Forecast and anomaly detection on time series with dynamic context : Application to the mining of transit ridership data

Kevin Pasini

Université Gustave Eiffel

Forecast and anomaly detection on time series with dynamic context

Application to the mining of transit ridership data.

Thèse de doctorat de l'Université Gustave Eiffel

École doctorale n°532 - ´Ecole Doctorale Mathématiques et STIC (MSTIC)
Spécialité de doctorat: Informatique
Unité de recherche : GRETTIA

Thèse présentée et soutenue à l'Université Gustave Eiffel, le 11/05/2021, par

Kévin PASINI

Composition du Jury

Fabien MOUTARDE Full Professor, Ecole Nationale Supérieure des Mines de Paris, CAOR	Rapporteur
Romain BILLOT Full Professor, IMT Atlantique, LUSSI	Rapporteur
Ahlame DOUZAL Associate Professor, Université Grenoble Alpes, LIG-AMA	Examinateur
Martin TREPANIER Director of research, Ecole Polytechnique de Montréal, CIRRELT	Président du jury
Marc DERUELLE Innovation manager, LAB Mass Transit, SNCF	Invité
Mostepha KHOUADJIA Research Associate, Institut de Recherche Technologique System-X	Tuteur en entreprise & Examinateur
Allou SAME Researcher, Université Gustave Eiffel, GRETTIA	co-Directeur de thèse
Latifa OUKHELLOU Director of research, Université Gustave Eiffel, GRETTIA	Directrice de thèse

Résumé français

Pour répondre aux enjeux liés à l'augmentation de la demande de mobilité, aux problématiques environnementales et économiques, les transports en commun se sont imposés comme une des composantes essentielles des politiques de mobilité urbaine durable. Ces systèmes de transport permettent en effet de transporter un grand nombre de passagers pour un coût économique raisonnable et une empreinte écologique maîtrisée. Cependant, les réseaux de transport en commun font face aujourd'hui à des défis inédits en vue d'augmenter leur attractivité.

Les travaux de cette thèse s'inscrivent dans un contexte général qui vise à valoriser des données collectées sur l'infrastructure de transport par la conception d'outils d'analyse permettant d'extraire des informations à haute valeur ajoutée à l'intention des passagers, des analystes de données et des opérateurs de transport. Une première analyse exploratoire des données réelles (données SNCF transilien, et données du métro de Montréal) a permis de mettre en exergue les verrous scientifiques auxquels cette thèse s'est attaquée.

Les travaux de thèse comportent deux principaux volets. Le premier porte sur la prédiction court-terme de la charge voyageur dans les trains. La thèse introduit les approches et modèles usuels de prédiction à base d'apprentissage automatique, puis identifie les spécificités du contexte applicatif. La principale difficulté est liée à la variabilité intrinsèque des séries temporelles des charges à prédire, induite par l'influence de plusieurs paramètres dont ceux liés à l'exploitation (horaire, retard, type de mission. . .) et au contexte (information calendaire, grand évènement, météo, ...). Une autre difficulté est liée à l'échantillonnage temporel irrégulier des séries temporelles à prédire. Formalisé comme un problème de prédiction de séries temporelles avec un échantillonnage irrégulier et évoluant dans un contexte dynamique, la thèse s'intéresse alors à la conception d'un modèle LSTM encodeur-prédicteur capable de résoudre la tâche de prévision en faisant face à ces difficultés. Le modèle proposé est comparé à plusieurs modèles d'apprentissage automatique en se basant sur les performances de prédiction à plusieurs pas de temps.

Le deuxième volet de la thèse concerne la détection d'anomalies contextuelles sur des séries temporelles. L'objectif porte sur la détection de l'impact des perturbations sur l'affluence en station. Une spécificité applicative concerne la forte variabilité des séries temporelles qui doit être prise en compte dans l'étape de détection. Les travaux formalisent une approche de détection d'anomalies basée sur l'analyse des résidus de prédiction normalisés par une variance contextuelle estimée par apprentissage automatique. Cette approche vise à construire un score d'anomalie contextuellement robuste permettant de qualifier la déviation dans les séries temporelles en tenant compte de leur variabilité contextuelle. Les travaux sont d'abord évalués sur des données synthétiques. Puis ils sont appliqués sur les données réelles d'affluences en station pour quantifier l'impact des perturbations sur l'affluence en station et de détecter des anomalies inconnues.

Abstract

To meet the challenges of the increasing demand for mobility, environmental and economic issues, public transportation has emerged as one of the main components of sustainable urban mobility policies. Indeed, these transportation systems can carry large numbers of passengers at a reasonable economic cost and with a controlled ecological footprint. However, public transit systems today are facing unprecedented challenges to increase their attractiveness.

The thesis is part of a general framework that aims at valorising the data collected on transportation infrastructure by designing analysis tools that enable the extraction of high value-added information for passengers, data analysts and transportation operators. A first exploratory analysis of real data (SNCF transilien data, and data from the Montreal metro) highlighted the scientific obstacles that this thesis has tackled. The thesis has two main components. The first concerns the short-term prediction of the passenger load in trains. The thesis introduces the usual forecasting approaches and models based on machine learning, and then identifies the specificities of the application context. The main difficulty is due to the intrinsic variability of the time series of the loads to be predicted, induced by the influence of several parameters including those related to the operation (schedule, delay, type of mission, etc.) and the context (calendar information, major events, weather, etc.). Another difficulty is related to the irregular temporal sampling of the time series to be predicted. Formalized as a problem of time series prediction with irregular sampling and evolving in a dynamic context, the thesis then focuses on the design of an encoder-predictor LSTM model capable of solving the forecasting task by dealing with these difficulties. The multi-step forecasting performances of the proposed model are compared to several machine learning models.

The second part of the thesis concerns the detection of contextual anomalies on time series. The objective is the detection of the impact of the perturbations on the station ridership. An application specificity concerns the strong variability of time series which have to be considered in the detection step. The work formalizes an anomaly detection approach based on the analysis of prediction residuals normalized by a contextual variance estimated by machine learning. This approach aims at building a contextually robust anomaly score capable of qualifying the deviation in time series considering their contextual variability. The work is first evaluated on synthetic data. The approach is then applied to the actual data of station inflows. The objective is to quantify the impact of disturbances on station ridership and to detect unknown anomalies.

Remerciement :

Cette thèse est le fruit de trois ans d'efforts, mais également de nombreuses discussions et échanges avec toutes les personnes que j'ai croisée et qui m'ont accompagné durant ces trois longues années. Il me faut donc au moins citer celle qui ont eu le plus grand impact dans mes travaux :

J'aimerais tout d'abord remercier mes examinateurs de thèse, Monsieur Fabien MOUTARDE, Monsieur Romain BILLOT, Madame Ahlame DOUZAL, Monsieur Martin TREPANIER et Monsieur Marc DERUELLE pour leurs commentaires sincères et leurs retours enrichissants qui me permettent enfin d'être fier de l'ensemble de mes travaux.

J'aimerais ensuite remercier mon équipe encadrante qui m'a guidé tout au long de cette thèse : Etienne qui malgré son passage bref dans mon encadrement, fut une source d'inspiration importante, Martin dont la gentillesse et l'accueil m'ont permis de vivre une expérience formidable à Montréal, Allou qui fut un pilier dont la rigueur mathématique et la bienveillance m'ont beaucoup soutenu pendant les rédactions, Mostepha qui m'a accompagné depuis 4 ans, m'a vu m'affirmer en tant que chercheur, en m'accordant toute l'autonomie dont j'avais besoin, tout en me prodiguant toujours de précieux conseils. Et enfin Latifa : Je serais redevable à vis de ton dévouement et de ton implication dans cette thèse, j'ai eu beaucoup de chance de t'avoir en directrice.

Mon passage à l'IRT n'aurait pas été si enrichissant sans de nombreuses personnes, Ingénieurs de recherche et partenaire de projet (Mostapha, Fereshteh, Ahmed, Fabien, PO et tant d'autres), stagiaires (Thibault, Jules, Paul, Jacques), doctorants (Adrien, Antoine, Elies, et toute la team DOCSX). Deux petites mentions spéciales à Johanna (nos échanges quotidiens était toujours un petit rayon de soleil en plein cœur de morne journée de rédaction) et Pascal (Indéniablement mon frère de thèse et une source infinie de soutien et de rire).

Je conclurais avec l'ensemble de mes amis et de ma formidable famille, qui sont toujours d'un soutien indéfectible en tout circonstance, avec une pensée particulière pour Bernard et Agnès, mes parents, qui m'ont toujours apporté tout en me permettant de devenir qui je suis, j'aimerais pouvoir vous en rendre ne serait-ce que la moitié.

Enfin, J'aimerais dédier cette thèse à Edouard, mon petit neveu parti trop tôt, mais qui est désormais une étoile qui nous guidera à jamais dans nos vies.

Contents

1 Introduction

1.1 Transportation Issues

In our contemporary societies, urban mobility is a central issue with strong societal implications. Our lifestyles revolve around numerous journeys related to work, leisure, shopping activities and many other things. The conditions of these trips in terms of duration and comfort have a direct impact on our quality of life. The subject of mobility is all the more important for large conurbations such as Ile-de-France, which has a high population density and generates significant amount of daily travel (11 million ticket validations per day). There are three main modes of transportation for daily travel: individual motorized modes (cars, motorcycles, etc.), public transportation (bus, train, metro, etc.) and soft modes (walking, cycling, etc.). The choice of transport modes at the individual level will depend on several factors related to mobility policies that structure the different transport offers and their costs. These policies today require complex thinking that must integrate many societal issues, including quality of life, environmental impact, the economy and urban planning.

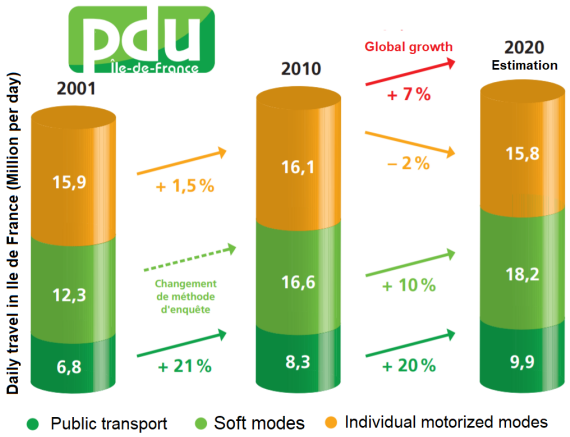

Figure 1.1.: Evolution of the number of annual trips in the Paris area.

Since the industrial revolution, a large part of economic and cultural activities have been centered in the metropoles. This has attracted a growing numbers of people to urban areas that have become increasingly crowded (30 percent of the urban population in 1950 and 55 percent in 2018), leading to a constant increase in the demand for daily mobility (Figure 1.1). To deal with the influx of population, cities are spreading out more and more to the suburbs. This peri-urbanization has gradually led to the creation of suburban areas that have evolved into true suburbs that now form a continuous dense urban fabric. This development often goes hand in hand with the transport infrastructure, since the quality of the transport supply is a major factor in the attractivity of the area, which subsequently generates a strong demand for transport in these new urban areas.

In addition, new thoughts are also emerging on the impact of urban planning on quality of life. The need for healthier living spaces (more spacious, less polluted, more vegetated and conducive to social interaction) in overcrowded areas also invites a rethinking of urban design. One of the main issues is the importance of the transportation infrastructure, which can be not only cumbersome but also a source of nuisance and pollution. Many major European cities such as Oslo, Madrid, Paris[1] and Copenhagen are implementing pedestrian policies coupled with incentive initiatives in the form of bonuses and subsidies for certain sustainable modes of transport, particularly cycling and public transport. (Figure 1.2).

Figure 1.2.: Pedestrianization of the right side of Paris Quai de Seine
*Image retrieved from the official website of Paris city [1].

Furthermore, today's ecological challenges are increasingly important in urban mobility issues, bringing with them new environmental constraints. These constraints aim to limit pollution (degradation of air quality, greenhouse gas emissions) and the nuisance produced by modes of transport in city centers. When considering the issue of global warming, the combination of high growth in urban populations and peri-urbanization is partly attributable to mobility [CGR02]. The transport sector, of which urban mobility is a significant part, is one of the main contributors to greenhouse gas emissions (23% according to the IPCC (IPCC) [Sim+14]). However, this sector has financial resources and more sustainable mobility solutions (car sharing, public transport, soft mobility) that can also partly replace the use of certain more pollutant modes, such as the private vehicle. This makes urban mobility one of the first levers of the ecological transition.

To meet the challenges of the increasing demand for mobility as well as the environmental and economic issues, public transportation (bus, train, metro, light rail) has emerged as one of the essential components of sustainable urban mobility policies. These transportation systems can carry large numbers of passengers at a reasonable economic cost and with a controlled ecological footprint. However, public transit systems today are

[1]https://www.paris.fr/pages/le-saviez-vous-le-trafic-automobile-peut-s-evaporer-4080/

facing unprecedented challenges to increase their attractiveness. The constant increase in demand for urban mobility is leading to situations of near saturation on portions of the transportation network. The combined spread of cities and transportation networks has also made infrastructures more and more complex. In this context, the planning of a transportation plan intended to optimize transportation supply to meet mobility demand comes up against the management of hazards inherent in the operation of any such complex system. In addition, the organization of human activities imposes a very high concentration of mobility demand over short time periods (peak hours), resulting in intense stress on the transportation network. An unforeseen event related to an operating incident can have severe consequences. The temporary suppression of part of the supply causes saturation that spreads throughout the networks on alternative routes, leading to cascading malfunctions in an already weakened system. The impact of this disruption is then very costly in time and money, leading to long waits and delays for hundreds of thousands of people.

In addition, the expectations of transit users, who today spend an average of one hour a day in transit, have also evolved. In addition to the demand for faster travel times, users also want to travel in greater comfort (avoid congested trains, have a seat) in order to make their travel time (reading, work, etc.) more profitable. They also want to be better informed in the event of a problem so that they can better plan their daily trips and choose better alternatives when they face a disrupted situation. In 2018, passenger information became a priority for many transportation operators. The rapid and effective dissemination of information can provide confidence in the operator's ability to manage the crisis, and can even encourage users to shift trips or adopt alternative routes to reduce the pressure on sections of the transportation network that are under stress.

The operational management of transportation systems has thus become a colossal challenge with high stakes. In conjunction with this ecological transition, the field of transportation has been undergoing a digital transition. The possibilities available in terms of data collection and storage allow for the renewal of modeling approaches in the field of transportation. The cross-exploitation of various data sources is aimed at creating services with high added value for the user. The decision-support tools developed can help to better understand people's mobility behavior, improve the diagnosis and management of saturation and disruption situations, and generally better monitor transportation networks. Transportation operators are currently setting up near real-time data collection infrastructures capable of providing information on train loads, station ridership, passenger flows in stations, and vehicle positions and delays. Finally, open source initiatives, encouraged by certain states and the European Union, are pushing transportation operators to open up access to mobility data, the use of which is intended to create new mobility services for citizens.

Figure 1.3.: Control center of the commuter train line 'RER B' of Paris
*Image retrieved from the official blog of the RER B line[1].

However, the analysis of these massive and structured data, impacted by numerous contextual factors and presenting spatio-temporal dynamics remains a complex academic issue. The design of advanced analysis tools will allow us to better consider the abundant contextual structuring information in machine learning-based approaches. In the public transport sector, the extraction of indicators that provide information on the current state and enable the future state of transport systems and networks to be anticipated will provide valuable information for the supply dimensioning, network regulation and dissemination of passenger information.

With the explosion of computing power in computers, combined with advances in data science and learning algorithms, it is now possible to develop advanced analysis tools capable of processing data that are both massive and complex. This is a topical issue known as 'Urban Computing' [Zhe+14] in which a central question in the field of mobility in public transport is how to design decision support tools that can be used to make the most of the massive data collected from the public transport infrastructure to help manage it better and thus increase its attractiveness.

The new generation of analysis and decision support algorithms will be integrated in the next few years in the supply control and planning centers. A few more years of research will be needed to apply the first intelligent infrastructure systems. Additional years will undoubtedly be needed to enrich and perfect the operation of these intelligent systems. However, these systems will gradually go become the key in the future to managing transportation systems and networks because of their anticipated ability to:

- understand the mobility behaviors of users in detail
- analyse *a posteriori* the impacts of disturbances and regulation strategies
- achieve a fine-grained diagnosis of the state of the transport network

[1] https://www.rerb-leblog.fr/centre-de-commandement-unique-la-tour-de-controle-du-rer-b/

- predict the evolution of this state in the event of a disturbed situation
- help assess the relevance of a regulation strategy or a modification of the transport supply.

These decision-support tools will not replace the operators in charge of network management, quite the opposite, they will lead to an upgrading of jobs in the field of transportation and mobility. On the one hand, new qualified personnel will have to supervise and evaluate the correct operation of these analysis systems. On the other hand, they will also have to focus on more complex tasks related to optimizing the transportation network and decision-making in crisis situations, where they have to reach a compromise that is difficult to measure or that involves human responsibility.

1.2 Context of the thesis

The thesis took place at the Technological Research Institute (IRT) System X. This is a scientific cooperation foundation that carries out numerous R&D projects around the themes of digital engineering of the future by involving industrial and academic partners. Within these projects, research engineers collaborate to carry out proofs of concept responding to the use cases defined by the industrial partners. These are also carried out on specific subjects in partnership with academic researchers. The Enhanced Passenger Information (IVA) project is an R&D project led by the IRT systemX on the theme of mobility. It brings together major players in the field of mobility, namely the SNCF transport operator, the transport authority Ile-de-France Mobilités (IDFM), companies providing digital solutions such as Kisio (intelligent browser), Spirops (multi-agent simulator) and the GRETTIA laboratory of Gustave Eiffel University (formerly IFSTTAR), an academic laboratory specializing in mobility studies.

