****  Programme Grande Ecole

MSc Business Analytics

**MSC DISSERTATION**

**Enhancing Public Transportation Services through Data Analytics and Artificial Intelligence**

PLATEAU Vincent

Head of MSc Programme: POPOVIČ Aleš

Dissertation Supervisor: POPOVIČ Aleš

Date: 30/06/2024

****

**DECLARATION OF ACADEMIC INTEGRITY**

I, undersigned, M. PLATEAU Vincent, student enrolled in the MSc Business Analytics certify that the data and information contained in the Dissertation entitled “Enhancing Public Transportation Services through Data Analytics and Artificial Intelligence” have not been plagiarised.

Paris,

Date 30/06/2024

Signature

Une image contenant croquis, dessin

Description générée automatiquement

**Acknowledgement**

I would like to express my gratitude to all those who supported me throughout my academic journey, especially during my time at NEOMA Business School.

I would like to thank all the teachers at the school and especially those of the MSc Business Analytics.

Lastly, I am grateful to my parents, my family and those who supported me.

Table of contents

[**Abstract** 5](#_Toc170675882)

[**1 - Introduction** 6](#_Toc170675883)

[**2 - Topic** 7](#_Toc170675884)

[**3 - Literature review** 10](#_Toc170675885)

[**A - Operational efficiency improvements through AI** 10](#_Toc170675886)

[1- Predictive Maintenance and Resource optimization 10](#_Toc170675887)

[2- Traffic Management and scheduling 11](#_Toc170675888)

[3- Efficiency in service delivery 12](#_Toc170675889)

[**B – Focus on DA and some innovative concepts** 13](#_Toc170675890)

[1- Understanding and learning: DA for decision-makers 13](#_Toc170675891)

[2- Innovative Concepts: Data Visualization, Dynamic Pricing and Mobility as a Service 14](#_Toc170675892)

[**C – Social media analytics: a tool to better understand user expectations** 16](#_Toc170675893)

[1- Decoding passenger behaviour via social media 16](#_Toc170675894)

[2- Analysing passengers’ sentiments 17](#_Toc170675895)

[**4 - Research Question and Hypothesis** 20](#_Toc170675896)

[**5 - Method** 21](#_Toc170675897)

[**A - Data collection** 21](#_Toc170675898)

[**B – Data description and preprocessing steps** 23](#_Toc170675899)

[**6 – Results** 24](#_Toc170675900)

[**A - Analysis of ticket validation data** 24](#_Toc170675901)

[1- Number and patterns of ticket validations 24](#_Toc170675902)

[2- Spatiotemporal analysis 26](#_Toc170675903)

[3- Clustering analysis 27](#_Toc170675904)

[**B – Sentiment analysis of Twitter/X data** 30](#_Toc170675905)

[1- Sentiment detection in tweets 30](#_Toc170675906)

[2- Topics covered in tweets 32](#_Toc170675907)

[**7 – Discussion** 35](#_Toc170675908)

[**A - Interpretation of the findings** 35](#_Toc170675909)

[**B - Conclusions** 36](#_Toc170675910)

[**8 - Recommendations** 37](#_Toc170675911)

[**References** 38](#_Toc170675912)

# **Abstract**

Public transport is at the heart of our societies and faces the challenges of our time: adapting to new post-covid transport habits or positioning itself as an alternative to the private car in cities. This study examines how Data Analytics (DA) and Artificial Intelligence (AI) can improve the operational efficiency and customer satisfaction of the Parisian public transportation network managed by RATP. By analyzing ticket validation and social media data, the study proposes a mixed approach, with several concrete applications of these technologies and the identification of key areas for improvement. Temporal patterns in ticket validation data highlight the need for dynamic resource allocation, while sentiment analysis of Twitter data and topic modeling reveal recurring areas of dissatisfaction providing actionable insights for RATP. Despite the promising findings, the study acknowledges limitations such as some difficulties in sentiment analysis and the potential bias in social media data.

**Keywords:** Artificial Intelligence (AI); Data Analytics (DA);Public Transportation (PT)**;** Sentiment Analysis**;** RATP

# **1 - Introduction**

Public transports (PTs) are the cornerstone of the economic vitality of our societies. These are the veins that carry thousands of people to their places of work every day. These transport systems are at the heart of some challenges facing our century such as sustainable development or social inclusiveness. Taxpayers and societies expect public transport systems to be efficient and effective, increasing the pressure on transport agencies. But the sector often struggles with issues of operational inefficiency, service reliability or customer dissatisfaction. Facing these challenges requires innovative solutions to manage and improve public transportations systems. In this context, Data Analytics (DA) and Artificial Intelligence (AI) emerge as transformative and disruptive technologies that have the potential to revolutionize passenger experience and service management.

The recent advancements in AI and machine learning (ML) coupled with the exponential growth of data generated is a real opportunity to enhance the transport industry and specifically public transportation systems. According to a recent study (Impact of AI as a Percentage of Industry Revenues, 2023), on an annual basis, generative AI as the potential to add between 1,2% and 2% of the total industry revenue in the transport sector. And another study claims that AI adoption will result in a 44% increase in profitability in the transport and storage industry and 27% for public services (Accenture Report, 2017). But increased revenue thanks to AI is often only a side-effect of its implementation. In very concrete terms, AI in PT (supported by more traditional data analysis techniques) can be primarily used for traffic monitoring and management, predictive maintenance, public safety, fare management, demand prediction or better understanding and consideration of user behaviour and opinions, just to name a few. Despite these advancements, it seems that AI applications for improving PT in the real world are still relatively rarely implemented, particularly if we are looking at certain areas of application such as analysing data from multiple sources in real-time (or near real-time), from social networks for instance.

The rise of social media platforms provides an undeniable source of data that is still under-used in the PT sector. Some studies like Gal-Tzur et al. (2014) have demonstrated the potential of these sources in capturing public opinions and experiences related to transport, which can inform long-term policy development and operational decisions. A platform such as Twitter, renamed X[[1]](#footnote-1), alone has more than 600 million active users per month, including 15 million in France (April 2024) (Biggest Social Media Platforms 2024; X/Twitter: Global Audience 2024). The capabilities of AI to process and analyse such important and voluminous sources of data could lead to more passenger-centric service delivery. Several studies have employed data from Twitter[[2]](#footnote-2) successfully: for instance Gu et al. (2016) have applied ML algorithms to detect traffic incident in real time, underscoring the feasibility of leveraging social media insights in the context of PT. While Mehri et al. (2023) used natural language processing (NLP) techniques to analyse tweets and understand user perception and satisfaction, which is a less expensive method and without the limitations associated with traditional surveys.

While there is a significant interest in the potential of AI and DA to transform public transportation, there is a need for empirical studies that explore their practical applications and impacts. This research aims to bridge this gap by investigating how advanced text mining, combined with the analysis of operational historical data about tickets validations, can improve customer satisfaction and operational efficiency in PT systems. It will not only contribute to the academic literature by providing a detailed analysis of AI and DA's practical impacts on this sector but will also offer concrete examples and actionable insights for policymakers and transport authorities looking to implement these technologies.

For study purposes, the public transport company RATP (*Régie Autonome des Transports Parisien* or the Paris Metropolitan transit operator) was chosen, and the area studied is the Ile-de-France region (which includes Paris and many surrounding areas). RATP is the historic transport agency in Paris and has acquired valuable knowledge and expertise since the opening of the first bus lines and the first metro line at the beginning of the last century. Over the decades, RATP has been at the forefront of numerous innovations in urban mobility. It pioneered the automation of the Paris Metro with the introduction of the fully automated Metro Line 14 in 1998, which represented a significant leap forward in transport technology and efficiency. The company not only operates the Paris Metro but also manages bus services, tram lines and parts of the regional express network (RER) that it jointly manages with the SNCF (*Société Nationale des Chemins de fer Français*). The company consistently invested in infrastructure upgrades, in the expansion of the network, and modernizing rolling stock to improve capacity, safety and environmental sustainability.

# **2 - Topic**

Today, the RATP group is present in 15 countries on 5 continents. The company manages 9 different modes of transport, from trams and buses to sightseeing, cable cars and autonomous vehicles. It is the world's third-largest urban transport operator, with a global market share of 4% among PT providers according to Public Transportation - Global | Statista Market Forecast, (n.d.). The same report describes a highly fragmented market, with most companies initially attached to one town or region. PTs are also sensitive to several macro-economic factors and trends that will influence their development. As the Statista analysts point out, one of the key trends that will influence the modal shift towards PT is investment in 'smart transportation systems', which will use technology to improve the service. The future also lies in multimodality, the ability for travellers to easily change modes of transport, and AI is part of the solution, as we shall see in this study. This change in consumer habits is also reflected in the way tickets are bought. By 2028, online sales are expected to contribute to 25% of the total revenue in the PT market, and this increased online activity is a promising source of data.

