and Spatial Distribution of Warehouses and Terminals in Metropolitan Areas”. In: *Transportation Research Record* 2609.1, pp. 67–76. DOI: 10.3141/2609-08. URL: https://doi.org/10.3141/2609-08.

— (2019). “Heterogeneity of logistics facilities: an issue for a better understanding and planning of the location of logistics facilities”. In: *European Transport Research Review* 11.5. DOI: 10.1186/s12544-018-0341-5. URL: https://doi.org/10.1186/s12544-018-0341-5.

Hesse, M. et al. (2004). “The transport geography of logistics and freight distribution”. In: *Journal of Transport Geography* 12, pp. 171–184. URL: https://api.semanticscholar.org/CorpusID:15397238.

Hesse, Markus (2008). *The City as a Terminal: The Urban Context of Logistics and Freight Tra*. Routledge. ISBN: 978-1-138-25520-3. URL: https://www.routledge.com/The-City-as-a-Terminal-The-Urban-Context-of-Logistics-and-Freight-Transport/Hesse/p/book/9781138255203 (visited on 05/13/2024).

IGN (2022). *L’IA pour une description plus rapide de l’occupation du sol*. URL: https://www.ign.fr/espace-presse/l-ia-pour-une-description-plus-rapide-de-loccupation-du-sol.

INSEE (2020). *Aire d’attraction d’une ville*. https://www.insee.fr/fr/metadonnees/definition/c1804. Accessed: 2024-09-03.

Institut Paris Région (2024). *Mode d’occupation du sol (MOS)*. Accessed: 2024-11-07. URL: https://www.institutparisregion.fr/mode-doccupation-du-sol-mos/.

Jacobs-Crisioni, C. et al. (2014). “The impact of spatial aggregation on urban development analyses”. In: *Applied Geography* 47, pp. 46–56. DOI: 10.1016/j.apgeog.2013.11.

Johnson, R. A. et al. (2014). *Applied Multivariate Statistical Analysis*. 7th. Pearson.

Jomthanachai, S. et al. (2022). “An application of machine learning regression to feature selection: a study of logistics performance and economic attribute”. In: *Neural Computing and Applications* 34, pp. 15781–15805. DOI: 10.1007/s00521-022-07266-6.

Kwan, Mei-Po (2012). “The Uncertain Geographic Context Problem”. In: *Annals of the Association of American Geographers* 102.5, pp. 958–968. DOI: 10.1080/00045608.2012.687349. eprint: https://doi.org/10.1080/00045608.2012.687349. URL: https://doi.org/10.1080/00045608.2012.687349.

Lefebvre, H. (1991). *The Production of Space*. Blackwell Publishing.

Li, Y. et al. (2018). “Deep learning for remote sensing image classification: A survey”. In: *WIREs Data Mining and Knowledge Discovery* 8.6, e1264. DOI: https://doi.org/10.1002/widm.1264. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1264. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1264.

Moraga, P. (2023). *Spatial Statistics for Data Science: Theory and Practice with R*. en. Data Science Series. Chapman and Hall. URL: https://freecomputerbooks.com/Spatial-Statistics-for-Data-Science-Theory-and-Practice-with-R.html (visited on 08/05/2024).

Oliveira, L. K. de et al. (2022). “An investigation of contributing factors for warehouse location and the relationship between local attributes and explanatory variables of Warehouse Freight Trip Generation Model”. In: *Transportation Research Part A: Policy and Practice* 162, pp. 206–219. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2022.05.025.

Oliveira, R. de et al. (2022). "Changes in warehouse spatial patterns and rental prices: Are they related? Exploring the case of US metropolitan areas". In: *Journal of Transport Geography* 104, p. 103450. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2022.103450. URL: https://www.sciencedirect.com/science/article/pii/S0966692322001739.

Openshaw, S (1983). *The modifiable areal unit problem.* Norwich: Geo Books.

Ouattara, Adama et al. (2021). "Big data econometrics: Logistic regression on Apache Spark". In: *arXiv.* URL: https://arxiv.org/abs/2106.10341.

Pajić, V. et al. (2024). "Strategic Warehouse Location Selection in Business Logistics: A Novel Approach Using IMF SWARA–MARCOS—A Case Study of a Serbian Logistics Service Provider". In: *Mathematics* 12.5. ISSN: 2227-7390. DOI: 10.3390/math12050776. URL: https://www.mdpi.com/2227-7390/12/5/776.