Launched in September 2017 for a period of four years, the IVA project aims to optimize passenger travel on the entire Île-de-France multi-modal transportation network and to provide analysis, modeling and visualization tools to enrich passenger information and assist regulators. In particular, it also aims to gain a better understanding of the state of the multi-modal transport system in the short and medium term, develop a mobility assistant based on Artificial Intelligence, and model passenger behavior in relation to passenger information in order to identify the most suitable scenarios in the event of a disrupted situation.

The project participants are composed of 4 research engineers and 2 PhD students, supported by two academic researchers and several resources made available by the industrial partners. It includes two theses which aim to analyze mobility behaviors from two different angles :

- Analysis of mobility practices and load prediction in public transportation; the case of Paris city (data science oriented analysis)
- Impact of Traffic Information on Transit Passenger Behavior (Social Sciences and Humanities Oriented Analysis of Mobility Behavior)

The thesis is in a semi-industrial and semi-academic context, which led to numerous exchanges with research engineers and industrial partners of the IVA project (feedback on data, regular presentation of work, exchanges related to computer developments). The work has directly fed into the various proofs of concept of the IVA project.

1.3 Thesis Outline

After a general introduction, the thesis is divided into 3 chapters.

- The first chapter presents the different types of data used in the research about public transport mobility and the dedicated issues. It then details, through exploratory analyses, the two sets of data, highlighting their specificities that need to be taken into account in further processing. The first set of data concerns the passenger load in the trains of a multi-branch line serving the northern suburbs of Paris, where the collection period of one and a half years is extends from January 2015 to June 2016. It contains data on passenger load in trains as well as theoretical and actual transportation plans. The second set of data concerns ridership in Montreal metro stations over a two-year time period between 2015 and 2016. It provides measures of ticketing data at station entrances aggregated by quarter-hourly intervals for a total of 15 stations. This second set of data also includes a disturbance database that provides information on some of the events and incidents that took place during this collection period.
- The second chapter deals with short-term prediction of train load. It starts by presenting standard approaches and models for prediction based on machine learning. The main difficulty in forecasting is related to the intrinsic variability of the time series of the loads to be predicted, induced by the influence of several parameters including those related to the operation (schedule, delay, type of mission...) and to the context (calendar information, major events, weather...). Another difficulty is related to the irregular temporal sampling of the time series to be predicted. We propose an encoder-predictor LSTM model to solve this forecasting task. Several experiments were conducted on real data. The forecast results are detailed in order to show the interest of feature engineering and to compare the performance of such a model at several time horizons with other more classical models used in forecasting.
- The third chapter deals with the detection of contextual anomalies in time series. After a general introduction to the problem of anomaly detection, we present the

issues related to this problem in the case of transport data. The objective is to detect the impact of disturbances on station ridership. An application specificity concerns the strong variability of the time series to be processed, which will have to be taken into account in the detection step. To this end, we formalize an approach to detect anomalies based on the analysis of prediction residuals normalized by a contextual variance estimated by machine learning. This approach aims at building a contextually robust anomaly score capable of qualifying the deviation in time series taking into account their contextual variability. The construction process of the anomaly score and the different models proposed for the estimation of the contextual variance will be presented. The work is first evaluated on synthetic data generated with the presence of anomalies. It is then applied to the actual data of station inflows. The objective is to quantify the impact of the disturbances on station ridership and to detect unknown anomalies. The results first illustrate the performance of prediction models, a 'limited quantitative' analysis of anomaly detections, then a more qualitative analysis also conducted on a few examples to illustrate the relevance of variance accounting for the detection of contextual anomalies.

1.4 Motivations & Thinking

The motivations for the thesis aim at valorising the different sources of data collected on the transport infrastructure that is now undergoing a process of maturation. The objective is to set up digital platforms driven by data collected in real time and related analysis algorithms in order to enrich passenger information and provide tools to help transport network management for operators in charge of planning and regulation. The work initially focuses focuses on the Transilien Line H operated by the SNCF company due to collection experimentation. Two operational objectives are targeted:

1. forecasting train/station ridership taking into account transportation supply and demand.
2. analyzing the impact of disruptions on ridership.

The indicators extracted in this way are intended to better anticipate changes in mobility demand and help regulation in disrupted situations. They also aim to enrich passenger information, thus providing public transit users with information enabling them to better plan their daily trips and to be well informed in the case of disruptions of possible impacts on their scheduled trips.

The task of exploratory data analysis raised a number of issues and challenges that led to frequent discussions with the industrial teams in charge of data collection and its analysis. It also made it possible to extract, build and consolidate a data set for the 'Short-term load prediction in trains' part. On the other hand, on the Parisian perimeter, problems of data capture in disturbed situations prevented the achievement of the work on the

'anomaly detection' part within the deadlines of the thesis. A 4-month international mobility opportunity in the Montreal laboratory for transportation research (CIRRELT) in partnership with the Montreal transportation company (STM) allowed work on this aspect to be pursued. This work was carried out in partnership with Professor M. Trépanier. Within this framework, we applied the anomaly detection methodologies to time series of ticketing data in the Montreal metro. This work is now being applied on the Parisian perimeter by the research engineers of the IVA project.

1.5 Contributions & publications

The work has made several contributions that combine mobility data analysis, data science and machine learning. It has led to several scientific publications and transfer valuations within the SystemX IRT.

The first part of the work focused on the **exploratory analysis of the data** made available, the collection of which is currently in the process of maturation. An in-depth analysis, merging and formatting were carried out. This enabled us to qualify the data, to identify certain weaknesses related to the collection infrastructure and to identify issues of interest. These analyses were carried out as the data were being delivered, and led to several exchanges with SNCF teams in charge of data collection or dedicated to the implementation of a platform for the analysis and visualization of the collected data.

In the second part of the thesis, we aim at the **short-term prediction of loads (the number of passengers) of trains** in public transport. The forecast time horizon is short term, i.e. 15mn, 30mn or 1h. This forecast exploits two heterogeneous data sources: passenger counting data and automatic train location data.

A large part of the work carried out in the field of transport demand forecasting concerns the forecasting of train station traffic or flows at an aggregated level (every 15 minutes, 30 minutes, etc.). The originality of the work presented here lies in the taking into account of the transport plan carried out for the forecast, which may differ quite significantly from the nominal[1] transport plan. In fact, we have chosen to develop a univariate prediction model for each station considering the following specificities :

- A variable sampling period due to train schedules and railway operations. Each station has its own train frequency.
- A specific time profile of each series. The profile is directly related to the use of the station and in particular to its spatial location and the geographical characteristics of the surrounding urban area (population density, employment density, leisure, etc.).

[1]The nominal operation of a system corresponds to an operation without unexpected events or anomalies.

- Train load series are influenced by calendar factors such as the type of day (weekday or weekend), public vacations, school vacations, etc.
- Train load series are also impacted by the characteristics of trains and their missions (multi-destination line, various rail services).

The forecasting task is seen as a multi-horizon forecasting problem over irregular time series impacted by several contextual factors. In order to take these particularities into account and by relying on the abstraction capabilities of neural networks combined with representation training, we propose an encoder-predictor LSTM model combined with representation training on contextual factors. The goal is to predict the train load from a station over several time steps, taking into account past load values and all the contextual factors characterizing the train operation. This work gave rise to several publications [Pas+19a; Pas+19b; Pas+19c].

The third part of the thesis deals with the theme of **anomaly detection in multivariate time series evolving in a dynamic context**. In terms of applications, the aim is to quantify the impact of perturbations linked to events or incidents on the time series of station ridership. One of the specificities of this problem lies in the inherent variability of the series which must be taken into account for the construction of an **anomaly score**. The presence of a dynamic context impacted by a set of influencing factors (calendar and spatial in particular) makes the modeling of normal ridership non-trivial. The work also focuses more specifically on the exploitation of the **prediction residuals** of the models developed in Chapter 3 for the construction of a robust anomaly score that takes into account the **dynamic context** in which the series evolve. Two questions have to be addressed: How to characterize and model the dynamic context? How to quantify statistical anomalies in time series taking into account the dynamic context?

After a formulation of the problem of contextual anomaly detection on time series, an approach consisting of two main steps has been proposed. The first step is based on the definition of the dynamic context in which the series evolves through the estimation of contextual statistics (means and variances). In particular, we propose to estimate the variance of the data by exploiting the prediction residuals of a learning model dedicated to this task or by extracting it directly from a set prediction model (Random forest type) or by deep learning. The second step is dedicated to deepening the approaches of anomaly detection based on the analysis of the prediction residuals by applying in particular the formalism and models developed in Chapter 3 of the thesis and by combining them with the estimated contextual statistics. The evaluation of all the approaches developed is carried out both on synthetic data and on real ridership data collected on about fifteen Montreal metro stations. The synthetic data made it possible to quantitatively evaluate the anomaly detection performance of the different approaches in a controlled experimental framework. On the real data set, we have an incomplete database of disturbances. The objective is to illustrate the impact of certain disturbances listed by the transport operator in the database or others not listed but highlighted by our approach. The aim is to better

understand the impact and propagation of disturbances on transport networks in order to extract information that can help refine regulation strategies, anticipate the evolution of disturbed situations and enrich passenger information in disturbed situations. Contextual statistics are used here to construct an anomaly score, but they can also be used to qualify the contextual variability of data for data mining purposes, or to quantify margins of error for prediction models through prediction confidence intervals.

While our work has mainly focused on the detection of anomalies, it is directly applicable to other purposes. It is currently being valorized with a review article submitted to an international journal [Pas+19d], a conference paper in preparation [Pas+21] as well as an oral presentation given during a workshop [Pas+20a].

The work accomplished during this thesis also provided the opportunity for industrial valorisation by the partners of the IVA project. This is currently being completed. The analysis bricks aim at extracting, from data collected in real time, forecasting indicators that inform on the evolution and abnormality of the train load and station ridership. They are part of the global architecture of the IVA project (Figure 1.4). This work is part of a multi-thematic research project aimed at designing network state analysis tools for the control system [Val+], and which can also be used in the 'intelligent' route planner [APK19].

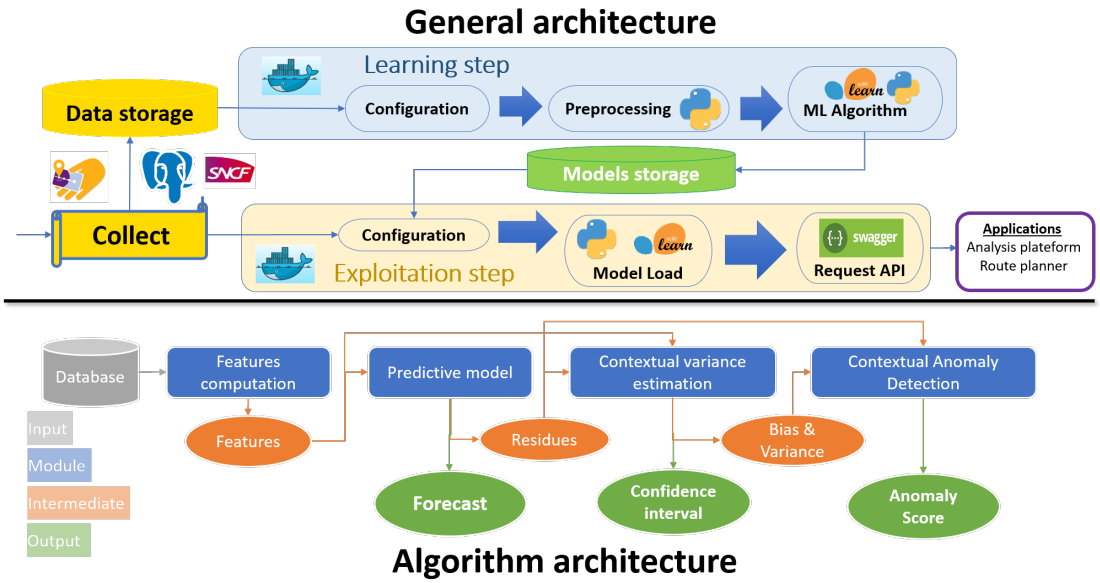

Figure 1.4.: Pipeline of IVA data analysis tool

Publications

International machine learning conference:

- Kevin Pasini, Mostepha Khouadjia, Allou Same, Fabrice Ganansia, and Latifa Oukhellou. "LSTM encoder-predictor for short-term train load forecasting". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham. 2019, pp. 535–551

International industrial conference:

- Kevin Pasini, Mostepha Khouadjia, Fabrice Ganansia, and Latifa Oukhellou. "Forecasting passenger load in a transit network using data driven models". In: *WCRR 2019, 12th World Congress on Railway Research*. 2019

National machine learning conference:

- Kevin Pasini, Mostepha Khouadjia, Allou Same, Fabrice Ganansia, Patrice Aknin, and Latifa Oukhellou. "Modèle LSTM encodeur-prédicteur pour la prévision court-terme de l'affluence dans les transports collectifs". In: *CAP 2019, Conférence sur l'Apprentissage Automatique*. 2019

Submitted:

- Kevin Pasini, Mostepha Khouadjia, Allou Same, Martin Trepanier, and Latifa Oukhellou. "Contextual anomaly detection on time series: A case study of metro ridership analysis". 2020
- Kevin Pasini et al. "Anomaly detection on time series evolving in a dynamic context. Application to smart card data analysis". To be submitted to ESANN. 2021

Workshop talk:

- Kevin Pasini, Allou Same, Mostepha Khouadjia, Fabrice Ganansia, and Latifa Oukhellou. "Representation Learning of public transport data. Application to event detection". In: *5th International Workshop and Symposium TransitData 2019*. 2019
- Kevin Pasini, Mostepha Khouadjia, Allou Same, and Latifa Oukhellou. "Modèle LSTM Encodeur-Prédicteur pour la Prévision Court-Terme de la Charge Passagers dans les Transports en Commun". Workshop virtuel - Nouvelles méthodes pour l'analyse descriptive et prédictive de données massives et structurées. 2020

Co-author contribution:

- Ahmed Amrani, Kévin Pasini, and Mostepha Khouadjia. "Enhance Journey Planner with Predictive Travel Information for Smart City Routing Services". In: *2020 Forum on Integrated and Sustainable Transportation Systems (FISTS)*. IEEE. 2020, pp. 304–308

2 Exploratory analysis of public transit data

Introduction

Transport operators have initiated a digital revolution with the deployment of numerous connected sensors and the implementation of data collection and storage infrastructures that are now maturing. Massive databases store historical measurements captured in real time providing information about the state of transportation supply and demand. Operators also want to identify and store all the hazards that could have an impact on mobility, whether they are related to regulation, maintenance, malfunctions or cultural, sporting and social events. All data relating to public transportation systems can be classified into three categories:

- data qualifying the transportation offer,
- data measuring mobility demand,
- data informing on hazards that can impact transportation demand and/or supply.

The analysis of mobility must consider the complex relationship between supply, demand and hazards that may impact them in order to better understand the dynamics governing the mobility system. The combination of these data initially aims at designing systems that analyze situations a posteriori to extract relevant indicators for the qualification and management of the transportation network. The design of data analysis algorithms is a major issue at the heart of much research and is a crucial step in the implementation of intelligent transportation infrastructures capable of analyzing the state of a network from data collected in real time.

The overall objective is to optimize the management of a transportation network in both nominal and atypical situations. This will involve providing rich indicators to qualify and anticipate the impact and propagation of a disruption on the network, facilitate decision-making and enrich the passenger information to be disseminated to resolve an anomaly.

In this chapter, we will briefly introduce the data from various sources collected in public transportation and the main issues studied in several scientific studies. In a second step, we conduct an exploratory analysis of mobility data from the two case studies on which

the thesis focuses: passenger load data on Parisian commuter trains and ridership data on the Montreal metro network.

2.1 Transit Data

Within the framework of mobility data analysis, there are several issues ranging from data enrichment, characterization of transportation supply and demand, prediction of short and long term indicators on supply and demand (delay, ridership, loads, ...), to the detection of anomalies (unknown events, impact of disturbances). These issues can be addressed by exploiting various data sources.

We will first describe the different data sources used in research on mobility and public transport, and then discuss the various problems that they allowed us to address.

2.1.1 Transportation supply data

The transportation offer corresponds to all the resources made available by the various transportation operators to meet the mobility needs of transit users while considering safety and technical constraints. A distinction is made between the theoretical offer and the realized offer:

- *The theoretical transportation offer* is planned in advance to set the theoretical transit schedules of all vehicles at each stop for each day of the week. There are several standard formats for specifying this offer, the most widely used of which is the General Transit Feed Specification (GTFS) format. This is a general specification of the transportation offer that takes the form of a set of data tables structuring all train passages at the stations according to the timetable, operators, missions and transportation network.