Overall, demand will increase. The Ile-de-France region is forecasting a 15% increase in journeys by public transport, with a lot of money being invested in modernising the existing network, financing new lines and developing cycle lanes (La Région Île-de-France arrête le projet de Plan des mobilités en Île-de-France 2030, 2024). In its latest annual report (Rapport Annuel 2023 | Groupe RATP, 2024, p.7), RATP itself expects the number of people living in urban areas to rise from 56% of the world’s population in 2020 to 68% in 2050. On a global scale, according to Statista Mobility Market Insights, (2023) in their market study “Public Transportation Worldwide”, despite a drop in PT use as a result of the Covid-19 pandemic, forecasts seem to indicate a return to pre-pandemic passenger numbers (4.16 billion users in 2019, 4.31 billion expected in 2023), followed by an increase to 4.49 billion users in 2027. And if we focus on specific transport mode such as the bus, the passenger travel by bus in cities around the world is expected to nearly double with some markets with a strong growth in demand such as Sub-Saharan Africa. For urban passenger travel demand for rail travel, global demand is expected to rise sharply, reaching 2.3 trillion passenger-kilometres in 2050 compared with 1.1 trillion in 2022, the South South-West Asian market could even see demand rise by 262% over the same period. All these figures take us to the conclusion that the future of this industry is promising, growing demand, but also increasing expectations from users. RATP, which has already exported its know-how internationally, could benefit even more from pursuing its development in promising markets.

Locally, however, the challenges are still considerable. RATP sometimes struggles to meet service requirements. There have been numerous service disruptions and other operational challenges which, amplified by the pandemic, have caused driver shortages and frequent service interruptions. For users, this means overcrowded transport, and therefore dissatisfaction. Maintenance, repair, and improvement projects also sometimes affect users in their daily lives: a concrete example is the extension of line 14 for the 2024 Olympic Games (OG) or work on the RER B lines. These projects are the result of several factors, including the ageing infrastructure of the Paris metro and preparations for peak passenger periods such as the OG with a forecast of 1 million additional passengers per day during the event (The 10 Problems with Paris Transport System France’s Ex-PM Must Deal With, 2022).

Another area of concern is energy consumption, which is part of resource optimization strategy. It accounts for a significant proportion of the operating costs of transport companies, and in 2023 the rise in energy prices had a negative impact on RATP's finances. In the 2023 annual report, the company emphasizes the "unfavourable macro-economic context", an inflationary environment and "a sharp rise in operating costs, in particular [...] energy bills (electricity, gas, fuel, etc.)". In a context of growing ecological awareness, RATP is seeking to be more economical and ecological, with an aim to reduce costs but also to meet users' expectations in terms of ecology. In 2023, the company will have saved 21 GWh of energy consumption on its rail networks and reduced energy consumption in its buildings by 13% compared to 2019.(Rapport Annuel 2023 | Groupe RATP, 2024).

RATP must therefore face up to multiple constraints to remain the leader in its own fiefdom, the Paris region. The satisfaction of Parisian transport users is now a priority for the group, which also places it at the top of the list of challenges facing its business model. However, again according to the 2023 annual report, in their "CSR scorecard", we note that user satisfaction is following a negative trend, with a satisfaction rate of 85.9% in 2023. It is therefore urgent for the company to adopt measures to reverse this trend, especially as PT in the Paris region is being opened up to competition (La mise en concurrence des lignes du réseau francilien, 2020). RATP is gradually losing the monopoly it had on bus routes in and around Paris, and on the metro and RER in the longer term. It means increased financial constraints and the need to continue to meet users' expectations, which can be done by improving the PT user experience and operational efficiency.

The solution to the importance of the issues at stake may lie in the formidable technological boom we have been experiencing for several years, particularly with AI. A broad definition of AI would be that it’s a technology that try to simulate human intelligence and problem-solving skills. The term AI encompasses various concepts and domains such as ML, Deep Learning (DL), NLP, and generative AI. These kinds of algorithms are made to analyse data, recognize patterns, and ultimately take decisions alone (What Is Artificial Intelligence (AI)?|IBM, 2021). In the other hand, the concept of DA means to extract some useful insights from raw data to solve business problems (Provost & Fawcett, 2013). It’s a set of techniques, it aims to transform data into actionable knowledge. We consider only certain types of analysis such as descriptive, predictive, or prescriptive analytics. Provost and Fawcett emphasize that sometimes the separation between the fields has blurred, and AI and DA are often complementary.

Research already provides us with numerous examples of the use of AI and DA. But before considering the use of more effective tools we also need to identify the drivers of satisfaction. Van Lierop et al. (2018) have conducted a literature review and identified several “service factors” that contribute the most to increase this satisfaction, in descending order: cleanliness, comfort, driver and personnel’s behaviour and attitudes, safety, frequency, punctuality, travel time, waiting conditions and the value/price were in the top ten. This highlights potential areas where AI and DA could and must be applied to improve user experience and operational efficiency.

# **3 - Literature review**

Scientific articles about improving public transport using data science, mathematical, and AI techniques already exist. As we have already mentioned, the impact of these technologies could revolutionize the sector, particularly in terms of how transport companies could improve their operational processes and related decision-making. The challenge is to rationalise workflows, make better use of resources and eliminate the unnecessary. Figure 1, taken from a paper reviewing the scientific literature on the applications of AI in PT, shows us that this area of research is particularly new, and that interest seems to focus primarily on improving travel services.

Une image contenant ligne, diagramme, texte, Tracé

Description générée automatiquement

Figure 1. Number of studies about AI applications in PT for different periods of years. Extract from Jevinger et al. ( 2024)(p.121).

## **A - Operational efficiency improvements through AI**

1. Predictive Maintenance and Resource optimization

One of the various methods and applications cases is predictive maintenance. It means using data mining algorithms to predict potential equipment failures before they occur, thereby reducing downtime and maintenance costs. Optimising maintenance is vital for transport operators, who are extremely dependent on the smooth running of their equipment (buses, coaches, metro systems, etc.). This approach relies on real-time data from sensors and historical maintenance records to identify patterns that precede failures. For instance, Massaro et al. (2020) have used Internet of Things (IoT) devices and AI techniques to predict maintenance needs for a bus fleet based on drivers behaviours and vehicles data. They analysed data such as the revolutions per minute, throttle position, or the fuel consumption to classify driver behaviours using k-means clustering algorithms. They also used an artificial neural network (ANN) to predict vehicle wear and thereby correlating driving style with vehicles deterioration. This holistic approach allows for a comprehensive maintenance strategy that extends the life of buses and enhances overall fleet efficiency.

“The heart of operational optimization of a transit agency lies in the challenge of precisely forecasting the electricity and fuel consumption of transit vehicles” - (Wilbur et al., 2023)(p.10). In the same logic of operational efficiency, it is the control of energy consumption which poses a major challenge and to which data analysis can provide solutions. The study by Lin et al. (2020) on the construction of analytical models for driving energy consumption offers a complete approach to optimising the use of energy in PT systems. The research investigated driving behaviour influencing energy consumption using Battery Management System data. By applying machine learning techniques, the researchers were able to identify eco-driving indicators, enabling them to determine the behaviour to adopt, such as controlling speed and the level of the bus battery. Implementing such models allows transit authorities to identify energy-saving opportunities, optimize schedules, and manage resources more efficiently. This not only reduces operational costs but also supports sustainability goals by lowering the carbon footprint of PT systems.

1. Traffic Management and scheduling

Scheduling and forecasting are key to the smooth operation of transport systems. It is essential for transport agencies to be able to adapt to demand, traffic and to the unexpected. The notion of dynamic scheduling is an interesting way of looking at AI in PT, allowing transit systems to adjust schedules in real-time based on current conditions and passenger demand. Studies have shown that it is possible to use different algorithms and techniques from the field of AI to find the optimum schedule for bus drivers. Systems enabling the real-time location of buses make it possible to have a flow of data in real time thanks to GPS. These data are then re-used for a wide range of applications: improving the design of future routes by predicting demand at a stop, or warning whether a change in schedule will meet demand on the network. And in some parts of the world, customised bus services are being imagined, as in China, combining the advantages of buses and taxis. Of course, autonomous buses equipped with computer vision technology are also appearing everywhere. One of the promises of these autonomous vehicles, if their use becomes widespread enough, is to reduce the cost of congestion (a saving that could amount to several billion euros for a city like Paris) (Abduljabbar et al., 2019).

PT systems are, by their very nature, multidimensional problems requiring real-time decision-making and high levels of computational capacity. And computer vision, which is the fruit of years of research in the field of DL, can also help to improve the management of passenger flows. A great example is provided by Velastin et al. (2020) who propose a video-based system using DL algorithms to accurately detect, track and count passenger boarding and alighting from metro trains. This system achieves a mean accuracy of 92%, allowing real-time counting of people inside the trains and on the platform. The study highlights that AI technologies could significantly reduce the time vehicles spend at stations by streamlining passenger boarding and alighting processes.

Another original source of data, but one that seems to have proved its worth in traffic management, is social media. For example, Gu et al. (2016) succeeded in developing a method for extracting and filtering tweets to identify incidents on highways and arterial roads. Here again, they used ML algorithms (classifiers) to analyse and process nearly 20,000 tweets combined for two cities, to determine whether the tweets were traffic-related or not and where the incident was located (if geolocatable). This method has proven to be a good additional source of information. In terms of spatial distribution, the detector allows for the identification of more incidents over extended areas. Therefore, it is also a cheaper alternative compared to traditional incident detection methods.