Raimbault, N. et al. (2012). "Understanding the Diversity of Logistics Facilities in the Paris Region". In: *Procedia - Social and Behavioral Sciences* 39. Seventh International Conference on City Logistics which was held on June 7- 9,2011, Mallorca, Spain, pp. 543–555. ISSN: 1877-0428. DOI: https://doi.org/10.1016/j.sbspro.2012.03.129. URL: https://www.sciencedirect.com/science/article/pii/S1877042812005964.

Rao, C. et al. (2015). "Location selection of city logistics centers under sustainability". In: *Transportation Research Part D: Transport and Environment* 36, pp. 29–44. ISSN: 1361-9209. DOI: https://doi.org/10.1016/j.trd.2015.02.008. URL: https://www.sciencedirect.com/science/article/pii/S1361920915000176.

Reda, A. K. et al. (2023). "Modelling the effect of spatial determinants on freight (trip) attraction: A spatially autoregressive geographically weighted regression approach". In: *Research in Transportation Economics* 99, p. 101296. ISSN: 0739-8859. DOI: https://doi.org/10.1016/j.retrec.2023.101296. URL: https://www.sciencedirect.com/science/article/pii/S0739885923000367.

Rodrigue, J-P. (2024). *The Geography of Transport Systems.* 6th. New York: Routledge, p. 402. ISBN: 9781032380407. DOI: 10.4324/9781003343196.

Russo, A. (June 2024). *Le projet Green Dock de Gennevilliers concentre la contestation contre les entrepôts logistiques.* URL: https://www.lenouveleconomiste.fr/le-projet-green-dock-de-gennevilliers-concentre-la-contestation-contre-les-entrepots-logistiques-114257/.

Russo, F. et al. (2011). "Measures for Sustainable Freight Transportation at Urban Scale: Expected Goals and Tested Results in Europe". In: *Journal of Urban Planning and Development* 137.2, pp. 142–152. DOI: 10.1061/(ASCE)UP.1943-5444.0000052.

Sakai, T. et al. (2015). "Locational dynamics of logistics facilities: Evidence from Tokyo". In: *Journal of Transport Geography* 46, pp. 10–19. ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2015.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0966692315000769.

Schorung, M. et al. (2022). *Atlas of Warehouse Geography in the US.*

Sheffi, Y. (2012). *Logistics Clusters: Delivering Value and Driving Growth.* The MIT Press. ISBN: 9780262526791. URL: http://www.jstor.org/stable/j.ctt5vjqnb.

Talwar, S. et al. (2021). "Big Data in Operations and Supply Chain Management: A Systematic Literature Review and Future Research Agenda". In: *International Journal of Production Research* 59.11, pp. 3509–3534. DOI: 10.1080/00207543.2020.1868599. eprint: https://doi.org/10.1080/00207543.2020.1868599. URL: https://doi.org/10.1080/00207543.2020.1868599.

Waller, M. A. et al. (2013). "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management". In: *Journal of Business Logistics* 34.2, pp. 77–84. DOI: https://doi.org/10.1111/jbl.12010.

eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbl.12010. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jbl.12010.

Woudsma, C. et al. (2008). “Logistics land use and the city: A spatial–temporal modeling approach”. In: *Transportation Research Part E: Logistics and Transportation Review* 44.2. Selected Papers from the National Urban Freight Conference, pp. 277–297. ISSN: 1366-5545. DOI: https://doi.org/10.1016/j.tre.2007.07.006.

Xiao, Z. et al. (Mar. 2021). “New paradigm of logistics space reorganization: E-commerce, land use, and supply chain management”. In: *Transportation Research Interdisciplinary Perspectives* 9, p. 100300. ISSN: 2590-1982. DOI: 10.1016/j.trip.2021.100300. URL: https://www.sciencedirect.com/science/article/pii/S259019822100004X (visited on 09/03/2024).

Zhang, S. et al. (2021). “Logistics service supply chain order allocation mixed K-Means and Qos matching”. In: *Procedia Computer Science* 188. CQVIP Conference on Data Driven Intelligence and Innovation, pp. 121–129. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.05.060. URL: https://www.sciencedirect.com/science/article/pii/S1877050921011455.