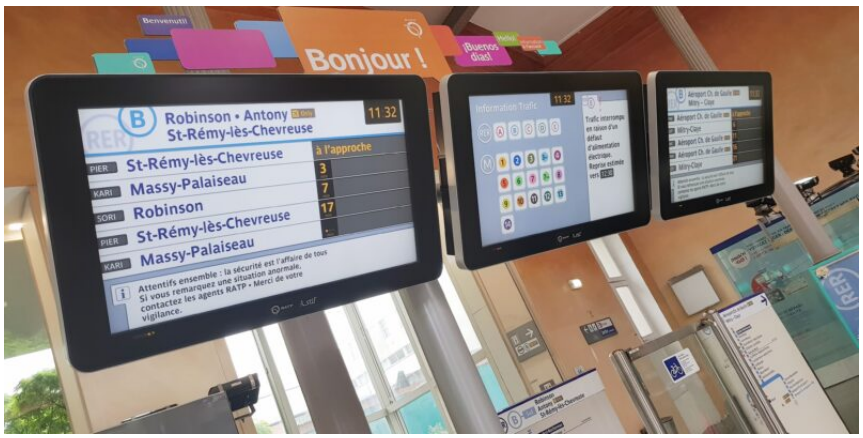

Figure 2.1.: Real time schedule of a Parisian commuter train line
*Image retrieved from the official web-blog of line RER B [1]

- *The realized transportation offer* corresponds to the set of trips actually made (Figure 2.1) collected in real time by several types of 'AVL' sensors (Automatic Vehicle Location). Some of these sensors provide the GPS position of the vehicles, while others record their passage through stations. All the information is collected and pre-processed in real time to be stored in tables containing details about the train (number, type of train, mission), the stop (line, station, platform) and the timetables. As the collection infrastructure can experience failure (sensor failure, communications or software issues), several pre-processing operations are necessary to ensure data quality.
- *Transportation plan alteration* corresponds to the difference between theoretical and realized planning. It is related to the hazards and changes that occur on the transportation network that may require some regulation strategy (delay, modification of train service, deletion or addition of trains). It is therefore one of the measures that determine the quality of operation of a transportation network and its resilience. The operational management of a rail network requires considering numerous safety and logistical constraints (track occupancy, equipment rotation, distance between trains) that facilitate the propagation and aggravation of a disruption. Operators have limited levers for managing disruptions, especially on a saturated network such as the one in the Paris region.

2.1.2 Transport demand Data

Demand is one of the most complex components of a transportation system. It is structured both by the transportation supply and by a set of exogenous factors such as the spatial distribution of activities, socio-demographic areas, temporal organization of activities, urban planning, cultural habits, ... The demand results from transit users who travel across the transportation network from an origin to a destination with time and accessibility constraints as well as preferences for comfort or trip characteristics.

Transport demand data are massive, spatially and temporally distributed in different forms (e.g., individual traces, aggregated flows into and/or out of a transportation network). It also has different levels of granularity (train, platform, station, line, network), spatial, and temporal aggregation.

Surveys

Survey data are declarative data that allow the observation and the analysis of mobility on a large scale. Travel surveys take the form of a survey of individuals' travel habits. They collect information on individuals (socioeconomic, demographic, etc.), households (size, structure, composition) and their trips (time of start and end of trip, location, mode

[1] https://www.rerb-leblog.fr/temps-dattente-saffichent-desormais-ecrans-dinformation-rer-b/

of transport used, purpose of trip). Major travel surveys are conducted in metropolitan areas, on average once every decade. They are sometimes replicated on a small panel to observe changes in travel behavior. In France, these include the national transport survey (ENT) and the household travel survey (EMD). Although these data have a number of advantages (knowledge of travel patterns, insight into users' socio-economic status, ...), they also have limitations related to poor temporal tracking of users, stereotype of average weekday trip study, survey bias, and costs involved in carrying them out. The emergence of digital survey systems makes it easier to conduct and analyze targeted and more precise surveys on small panels, but with the bias of representativeness linked to the use of digital technology.

Automatic Fare Collection Data

Ticketing data measure the flow of users (entry, exit, transfer) moving through the transportation network. These data are collected through station access gates which record the validation of ticket or transportation cards.

Figure 2.2.: Validation gate of public transport
*Image retrieved from the official web-blog [1] of line transilien J.

A distinction is made between entrance validation (TAP-IN) and exit validation (TAP-OUT). The accuracy of these ticketing traces are linked to the arrangements inside the stations. On some networks, transportation line delimitation offer rich ticketing traces that capture the transportation connections. Other networks such as the Parisian network only collect the entrances to the network (TAP-IN) in order to fluidify traffic within the stations and in relation to pricing policies linked to price zones rather than distance traveled. Ticketing data can be used in the form of :

[1] https://malignej.transilien.com/2013/09/16/mise-en-service-du-passage-elargi-controle-aux-claireres-de-verneuil

- *Individual trajectories*. The traces of individual ticketing validation can be directly analyzed using the anonymized identifier of transportation cards which changes several times a year for privacy reasons (every three months on average). These data have been studied in several scientific studies (Section 2.2.2) aiming at characterizing travel behaviors through notions of travel frequencies, patterns and regularities.
- *Aggregated flows* In order to analyze mobility flows on a large scale, all individual input/output validations are aggregated for a temporal granularity (from a whole day to a few minutes depending on the infrastructure) over a predetermined perimeter (line, station, network portion). The result is a set of time series that can be used to analyze the evolution of ridership on the public transit network. Research investigations are mainly interested in the modeling of series dynamics in order to better understand their evolution and be able to make reliable predictions (Section 2.2.3).
- *Origin/Destination (OD) matrices*. From the pair of records (TAP-IN,TAP-OUT), it is possible to build OD matrices that synthesize the trips by aggregating those with the same origins and destinations. The first OD matrices, reconstructed from survey data and concerning all modes, were static and did not include the temporal dimension. Ticketing records (TAP-IN, TAP-OUT, connection) allow the construction or estimation of dynamic OD matrices for public transport. These matrices aggregate trips with the same origin and destination for regular time periods. These periods can range from a whole day to a few minutes, depending on the temporal granularity of the data collected. Much work has been done on the estimation, prediction, or analysis of these OD matrices (Section 2.2.1).

Passenger Load Data Passenger load refers to a quantity of passengers on a platform or in a train at a given time. It is a measure that makes it possible to estimate the saturation on a portion of the network. There are several techniques for estimating load from different data sources. We can cite the following:

Figure 2.3.: Train door equipped with load sensors, ©transportparis [1]

- *Train weighing*. This is the basic technique to estimate the load in trains. Piezo-electric load sensors are placed at several locations on the railway infrastructure to measure the mass of the trains. By calculating the ratio between the passenger mass (Measured Mass - Empty Mass of Trains) and the average weight of a passenger, this gives us a estimation of the average number of passengers in a train.
- *Counting at the Doors* which aims to count the number of passengers boarding and alighting at the doors of a train using several types of sensors (laser, radar). The load of the train is then estimated over the entire train service.
- *Video sensors* which now make it possible, thanks to the progress of video surveillance and pattern recognition algorithms, to carry out automatic counts in real time using surveillance cameras. These cameras can estimate the number of people on a platform or the boarding and alighting traveler on the train. However, there are privacy issues which may limit the use of this type of counting for analysis purposes.

Digital traces of mobility

Cell phone and internet connection infrastructures can also enable the collection of rich spatio-temporal data by exploiting the almost permanent use of cell phones. However, this additional source is impacted by issues of representativeness and depends on delicate re-calibration due to the partial observation of transportation users. Within this category, several sources of information can be distinguished:

- Mobile phone data. Telephone operators collect a large amount of information related to passive (network roaming) and active (mobile data consumption, phone calls) connections to phone relay antennas. This provides approximate information on the location of the user, particularly in underground transport with antennas dedicated to transportation users. These data can then be studied by aggregating anonymized trips to extract flows.
- *GPS data*. GPS location data from phones can be collected in agreement with the user by various applications related to mobility (Movit, Google maps, Transport Operator, ...).
- *WIFI Data*. Several experiments aim to install WIFI terminals on the transportation network to provide free Internet access to users. It is then possible to collect information on telephone connections to the WIFI network. Individual trajectories can be aggregated to extract indicators related to waiting time or serve as complementary information for the reconstruction of OD matrices. Several massive WIFI data collection experiments are currently being conducted by the operator TFL (Transport for London) on the London transportation network [2].

[1] http://transportparis.canalblog.com/tag/Francilien/p10-0.html
[2] https://tfl.gov.uk/corporate/publications-and-reports/wifi-data-collection

Route planners

There are many online route planners that offer to provide a detailed route from an origin, a destination and a time constraint. They are based on operational search algorithms that optimize routes based on transportation schedules provided by operators. There are various calculators that can be linked to transport operators (Transilien), transport organizing authorities (Vianavigo), or independent companies (Citymapper, Google Maps, Move-it). These online applications can be based on internal tools or commercial solutions. The calculator of the London transport regulator TFL is powered by Google, while the calculator of the Paris public transport authority IDFM is powered by Navitia, a route calculation engine developed by KISIO-DIGITAL, a subsidiary of the French rail operator SNCF.

Route planners are becoming more and more powerful. They can be consulted from a smartphone to respond almost instantaneously to increasingly complex requests, considering the state of the network in real time. They have become one of the main sources of information related to mobility. However, route planner queries are an almost unexploited data source until now.

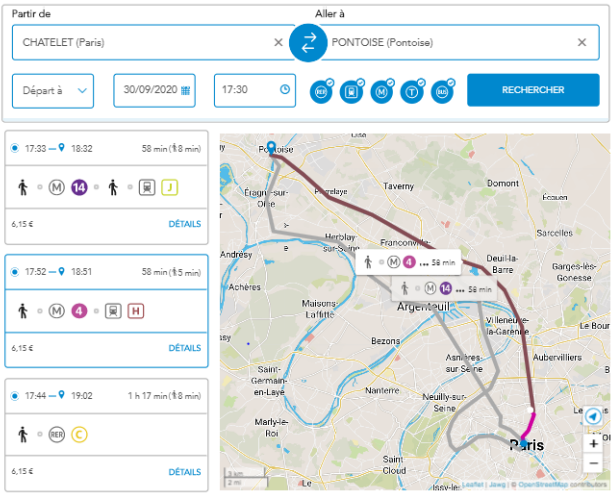

Figure 2.4.: *Transilien* route planner application [1]

In 2005, the study by [TCA05] showed that the analysis of these route requests could be a rich source of additional information and would provide a better understanding of user habits and the impact of transport disturbances. The limitation raised in the article about use constraints no longer holds, with the emergence of the Internet and smartphones. A more recent study [SSC16] illustrated with a particular case the interest that the storage and analysis of route requests can represent. Their conclusions show that it would be possible to extract mobility trends that could help in planning the transportation. As these two studies indicate, route queries provide information on travel intentions that differ from the actual mobility demand. These significant biases (Multiple trip requests, Specific

[1] https://www.transilien.com/

sample) restrict this source of information to being only a complementary indicator that can illustrate mobility trends.

2.1.3 Exogenous data

The operating state of a transit system depends on many factors, contexts and hazards that are sometimes difficult to capture. There are a number of data that can provide valuable insights to better understand the global context around operations (supply and demand). We can distinguish between endogenous hazards (technical/human incidents, maintenance work) directly related to operations, which often impact the transportation supply, and exogenous factors (cultural or sporting events, demonstrations, strikes, climatic conditions) that do not depend on the network but can significantly impact the mobility demand. In addition to their own data, transport operators are beginning to collect a large amount of information from a variety of sources in order to facilitate the analysis and use of mobility data:

- *Meteorological data*. Information related to weather conditions can impact transportation demand by changing users' habits. For example, walking and outdoor trips are reduced in rainy or low temperature conditions. Such data can be obtained from various public and private weather agencies.
- *Operator incident data*. A lot of information related to the regulation of the transportation plan and the management of technical problems is fed back into internal information infrastructures at the various operators. This information is sometimes collected to contextualize the hazards that may have an impact on the transportation network. Operators in charge of passenger information send relevant information to users in disrupted situations through various passenger information channels (station displays and announcements, social networks, telephone applications). The incident databases are either directly derived from the internal infrastructure, which implies rich but sometimes unstructured information, or collected from passenger information system logs. In a cross-source data analysis objective, operators are working to consolidate 'disruption' databases. These databases lists event detailing the type, start and end time of the incident, the spatial area of impact, and the characteristic elements related to the hazard.
- *Event data*. Events (musical, sporting, social) can significantly modify the transportation demand within a defined spatial and temporal perimeter. There are many online sources (Tourist office, Concert Hall, ...) that list some events. However, it is difficult to exhaustively collect all the events taking place in a metropolis. Several academic studies analyze the impact of events on transportation networks by exploiting event databases collected manually or recovered by Web scrapping.
- *Social network data*. Transport operators are increasingly investing in social networks (Facebook, Twitter, ...) to have channels for the dissemination of passenger

information that are particularly well followed by users. Official transit line accounts now broadcast real-time information on the line's operating status on Twitter. Social networks can also be used as a source of real-time feedback through reports of anomalies made by network users or by analyzing their feelings about disruptions. It can also be used more widely as a complementary source for the collection of exogenous factors, particularly for the collection of event data.

2.2 Problems of interest

Several objectives are targeted by the analysis of mobility data. In the following, we will detail 4 categories of application purposes related to the exploitation of mobility data.

2.2.1 Data enrichment

One of the first problems in analyzing mobility concerns the enrichment, completion, reconstruction or cross-checking of the information collected through these different sources.

Some studies have focused on the cross-referencing of data from different sources. For example, the work of [Kon+18] focused on the cross-checking of traces from telephony and ticketing validation. The authors proposed a technique for cross-referencing trajectories from digital traces, based on spatial and temporal considerations in order to find concordances between the two sources. Other authors seek to extract mobility indicators from complementary information sources. The work by [Lu+17] showed that the use of GPS data to estimate of mobility indicators at the individual scale (total travel distance, movement entropy) must consider the estimation biases related to hourly factors. The work by [El +17] aimed at extracting, from WiFi data sources, indicators on waiting time in stations. Similarly, the work of [SKO16] sought to extract, from WiFi data, several types of indicators such as station density, connection time or the distribution of origin/destination flows.

The estimation/reconstruction of OD matrices is a particularly well-studied theme. The work by [TTC07] proposed to estimate the destination of users from individual ticketing data by logical inference using the history of users' trips, combined with probabilistic travel estimation in the absence of more reliable information. The authors in [MP12] presented a methodology for estimating OD matrices by combining ticketing validation data, transportation infrastructure data, and realized transportation plan. The main objective is the reconstruction of travel chains by estimating destinations and connections using decision rules. The work by [Toq+16] proposed to predict dynamic OD matrices in

the short term using recurrent neural network models. The OD matrices are previously estimated from aggregated ticketing data (AFC).

The work by [EB20] focused on the relevance of OD matrices estimated from ticketing validation data. The comparison of OD surveys with ticket validation data revealed several biases. While the surveys seem to underestimate the number of trips, the difficulty of inferring destination from ticket validation and the presence of fraud are the main parameters that affect the representativeness of OD matrices estimated from ticketing. The authors recommend combining ticketing validation data with complementary sources, such as automatic train load count data, which allow the adjustment of the OD matrices.

2.2.2 Characterization of mobility behaviors

The characterization of mobility behaviors is a major issue in the analysis of mobility data. It aims at analyzing mobility behaviors through the travel habits of users in order to improve the planning of the transportation supply.

To explore the contribution of survey data, [Zmu+13] grouped a set of studies aimed at deepening survey approaches and methodologies for the study of mobility. These surveys generally cross-reference the geographical and socio-economic information of people with their travel habits in order to extract and identify typical behaviors associated with human profiles or geographical areas. In the same vein, the authors of [Zha+15] exploited a new form of survey carried out via a mobile application. Daily trip data were collected to automatically group travel habits.

Other authors use cell phone data to characterize mobility behaviors. The study by [JFG17] extracted from these data regular daily travel patterns between the different locations visited. By adjusting the mobile activity data using spatial census data, the authors conducted an analysis of the distributions of the identified behaviors. The work by [Bac+19] focused on the assignment of mobility flows measured from cell phone data to the different transport modes. This assignment is carried out using a Bayesian model that cross-references these data with mobility survey results and spatial characteristics extracted by data mining. The assignment is then used to infer dynamic OD matrices by mode which are then matched with ticketing validation data.

Numerous studies have studied individual ticketing validation traces to identify typical mobility behaviors. [LC11] proposed a subscription recommendation algorithm aimed at reducing overspending due to package errors by combining a travel prediction model and a clustering model. [ZKZ18] proposes a model dedicated to predicting all the daily trips of an individual. They proposed a Bayesian N-gram model based on a Prior learned over a large set of trips and refined on the individual trips. The work by [Bri+17] characterized

the behaviour of public transportation users based on their temporal habits using a two-level clustering model. The proposed modeling is based on Gaussian mixture models. [Ma+17] is also explored this issue on the transportation perimeter of the city of Beijing based on a spatial clustering combined with a multi-criteria analysis. [Pou+15] formalized a robust approach for the characterization of users and station ridership patterns. The approach is based on a multi-scale representation of user movements combined with a spectral matrix decomposition (Non-negative Factorization). On aggregated validation data, one can cite the work by [Zha+19] which sought to classify stations according to their ticketing ridership profiles cross-referenced with data on the activities and uses of the different zones.

2.2.3 Mobility forecasting

Forecasting consists in anticipating the evolution of the variable of interest in the short or long term using a set of explanatory factors. In the context of mobility, there is operational interest in the forecasting of several variables (such as train delays, train loads, mobility flows, etc.) related to transportation supply or demand. These variables correspond to indicators that provide information on user mobility or on the state of the transportation network. The forecasting of these indicators aims at analyzing and anticipating the evolution of nominal situations but also problematic situations. There are generally two forms of forecasting:

- Long-term forecasting, which consists in forecasting a variable based on contextual, mainly calendar, attributes. Its purpose is to identify trends that can help to better size and the plan transportation supply.
- Short-term forecasting, which involves predicting the evolution of a variable over short time horizons through the short-term dynamics. It provides valuable information for analyzing and regulating the transportation network.

The first part of the thesis will focus on the short-term forecast of the train load. This is a new variable of interest resulting from recent collection experiments conducted by the transport operator SNCF on some pilot lines. We will detail the subject in the chapter dedicated to prediction.

Without being exhaustive, we will cite some work on the prediction of public transit mobility. The authors of [Din+16] proposed a short-term prediction of passenger ridership based on historical values and calendar attributes using standard machine learning approaches. [Cui+16] used a non-linear autoregressive statistical model with exogenous factors (NARX) using historical and climatic attributes to predict demand. [RBG16] performed the short-term prediction of train load from a dynamic Bayesian model that models the flow exchanges between platforms, trains and connections. For this purpose, train weight measurements and ticketing ridership data were exploited. In the same

perspective, the authors of [Jen19] carried out load prediction in stations or in trains from information on the mobility offer, ticketing validation, and passenger load sensors of trains. They evaluated several prediction models based on statistical regressions or machine learning algorithms.