1. Efficiency in service delivery

Other less common but equally relevant applications of these technologies can improve service delivery. And one of the most persistent problems in PT is fare evasion, when stowaways don't validate their tickets. It's a real problem because it means less revenue and it's unfair for other passengers. Some researchers have used innovative approach to optimize control agent shifts in PT by using reinforcement learning techniques (when the machine learn by receiving rewards or penalties) to find the stations where the risk of fraud is highest and adjust the controllers' planning accordingly. This method, applied in a bus network (called TICE) in an area near Paris, has proved its effectiveness, as the validation rate has increased by 6% and the number of fines has doubled over the last 4 months of 2019 (Datategy, n.d.; Delfau et al., 2018). It's a real improvement in the way the service is carried out, with controllers being less of a nuisance to loyal travellers who validate, and more efficient because more accurate.

Delivering a decent service also means fighting insecurity, and this is a real concern in the Paris metro for example. A common but useful application of AI is the recognition of 'suspicious' activity using ML techniques. Applications to public transport already exist. Researchers have, for example, used security cameras (sometimes with low-quality images) to prevent theft. Using a classifier based on a convolutional neural network (CNN), they were able to determine whether the images from the bus cameras reflected a 'normal' or 'abnormal' situation (Affonso et al., 2021; Alexandre et al., 2023). To take this a step further, some researchers wanted to explore other types of data in addition to images and try to take into account the emotions that might be expressed in speeches during abnormal situations in transport. Eleonora Mancini et al. (2024) used speech datasets with native English speakers to simulate PT scenarios. Their model proved to be fairly accurate, robust to environmental noises and gender-neutral in recognition while being able to identify disruptive and non-disruptive emotions. Here again, note that CNN technology was used and produced better results than alternative solutions.

As we have seen, most of these examples of operational improvement involve AI. This is not surprising, as the sector is following the global trend. The next section complements this and aims to present interesting and potentially emerging concepts.

## **B – Focus on DA and some innovative concepts**

1. Understanding and learning: DA for decision-makers

Before reaching an advanced stage where organisations can develop and use tools linked to the field of autonomous analytics, most of them will first have to gradually familiarise themselves with descriptive, predictive, or prescriptive data analysis, which does not necessarily require the same level of sophistication in tools as the very recent AI applications we have just explored in the previous section. Ensuring that your organisation gains a competitive advantage through the use of data requires multiple conditions, including the involvement of top managers (e.g. executive committee) (Davenport & Harris, 2017). One of the major issues will therefore be to encourage these people to make decisions, and it can be done by proposing the right tool for instance. In the case of a public company or one whose activity is in the public interest, as it is generally the case in the PT sector, these may also be political decision-makers from outside the company.

Descriptive analytics involves analysing historical data to uncover patterns and trends. This type of analysis, which focuses on questions such as what is happening, what action should be taken, or the occurrence of certain problems, is fundamental to understand passenger flows and usage patterns. For instance, Van Oort et al. (2015) highlighted how Dutch PT agencies can use DA to monitor and enhance network performance. By examining historical open data, they were able to identify bottlenecks to optimize bus schedules and behaviours, accordingly, leading to more efficient and reliable services. Their study is designed so that transit agencies (and their decision-makers) can use the data to identify areas for improvement, thanks to a software that "translates" the data into information that can be easily interpreted and used, for example, with graphs showing the waiting time at each stop for a specific bus line. It's a very comprehensive work, providing insights into the performance of transit lines, making it possible to visualise this performance and providing indicators such as "additional travel time due to unreliability", to understand the impact on passengers.

In another study whose interest is limited to a particular context, Roșu & Blăgeanu (2015) use a quantitative spatial approach to evaluate the performance of the PT network of post-communist cities. The analysis involves using a Geographic Information System (GIS) to visualize and assess various performance indicators. These indicators include network coverage and accessibility, which are crucial the strengths and weaknesses of a transport system. More broadly, the authors have succeeded in identifying areas (stops) where the public transport network appears to be deficient, they have gathered indicators (functionality, closeness, demographic pressure, nodal connectivity, and others...) to create a composite index and then applied a clustering technique. This has resulted in maps showing that the city's historical heritage is at the root of certain patterns that mean that PT in the city is no longer fully adapted to demand. These results (which can be seen as a form of prescriptive analysis), as the paper points out, can be used by decision-makers to develop a more holistic view of the network’s performance, leading to more informed policy and investment decisions.

Finally, the integration of big data and DA into PT planning provides decision-makers with unprecedented opportunities to enhance this overall service. Big data offers comprehensive and continuous monitoring capabilities that traditional data sources lack. Their use supports the development of predictive models that can anticipate future trends and challenges, thus facilitating proactive rather than reactive planning. Continuous monitoring can enable a trial-and-error approach to policy development, rather than pre-testing and sudden implementation. Big data therefore means an increased flow of information, giving greater insight into transport activities, but also requiring the resources needed to manage them (Milne & Watling, 2019).

1. Innovative Concepts: Data Visualization, Dynamic Pricing and Mobility as a Service

This sub-section aims to explore concepts individually to explain the importance, and possible future role, of some concepts the PT sector. As we have already seen from previous examples, data visualisation plays a very important role when it comes to translating the complexity of a (sometimes colossal) dataset into usable and readable information for decision-makers. It can also be useful to passengers, on mobility applications. Effective data visualization is essential for a transit agency to enhance operational fluidity, and indeed in the world of mobility, this is a very important area for revealing trends, behaviours or problems. Sobral et al. (2019) explore how urban mobility data visualisation can aid in decision-making processes, they show examples of how PT systems data can be used to analyse (and visualize) passenger behaviour, usage and service reliability. According to them, the most widely used data sources are vehicle location data and data from smart cards (ticket validation).

Similarly, Dimanche et al. (2017) are proposing to help RATP analysing the vast quantities of operational data from the railways by creating a tool. The idea is to use a ‘Visual Analytics’ process that aims to integrate human intuition and mathematical model analytical power in the day-to-day operations. The authors of the study introduce the concept of VA, which they describe as a means of displaying massive quantities of data, with such a level of detail and complexity that it would otherwise be impossible to quickly identify relevant information. Therefore, RATP operators can better understand the patterns and discrepancies in rail operations such as delays or anomalies in train schedules, particularly for the RER and metro, by visualizing them. One of the graphs (fig. 4, p.3) shows, for example, 33 "operational days" on the RER line A, and we can clearly see that the most frequented hours create recurrent stress zones (where trains accumulate delays).

In another, even more concrete area, there are innovations that look at optimising pricing strategies and how to increase PT agencies revenues. Dynamic pricing strategies are one of those cutting-edge approach, leveraging AI and DA to optimize fare prices based on real-time demand and various influencing factors. Branda et al. (2020) developed a model for predicting ticket purchases with a high accuracy (95%). This precision supports dynamic pricing strategies that increase both number of ticket sales and revenue (respectively increased by 6 and 9% compared to other pricing strategies). The use of ML models in this study (trained on historical data) helped in identifying key factors that influence purchasing decisions such as fare costs, occupancy rates and booking time relative to the travel date. It’s interesting to note that this is one of the few studies reviewed whose primary aim is to increase revenues for PT companies. Kaddoura et al. (2015) further expand on pricing strategies by exploring an advanced simulation technique “that considers microscopic user-by-user bus fares, calculated with the objective to maximise social welfare”(p.2). Their research underscores the importance of setting prices that reflect the true social cost of transport services, such as such as delays from boarding and alighting. The model supports more informed decisions about bus frequencies, routes, and operational strategies. Finally, pricing strategies in the sector can therefore be reinvented, to the benefit of both transport companies and users, who expect a fair public service. Other approaches could be explored, such as distance-based payment.

Among the emerging concepts, Mobility as a Service (MaaS) represents the future of urban mobility. This is a matter of concern for public transport agencies (like RATP in France) who are trying to offer a multimodal service, bringing together several transports offers (bike, metro, bus, car sharing...) on the same platform. This concept must be supported by AI and DA, to propose the ultimate user-centric offer. Rajabi et al. (2023) discuss a knowledge-based AI framework for MaaS, which synthesizes data from multiple sources to deliver customized and sustainable mobility solutions. A MaaS system allows users to plan, book and pay for a seamless journey across different transport modes through a single platform. A commuter can receive a single journey plan that includes a bus ride, a train transfer and a bike-sharing option, all coordinated to minimize waiting times and travel disruptions. In this task of assisting mobility, the AI must, according to the authors, remain comprehensible in its recommendations to users. By clearly explaining how travel options are selected and optimized, MaaS platforms can therefore build user confidence, fostering greater adoption and satisfaction.

The adoption of dynamic pricing, MaaS, and data visualization significantly enhances the efficiency and quality of public transportation services. These innovative approaches optimize fare structures, integrate multiple transport modes for seamless travel, and provide actionable insights for better decision-making. As we move to the next section, we will explore how social media serves as a vital tool for customer engagement and satisfaction analysis, offering real-time data to improve service responsiveness and overall passenger experience.

## **C – Social media analytics: a tool to better understand user expectations**

1. Decoding passenger behaviour via social media

Social media has become a crucial tool for PT agencies to engage with passengers, gather feedback and enhance service quality. The vast amount of data generated and users on platforms like Twitter, Facebook, and Instagram is a unique opportunity to involve passengers more closely in the design of PT systems, but also to better understand their satisfaction levels and expectations, thus giving indications on how to improve the service.