Other authors have constructed intermediate indicators (load class, saturation level, ...) that may be easier to predict and/or interpret. The authors of [CSC12] tackled the prediction of station overload indicators from reference histograms based on historical load data. More recently, the work by [Hey+18] focused on the prediction of overload in trains. The authors considered the problem as a load classification task and used standard classification algorithms to solve it (nearest k neighbors, Support vector machine, neural network).

Finally, recent work focuses more specifically on the prediction of atypical situations related to events or incidents. In [Li+17], the authors compared the performances of several machine learning approaches in disrupted situations. The authors of [Toq+18] proposed to integrate anomaly information to perform prediction in a disturbed situation with a recurrent neural network model.

2.2.4 Anomaly detection

Anomaly detection is an academic field that aims to discriminate, within a data set, the elements having an atypical behavior. This is a complex problem in the case of structured data because the notion of atypical will depend on structures and contexts. Anomaly detection consists in identifying and extracting normal behaviors which can be complex because of the many influential factors, then discriminating elements presenting strong deviations from ‘normality". Lastly the characterization of anomalies aims to quantify the deviation from normality, which reflects the impact of an anomaly on the variable of interest.

Operating public transport is a complex task due to the spatial extent of these systems. Numerous disturbances and atypical situations of various magnitudes and causes (incidents, technical failures, logistical problems, external events) will impact transportation supply and/or demand. Meticulous analysis must be carried out on the different sources of data collected to better understand these phenomena.

Thus, the objective is to extract knowledge from the data in order to help improve the management of public transportation systems. This knowledge can be used to improve planning by identifying mismatches between supply and demand and to better anticipate the impact of a disrupted situation on demand (maintenance work, event). It can also

be used to assist regulation by reporting anomalies and to better predict the impact of incidents on network operations.

The work of [Ton+18] tackled the detection of anomalies in demand in metro stations based on ticketing data. The method exploited a 'Non-negative matrix factorization' decomposition to decompose the ridership data in order to extract references qualifying the nominal behaviors. [Zha+17] analysed traces of individual mobility on the Shenzhen metro (China) to identify hidden regularities and anomalies in travel patterns. They used a statistical method to detect passenger travel patterns that deviate from the normal distribution and then used clustering methods to classify passengers based on the similarity of their travel behavior. [Bri+19] also proposed to identify atypical events using hierarchical clustering on average daily profiles with specific calendar contexts. The anomaly score considers the internal variability through the difference between the ridership value and the cluster mean normalized by its interquartile range. The work by [He+19] is also investigated the detection of anomalies in mobility flows. The authors constructed a representation of the flow state from an aggregation of flow graphs combined with a dimension reduction method. They then applied probabilistic clustering based on a Gaussian mixture model combined with a statistical test to detect anomalies.

The second part of the work focuses on the question of contextual anomaly detection. It deals with the detection of the impact of an incident or an event on station attendance. In the transportation literature, this is a topical issue that is beginning to be studied. Table 2.1 synthesizes the positioning, and data from scientific articles related to mobility and public transport in particular.

Table 2.1.: Summary of mobility reference work

Author	Task	Approach	Data sources
[Zmu+13]	Mobility characterisation	Survey enrichment	Survey
[Zha+15]	Mobility characterisation	Clustering	Survey + Mobile GPS
[LC11]	Fare recommendation	Clustering	Survey
[TCA05]	Mobility characterisation	Theoretical analysis	Trip planner log
[SSC16]	Mobility characterisation	Use case analysis	Trip planner log
[El +17]	Dwell time estimation	Heuristic	WIFI
[SKO16]	OD reconstruction	Heuristic	WIFI
[Lu+17]	Users characterisation	Data-mining & linear regression	Mobile phone
[Kon+18]	Users characterisation	Spatial overlapping	Mobile-Phone & AFC traces
[JFG17]	Users characterisation	Data-mining	Mobile-Phone
[Bac+19]	Modal assignment	Data-mining & Bayesian	Mobile-Phone
[TTC07]	OD reconstruction	Heuristic	AFC traces
[MP12]	OD reconstruction	Rule based	AFC traces + AVL
[Toq+16]	Short term OD forecasting	Deep learning (RNN)	AFC traces
[EB20]	OD evaluation	Source comparison	Survey & AFC traces
[Ma+17]	Mobility characterisation	Spatial Clustering	AFC Traces
[Bri+17]	Mobility characterisation	Gaussian mixture clustering	AFC traces
[Zha+17]	Mobility characterisation	Statistical clustering	AFC traces
[Pou+15]	Users characterisation	Spectral decomposition	AFC traces
[Zha+19]	Station characterisation	Series clustering	AFC traces
[CSC12]	Long term Prediction	Histogram reference	AFC
[Hey+18]	Short term prediction	ML Classification	AFC
[Din+16]	Short term prediction	ML prediction model	AFC
[RBG16]	Short term prediction	Dynamic Bayesian model	AFC + Train Load + AVL
[Li+17]	Short term prediction	Radial basis function	AFC + Weather
[Toq+18]	Short term prediction	Deep learning (RNN)	AFC + Event
[Jen19]	Short term prediction	ML prediction model	AFC + AVL
[Pas+19b]	Short term prediction	Deep learning (RNN)	Train Load + AVL
[Ton+18]	Anomaly Detection	Spectral decomposition	AFC + Twitter
[Bri+19]	Anomaly detection	Gaussian mixture clustering	AFC
[He+19]	Anomaly Detection	Gaussian mixture clustering	AFC

AFC: Automatic Fare collection AVL: Automatic vehicle localisation
RNN: Recurrent neural network ML: Machine learning

2.3 Case Study 1: Passenger load on Paris commuter trains

2.3.1 Issues

The IVA (Enhanced Passenger Information) project, of which the thesis is part, aims to use the data collected on transportation infrastructure to better understand the mobility behavior of public transportation users in disrupted situations in order to:

- enhance passenger information to enable transit users to better plan their trips.
- assist real-time regulation by providing indicators on the state and evolution of the transportation network.

This thesis contributes to the analysis and valorization of mobility data. It first aims to explore different sources of data collected around public transportation infrastructures and made available by the transport operator SNCF and the organizing authority Ile-de-France-Mobilités (IDFM). More specifically, the work focuses on a new source of data for automatic counting of the load on board trains. The thesis focuses primarily on the exploration and consolidation of passenger load data with structural information on the actual transport offer and calendar contexts. This first step was motivated by the construction of learning datasets based on train loads enriched with information from the transportation plan and the calendar context. These data will be used as a learning set to develop and compare several short-term prediction models. This is an interesting issue of prediction on highly structured sequential data in a rich dynamic context.

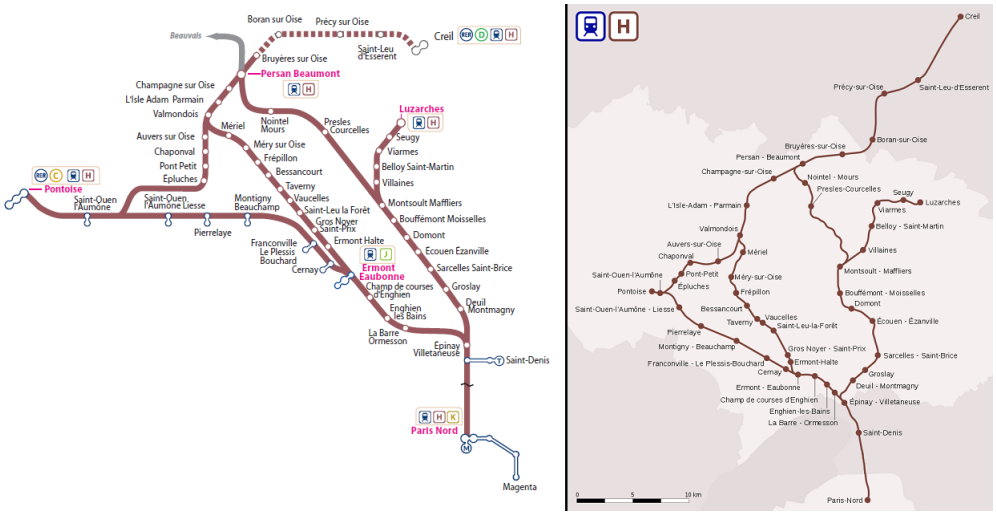

Figure 2.5.: Scope of study: Transilian line H

The scope of the study is the Transilien H line, a commuter train line serving the north of Paris (Figure 2.5). This line is the field of experimentation for several studies including the one collecting passenger load data that we use for valorization purposes in this thesis.

The H-line is an interesting framework for study. As can be seen in Figure 2.5, it is a multi-branch line which is split into 4 main sections connecting the suburbs to Paris and a secondary section (suburb-suburb) connecting Creil to Pontoise which is outside the data collection perimeter. A field of study like this one raises interesting issues related to the variability of services and destinations, and the interactions between trains that share only some of the stations they serve. However, this adds a source of complexity to the modeling and formalization of the prediction problem.

2.3.2 Data description

The databases used for the thesis required the cross-mining of passenger load data and data on the theoretical and realized transportation plan.

- ***The database of the theoretical and realized transport offer*** was provided by SNCF on line H between January 2015 and June 2016. It includes the timetables of all the trains running on this line. The theoretical transportation plan is scheduled 3 weeks in advance, on a flexible basis to meet demand according to the season and events. Theoretical planning is then adapted in real time by the regulating operators to deal with various contingencies in order to best meet real passenger demand. In the event of disruption, the dispatching team has at its disposal some measures (slowdown/acceleration, service modification, addition/deletion) that alter the theoretical transportation plan. These data are contained in a database. An entry is associated with the passage of a train at a station. It contains information about the station concerned (name, identifier, line), the date (date of passage), the train concerned (number and type of vehicle), the theoretical transportation plan (theoretical arrival and departure time) and the actual transportation plan (actual arrival and departure time). This database is consolidated to deal with the a majority of inconsistent data feedbacks.
- ***Load counts*** come from an experiment with train load sensors conducted by the SNCF on line H for several years. The train gates are equipped with several radar sensors to count the number of passengers using the door. The information is then aggregated at the scale of the train and provides for a given stop, the number of boarding and alighting passengers. The train load is then deduced from these measurements by summing the boarding and subtracting the alighting for all stations upstream of the position in question. Several lines of the SNCF operator are partially equipped by this system, including line H, the main experimental line which has both a high rate of equipped train sets (90%) and an interesting time depth (1 year and a half). These counting data were provided to us in the form of a database named "*Châtelet*". An entry is associated with the passage of a train at a station and contains information about the station (name, identifier, line), the train (vehicle number, mission, status, train number), the date and time of recording and the counts recorded (boarding, alighting and estimated loads).

2.3.3 Exploratory statistics

Figure 2.6 shows the percentage of missing data according to the type of information and the station concerned. The proportion of missing data related to the theoretical schedule is very low (less than 1%). On the other hand, 12% of the information related to the mission of the trains is missing, which can however be partially reconstructed. We note that 17% of the train loads are missing on the totality of the data of line H (including the

transverse branch) for the studied period. This can be explained by the fact that some of the running trains are not equipped with sensors and that there may be punctual failures in data collection. For the stations connecting Paris (excluding the transverse branch), 10% of trains are not equipped, according to the SNCF operator. This percentage is higher for the transverse branch, which is outside the scope of the experiment. This order of magnitude is clearly reflected in the proportion of missing data per station. The portion induced by data feedback problems is therefore quite low (around 1%).

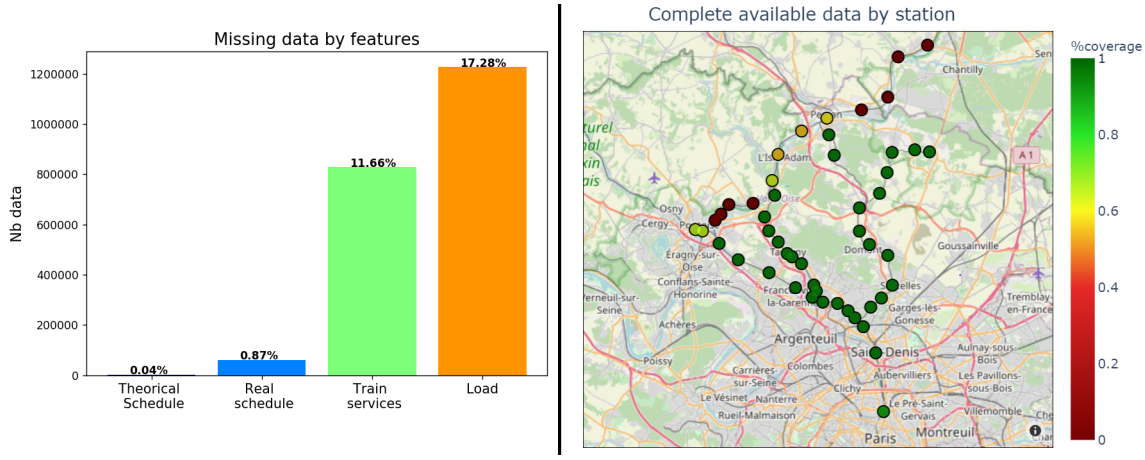

Figure 2.6.: Missing data by feature and by station

By extracting statistics on the distribution of data and load values as a function of time for all the stations studied (Figure 2.7), we observe classical patterns in the mobility data with higher load values (Figure 2.7.a) in peak hours (8am-10am and 6pm-9pm) associated with greater variability (Figure 2.7.d). There is also a relatively homogeneous overall distribution of data (Figure 2.7.b) with a slightly lower average load for the summer vacation months (Figure 2.7.e). Lastly, we observe a greater train distribution for working days (Figure 2.7.c) with a higher average load (Figure 2.7.f).

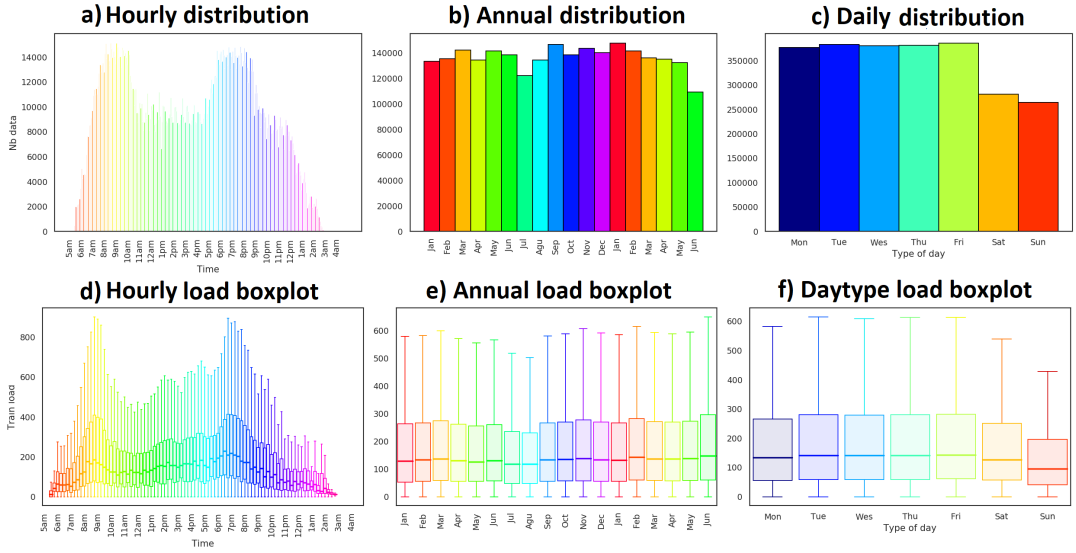

Figure 2.7.: Exploratory temporal statistics for all the stations studied.

Figure 2.8 shows the average train load per station. As expected, the train loads are correlated to the distance of the station from Paris.

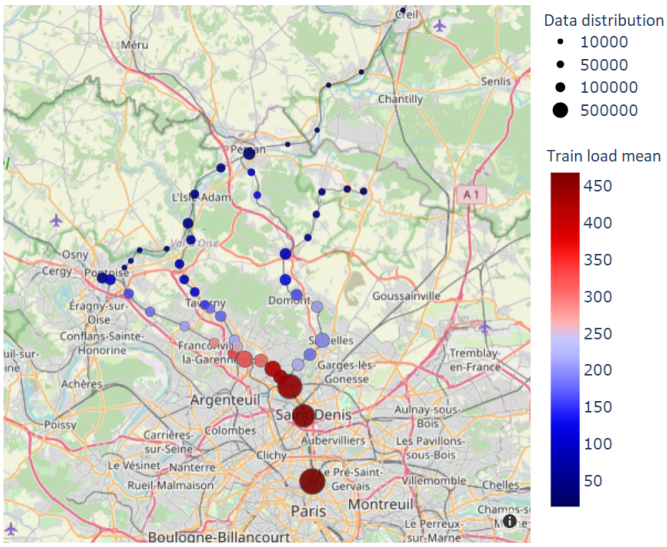

Figure 2.8.: Space exploratory statistics

2.3.4 Train passenger load data specificities

Influences factors in mobility time series

The thesis focuses on the prediction of the train load at a station based on its calendar context, the characteristics of the transportation plan, and the history of loads at the station. The following Figure 2.9 illustrates the average and distribution of the load as a function of time and type of day for 4 distinct stations.