Twitter is a micro-blogging social network, users are limited in the number of characters they can write (except for premium accounts), these messages are commonly called “tweets”. In general, Twitter users mainly comment on the news and a proportion use it to react to everyday events. It can provide PT agencies with immediate access to passenger feedback, allowing any service adjustments, and informing passengers in real time by interacting directly with them, for example. With a view to providing decision-makers with relevant feedback on transport services, Gal-Tzur et al. (2014) are specifically investing in the potential of social media in the design of PT companies strategy and policy in a precursory study. They emphasize the value of harvesting information from social media to complement traditional data sources. They were able to identify three purposes for tweets: the expression of a need to travel (and therefore for a transport service), an opinion about a service and reports of incidents (planned or unforeseen). By choosing to focus on several sporting events (football matches), they extracted a vast amount of tweets, then applied a classification (supervised learning) of the tweets into different categories (for example: density, availability, security and cost are all part of quality of service). This work shows the relevance of using social media in the construction of PT policies, making it possible to see what themes are recurring in tweets, getting people more involved (even if indirectly) in improving these services.

Other researchers have added a spatial dimension to the collection of tweets, associating a location to each one. Raczycki et al. (2021) focused on the PT in Wrocław (Poland), they analysed the city's PT operator posts from Facebook and Twitter, focusing on alert messages, with their related comments. The posts were classified according to the type of incident they described (malfunction, traffic accident, etc.), and the location of these incidents was determined using the names of the stops in the posts. This research also made it possible to determine which Tram and Bus lines were most frequently affected by operating incidents, and how many passengers were impacted per region. We can see that social network users react the most to malfunction-type incidents, and obviously the analysis of sentiments reveals that most of these reactions express something negative.

Finding the geolocation of tweets is sometimes a difficult task, unlike the previous study where messages were structured in such a way as it is easy to identify this information. As we have seen, the spatial dimension is important for making targeted improvements to transport. Osorio-Arjona et al. (2021) have implemented a geocoding process to retrieve the name of metro stations from the tweets. To do so, they used a Python dictionary with all the metro stops, then tweets containing abbreviations or specific terms related to these stations were identified and geocoded accordingly. In this same study, which looks at the Madrid region, the authors were able to take advantage of several analysis techniques (e.g. Geographic Weighted Regression) to look at the demographics of metro users. They conclude that high-income workers use cars more to get to work, or that Twitters users seem to be mostly mid-income workers living in peripheral areas, and that punctuality problems seem to affect them more (i.e. higher negative sentiment scores for this topic).

By using advanced techniques such as topic modelling valuable information can be extracted from the data. By analysing transport agency publications and user reactions, it is possible to identify the motivations for travel and the elements of (dis)satisfaction in real or near-real time. Malfunctions or high levels of use will obviously cause dissatisfaction. This immediate feedback loop could help maintain a higher level of service and meet passengers' needs.

1. Analysing passengers’ sentiments

As well as analysing the topics raised about PT on social networks, it is also the analysis of feelings that will enable us to gauge the level of user satisfaction. Sentiment analysis, which involves using NLP techniques to determine the emotional tone (generally: positive, negative, or neutral) behind social media posts, is particularly relevant for PT because people use the metro, bus or tram every day and the amount of tweets (i.e. data) generated is constant. Haghighi et al. (2018) developed a framework using twitter data to evaluate transit riders’ opinions about the quality of service, they provide tools for topic modelling and sentiment analysis. Their work reveals some interesting trends: it highlights the importance of transport agencies to respond to users' criticisms and requests (this would generate more positive feelings). Twitter users post more negatively about public transport services than about other public services. In fact, they've found that when people complain about a service, they do so by replying to the transport operators' Twitter accounts or by including their name in the comment, hoping for a response. The user experience insights reveal negative sentiment mostly related to train and bus service performance issues in the city analysed, in Utah. One of the insights from these type of studies is that although data mining on social media is cheaper and less time-consuming than traditional techniques (e.g. surveys), twitters' data (and data from other social media) is also likely to be biased: low-income or elderly people may not have access to these kinds of platforms, and the users are younger (between 20 – 39 years old) and more educated than the total population, it is therefore difficult to generalize based on these data. (Haghighi et al., 2018; Lock & Pettit, 2020; Osorio-Arjona et al., 2021).

Of course, multiple case studies exist and offer different points of view on how to go about analysing feelings (and modelling topics, as they often go together). Many studies converge on the same conclusion that social media mining provide a low-cost alternative to traditional surveys methods when it comes to measuring satisfaction levels in PT. Lock & Pettit (2020) see these two approaches as complementary, given that data collection on social media is imperfect and partly biased. In their study, they combined and compared both methods, even by asking Twitter users who had posted on PT about their motivations. Even when using highly advanced ML and NLP techniques to analyse subjects, tones and feelings, barriers emerge, such as the detection of sarcasm (more frequent during periods of disruption), or mixed motivations. Traditional surveys, which have their weaknesses such as limited sample sizes, still play a crucial role in understanding the detailed intentions and preferences of users.

Still based on the same consensus about the effectiveness of AI for analysing social media, all over the world we can see that researchers have been trying to apply sentiment analysis methods. It's often when used in association with another data source that tweets reveal their true value. For example, (Mehri et al., 2023) collected several thousand tweets over 5 years, created their own web crawler[[3]](#footnote-3) in Python to bypass the limitations of the Twitter API and search for multiple keywords. Then they cross-referenced the Twitter data with incident data from an official transport agency in the city studied, Montreal (Canada), adding an emotional explanatory dimension to the incident analysis. Their study clearly shows that Twitter users tend to be more active during commuting periods, and vice versa, when the metro is most used. In China, a country whose cities are densely populated but where PT infrastructure is more recent, Luo & He, (2021) remind us how important it is to take into account the spatial and temporal dimensions in users' perception of the transport service, and the specific characteristics of the populations studied. By identifying Service Quality Attributes and geolocatable user-generated content in the city of Shenzhen, they were able, for example, to precisely identify that the behaviour of staff on certain metro lines was more appreciated than elsewhere (and this is one of the most important satisfaction factors). In Shenzhen, crowdedness seems to be a critical issue (a source of mostly negative sentiments), and other attributes specific to the country are discussed, such as security checks prior to entry. These are valuable insights for policymakers that can think about and design “spatial-temporal strategic interventions” to make the PT system more customer-centric.

As we have seen, the research provides us with numerous examples of the use of social media as an alternative or complementary source for analysing the satisfaction of PT users. Social networks are an excellent way of studying populations and socio-human issues. However, using data from these sources is tricky and sometimes complex, they are also unreliable sources that may be subject to restrictions or changes in policy by the companies that own them. Many of the studies presented in this section use Twitter as a data source, because the data is often text-based, which is easier to use, and until recently it was easy to retrieve tweets. But recently the platform has changed its API policies (when Twitter became X), raising a major challenge for the research community. According to the Coalition for Independent Technology Research (2023)[[4]](#footnote-4), these new policies could severely limit the ability of researchers (and therefore PT agencies) to collect and analyse large volumes of tweets, hindering the development of innovative solutions based on social media data. The first price is $100 per month and doesn't allow you to extract as many tweets as before, and several thousand dollars for businesses. The company has also put an end to several projects aimed at bypassing the API. This change means that we need to find alternative ways of extracting tweets for analysis.

# **4 - Research Question and Hypothesis**

This study seeks to take a multidimensional approach to the issue of improving PT, the only way to address solutions to complex systems like these. The extent to which AI and DA can improve them, and therefore be effective tools, is at the heart of this research. Fundamentally, we're interested in the quality of the insights and decisions that can be derived from the analyses made with these kinds of tools. As part of this study, we will focus our analyses on the company RATP, which, as detailed previously, faces multiple challenges to maintain its leadership.

To test and prove the relevance of the tools, the proposed research question is as follows:

To what extent can the analysis of ticket validation and social media data improve RATP's operational strategy and the customer experience of its users on Parisian transport?

We will begin by studying transport users through the analysis of ticket validation data, which enables us to understand user behaviour and provides essential information for decision-makers. This approach is supported by the examples cited in the previous section, showing that historical data can be a surprising, exploitable and useful source of information. Given the importance of user satisfaction in this sector, it is clearly a key variable for measuring transport efficiency and service quality. This study will therefore focus on the analysis of social networks, which, as the examples provided in the literature review have shown, can prove to be an alternative to traditional data collection processes. We will then be able to draw cross-referenced conclusions from these two sources of data. With the objects of study clearly defined, several hypotheses were made to guide the analysis and answer the research question:

**H1**: There is a significant difference in the number and patterns of ticket validations between weekdays, weekends and public holidays as observed through data analytics.

**H2**: Ticket validation data analysis can identify temporal and spatial patterns to optimize the allocation of resources and services.

**H3**: Natural language processing is efficient in measuring negative and positive feelings.

**H4:** Topic modelling for a chosen period allow to identify problems and causes of dissatisfaction on the RATP network.

# **5 - Method**

The research design for this dissertation is a mixed-methods approach that combines quantitative analysis of tickets validation data with the analysis of social media data, which is an in-situ observation that can be considered more qualitative, although quantifiable in many respects. This design allows for a comprehensive understanding of both the operational aspects of RATP’s services and perceptions and sentiments of its Paris PT customers. Choosing to analyse two different data sources also allows us to use different techniques, the aim is also to discuss the effectiveness of the tools we use, by noting the quality of the information we find. In this way, we will verify the assumptions made upstream, using the lessons learned from the two analyses. Figure 2 summarizes the whole process without going into the details of data collection.