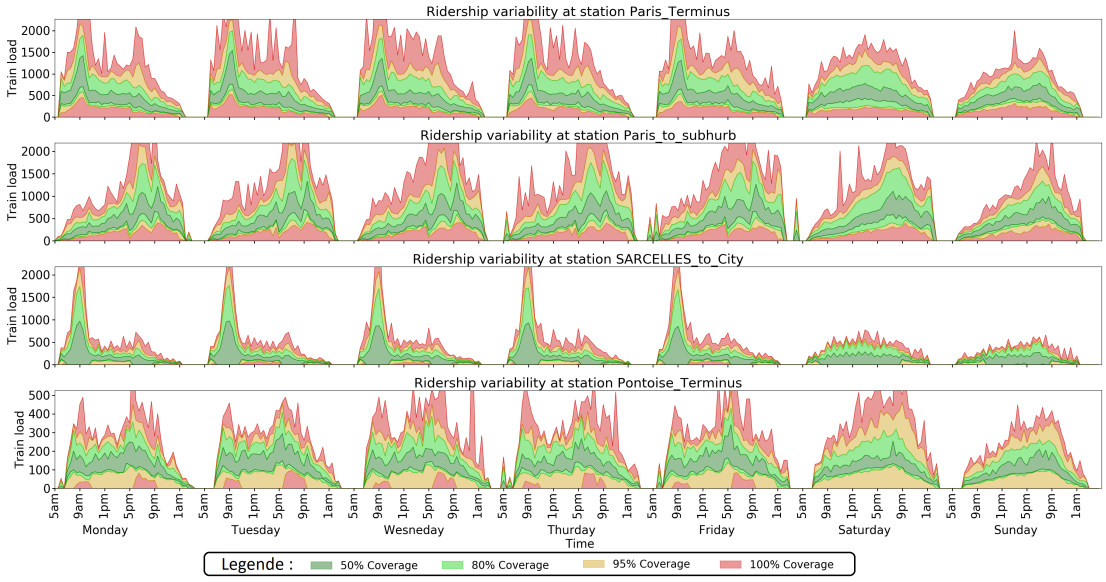

Figure 2.9.: Daily load profiles and variability for 4 stations

Firstly, there is a large variability in load profiles between stations. There is also a large variability within the same station illustrated through the spread of the coverage envelopes which represents the statistical dispersion of the data. The 'most normal' 50% of passenger loads are contained in the dark green envelope. The sequences of train loads at a station behave like time series structured by a set of observed and non-observed factors whose 'entangled' influences generate a specific variability for each station. Capturing the influences of the set of contextual attributes is a major issue in train load forecasting.

On mobility data, **calendar** contexts are major influential factors that strongly structure the data. The influences can be: **hourly** structuring in particular the dynamics of peak/off-peak hours, **daily** influences that generate specific patterns related to types of day, **holidays** that impact transportation demand during the vacation period, **seasonal** influences that modify travel habits, or **annual** influences through growth or decline related to urban dynamics.

In the frame of prediction at the train level, one must also consider **transportation supply** with the influences of **train missions** (service, destination), **past load values** on the platform and in the trains and **alteration of the transportation plan** causing delays and train cancellations.

In conjunction with these calendar and transportation supply factors, one must also consider the presence of numerous **unobserved influential factors** that may impact supply or demand and thus alter the unexplained variability of the data. These include, but are not limited to, events (sports, cultural, etc.), demonstrations, strikes, work on the network, disturbances that degrade the transportation supply and cause load transfers, but also regulation choices, weather conditions, and the impact of passenger information on transportation user behavior. The collection of additional data aims at better estimating the influence of latent factors and improving the performance of prediction models.

The complexity of the prediction increases with the numerous, partly unobserved and entangled influential factors. To make a good prediction, the model must *a posteriori* disentangle and extract the influence of the maximum number of factors based on contextual attributes and the data structure. It thus infers the behavior of the data by approximating the generative function that links the influential factors to the variable of interest. Not knowing some of these influential factors complicates the prediction task by adding unexplained variability in the data.

Data with irregular temporal structure

The series of train loads have the particularity of being structured by the transportation plan, which introduces several forms of variability in the data. The load will be strongly

influenced by the mission of the trains. The frequency and regularity of time series observations will also be affected by the specificities of the transportation plan. Figure 2.10 illustrates these phenomena. The train load series for one day is observed on two different stations by representing the load as a function of the time on the ordinate for several train missions (observation color).

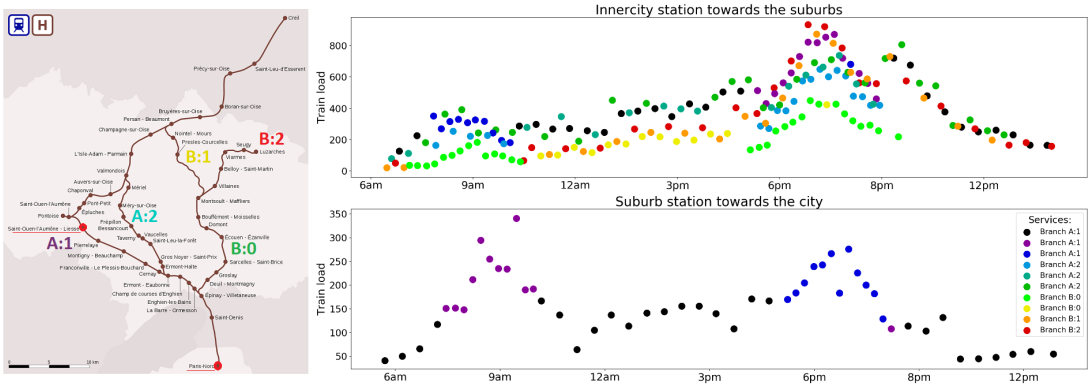

Figure 2.10.: Variance of train frequency on 2 platforms

The temporal irregularity related to the transportation plan is accentuated by the presence of missing data (trains not equipped with sensors and collection failures). On regular time series, it is possible to estimate the influence of known and unknown factors by analyzing the regular patterns that structure the data. With temporal variability on observations, algorithms can only extrapolate it from contextual attributes, which makes the inference task all the more complex.

2.3.5 Synthesis

The case study is based on a recent data source from an experiment piloted by the SNCF operator. It studies the variable of interest of passenger load in trains on a multi-branch commuter line. Within the framework of the thesis, exploratory work was carried out in order to analyze these new data to extract relevant information that can enrich the passenger information and facilitate the task of the operators in charge of the regulation of the network.

The main problem concerns the short-term forecasting of the passenger load in trains based on explanatory variables. It aims at capturing aggregated mobility behaviors based on historical passenger load data, characteristics of train time and missions, and contextual attributes that define the calendar context. The data are derived from the complex interaction between demand (social, economic and demographic factors in the different areas served by the stations) and transportation supply.

The series studied are impacted by several observed (calendar, transport offer) and unknown (latent dynamics, events, incidents) influencing factors. The series are temporally

structured by the transport offer, which induces an irregularity both in the temporal sampling of observations and in the nature of the observations that depend on the characteristics of the trains. They are also spatially impacted by the location of the station and the related social activity. A station located in an employment area will not have the same temporal profile as a station located in a residential area. The task of short-term prediction is to build a model that will model the impact of the influential factors and infer the short-term dynamics to predict the evolution of the train passenger loads.

The thesis will focus on:

- the use of standard machine learning models based on ensemble learning. These models are capable of providing good short-term prediction performance and will be used for comparison with more advanced models.
- the extraction and construction of rich representations from the available structured data set to facilitate the learning of prediction models.
- the design of advanced models based on recurrent neural networks capable of exploiting the specificities of the structures in the data.

2.4 Case Study 2: Ticketing data of Montreal metro

2.4.1 Issues

The second use case concerns the **analysis of the impact of disruptions on metro network ticketing data**. The analysis of this database covers both the prediction and detection of anomalies. The disruptions linked to transport will cause various impacts (saturated station, modal shift, propagation and cascading delays) which deteriorate the supply and the conditions of transport. These disruptions generate a cost related to the limitation of the mobility of many users. Reducing the impact of disruptions is an important issue for transport operators.

The posterior analysis of anomalous situations can provide a better understanding of phenomena linked to the propagation of disturbances. The real-time extraction of indicators about the state and the evolution of the network can be a valuable guide for operators in their choice of regulation. This is a complex research problem linked to the analysis of data structured by many contextual influences. It requires identifying normal behaviors that change according to the context. These behaviors can then be used as a reference to estimate the deviation caused by the disturbance.

This problem is at the core of an exchange carried out linking the thesis with a Montreal research laboratory: Centre interuniversitaire de recherche sur les reseaux d'entreprise la logistique et le transport (CIRRELT). It aimed to capitalize on previous work carried out

by applying short-term predictive models to the Montreal metro ridership data provided by the STM [1] , an organization in charge of public transportation in the city of Montreal. The research problem concerns **the detection of unsupervised contextual anomalies on regular multivariate time series linked to mobility**. Specifically, the work adopts an anomaly detection approach based on the study of the residuals of prediction models. It aims to formalize a method of building an **anomaly score that is robust to the context** by exploiting the unexplained variability of the data.

This makes it possible to detect statistically significant anomalies in the series of ticketing ridership observations related to the context. These anomalies can then be confronted with an operator disturbances database to reliably discern any possible impact such as under-crowded/overcrowded situations and spatio-temporal propagation patterns.

Indeed, in many applications, including mobility, the information available on anomalies is imperfect:

- Non-exhaustive because it does not reference all of the anomalies impacting the network.
- Indirect because it does not necessarily indicate disturbances having a real impact on the network.

Therefore, it is not straightforward to assess the performance of detection approaches quantitatively without a perfectly reliable anomaly base.

2.4.2 Exploratory analysis of real Data

Ridership series

The Montreal Transit Corporation (STM) provided us with smart card ticketing data from the TAP-IN logs of the automatic fare collection (AFC) recorded at 50 metro stations in the city for three years from 2015 to 2017. The data were pre-processed for previous work aimed at proposing long-term prediction approaches. For each Montreal subway station, smart card tap-in logs are aggregated with a temporal step of 15 minutes covering the whole exploitation day from 5:00 am of each day until 1:00 am of the following day.

To facilitate the study, we applied the proposed methodology to a selected number of network stations. The study perimeter focused on fourteen stations mainly located in downtown Montreal (illustrated by Figure 2.11. It mostly covers the Green and Orange lines, which serve the downtown in a similar and sufficiently close way to observe passenger transfer phenomena in a disrupted situation. We also included multi-modal multi-line hub stations to monitor the possible transfers between lines at these hubs. Lastly,

[1] STM: Société de transport de Montréal

we completed the perimeter with two stations located near a major event infrastructure to highlight the influence of these major events on station ridership.

Figure 2.11.: Spatial perimeter with study stations

Disturbance database

The Montreal Transit Corporation (STM) also provided us with a disturbance database containing some events and incidents that occurred within the studied period and that might impact the station's ridership. Information about the events was posterior and was manually collected by an STM employee. Data about incidents were collected from the internal incident report infrastructure. Each disturbance is characterized by the date, the starting and the ending time, the impacted station, and the class of underlying disturbance.

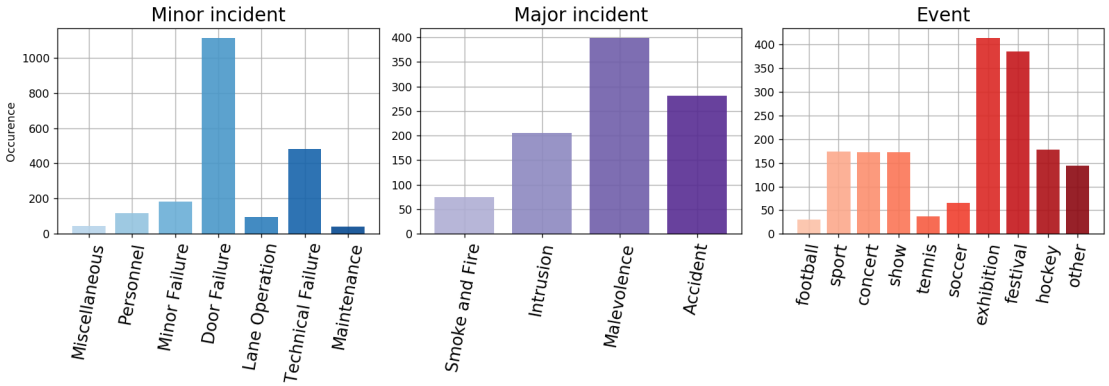

Figure 2.12.: Number of disturbances by category (Montreal Dataset)

Figure 2.12 reports all the disturbances of the database. Events, minor, and major incidents are distinguished. The 2076 **minor incidents** have an average duration of 35 minutes and are divided into seven categories: door failure (1123), technical failure

(482), minor failure (184), track operation (96), works (42), and miscellaneous (43). The 960 **major incidents** have an average duration of 45 minutes. They may be associated to a variety of causes, including malignancy (299), accident (281), intrusion (202), and fire (75). The **event data** include 1772 events with 10 event categories including exhibition (414), hockey match (385), festival (365), concert (178), sport (174), show (172), tennis (37), football (30) and other (144). The duration of the events is highly variable, ranging from a few hours for soccer events to an entire day for exhibitions.

The temporal distribution of disturbed timesteps is illustrated in Figure 2.13. First of all, the percentage of timesteps impacted by a disturbance has to be put into perspective. These graphs illustrate the temporal distribution of anomalies, but the impact of an anomaly is spatially and temporally limited. It is difficult to define the spatial area of a disturbance's effects without careful analysis of the data. Moreover, the event database informs on the schedules of events, which is different from the periods of their possible impact on the transportation network. So for 50% of timesteps between 8 pm and 10 pm, there is an event that takes place somewhere and may impact ridership series at some stations.

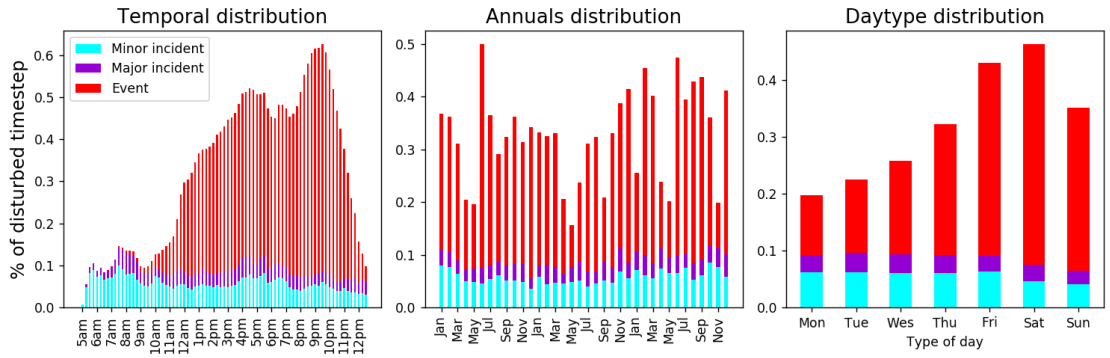

Figure 2.13.: Temporal distribution of disturbed timesteps

The minor and major incidents are relatively well distributed across different calendar contexts. They cover respectively about 5% and 2.5% of all timesteps. The events are, on the other hand, less homogeneously distributed. They never take place in the early morning but rather at the end of the day between 8 pm and 10 pm. Events are often held at the end of the working week or over the weekend. The seasonal distribution of events is rather chaotic, but April and May seem to have fewer events than the other months due to the regular scheduling of certain events.

However, like most anomaly datasets, the operator disturbance database does not constitute a reliable and full dataset of anomalies. It is an incomplete information source that cannot be considered as a ground-truth reference for the anomalies impacting ridership series. Consequently, the application goal is not to detect all disturbance database elements but rather to evaluate which disturbances have a significant impact on ridership series.

Influential factors of ridership series Once a pre-processing step has been carried out, the data are composed of a multivariate time series of 14 dimensions (14 stations) and 87860 timesteps corresponding to 1096 days with 80 daily time steps of 15 minutes, a set of contextual attributes and a database of disturbances.Figure 2.14 show the multivariate ridership series for the 14 metro stations. The different series have their own behaviour and magnitude which are mainly due the social functions of the area covered. There is also a huge variability within each series due to numerous known (Temporal, calendar, ...) and latent (Anomalies or unknown phenomena) influential factors.

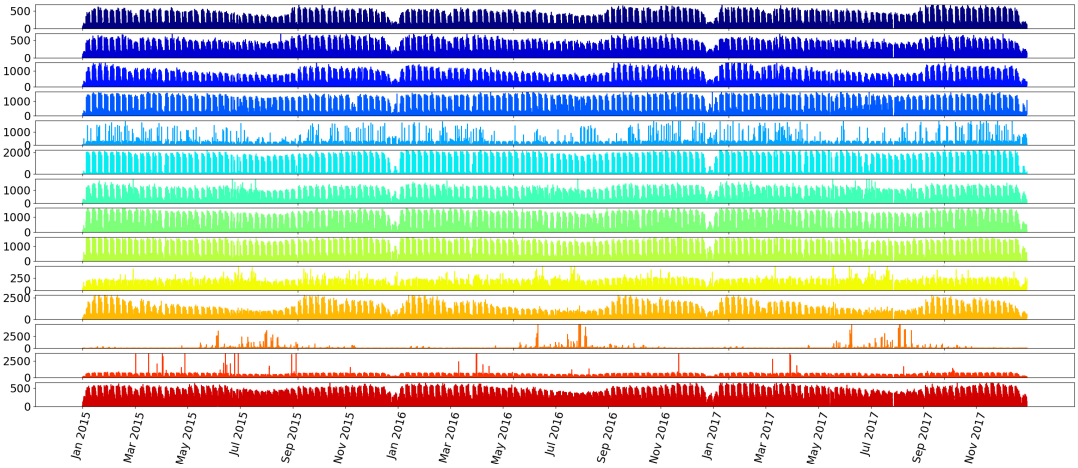

Figure 2.14.: Ridership series of the 14 stations studied

Some basic statistics about the distribution of ridership by hour, by type of day, and by month are represented in Figure 2.15. As for the previous case study, we observe two ridership peaks linked to the peak hours of transportation network usage. The study perimeter covers downtown areas rather than residential areas. This explains the low magnitude of morning peaks and the substantial magnitude of afternoon peaks. The annual ridership boxplot shows a seasonal pattern, with fewer trips in summer and in January. On working days, there are fewer trips at the beginning of the week than at the end. There is also a significant decrease in ridership on the remaining days.