Une image contenant texte, capture d’écran, diagramme, Police

Description générée automatiquement

Figure 2. Simplified overview of the study methodology

## **A - Data collection**

Firstly, for ticketing data, we have taken the data from the website of Île-de-France Mobilités (IDFM), the organization representing the regional transport authority. The data is open source[[5]](#footnote-5), making it easily accessible to everyone, and it's our intention to have chosen this data so that the process remains as transparent and replicable as possible. On this platform, several types of data and datasets are freely accessible, with even pre-designed interfaces for easy data visualization. The selected quantitative data are focused on the rail network and covers the first half of 2023 (from January 1 to June 30 of that year). The choice of this period is explained by the fact that the IDFM platform offers two sets of data: the first and larger one is the “Number of validations per day (1st semester 2023)” [[6]](#footnote-6). The other dataset is “Hourly profiles by day type (1st semester 2023)” [[7]](#footnote-7). These two tables provide different levels of information over the same period, and are complementary, which is why they were chosen for the study.

Secondly, data from the Twitter social network was systematically observed and collected. Over a two-week period (June 10 to 24, 2024), every evening at a fixed time, tweets and replies to these tweets were collected using a web scrapping tool called "Data Miner" [[8]](#footnote-8), which takes the form of a web extension that can be activated to scrape data from any website. Data harvesting was very challenging and given that Twitter's API had to be paid for, alternative free solutions had to be explored, such as databar.ai[[9]](#footnote-9), which deserves to be mentioned here but couldn't be a reliable tool for very long. Data Miner was chosen for its ease of use and flexible customization. It works by allowing users to create custom data extraction rules or use pre-defined templates to scrape data from websites. Users can select the data they want to extract by pointing and clicking on the elements (HTML tags) on a web page. The software then converts the extracted data into a structured format, such as CSV or Excel, which can be downloaded for further analysis. The tool was configured on a Twitter page, and after a few tries, it was possible to extract the usernames, the time since the message was posted (since it was not possible to extract the exact date and time), and the content of the message. Then, using the advanced tweet search tool, it was possible to search for tweets with a keyword and to use some filters (language, links in the content, etc.). The combination of words in the Twitter search bar therefore looked like this: "RATP lang:fr -filter:links", which made it possible to extract tweets in French containing the keyword "RATP" and without any link. Replies to tweets mentioning RATP have been deliberately kept, to observe users' responses to official RATP accounts (i.e. all metro lines from 1 to 14 plus an account dedicated to customer service: "Service Client RATP"). The overall procedure is justified by the various studies presenting the retrieval of Tweets, using one or several keywords to target data extraction. The use of a single keyword in this case can be seen as a limitation reducing the relevance of extracted tweets, as mentioned by Haghighi et al. (2018), but this deliberate choice is explained by the fact that the RATP company is strongly associated with PT in Paris, effectively targeting a geographical area. Moreover, SNCF is also very present in the Paris region, managing several intra-regional rail lines, so it was necessary to target the company chosen for the case study.

## **B – Data description and preprocessing steps**

Ticket validation data is therefore divided into two tables. The number of validations per day is made up of 1 096 209 rows and 8 columns: JOUR (days), CODE\_STIF\_TRNS (transporter identifier, RATP is represented by the code “100”), CODE\_STIF\_RES, CODE\_STIF\_ARRET, LIBELLE\_ARRET (stop name), lda, CATEGORIE\_TITRE (ticket category), and NB\_VALD (number of validations). Similarly, the second table, Hourly profiles by day type, is made up of 84 137 rows and also 8 columns: CODE\_STIF\_TRNS, CODE\_STIF\_RES, CODE\_STIF\_ARRET, LIBELLE\_ARRET, lda, CAT\_JOUR (day type), TRNC\_HORR\_60 (a 1 hour slot) and pourc\_validations ( the percentage of validation is a ratio between the number of validations at a station, over a time slot, and the number of validations over the whole day at this station). For the purposes of the analysis, not all the columns were used, but the names of the stops, the categories of ticket and day, and the number and percentage of validations at a station, were the focus of the analysis.

For tweets, as mentioned above, the data collection tool enabled us to obtain tables with 4 columns for each day, which were then aggregated. The name of the author, the date of extraction, the content of the tweet, and the time since publication were the 4 essential data collected for each message published on the social network, to this we added the sentiment and confidence score after the analysis. We were able to collect more than 3,860 tweets over the whole period. Once the data had been collected, several cleaning and formatting operations had to be carried out using Excel. These included removing any odd characters that might affect the quality of the analysis and finding the approximate time of publication based on the time elapsed since publication. The approximate time of publication of the tweets is a limit to the accuracy of the analysis, but the method used simply consisted of subtracting the time of extraction of the data, generally around 11pm, from the time since the tweet was published, so a tweet published 13 hours ago was automatically classified in the 10am time slot (23 – 13 = 10). Relevant tweets were also first filtered by hand on a day-to-day basis, using objective criteria such as message size, clarity and readability, and relevance to the theme (whether the tweet referred to services provided by RATP or not).

Tweets are very special data. People use a lot of abbreviations, slang terms, expressions, and so on. The writing style is sometimes very 'spoken', and the message can be punctuated with special characters, emojis etc. All this complicates the textual analysis. In addition, as highlighted in other studies, the use of sarcasm, such as in this tweet: *"Gold medal for the RATP! What a symbol! Station inaugurated less than 12 hours ago and already an incident at St Denis-Pleyel...Ah champions...."* complicates the analysis of sentiment. In this example, it is obvious to a human reader that the comment expressed something negative, but the task is more difficult for a language model. Moreover, there are many limitations to analyzing tweets: they are a fairly unreliable source of data, requiring several stages of transformation before information can be extracted and users are not necessarily a representative sample of the population.

# **6 – Results**

The aim of this section is to validate or invalidate the hypotheses and answer the research question. The data was analyzed using the Python programming language, which is very popular in the field of data analysis and ML. First, we will look at ticket validation data to study the behavior of Paris metro users. Then we will try to take advantage of data from the social network Twitter to gain a better understanding of how these users perceive RATP services and Parisian PT in general. The all code is available in the Appendix A of this study.

## **A - Analysis of ticket validation data**

1. Number and patterns of ticket validations

To understand how and when people travel, ticket validation data is relevant for observing habits. Hourly profile data can be used to distinguish between different times of the week or day. To clearly visualize the differences in validation behavior, the days were sorted and grouped into two broad categories: working days (Monday to Friday) and weekends and public holidays. As shown in , at the RATP network stations, it is rather evident that during weekdays there are when two distinct periods of intense traffic that can be observed, characteristic of a bimodal distribution (a classic and well-known pattern for this type of data). Workers and students heavily use the Paris metro: in the morning, primarily between 8 and 9 AM, to get to their workplaces or schools, and in the evening, more diffusely between 5 and 7 PM. This helps identify the times when the company's resources need to be maximized (trains, drivers, support staff to inform travellers, tickets controllers, etc.). In contrast, a unimodal distribution characterizes travel on weekends and public holidays, Parisians seem to travel mostly in the afternoon.

To understand the extent of the difference in ridership between working days and weekends, and the way in which certain events can influence metro ridership, the number of validations per day in the first half of 2023 is visualized in Figure. We can clearly see each week of the first 6 months of the year 2023, where each peak represents a week. For weeks excluding holidays and working days, the number of validations is around 3.5 million every day. Saturdays and Sundays are clearly lower in terms of validations and distinctly indicate the end of the week. From 18 February to 6 March and from 22 April to 9 May 2023, the school holidays cause the number of validations to drop for 2 weeks each time. More sporadically, some days appear to be anomalies. Annotated here on the graph, Valentine's Day and the Festival of Music are days when people go out more in the evening and increase the total number of validations. By contrast, some weeks appear to be "broken". Here, for example, on 19 January 2023, a large-scale mobilization against a political reform took place in Paris, bringing transport to a virtual standstill (only half a million validations), almost the same as on New Year's Sunday 2023.

Une image contenant texte, Tracé, ligne, diagramme

Description générée automatiquement

Figure 3. Average percentage of validations by time slot and day category at RATP stationsUne image contenant texte, ligne, Police, Tracé

Description générée automatiquement

Figure 4. Number of validations at RATP stations in the first semester of 2023

Figure 4. Number of validations at RATP stations in the first semester of 2023

Une image contenant ligne, Tracé, diagramme, texte

Description générée automatiquementTo understand the profile of PT users, it may also be useful to distinguish between validations by type of ticket used, detailed in Figure 5. It's interesting to see that, over the same period, if we keep only the categories where the average per day is over 250,000 validations (left graph), 2 categories emerge as being the most used. Navigo passes (which include annual, monthly and weekly passes) are the most popular, and are reimbursable up to a minimum of 50% for working people by their employers. Imagine R passes are intended for schoolchildren, apprentices and students.

Figure 5. Number of validations at RATP stations by ticket category: focus on the most important ones and on a week in June

On the chart on the right, focusing on the week from 19 to 25 June reveals some interesting behaviour. Imagine R passes, in red, appear to be more active on 21 June, the day of the Festival of Music, than other passes. This can be explained by the fact that this is a popular event for young people, which translates into the metros closing later, for example. We can also see that most of these passes are only used by a small proportion of the total passenger population. The AMETHYSTE pass, designed for the elderly and disabled, barely reaches the 90 000 validations threshold and is therefore invisible on this graph. The FGT and TST passes are reserved for social welfare beneficiaries.