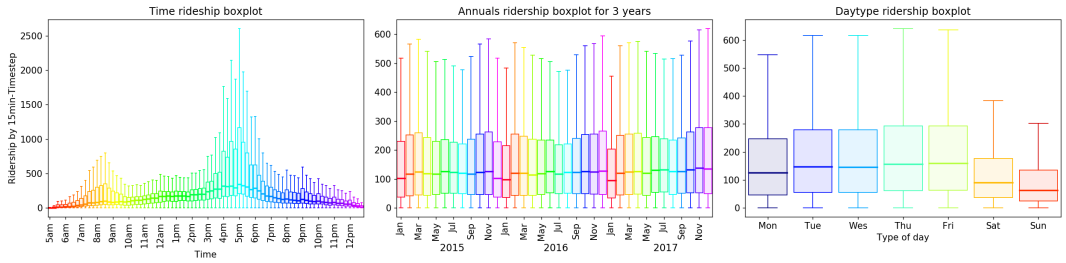

Figure 2.15.: Boxplot of 15min ridership by calendar information

Figure 2.16 shows the sum of daily ridership by station. The busiest stations are located at the terminus of the lines and downtown which is the main focus of the study perimeter.

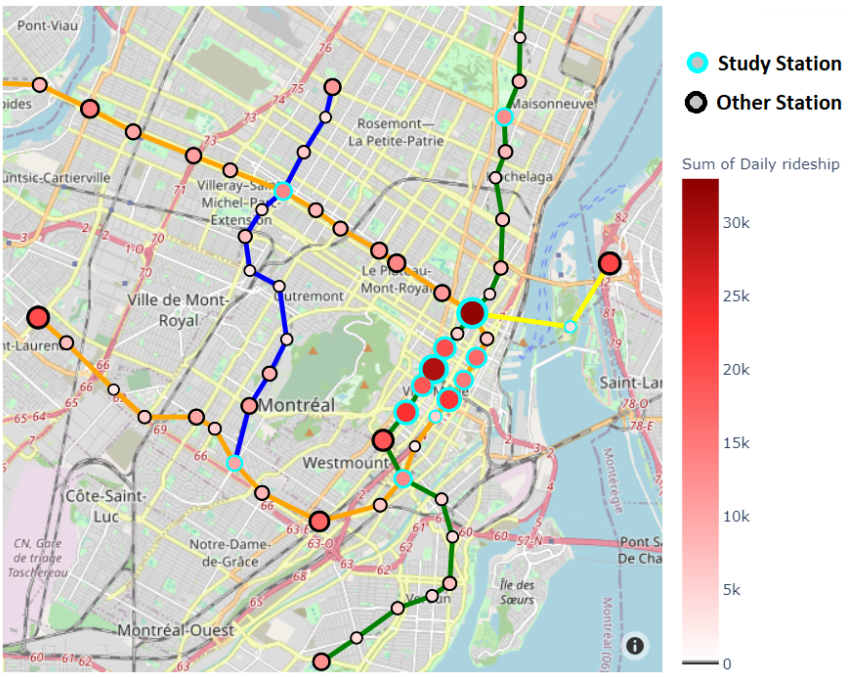

Figure 2.16.: Spatial ridership statistics

2.4.3 Synthesis

The second case study of the thesis concerns ridership data of Montreal metro stations collected from ticketing validations for two years. These data are aggregated per station and 15-minute periods. The goal is to build data mining and machine learning approaches to analyze these regular and multivariate time series. Contrary to the previous use case, we have chosen to analyze station ridership on a set of 14 stations, forming a perimeter of interest around Montreal city center by adding the main hubs. The transportation regulator (STM) also provided us with a disturbance database that contains events and incidents over the period. The disturbances are characterized by time bounds, the main station impacted, and the type of disturbance classified in three categories: event, minor Incident, and major Incident. This dataset will allow us to quantify the impact of the disturbances on the station's ridership.

The aim is to design an analysis tool to finely quantify the impact of these disturbances by considering the contextual variance. Academic issues concern the detection of a contextual anomaly on regular multivariate series. These ridership series are structured by a dynamic context linked to the influence of contextual factors (mainly calendar), short-term dynamics, and unexplained variability linked to unobserved factors. Therefore, we aim to disentangle and extract as much variability as possible to reveal the real impact of disturbances. Our contribution is to use machine learning models based on contextual and historical attributes to forecast the ridership series and to estimate contextual variances through these prediction residues. The aim is to build a context-invariant anomaly score expressing the contextual deviation of ridership series observations.

3

Short-term prediction of mobility demand. The case of public transport.

Introduction

In recent years, the population growth in metropolitan areas has led to overcrowding on trains. Transport operators are working on enriching real-time passenger information systems by providing passengers with trainloads in addition to train schedules. This information can allow passengers to plan their daily trips better, improving overall comfort and avoiding overcrowding on trains. Moreover, such forecasting can be used by public transport authorities and transport operators to enrich public transport route planning or improve the synchronization of train traffic and passenger flows. Transport operators will increasingly need to evaluate and predict network passenger load to improve train regulation processes and service quality levels.

The development of collection and streaming technologies and the rapid growth in data storage abilities have increased the availability of massive transport data, such as passenger ridership, trainload, real-time train schedules, and so forth. This data availability contributes to leveraging data mining and machine learning approaches for processing such spatio-temporal data to extract valuable information to provide better services to passengers or match the transportation supply with the demand. This chapter addresses the forecasting of trainload at a railway station considering a historical dataset that includes two data sources: train load data and automatic vehicle location. The latter source contains all information related to train operation (delay, time of arrival/departure from vehicles, and so on). Most of the prediction problems in this domain address passenger ridership's prediction at an aggregated level (per 15 minutes or 30 minutes time horizon) [Toq+16; ZZQ17; Zia+17]. In contrast to these studies, we focus on prediction at the non-aggregated level considering real-time train schedules. This induces variability in the time step of the time series that we should predict. Furthermore, the prediction model has to consider the contextual factors that impact the trainload, such as calendar information (day, time, holiday, and so forth) and train operations.

We address this prediction task as a multi-step short-term forecasting problem on irregularly structured time series influenced by several contextual factors. We work at

the station level for each train passage, which involves temporal variability, making it difficult to apply techniques that usually exploit the structural regularity of time series. To handle these specificities, we rely on the abstraction capabilities of neural networks linked to the concept of representation learning [BCV13]. The underlying idea is to build a mobility representation of our known influential factors. The model takes the form of an encoder-predictor neural network architecture associated with representation learning on contextual factors. It aims to predict the next trainload at the moment of its passage through the station from the values of the last trains and all the contextual features characterizing these trains.

This chapter will first review the scientific literature related to the prediction of public transport ridership and then discuss the different types of models that are mainly used in time-series prediction. Then, we will detail a prediction formalism for time series structured by a dynamic context. Finally, we will conduct experiments on the real dataset described in Chapter 3. This dataset provided by the SNCF (French National Railway Company) was collected on a commuter train line.

3.1 Related work on short-term mobility prediction

The first work on modeling mobility began in the 1950s and aimed to better size the transportation system. These studies were based on survey data identifying mobility behavior, combined with urban areas' socio-geographic characteristics. The four-step model [McN07] became a standard model that consists of step (1), quantify the number of displacements by area, step (2), infer the destination distribution of the shift by their origin, step (3), set a transport mode for each journey, step (4), assign a route for each journey.

Mobility data analysis appeared in the 2000s with the first automatic fare collection data to better understand mobility behavior. The forecasting task has emerged as one of the main subjects of study with the clustering of mobility behaviors. The aim is to extract valuable information that can help the sizing of the transportation supply. It seeks to forecast passenger flows from ticketing data on long and short term time horizons. Furthermore, the rise of collection infrastructures allows the development of short-term prediction models concerning data analysis. These models seek to exploit the historical context over the previous period to better predict the next step by capturing the dynamic around the mobility demand flows. These more reliable predictions can be used as predictive indicators allowing regulators and network users to better understand the forthcoming evolution of the demand over the transportation network and thus could serve as a decision making tool to support the daily trip planning and regulation operations.

In parallel with the latest advances in machine learning techniques, research studies in the transport domain have explored their application with the help of additional data sources to improve prediction accuracy. We focus our literature review on the studies related to the public transport domain. This review shows for instance that they try to assess the importance of the weather [Yao+18; Cui+16], or measure the connections and modal shifts [RBG16; Din+16], as well as the information events and incidents [Toq+18; LPJ17] in order to model the impact of these complementary attributes on the mobility demands.

Short-term prediction consists in forecasting passenger flows at the next time steps based on previous observations. By analogy, we can consider the forecasting models for the short term horizon as those we can exploit in real-time conditions. The relevant temporal aggregation and prediction horizon depends on the network and the task of interest. Still, many studies are particularly interested in prediction with time steps around a quarter of an hour. The literature related to the short-term prediction of mobility flows explores different types of models. Table 3.1 summarizes some prediction work:

Table 3.1.: Summary of some mobility data studies

Author	Data	Main Approach	Main Model	Temporal Aggregation	Spatial Dimension
[TLW09]	Train ridership	Neural network	Multilayer Perceptron	Day	
[CSC12]	Train ridership	Historical	Histogram	Day	
[Din+17]	Metro ridership	Statistical	ARIMA-GARCH	15min	
[Cui+16]	Metro ridership	Statistical	Autoregressive	1h	
[Din+16]	Metro ridership	Machine learning	Gradient Boosting Tree	15min	
[NHG16]	Metro ridership	Statistical	ARIMA	4h	
[LPJ17]	Metro ridership	Neural network	Radial Basis function	15min	✓
[RBG16]	Metro ridership	Bayesian	Bayesian dynamic	2min	✓
[Ke+17]	Taxi demand	Neural network	C+RNN	1h	✓
[ZZQ17]	Taxi & Bike	Neural network	C+RNN	30min	✓
[Yao+18]	Taxi demand	Neural network	C+RNN	30min	✓
[WT16]	Traffic	Neural network	C+RNN	5min	✓
[YYZ17]	Traffic	Neural network	GCNN	15min	✓
[Toq+18]	Metro ridership	Neural network	LSTM	15 min	✓
[Hey+18]	Train ridership	Machine learning	ML-Classification	15min	
[Jen19]	Train Load	Neural network	Machine learning	Gradient boosting	15min
[Pas+19b]	Train Load	Neural network	LSTM Encoder decoder	15min	

Numerous studies have addressed the mining of large-scale mobility data for exploration, clustering, or prediction purposes. Depending on the available data, the scale of analysis, and the targeted goal, different methodologies can be distinguished. Most of studies tackle temporally regular series forecasting through the temporal aggregation of regular time steps. The first studies aimed to extract basic statistics synthesizing the typical behaviors of the studied series. For example, the authors in [CSC12] proposed to predict crowding levels from automated fare collection data using simple techniques based on historic aggregates. Subsequently, more complex techniques dedicated to time series analysis were applied. We have categorized them as follows:

Statistical methods : Several studies propose to model the evolution of mobility demand flows using statistical models. The work by [Din+17] suggested using an autoregressive integrated moving average combined with a generalized autoregressive conditional heteroskedasticity (ARIMA-GARCH) model based on auto-regression to predict and estimate the variability on series of metro passenger traffic. [NHG16] also proposed to employ seasonal auto-regressive moving average (SARIMA) combined with an event detection based on Twitter data to make short term prediction of subway ridership. The authors in [RBG16] used a model based on explicit modeling of mobility flows by analyzing causal relationships between the adjacent flows on a public transport station with transport service features. The proposed methodology, based on a dynamic Bayesian network, highlights causalities and performs the prediction. [NHG16] also proposed employ seasonal auto-regressive moving average (SARIMA) combined with an event detection based on Twitter Data in order to forecast ridership under event occurrences.

Machine learning models: Machine learning approaches are proving to be powerful and robust approaches to solve many problems. The authors in [Din+16] compared the performance on short-term subway ridership of several prediction models (Support vector machine, Random forest, Gradient Boosting Decision Tree and Neural network). Based on smart card data, the authors built prediction models using both temporal features (time and calendar) and historical data related to subway activities and bus transfer activities. Recently, the authors in [Hey+18] formalized the problem of tram load passenger prediction as a classification task, where the passenger load was labeled in different classes depending on the percentage of occupied seats. Once this labeling had been performed, the authors built classical machine learning classifiers to predict the level of crowding in the transport using temporal and historical data as model inputs.

Neural network models: Among machine learning techniques, neural networks have proved to be potent models used in prediction problems. The work by [TLW09] was among the first uses of a neural networks architecture using feature engineering capturing daily and monthly trends to perform daily passenger demand forecasting. Many advances of deep neural networks on the prediction task have made this approach very popular in the mobility forecasting field. Recurrent neural networks (RNN) [GSC99] are tools potentially capable of capturing the time-series dynamics. These models have even been extended to handle regular spatio-temporal data in [Xin+15] with a convolutional LSTM for weather prediction. Several approaches use a combination of convolution layers (CNN) and recurrent layers (RNN/LSTM) to consider spatial, temporal, and exogenous dependencies, which are often predominant in mobility data.

An LSTM recurrent neural network was proposed in [Toq+18] to address the short-term forecasting of passenger flows in a transportation network considering event data. The Authors in [WT16] proposed a hybrid deep neural network parallelizing RNN and CNN layers to perform traffic load predictions. In [Ke+17], the authors used a deep neural

network based on Convolutional-LSTM layers to forecast passenger demand related to an on-demand ride service station with calendar and weather information combined with real-time measure of travel time and estimation demand. Similarly, [Yao+18]'s work proposed an approach based on a deep convolutional neural network (CNN) and recurrent (LSTM) to achieve a short-term prediction of taxi demand. A first component synthesizes spatial information by a CNN for each time step, whereas a second captures spatial dynamics (LSTM) by analyzing the spatial component's outputs. The third component explores the similarities between areas based on socio-demographic information.

With a different approach, the authors of [ZZQ17] used a deep-learning-based model to forecast the flows of crowds in all regions of a city. The methodology uses a combination of Convolution (CNN) and Residual (Res-net) layers to capture spatial and temporal dependencies. Dependencies are captured over long, medium, and short term horizons to determine trends, periodicity, and short term dynamics. Historical trajectory data, weather, and events are used to build the model. Considering the spatio-temporal dependencies in the forecasting, [Zia+17] proposed a dynamical spatio-temporal neural network to forecast the time series of spatial processes. The idea investigated in this model is to learn both temporal and spatial dependencies between the series to be predicted through the combined use of a latent embedding structured by the temporal dynamics of the series and a decoder mechanism to make the prediction. In [LPJ17], authors implemented a multi-scale Radial Basis Function Network (MRBF) to predict outgoing ridership of several Pekin stations with 15min-aggregated smart card data by using in-going ridership as features. The study also focused on analyzing the results in disturbed situations with some major events and incident information.

Positioning and contribution

The rise of connected sensors allows for collecting new data sources, such as 'real-time' train loads. However, with fine granularity data, we cannot ensure aggregation without loss of information. The irregular shape of these data makes it difficult to apply approaches based on the regular temporal structure. Time-structure independent approaches can be applied, such as machine learning regression [Din+16] or classification models [Hey+18], but they require some improvements to exploit all the available information fully.

In this thesis, we aim to formalize the forecasting task so as to better exploit the temporal structure and transport scheduling that fuel the data dynamics. We will pay particular attention to the extraction of a relevant space representation of features and the building of advanced machine learning models. In particular, the work focuses on designing an advanced neural network model able to exploit sequential structures with heterogeneous attributes. It aims to perform short-term multi-step prediction on train load data at the

station scale, a relatively recent and little-studied data source. The approach is specifically based on:

1. A recurrent neural network designed to handle sequential data.
2. A recurrent Encoder-Decoder architecture popularized by [Cho+17] dedicated to the capture of complex semantics.
3. Contextual representation learning [BCV13] of temporal, calendar and transportation plan structures.

3.2 Standard forecasting approaches

3.2.1 Contextual average model (CA)

The contextual average (CA) forecasting model is a basic long-term prediction model based on a prior sampling linked to observed contextual attributes. In the mobility field, contextual attributes are often linked to calendar information (hours, day type, season, year). The aim is to capture periodical influences and specific patterns of calendar factors structuring ridership series by averaging the sets of observed values that belong to a sub-sample that define a particular calendar context (for example, ridership values on regular winter Mondays between 10:00am and 10:15am). The precision and granularity of the sampling must consider the size of the data to guarantee the representativeness of the average values extracted.

3.2.2 Last Observation Carried Forward model (LOCF)

This statistical baseline serves as a reference for the most basic short-term prediction models. It consists in predicting a step t of a time series using the last observed values (often $t-1$). It is used as a minimum benchmark value for the performance of a short-term prediction model. In the framework of strong context-dependent data, the LOCF performance can be much lower than a contextual mean long-term model.

Table 3.5 lists various mobility prediction studies using naive approaches as competitors.

Table 3.2.: Studies using naive models

Methods	Authors
CA	[RBG16; Ke+17; ZZQ17; Yao+18]
LOCF	[RBG16; Toq+18; Pas+19b]

3.2.3 Statistical models

Some statistical methods have been created in order to solve forecasting prediction on time series. The abundant literature has led to many successive improvements of standard models. We will briefly mention some standard approaches that can be found in reference books of statistical forecasting [AL09; Ham20].