Based on these initial insights, a highly descriptive analysis of the data enables us to validate **H1**. As we might have expected, most PT users use this mode of transport to get to their place of work, revealing a gap between working days and the other days. Events that increase ridership are easily identifiable, making it possible to allocate more resources and possibly target communication in advance.

1. Spatiotemporal analysis

Following on from the analysis of ticket validation data, an in-depth look of ticket validation behaviour is required to understand population flows. For the purposes of the analysis, metro line 8 operated by the RATP was chosen. It connects very different areas, crosses Paris from west to east, then heads south, reaching its terminus in Créteil. Inspired by Sobral et al. (2019), who present a "flow-comap" visualisation technique for identifying major passenger flows over a time period, Figure 6 shows how passengers use metro line 8, with each point being a station and the diameter of the point increasing in proportion to the percentage of validations at the station during a time slot. Wider time slots have Une image contenant ligne

Description générée automatiquementbeen made and the percentages have been averaged.

Figure 6. Visualization of spatiotemporal mobility patterns on metro line 8

The economic activity and therefore jobs in the Paris region are concentrated in Paris itself. Many people live in the suburbs, in more modest towns (like Créteil) where housing is cheaper, and commute to Paris every morning for work, or to access the best universities and schools for students. These maps[[10]](#footnote-10) clearly show that the early morning and late afternoon timetables are the busiest for metro 8, creating inconvenience due to overcrowding (a strong dissatisfaction factor). In the morning, stations outside the centre of Paris seem to be very busy between 7h and 9h when people go to work or school, and inversely a movement of population from Paris to the suburbs takes place between 15h and 19h when people go home, with stations in the centre of Paris recording high rates of validations at this time. In the off-peak period, it is interesting to note that in the evening, from 20h to 23h, it is the Paris stations that record the most validations, suggesting a more active nightlife in the capital than in the suburbs.

1. Clustering analysis

A third level of analysis was carried out on these data using k-means clustering. This is a widely used method of unsupervised ML, particularly suited for partitioning a dataset into distinct groups or clusters. The core idea of k-means clustering is to divide the data into *k* clusters, each cluster consists of data points that are more similar to each other than to those in other clusters, based on a specified distance metric, typically Euclidean distance. The aim, for this study, is to identify patterns and clusters in metro station use based on validation activity (percentage of validation) throughout the day (here specifically the working days). Once again, this can help to understand commuter behaviour, optimise resource allocation and improve service planning.

A pivot table was created grouping the stations in rows and the time slots in columns, which was then normalized and used to apply the KMeans function from the scikit-learn python library. Using different methods and after several trials, an optimum number of clusters was set at 5. The 309 stations identified as belonging to the RATP in the data were then assigned to a cluster (as shown in Table 1 and Figure 8) based on the behaviour of validations at that station over the course of the day.

|  |  |
| --- | --- |
| Cluster number | Number of stations |
| 0 | 103 |
| 1 | 84 |
| 2 | 45 |
| 3 | 44 |
| 4 | 33 |

Une image contenant texte, diagramme, ligne, écriture manuscrite

Description générée automatiquementTable 1. Number of stations assigned to each cluster

Figure 7. Visualisation of clusters by average percentage of validations per time slot for working days

Figure 7 clearly shows the results of the clustering and how the algorithm grouped the stations. The percentage of validations was grouped by time slot and averaged for all the stations belonging to the same cluster. Clusters 0 and 1 are the largest categories of stations, with two significant peaks in validations, one in the morning and one in the afternoon, which remain in the same proportions. These are typical home to work/school trips. Clusters 2 and 3 are much busier in the evening, as most of these stations are in the heart of Paris, or in areas where there is a concentration of jobs but little residential activity. These stations are used to get the metro and make the journey back to home. Finally, in cluster 4, there are more validations in the early morning than in the late afternoon, indicating, for example, more residential areas. Taking metro line 8 as an example, Figure 9 supports our analysis by showing Une image contenant texte, carte, diagramme, atlas

Description générée automatiquementUne image contenant diagramme, capture d’écran, ligne

Description générée automatiquementthe different types of stations on this line.

Figure 8. Cluster number for stations on metro line 8

Figure 9. 3D Principal component analysis (PCA) visualization of stations clusters

Spatiotemporal analysis and clustering are good demonstrations of the relevance of analytical tools applied to the PT sector. The analysis of station occupancy makes it possible to identify stations and timetables with high volumes of incoming passengers, but also to understand and identify more clearly the times when occupancy is lowest, optimising the allocation of resources. For decision-makers, it's also a way of understanding how and where to improve infrastructure: improving services at busy stations, considering expansions or increasing connections with other lines. The results of these two analyses therefore validate hypothesis **H2**. DA is a set of tools with no real equivalent, and the quality of the information extracted is evidence to its effectiveness.

## **B – Sentiment analysis of Twitter/X data**

1. Sentiment detection in tweets

To explore in greater detail how customers perceive a company's services, it is common practice to conduct perception surveys. Time-consuming, costly and not necessarily meeting all the criteria in terms of sample size, the alternative of social networking appears to be a serious competitor.

To carry out this analysis, we used Camembert, a French language model specially created and trained on French text corpora (Muller, n.d.). This approach uses the power of an advanced ML technique to ensure accurate sentiment classification. The tweets analyzed are first tokenized, the text is divided into units such as words, these words are normalized/simplified, and then converted into numeric IDs. From an openly available GitHub repository, the 'tblard/tf-allocine' model (Blard, 2020/2024)[[11]](#footnote-11), fine-tuned for sentiment analysis tasks, was loaded. For the purposes of analysis, a model specifically trained on French language was required, the model used was trained on a large dataset of film reviews, so it has limitations for processing text such as tweets. The output is only 'POSITIVE' or 'NEGATIVE', which doesn't allow us to capture all the nuances, with certain reactions or messages appearing more neutral.

Une image contenant diagramme, Police, cercle, ligne

Description générée automatiquementAs expected, the results of the analysis showed a dominance of tweets expressing negative sentiments. Indeed, it is a known trend, observed in the studies cited in the literature review, that twitter users express themselves mainly to criticize PT services negatively. Figure 10 shows the different proportions of sentiment detected in tweets, excluding tweets from official RATP accounts.

Figure 10. Proportion of sentiment in tweets by confidence score, excluding tweets from official RATP accounts.

The confidence score returned by the model is used to assess the reliability of the model's prediction. For 3 different confidence score thresholds, we can see that the proportion of negative tweets detected exceeds 80%.

Une image contenant diagramme, Tracé, ligne, texte

Description générée automatiquementIf we look at the temporal dimension, Figure 11 reminds us of the shape of (bottom chart in red), with two distinctive peaks in the morning and afternoon. In fact, the curves for the percentage of validations per hour and the average number of tweets per hour (although over different periods) seem to be correlated, and we can guess the causal link here: Twitter users comment live on their journeys on Paris PT and mention the RATP to report incidents, interact with RATP accounts, or simply complain about the services.

Figure 11. Average number of tweets per sentiment per hour

Une image contenant diagramme, ligne, Tracé, texte

Description générée automatiquement

Figure 12. Evolution of the number of tweets per sentiment per day over the period studied

More generally, if we look at the number of tweets published per day over the 14 days, Figure 12, we can see that certain events cause an explosion in the number of reactions on Twitter, for example the extensions of metro lines 11 and 14 on 13 and 24 June 2024. All these results show that better communication prior to line extensions is probably needed. More generally, an ambitious objective could be to reduce the average number of negative tweets over the long-term during periods of peak transport use.

1. Topics covered in tweets

To complete the detection of feelings, an analysis of the subjects was carried out. This is a complex task, and the quality of the result depends on certain pre-processing steps. For example, a comprehensive list of French stop words was defined to remove common words that do not contribute to the semantic meaning of the text. Tweets are also tokenized. Once again, we used the scikit-learn library, which provides a turnkey way of using Latent Dirichlet Allocation (LDA), even without being an expert. This is a popular method widely used in topic modelling. LDA assumes that documents (tweets) are made up of a set of topics. The tweets are converted into a matrix (vectorization), where each tweet is represented by a vector of word frequencies. The LDA algorithm analyses this matrix to identify patterns of word occurrence and groups frequently associated words to form themes. In this way, each tweet can be seen as a combination of these topics, and each topic is defined by a set of characteristic words.

The number of subjects was fixed at 5 after several trials. The main themes discussed were then displayed in the form of a "word cloud". clearly shows the LDA results. Let's look at each theme to identify the causes of dissatisfaction and the issues addressed, Figure 13 shows the results of the LDA only applied on negative tweets. In topic 1, the tweets are about station-specific problems. Words like "thank you" appear, suggesting responses to official RATP tweets. The word "orly" refers to the extension of metro line 14 to Orly airport, providing a connection from Paris. The acronym "jo" appears, referring to the forthcoming Paris 2024 Olympic Games (“Jeux Olympique” in French). The extension of lines 11 and 14 is a way for the city of Paris to prepare for such an event. Topics 2 and 3 are tweets expressing very specific problems such as delays (“retard”), traffic problems or breakdowns (“panne”) at stations or on metro lines. In topic 1 and 3 the words "contrôleurs” (controllers) and "conducteur” (drivers) are present, implying that there are potential problems with RATP staff (some Tweets even directly accuse the controllers of abuse of power or unjustified fines). Topic 4 is more likely to focus on the problems surrounding tickets and passes. "Navigo", "Navigo easy", "billet", "payer" (to pay), “zone” and "prix" (price) relate to fare systems, fare zones and financial aspects. Topic 5 is characterized by strong negative sentiments and criticism of the RATP services. Terms like "mer\*\*," "pire," "honte," express frustration and dissatisfaction. The name of the head of the RATP (and former French Prime Minister), Mr Castex, stands out, and it is interesting to note that several tweets use the " JO " (OG), to criticise the RATP, stating that the company would not be ready for such an event.