Linear regression Linear regression is a standard statistical model that predicts a value of interest from a linear combination of input attributes. To avoid overfitting, a regularization taking the form of a penalty function is often achieved. There are various kinds of regularization, such as Lasso (L2), Ridge (L1), elasticnet (L1+L2) that aim to reduce the complexity of the model through variance reduction and a sparsity constraint. There are many relaxing and log-linear regression variations or suppressing linear constraint regression such as the nonlinear regression using a Neural network or Kernel approach.

Auto-regressive Models The auto-regressive approach (AR) consists of a linear regression based on the last observed values to capture a time series's temporal dependencies. As with linear regressions, there are many variations of this model, for example, to correct the regression bias (ARMA), to better model the evolution of shifts (ARIMA), to consider seasonal patterns (SARIMA), or to introduce explainable factors (SARIMAX). There is also the Vector auto-regressive models (VAR) that can handle multivariate series.

Table 3.3 lists various mobility prediction studies using statistical models.

Table 3.3.: Studies using statistical models

Methods	Authors
Linear regression	[NHG16; Yao+18]
Auto-regressive model	[NHG16; Ke+17; ZZQ17; YYZ17; Toq+18; Din+17]
Bayesian Network	[RBG16]

3.2.4 Machine learning approaches (ML)

Machine learning (ML) techniques consist in a diverse set of methods based on statistical learning [HTF09]. Numerous studies have contributed to this now very popular field of research. The fundamental principle of these different techniques is approximating a function based on probably approximately correct learning (PAC-Learning). The learning is carried out by minimizing, on the set of the training samples, an objective function synthesizing the task to be solved. There are several forms of learning algorithms based on various techniques. We will only detail two of the many learning techniques of the Machine learning field.

Support Vector Machine & Regressor

The support vector machine (SVM) is a binary classification algorithm proposed by [BGV92] that exploits the kernel tricks to define the best separator hyperplane based on the margin formed between the two classes. It has been extended [Dru+97] for regression problems by looking for the hyperplane which, for a given margin ϵ, minimizes the set of spring variables of the training set $\mathcal{E}_i$ (the out-of-range prediction error). The hyperplane then corresponds to the prediction function. It is defines as a function of the harder elements to predict (The supports) on the training dataset, which have a non-zero value according to the Karush-Kun-Tucker conditions of optimization under nonlinear constraint. This ML model is used as a competitor in several academic studies, including flow mobility prediction.

$$
\begin{aligned}
&Min_w\ C * \sum_{n=1}^{N}(\mathcal{E}_n^+ + \mathcal{E}_n^-) + \frac{1}{2} * ||w||^2 : \begin{cases} y_n \geq f(x_n) + \epsilon + \mathcal{E}_n^+ \\ y_n \leq f(x_n) - \epsilon - \mathcal{E}_n^- \\ \mathcal{E}_n^-, \mathcal{E}_n^+ \geq 0 \end{cases} \\
&With\ f(x) = \sum_{i=1}^{N}(a_i^+ - a_i^-) * \kappa(x_i, x_n) + b \\
&w = (a^+, a^-, b)\ and\ \kappa\ a\ kernel\ function
\end{aligned}
\tag{3.1}
$$

Ensemble learning

Decision tree-based models [Bre+84] are classical classification or regression machine learning algorithms belonging to the set of ensemble learning that contains powerful algorithms such as Random forest (RF), or gradient boosting on decision tree (GBDT). Decision tree is the basic block which consists in storing element y_i from their attributes x_i on a binary tree structure (Figure 3.1). The tree is made up of nodes and leaves.

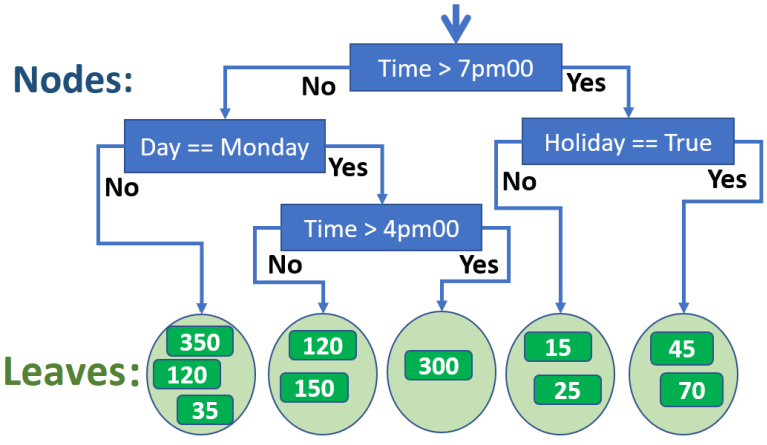

Figure 3.1.: Illustration of Decision Tree

The tree is traversed by crossing nodes which according to a binary constraint based on the attributes, indicate a path leading to a leaf. Training consists in choosing iteratively the node constraints according to the diminution of a loss function (Cross-entropy for classification and Quadratic error for regression) that will best discriminate elements, i.e form homogeneous sets within leaves.

The prediction of a Tree T_0 consists for an element t to fall in a leaf L^a according to the tree walk operator $F^0(x_t)$ based on these attributes x_t. Then, the result is determined from the predominant class (classification), or the average is taken of the elements y_j (regression) of all the elements of the learning set that have also been assigned to this leaf L^a.

$$T_0(x_t) = \sum_{j \in F^0(x_t)} \frac{y_j}{\#F^0(x_t)} \tag{3.2}$$

Random Forest (RF)

The model called Random Forest is a well-known machine learning model for solving nonlinear classification or regression problems. The introductory model by Breiman [Bre01]] is a bagging-type learning algorithm that combines the predictions of several decision trees. Each tree is built on different parts of the data, which are created by applying two sampling methods: random sampling with replacement, also known as the bootstrap aggregation method or bagging (Figure 3.2), and a random selection of characteristics. The average of results of each Tree's prediction based on diversified samples thanks to bagging methods makes the RF models more robust and precise than a simple decision tree.

Let M be a random forest composed of $(T^1, .., T^n)$ binary trees. Each tree T^k is composed of a set of leaves L^k. Values j are assigned to each leaf during the learning phase according to their attribute modalities x_i. We define a tree walk operator $F^k(x_t)$ that takes attributes x_t and returns, for the associated leaf L_i^k, the set of assigned values.

$$M(x_t) = \frac{1}{n} * \sum_{k \in [1,n]} \Big(\sum_{j \in F^k(x_t)} \frac{j}{\#F^k(x_t)} \Big) = \widehat{y}_t \tag{3.3}$$

Gradient Boosting Decision Tree (GBDT)

The Gradient Boosting model introduced by [Fri01] is a machine learning model for regression or classification tasks, which uses a set of weak learner (Basic forecast models) decision trees in our case to create a robust forecasting model. Unlike random forest, which builds a tree forest in a parallel and independent way, the GBDT builds a forest iteratively. The idea of Boosting (Figure 3.2) is to successively add weak learners to correct errors of the current model that combine the weak models learned so far. Subsequently, the new tree is learned on a weighting of the learning set. Weighting gives more importance to the training samples on which the current model makes more substantial errors. It acts as a corrective gradient of the current model.

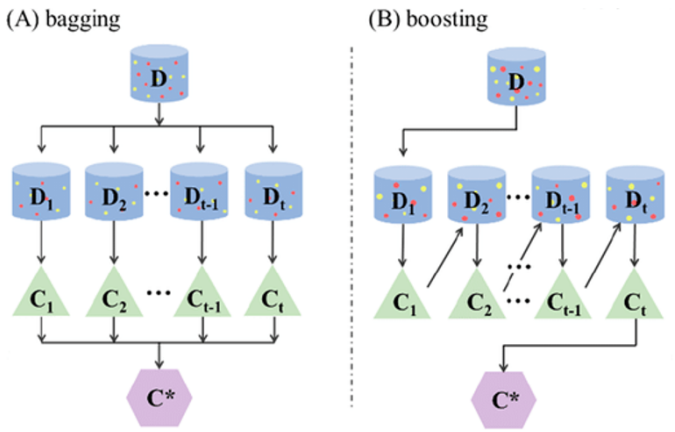

Figure 3.2.: Illustration of Bagging vs Boosting from [Yan+19]

Let M^n be a gradient boosting composed of $(T^1, .., T^n)$ binary trees. Each tree T^k is composed of a set of leaves L^k. Elements j are assigned to each leaf during the learning phase according to their attribute modalities x_i. We define a tree walk operator $F^k(x_t)$ that takes attributes x_t and returns, for the associated leaf L_i^k, the set of assigned values.

$$
\begin{aligned}
M^1(x_t) &= \sum_{j \in F^1(x_t)} \frac{y_j}{\#F^1(x_t)} = \widehat{y}_t \\
M^k(x_t) &= M^{k-1}(x_t) + \sum_{j \in F^k(x_t)} \frac{y_j - M^{k-1}(x_j)}{\#F^k(x_t)} = \widehat{y}_t^k
\end{aligned}
\tag{3.4}
$$

Table 3.4 lists various mobility prediction studies using ensemble learning approaches often as robust competitors.

Table 3.4.: Machine learning approaches

Methods	Authors
SVM	[NHG16; Din+16; Hey+18; LPJ17; YYZ17]
RF	[Din+16; Toq+18]
GBT	[Din+16; Toq+18; Ke+17; Yao+18; WT16; LPJ17; Jen19]

3.2.5 Neural network models

Neural networks perform nonlinear combinations of attributes to achieve various tasks (Classification, Clustering, Regression, Dimensional Reduction). Nonlinear combinations (achieved by neurons) are defined by a chosen nonlinear function parametrized by weights. Weights are optimized through an iterative stochastic gradient descent to minimize a loss function (quadratic errors for regression) on the learning set. Neural networks have emerged as a very efficient method, which can be rather expensive and

slow to learn but have proved to be particularly suitable for handling rich and structured data [LBH15].

Many architectures of neural networks have been developed to manage certain specificities of the data. The basic architecture known as the multi-layer perceptron (MLP) is often combined with different structures such as:

- Convolutional neural networks (CNN) aiming to synthesize spatial dimensions,
- Encoder-Decoder (AE) architectures aiming to compress information via a bottleneck,
- Recurrent Neural Networks (RNN) aimed at exploiting sequential structures of data,
- Generative Adversarial Networks (GAN) that build a synthetic generator from real data using a zero-sum game between a discriminator and a generator.

For several years, neural networks have been used extensively in studies focusing on urban mobility analysis and particularly in studies related to forecasting mobility flows. Among the techniques mainly used, the Multi-layer perceptron (MLP) captures dependencies of contextual attributes, the convolutional neural network (CNN) synthesizes the spatial dimension of mobility data, and recurrent neural networks are ideal for sequence and time series analysis. We will quickly detail recurrent neural networks and an extension called Long short term memory (LSTM).

Recurrent neural networks (RNN)

In recent years, Recurrent Neural Networks (RNNs) have been applied in different fields, from natural language processing to vocal speech analyses. These models are mainly used to process historical data covering long periods, making them more efficient to forecast time series. Unlike classical neural networks, RNNs consider that the outputs depend on previous forecasts by exploiting the sequential structure. To do this, they keep in 'memory' the previous observations in the form of a layer containing the hidden state that is constantly updated (Figure 3.3).

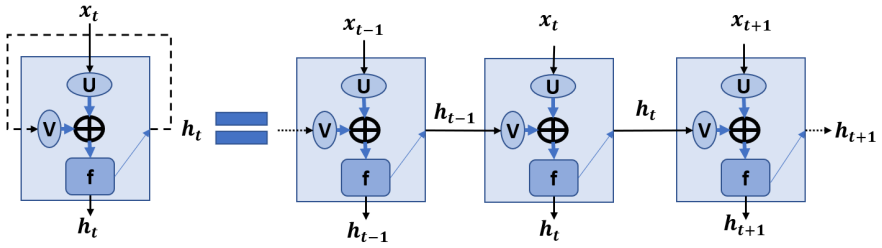

Figure 3.3.: RNN architecture (based on C. Olah[1])

The hidden state h_t of time step t is calculated by a linear combination f of the previous state h_{t-1} and input attributes x_t with the respective weight matrix U, V defined in

[1] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Equation 3.5. The output of recurrent layers is often interpreted by dense layers g associated with a weight matrix W.

$$\begin{aligned} h_t &= f(U * x_t + V * h_{t-1}) \\ y_t &= g(W * h_t) \end{aligned} \tag{3.5}$$

Long short term memory Neural networks (LSTM)

In practice, RNNs suffer from their inability to memorize information over long periods. Other recurrent neural network architectures have been proposed to overcome this problem, particularly the Long Short Term Memory (LSTM) neural network developed by Hochreiter and Schmidhuber [HS97] in 1997. To be able to retain information for more extended periods, the LSTM introduces a memory cell having a gate system. The purpose of these gates is to regulate the spread of information to counter the vanishing gradient phenomenon. The equations of the LSTM models are defined in the set of Equations 3.6.

$$\begin{aligned} F_t &= \sigma(x_t * U_f + h_{t-1} * V_f \\ I_t &= \sigma(x_t * U_i + h_{t-1} * V_i) \\ O_t &= \sigma(x_t * U_o + h_{t-1} * V_o) \\ c_t &= \sigma(F_t * c_{t-1} + I_t * tanh(x_t * U_c + h_{t-1} * V_c) \\ h_t &= tanh(C_t * O_t) \\ y_t &= g(h_t * W_y) \end{aligned} \tag{3.6}$$

There are three operation gates (Figure 3.4) parametrized by their own weight matrices which update information from current attributes x_t, past hidden states h_{t-1} and memory cells C_{t-1}. The 'forget gate' deletes information of the memory cell c_{t-1}. The input gate I_t allows to updates the memory cell c_t, and the output gate O_t allows to updates the hidden state h_t. As for the RNN, the output of the LSTM is often transformed by one or more output layers g.

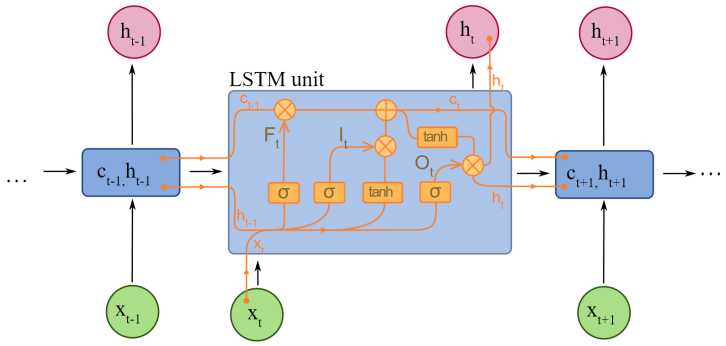

Figure 3.4.: LSTM architecture (Figure from Wikipedia, based on C. Olah[1])

[1] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Table 3.5 lists various mobility prediction studies using neural networks.

Table 3.5.: Studies using neural network approaches

Methods	Authors
MLP	[Zha+15; Cui+16; Din+16; WT16; Yao+18]
LSTM-ED	[Cho+17; SMS15; Pas+19b]
C+RNN	[Xin+15; Yao+18; YYZ17; Ke+17]
*Other	[Zia+17](STNN), [ZZQ17](RES-NET),[LPJ17](RBF)

3.3 Prediction on series with dynamic context

We are going to formalize the 'prediction with dynamic contexts' first on regular multivariate time series, then in the particular framework of irregular time series.

3.3.1 Regular time series with dynamic context

For the regular framework, we have a multivariate time series structured by a dynamic context which can be denoted as follows:

$$y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T) \text{ with } \boldsymbol{y}_t \in \mathbb{R}^d. \tag{3.7}$$

The time series evolution is structured by a dynamic context linked to the interactions among m known influential factors and several other latent factors. Each known factor i takes a state $c_{i,t}$ at each time-step t in a continuous or discrete set E^i. We can define $\mathbf{c}_t$ as the contextual vector as follows:

$$\forall\, t = 1, \ldots, T \;\; \mathbf{c}_t = (c_{i,t})_{i=1,\ldots,m} \text{ with } c_{i,t} \in E^i \tag{3.8}$$

Jointly with these known factors, another set of latent factors ℓ also evolving over time can be considered. The evolution of these known and hidden factors defines the "dynamic context" concept. Our goal is to infer the impact of the dynamic context on the time series (y_t) by obtaining better knowledge of the contextual mean (prediction task) and contextual variability (variance analysis task). In this chapter, we focus on the estimation of contextual mean. We will discuss the notion of contextual variability in the following chapter dedicated to anomaly detection.

Conceptual decomposition of series with dynamic context

We assume that the time series (y_t) is composed of a signal $\boldsymbol{M_t}$ and a noise $\boldsymbol{\epsilon_t}$. The signal $\boldsymbol{M_t}$ is structured by the dynamic context and can be split into several components linked to specific sets of known and latent factors (Equation 3.9).