Une image contenant texte, Police, écriture manuscrite, typographie

Description générée automatiquement

Figure 13. Cloud-of-words of the 35 most frequent terms in each topic, only on negative tweets

Topic analysis can also be used to see which metro lines are most frequently cited and thus targeted over a given period. Figure 14 shows that lines 14 and 11 are the most frequently mentioned because they were the focal points of interest for 2 days, and the inauguration of the extensions caused some inconvenience on the same days. Also in the top 5 are lines 7, 8 and 6, which are ageing lines, or which may have suffered significant disruption over the period studied.

Une image contenant texte, capture d’écran, affichage, nombre

Description générée automatiquement

Figure 14. Number of mentions of each line in negative tweets

In this section, we have demonstrated the effectiveness of advanced NLP tools. Firstly, sentiment analysis using a high-performance model has revealed trends in messages published on Twitter, the majority of which are negative or critical of RATP services. This is a good source of data for understanding the problems on a PT network as vast as that of Paris. Despite its good performance, sentiment analysis is a process that requires transforming and analyzing data that is not designed for that purpose. As the model used was not specifically trained on tweets, it seemed to fail to detect sarcasm and irony in many tweets, falsely classifying them as positive. Hypothesis **H3** is partially validated. By using a topic modelling technique, we were also able to determine the topics addressed in the tweets, which made it possible to put words to feelings. However, it is still difficult to extract reliable information using the method applied here. Although many issues were raised in the tweets, there is certainly a lack of detail to fully exploit the information. These findings provide partial support for hypothesis **H4**. Sentiment analysis should be seen as a complementary tool to traditional perception surveys.

# **7 – Discussion**

## **A - Interpretation of the findings**

The primary aim of this study was to assess how AI and DA can be used to improve the operational efficiency and customer experience of RATP, the public transport operator for the Paris region, and therefore to demonstrate the positive contribution of these tools, through a concrete application, on the PT sector. With this in mind, the method has been designed to draw conclusions from different sources of data and apply relevant techniques that are applicable in a business context. Through an extensive analysis of ticket validation data and sentiment analysis from Twitter data, several key findings emerged that underline the potential and limitations of these technologies in the PT sector.

The analysis of ticket validation data revealed distinct temporal patterns, with clear peaks during morning and evening rush hours on weekdays, and a more spread-out distribution on weekends and public holidays. This bimodal distribution underscores the need for RATP to allocate resources dynamically to match the demand. Additionally, specific events such as public holidays, strikes, and cultural happenings significantly impacted ridership patterns. Understanding these fluctuations can help RATP better plan and manage services during such events. The use of these indicators integrated into long-term demand forecasting models or simply for daily monitoring of validations would be a good way of using data analytics. In addition, the analysis of station clusters by hourly occupancy has enabled the identification of stations with similar usage, enabling more precise allocation of resources and targeted infrastructure improvements. These reliable results enabled us to validate hypotheses 1 and 2.

On the other hand, analysis of the sentiments and their underlying subjects showed a predominance of negative sentiments in the tweets. It seems that the number of tweets (and therefore also of negative tweets) is also closely linked to public transport use. The topic modelling analysis enabled us to identify several recurring topics in messages expressing something negative: material incidents, punctuality, prices of passes/tickets, and RATP staff are at the heart of the subjects discussed. The results of these two techniques show that, if used, they would enable citizens to be much more involved in the design of PT services. Decision-makers should seize this opportunity to act on key points that will increase satisfaction.

However, despite the promising results, a series of limitations prevent us from fully validating hypotheses 4 and 5. The lack of precision in the sentiment analysis remains a real problem. This study did not apply any fine-tuning to the model to improve the classification results and the measurement of its accuracy on data extracted from Twitter is not provided but simply deduced by a partial visual check of the data. The tweets themselves represent a challenge for the analysis. Sentiment analysis based on tweets can therefore be used to measure sentiment but remains limited because it is not very nuanced, it should be seen as a complementary tool to existing classic satisfaction survey. Similarly, the analysis of topics, while effective, could be improved or expanded to target relevant themes even more effectively.

As we already mentioned, there are also a series of limitations to the data source: social media users do not represent the entire population of RATP users. Younger individuals are overrepresented, which could bias the findings. Furthermore, recent changes in Twitter’s API policies have made it more challenging to collect large volumes of data, potentially limiting the scope of future analyses. That's also why this study is intended as an example of an alternative solution to data collection on social media, although the method used here is much more time-consuming and allows much more limited quantities of data to be extracted.

## **B - Conclusions**

Overall, this study offers valuable insights into how AI and AD can be applied to the PT sector and drive concrete improvements, making their use more comfortable. The benefits in terms of decision-making are substantial, and to do without these technologies would be a nonsense. AI and DA should enable RATP to consolidate its position as a leader in the Paris region and worldwide. Using AI and DA can not only save time in understanding the needs of PT users, but it can also save money, giving the company a competitive advantage at a time when PT is opening up to competition in Europe.

As we have seen and demonstrated with an example of a practical application, given the scale of the complexity represented by these transport systems, AI and more generally the field of DA can play a key role in improving these services. But their implementations require real in-house skills to maximize the benefits of their integration. On the whole, this is a promising area of research that addresses real issues for our societies, which are looking for greener ways of getting around.

# **8 - Recommendations**

These technologies must be applied with an awareness of their limitations, and with a clear understanding of the added value that these tools can bring. The range of applications is vast, as we have seen, and the choice of tool must be carefully considered. Future studies in this area can continue to explore the possible improvements that AI and DA can bring to this sector. The use of data from social networks is also a subject on which new disruptive technologies or techniques could greatly facilitate the understanding of interactions on these kinds of platforms and their exploitation.

More generally, the perspective of a General Artificial Intelligence, which could, for example, be specialized in the management of a PT system, offers immense possibilities for improvement in the long term.

# **References**

Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of Artificial Intelligence in Transport: An Overview. *Sustainability*, *11*(1), Article 1. https://doi.org/10.3390/su11010189

*Accenture Report: Artificial Intelligence Has Potential to Increase Corporate Profitability in 16 Industries by an Average of 38 Percent by 2035*. (n.d.). Retrieved from https://newsroom.accenture.com/news/2017/accenture-report-artificial-intelligence-has-potential-to-increase-corporate-profitability-in-16-industries-by-an-average-of-38-percent-by-2035

Affonso, G. A., De Menezes, A. L. L., Nunes, R. B., & Almonfrey, D. (2021). Using Artificial Intelligence for Anomaly Detection Using Security Cameras. *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 1–5. https://doi.org/10.1109/ICECCME52200.2021.9591068

Alexandre, T., Bernardini, F., Viterbo, J., & Pantoja, C. E. (2023). Machine Learning Applied to Public Transportation by Bus: A Systematic Literature Review. *Transportation Research Record*, *2677*(7), 639–660. https://doi.org/10.1177/03611981231155189

*Biggest social media platforms 2024*. (n.d.). Statista. Retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Blard, T. (2024). *TheophileBlard/french-sentiment-analysis-with-bert* [Jupyter Notebook]. https://github.com/TheophileBlard/french-sentiment-analysis-with-bert (Original work published 2020)

Branda, F., Marozzo, F., & Talia, D. (2020). Ticket Sales Prediction and Dynamic Pricing Strategies in Public Transport. *Big Data and Cognitive Computing*, *4*(4), Article 4. https://doi.org/10.3390/bdcc4040036

Datategy, C. de. (n.d.). *L’opérateur de transport public TICE (Transport Intercommunaux Centre Essonne) réduit la fraude grâce à une application basée sur l’IA*. Decideo - Actualités sur le Big Data, Business Intelligence, Data Science. Retrieved from https://www.decideo.fr/L-operateur-de-transport-public-TICE-Transport-Intercommunaux-Centre-Essonne-reduit-la-fraude-grace-a-une-application\_a12434.html

Davenport, T., & Harris, J. (2017). *Competing on Analytics: Updated, with a New Introduction: The New Science of Winning*. Harvard Business Press.

Delfau, J.-B., Pertsekos, D., & Chouiten, M. (2018). Optimization of Control Agents Shifts in Public Transportation: Tackling Fare Evasion with Machine-Learning. *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, 409–413. https://doi.org/10.1109/ICTAI.2018.00070

Dimanche, V., Goupil, A., Philippot, A., Riera, B., Urban, A., & Gabriel, G. (2017). Massive Railway Operating Data Visualization; a Tool for RATP Operating Expert. *IFAC-PapersOnLine*, *50*(1), 15841–15846. https://doi.org/10.1016/j.ifacol.2017.08.2324

*Documents de référence | Groupe RATP*. (n.d.). Retrieved from https://www.ratp.fr/groupe-ratp/presentation-du-groupe/documents-de-reference

Eleonora Mancini, Andrea Galassi, Federico Ruggeri, & Paolo Torroni. (2024). Disruptive situation detection on public transport through speech emotion recognition. *Intelligent Systems with Applications*, *21*(200305-). https://doi.org/10.1016/j.iswa.2023.200305

Gal-Tzur, A., Grant-Muller, S. M., Kuflik, T., Minkov, E., Nocera, S., & Shoor, I. (2014). The potential of social media in delivering transport policy goals. *Transport Policy*, *32*, 115–123. https://doi.org/10.1016/j.tranpol.2014.01.007

Gu, Y., Qian, Z. (Sean), & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, *67*, 321–342. https://doi.org/10.1016/j.trc.2016.02.011

Haghighi, N. N., Liu, X. C., Wei, R., Li, W., & Shao, H. (2018). Using Twitter data for transit performance assessment: A framework for evaluating transit riders’ opinions about quality of service. *Public Transport*, *10*(2), 363–377. https://doi.org/10.1007/s12469-018-0184-4

*Impact of AI as a percentage of industry revenues 2023*. (n.d.). Statista. Retrieved from https://www.statista.com/statistics/1012325/worldwide-artificial-intelligence-impact-as-percentage-industry-revenue/

Jevinger, Å., Zhao, C., Persson, J. A., & Davidsson, P. (2024). Artificial intelligence for improving public transport: A mapping study. *Public Transport*, *16*(1), 99–158. https://doi.org/10.1007/s12469-023-00334-7\*

Kaddoura, I., Kickhöfer, B., Neumann, A., & Tirachini, A. (2015). Optimal Public Transport Pricing: Towards an Agent-based Marginal Social Cost Approach. *Journal of Transport Economics and Policy (JTEP)*, *49*(2), 200–218.

*La mise en concurrence des lignes du réseau francilien*. (2020, January 14). Île-de-France Mobilités. https://www.iledefrance-mobilites.fr/mise-en-concurrence

*La Région Île-de-France arrête le projet de Plan des mobilités en Île-de-France 2030: Hausse de 15% des déplacements en transports collectifs | Région Île-de-France*. (n.d.). Retrieved from https://www.iledefrance.fr/presse/la-region-ile-de-france-arrete-le-projet-de-plan-des-mobilites-en-ile-de-france-2030-hausse-de-15-des-deplacements-en-transports-collectifs

Lin, K.-C., Lin, C.-N., & Ying, J. J.-C. (2020). Construction of Analytical Models for Driving Energy Consumption of Electric Buses through Machine Learning. *Applied Sciences*, *10*(17), Article 17. https://doi.org/10.3390/app10176088

Lock, O., & Pettit, C. (2020). Social media as passive geo-participation in transportation planning – how effective are topic modeling & sentiment analysis in comparison with citizen surveys? *Geo-Spatial Information Science*, *23*(4), 275–292. https://doi.org/10.1080/10095020.2020.1815596

Luo, S., & He, S. Y. (2021). Using data mining to explore the spatial and temporal dynamics of perceptions of metro services in China: The case of Shenzhen. *Environment and Planning B: Urban Analytics and City Science*, *48*(3), 449–466. https://doi.org/10.1177/2399808320974693

Massaro, A., Selicato, S., & Galiano, A. (2020). Predictive Maintenance of Bus Fleet by Intelligent Smart Electronic Board Implementing Artificial Intelligence. *IoT*, *1*(2), Article 2. https://doi.org/10.3390/iot1020012

Mehri, B., Trépanier, M., & Goussard, Y. (2023). *A methodology to assess passengers’ perceptions of transit services and impact of incidents*. Bureau de Montreal, Université de Montreal. https://www.cirrelt.ca/documentstravail/cirrelt-2023-07.pdf

Milne, D., & Watling, D. (2019). Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography*, *76*, 235–244. https://doi.org/10.1016/j.jtrangeo.2017.11.004

Muller, B. (n.d.). *CamemBERT*. CamemBERT. Retrieved from https://camembert-model.fr/

Osorio-Arjona, J., Horak, J., Svoboda, R., & García-Ruíz, Y. (2021). Social media semantic perceptions on Madrid Metro system: Using Twitter data to link complaints to space. *Sustainable Cities and Society*, *64*, 102530. https://doi.org/10.1016/j.scs.2020.102530

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O’Reilly Media, Incorporated. http://ebookcentral.proquest.com/lib/neoma/detail.action?docID=1323973

*Public transportation worldwide*. (n.d.). Statista. Retrieved from https://www.statista.com/study/36464/public-transportation-statista-dossier/

*Public Transportation—Global | Statista Market Forecast*. (n.d.). Statista. Retrieved from https://www.statista.com/outlook/mmo/shared-mobility/public-transportation/worldwide

Raczycki, K., Szymański, M., Yeliseyenka, Y., Szymański, P., & Kajdanowicz, T. (2021, October 11). *Spatial Data Mining of Public Transport Incidents reported in Social Media*. arXiv.Org. https://arxiv.org/abs/2110.05573v1

Rajabi, E., Nowaczyk, S., Pashami, S., Bergquist, M., Ebby, G. S., & Wajid, S. (2023). A Knowledge-Based AI Framework for Mobility as a Service. *Sustainability*, *15*(3), 2717. https://doi.org/10.3390/su15032717

Roșu, L.-I., & Blăgeanu, A. (2015). Evaluating issues and performance of a public transport network in a post-communist city using a quantitative spatial approach. *Urbani Izziv*, *26*(2), 103–116. https://doi.org/10.5379/urbani-izziv-en-2015-26-02-002

Sobral, T., Galvão, T., & Borges, J. (2019). Visualization of Urban Mobility Data from Intelligent Transportation Systems. *Sensors (Basel, Switzerland)*, *19*. https://doi.org/10.3390/s19020332

Taylor, B. (2023, April 3). *Letter: Twitter’s New API Plans Will Devastate Public Interest Research*. Coalition for Independent Technology Research. https://independenttechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/

*The 10 problems with Paris transport system France’s ex-PM must deal with*. (2022, November 23). The Local France. https://www.thelocal.fr/20221123/the-10-problems-that-frances-ex-pm-faces-in-his-new-job-running-paris-transport

van Lierop, D., Badami, M. G., & El-Geneidy, A. M. (2018). What influences satisfaction and loyalty in public transport? A review of the literature. *Transport Reviews*, *38*(1), 52–72. https://doi.org/10.1080/01441647.2017.1298683

van Oort, N., Sparing, D., Brands, T., & Goverde, R. M. P. (2015). Data driven improvements in public transport: The Dutch example. *Public Transport*, *7*(3), 369–389. https://doi.org/10.1007/s12469-015-0114-7

Velastin, S. A., Fernández, R., Espinosa, J. E., & Bay, A. (2020). Detecting, Tracking and Counting People Getting On/Off a Metropolitan Train Using a Standard Video Camera. *Sensors*, *20*(21), Article 21. https://doi.org/10.3390/s20216251

*What is Artificial Intelligence (AI)? | IBM*. (2021, October 6). https://www.ibm.com/topics/artificial-intelligence

Wilbur, M., Sivagnanam, A., Ayman, A., Samaranayeke, S., Dubey, A., & Laszka, A. (2023). *Artificial Intelligence for Smart Transportation* (arXiv:2308.07457). arXiv. http://arxiv.org/abs/2308.07457

*X/Twitter: Global audience 2024*. (n.d.). Statista. Retrieved from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

1. <https://x.com> [↑](#footnote-ref-1)
2. For the sake of clarity, we'll refer to "X" by its former and better-known name: "Twitter". [↑](#footnote-ref-2)
3. The library used ([GetOldTweets3 · PyPI](https://pypi.org/project/GetOldTweets3/)) no longer works [↑](#footnote-ref-3)
4. [Letter: Twitter’s New API Plans Will Devastate Public Interest Research - Coalition for Independent Technology Research (independenttechresearch.org)](https://independenttechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/) [↑](#footnote-ref-4)
5. The open data platform for the Greater Paris region: [https://data.iledefrance-mobilites.fr/](https://data.iledefrance-mobilites.fr/%20%20%20)  [↑](#footnote-ref-5)
6. <https://data.iledefrance-mobilites.fr/explore/dataset/validations-reseau-ferre-nombre-validations-par-jour-1er-semestre/> [↑](#footnote-ref-6)
7. <https://data.iledefrance-mobilites.fr/explore/dataset/validations-reseau-ferre-profils-horaires-par-jour-type-1er-semestre/> [↑](#footnote-ref-7)
8. Selected tool for social media data mining: [Scrape data from any website with 1 Click | Data Miner](https://dataminer.io/) [↑](#footnote-ref-8)
9. <https://databar.ai/> [↑](#footnote-ref-9)
10. It is important to note that validations at stations are only carried out at the station entrance. Passengers do not validate their tickets at the exit when they get off and leave a station (with rare exceptions). [↑](#footnote-ref-10)
11. [TheophileBlard/french-sentiment-analysis-with-bert: How good is BERT ? Comparing BERT to other state-of-the-art approaches on a French sentiment analysis dataset (github.com)](https://github.com/TheophileBlard/french-sentiment-analysis-with-bert?tab=readme-ov-file) [↑](#footnote-ref-11)