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{M}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{M}_t &= \boldsymbol{f}^c(\mathbf{c}_t) + \boldsymbol{f}^d(\mathbf{c}_t, \boldsymbol{y}_{P_t}) + \boldsymbol{f}^a(\mathbf{c}_t, \boldsymbol{y}_{P_t}, \boldsymbol{a}_t) \\ \boldsymbol{\epsilon}_t &\sim \mathcal{N}(B_t(\mathbf{c}_t, \boldsymbol{\ell}_t), \sigma_t(\mathbf{c}_t, \boldsymbol{\ell}_t)) \end{aligned} \tag{3.9}$$

where

- $\boldsymbol{y}_{P_t} = (\boldsymbol{y}_{t-1}, .., \boldsymbol{y}_{t-p})$ is the previous temporal horizon.
- $\boldsymbol{f}^c$ is the long-term contextual component linked to the known influential factors (contextual attributes).
- $\boldsymbol{f}^d$ is the short-term dynamic component resulting from the mixture between some of the known and latent factors. We want to infer this component through the short-term dynamics induced during the past temporal horizon.
- $\boldsymbol{f}^a$ is the abnormal component linked to anomalies that significantly impact the dynamics of the series over a short range. $\boldsymbol{a}_t$ is a characteristic series of anomalies that encode the presence of anomalies at a time step t for each dimension.
- $\boldsymbol{\epsilon}_t$ is the unexplained variability in the components f^c, f^d, f^a of $\boldsymbol{M}_t$. This variability is structured by known and latent influential factors $(\boldsymbol{\ell}, \mathbf{c})$ and can be represented as noise with a dynamic mean $\boldsymbol{B}_t$ and variance $\boldsymbol{\sigma}_t$. The use of a non-zero mean $\boldsymbol{B}_t$ makes it possible to consider any bias in the prediction model $\boldsymbol{M}_t$.

Feature-based prediction

A multivariate forecasting model F aims to predict a multivariate series (y_t) through a set of attributes x_t: $F(x_t) = \widehat{\boldsymbol{y}}_t \approx \boldsymbol{y}_t$. Generally, models capture the variability explained by the attributes and take the mean of the unexplained variability. In the proposed framework, we can define several prediction targets driven by the attributes provided to the model as follows:

Contextual prediction: $F^c(\mathbf{c}_t) \approx f_t^c$ is intended to make a long-term prediction from the contextual attributes. The purpose of this type of prediction is to capture the cross influences of factors observed over the time series. In mobility prediction tasks, the contextual factors are mainly linked to calendar information (types of days, months, years, season, holidays). A contextual model can then often be associated with a kind of historical average whose finesse depends on the contextual attributes considered.

Dynamic prediction: $F^d(\mathbf{c}_t, \boldsymbol{y}_{P_t}) \approx f_t^c + f_t^d$ is intended to make a short-term prediction by inferring the influence of the context and nominal dynamics from the contextual

attributes and latest historical values. The purpose is to infer the influence of latent factors through real time measurements. This prediction approach can significantly improve the performance provided that measurements informing us about the studied signal on a recent past horizon (Near-real-time information). In the study of mobility data, the development of real-time collection infrastructures has opened up the application of these types of models.

Dynamic prediction with anomalies: $F^a(\mathbf{c}_t, \boldsymbol{y}_{P_t}, \mathrm{a}_t) \approx \boldsymbol{y}_t$ is intended to make a short-term prediction in an abnormal context related to known anomalies. Data anomalies can be assimilated to contextual attributes but their specificities (rarities, heterogeneity, severe impact) make them more difficult to study and understand. Specific models paying more particular attention to the periods impacted by events can be implemented on dedicated prediction models.

However, the forecasting task is always performed on real-time series. It involves learning about data that are slightly different from the theoretical target, which generates some bias that may or may not be negligible. The data are assumed to follow certain assumptions to ensure the relevance of the learning:

1. the contextual factors mainly structure the time series evolution,
2. the nominal dynamic and unexplained variability follow a Gaussian distribution,
3. anomalies are rare events.

Our work takes place within the framework of dynamic prediction without knowledge of anomalies. We will compare the performance of dynamic forecasting models based on standard machine learning approaches such as Ensemble models or recurrent neural networks. Contextual prediction models are used as benchmarks in performance evaluation.

3.3.2 Times series with underlying structure

The data are often the result of complex interactions that combine several sources of influence. Some of these influences are observed directly through explanatory variables that can be exploited in the form of contextual attributes, as previously mentioned. However, others are due to more complex underlying structures and do not fit well with regular time series. These structures contain valuable information characterizing the nature of the data. It is essential to find a way to represent or synthesize the structural information to perform good predictions. As an illustration, our first case study concerns the task of train load forecasting at a railway station. We studied sequences of time series structured by the transportation schedule, human activities, and landscape use inducing specificities in the time series to be analyzed.

The most efficient way consists in learning directly about structures by manipulating structured data representations. However, translating structured data into a usable structured representation is a challenging task requiring good knowledge of these underlying structures. This representation must be manipulable by complex algorithms based on regular input forms to extract information.

Nowadays, extensive work has made it possible to make progress on these questions for certain types of structures such as images, texts, or graphs. This work has often combined a relevant representation and an algorithm designed to capture the underlying structure. In computer vision, convolutional neural networks [LB+95] exploit spatial structure through spatial neighborhood information on the matrix image. The NLP field often uses approaches that exploit both the sequential structure through recurrent networks (RNN) and the linguistic structure through word embedding representations [Mik+13]. More recently, studies on graph structures have built rich representations with embedding methods such as graph-embeddings or Node2vec [GL16].

However, this structural information is not always available on real data and sometimes has to be reconstructed in a fragmented form. For example, it might be possible to construct a relevant representation that explicitly incorporates transportation plan information in our case. The necessary information on the nature of the theoretical plan is not yet available because of storage and modeling issues. Nevertheless, it is still possible to exploit the available information by enriching the data representation through additional attributes. These other attributes contain information of the underlying structures that cannot be expressed directly.

Among the range of impacts that underlying structures can have on time-series data, our real-world case allowed us tof confront the following issues:

- **Variability in the time structure** may be caused by the influence of an underlying structure on the measurements or observed phenomenon. For example, our sequence of train stops has a temporal structure based on the schedule fixed by the transportation plan. This factor induces temporal variability in the time series. It makes it more challenging to analyze the temporal influence and regular patterns of data. Nevertheless, it is possible to recover part of the information scrambled by temporal variability through temporal attributes. These attributes will serve as support allowing a model to reconstruct the temporal influence. Since this influence is often linked to periodic phenomena, it will be relevant to construct a temporal representation through cyclic attributes (Section 3.3.3). These cyclic attributes can facilitate the capture of part of the dynamic context of the time series.
- **Heterogeneous observations** can be induced by the underlying structure. Each time series observation can be structured in a specific way. Therefore, it requires knowledge and in-depth analysis of the influence of the underlying structure to

extract, synthesize and format meaningful attribute representations that facilitate the prediction task. For example, in our application, each element of our trainload series related to a station will depend on the train's mission linked to the transportation plan. Providing a representation of information contained in this underlying structure through contextual attributes is decisive to produce a reliable prediction.

- **Asynchronous observation between stations** is a tricky problem that we faced but did not directly address. As the train loads of different stations are asynchronous series, simultaneous analysis becomes a challenging issue, particularly with machine learning models that are very dependent on regularity and face difficulties in handling non-synchronous time series. We therefore decided to focus on univariate prediction with a model predicting the passenger loads at a station level.

To synthesize, we propose to tackle the prediction task on time series with underlying structures by using contextual attributes. This underlying structure induces temporal variability and heterogeneity on the time series to be predicted. We propose to solve the temporal variability by relying on a temporal representation taking the form of cyclic attributes. In the same way, heterogeneity is solved by exploiting available information on the underlying structure to construct heterogeneous structural representations. Both representations enrich the set of contextual attributes to facilitate the capture of the dynamic context influence. The dynamic context is formed by the cross-influence of all the latent and observed factors from which the underlying structures evolve. We propose a short-term prediction (Equation 3.10)based on both short term attributes s_t (aiming to capture mainly part of the influence of latent factors) and contextual features c_t (aiming to capture the influence of observed factors and the extracted underlying structure).

$$F^d(c_t, s_t) \approx f_t^c(c_t, l_t) + f_t^d(c_t, l_t) \approx y_t \tag{3.10}$$

with c_t the combination of contextual representations.

Regarding the spatial aspect, we propose to tackle the prediction task on a univariate regular time series that corresponds to the sequence of trainload' values when passing through a station. Short term attributes are composed of past loads and delays. Contextual features consist of the calendar, temporal, and train service representations.

3.3.3 Cyclic encoding for temporal representation

The temporal structure is a valuable support for the analysis and understanding of time series behavior. Face to non-regular observation can make these regular patterns disappear in the other influential factors' variability. However, the irregular temporal structure can be extracted as additional contextual attributes that should retain as much information as possible about the initial structure.

The most direct way to extract the structure consists in using the form of 'one-hot' attributes. With periodic patterns, it is possible to cut each period into c homogeneous classes according to the temporal structure. The number of classes results from a trade-off between precision and representativeness. We can then associate each element to a binary vector of c dimension having a unique non-zero value for the corresponding class. However, considering our application problem, a one-hot representation leads to bulky attributes without continuity omitting the proximity between the time steps.

$$\forall\, t, \forall\, k \in [1, c],\ X_t^{oh} \mid\ X_{t,k}^{oh} = \mathbb{1}_{((t\ \%\ p) \div (p/c))}(k) \tag{3.11}$$

with p periodicity and c number of classes

There are more suitable representations providing compact and meaningful transcriptions. Cyclic encoding aims to encode continuous cyclic attributes by preserving their cyclic structure. Instead of having a sizeable one-hot vector per feature, the sine and cosine encodings project each attribute on a two-dimension plane. A more complex structure with several periodicity patterns, combining several pairs of sines and cosines with different frequencies, can better express meaningful and compact structures. This representation can be suitable for transcribing temporal structures .on irregular time series, particularly on mobility applications with several layers of periodic influences (hourly, daily, weekly, monthly, seasonally).

$$\forall\, t,\ X_t^{cycl} = \bigoplus_{f \in F} (cos(\frac{2\pi * f * t}{p}) \oplus sin(\frac{2\pi * f * t}{p})) \tag{3.12}$$

with p periodicity and f frequencies

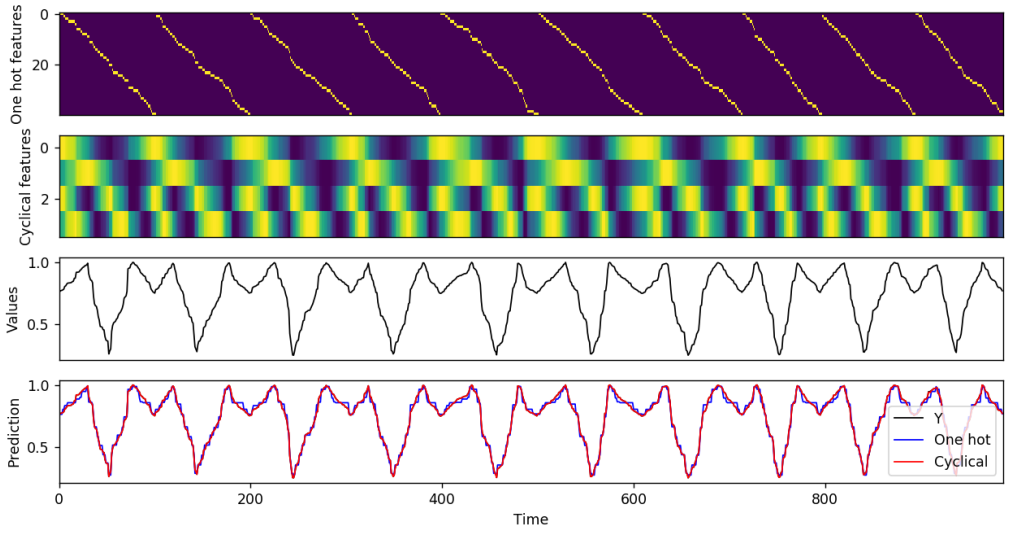

Figure 3.5.: Toy prediction using one-hot and cyclic features

The following Figure 3.5 illustrates the benefit of using cyclic features to forecast a periodic time series toy example. The task consists of a forecast by random forest models using cyclic or one-hot attributes with greed-search tuning (N-estimator, Max-depth) for each attribute set.

The time series are based on a random draw (5%) of 20000 regular measurements on several periods of regular ($p - periodic$) patterns to which we have added a low noise. Figure 3.5 illustrates the cyclic features, the one-hot features, the observations, and the predictions. one-hot encoding does not form perfectly regular streaks due to random sampling, which simulates the irregular temporal structure. For one-hot features, we split a period into 40 regular sections. We combine two pairs of sine and cosine with frequencies of (1/p) and (1/2*p) for cyclic features. Cyclic (4-dimension) attributes are more compact than one-hot (40-dimension) encoding. Yet, the prediction based on cyclic encoding is much better. It avoids the artifacts due to the loss of the continuity information in one-hot attributes.

The cyclic encoding allows a better transcription of time-periodic structures, provides more compact features and facilitates forecasting the learning of the models by reducing their complexities. This encoding is particularly efficient to transcend the impact of periodic temporal structures often present in mobility data. The one-hot encoding is voluminous and unsuitable for transcribing the notion of continuity in the time series.

3.4 Proposed model: LSTM Encoder-Predictor

3.4.1 Prediction models for Temporal Data with underlying structure

In this section we formalize the application of the recurrent encoder-predictor architecture to our particular structural constraints: a train sequence with variable time steps and heterogeneous attributes.

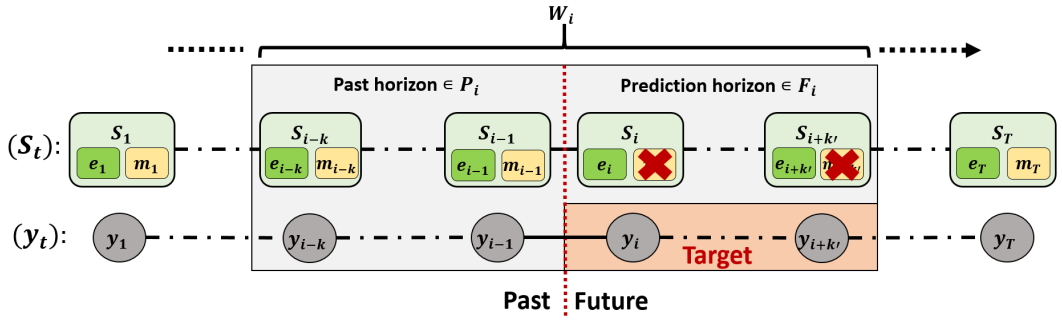

Figure 3.6.: Illustration of notations used in the forecasting model

Let (y_1,..,y_t) denote the sequence to be predicted, where y_t corresponds to a passenger train load in the same station, t referring to the trains' arrival order. It is assumed that each realization y_t is associated with an observation S_t which includes contextual features e_t and past measures m_t. We have tackled the issue of the variability of the time between consecutive trains by encoding it as a contextual feature associated with

specific coefficients in the model. We also use the notation y_I, S_I, e_I and m_I to designate the sub-sequences $(y_t)_{t\in I}, (S_t)_{t\in I}, (e_t)_{t\in I}, (m_t)_{t\in I}$ with $I \subset [1;T]$. Given a time window $W_i = [i-k, i+k']$ composed of a past horizon $P_i = [i-k, i[$ and a future horizon $F_i = [i, i+k']$, the goal of our multi-step forecasting approach is to infer a realization on the horizon y_{F_i} from information available on S_{W_i} as shown in Figure 3.6. Table 3.6 summarizes the notation used in this article.

Table 3.6.: Notations and variables

Notation	
t	A time step $t \in [1, T]$
$y_1, ..., y_T$	(y_t) Realization series
$S_1, ..., S_t$	(S_t) Observation sequences
$e_1, ..., e_T$	(e_t) Sequence of feature contextual vectors
$m_1, ..., m_T$	(m_t) Sequence of feature measure vectors
	Windows
W_i	$[i-k, i+k']$: Window associated to the ith observation
P_i	$[i-k, i[$: Past horizon of window W_i
F_i	$[i, i+k']$: Prediction horizon of window W_i
x_i	$(m_{P_i}, e_{P_i}, e_{F_i})$ Input features from the window W_i
	Latent space (see subsection 3.1)
$u_1, ..., u_T$	(u_t) Contextual representation
$h_1, ..., h_T$	(h_t) Latent past dynamic
$r_1, ..., r_T$	(r_t) Latent reconstruction state
$z_1, ..., z_T$	(z_t) Latent prediction state
	Model and sub-part
$LSTM_{EP}$	Neural network model
$Fact$	MLP factoring contextual features
Enc	Recurrent encoder of past observation
Dec	Recurrent decoder of past observation
$Pred$	Recurrent predictor of future observation
$Reconst$	MLP to reconstruct past realizations
$Predict$	MLP to predict future realizations

This forecasting is particularly challenging because it requires understanding the laws behind realizations (y_t) considering the multiple influential factors. The model must be able to dissociate the influential factors on a structurally irregular sequence by exploiting the contextual attributes.

3.4.2 Inspiration for the Model

The RNN encoder decoder architecture is a relatively recent predictive neural network architecture that emerged from research accomplished by [Cho+17]. The authors propose a recurrent neural network encoder-decoder for a statistical machine translation system. This model is capable of capturing both semantic and syntactic structures of phrases. This type of architecture is used in several fields, such as for example in video sequence

prediction [SMS15] with an LSTM encoder layer to encode an image sequence, an LSTM to reconstruct the sequence, and finally, an LSTM predictor layer to predict the sequence.

To address the structural variability in the passenger load series and influential factors, we relied on the abstraction capabilities of deep neural network models linked to the concept of representation learning [BCV13]. The underlying idea is to learn a meaningful representation of mobility flows taking contextual factors into account. The proposed model takes the form of an RNN encoder-decoder neural network associated with the representation learning of contextual factors. It aims to predict the passenger load of the next trains at a station from measures of the last trains and all the contextual features characterizing all of these trains at the same station.

3.4.3 Method description

Given observations on a time window S_{W_i}, the method aims to reconstruct the k last realizations $\widehat{y}_{P_i}$ and to predict the k' next realizations $\widehat{y}_{F_i}$ considering contextual information e_{P_i} and measure information m_{P_i} on the past horizon and contextual information e_{F_i} on the future horizon. It is a deep neural network that can be decomposed into sub-parts with specific roles. A general illustration of the proposed model is given in Figure 3.7.

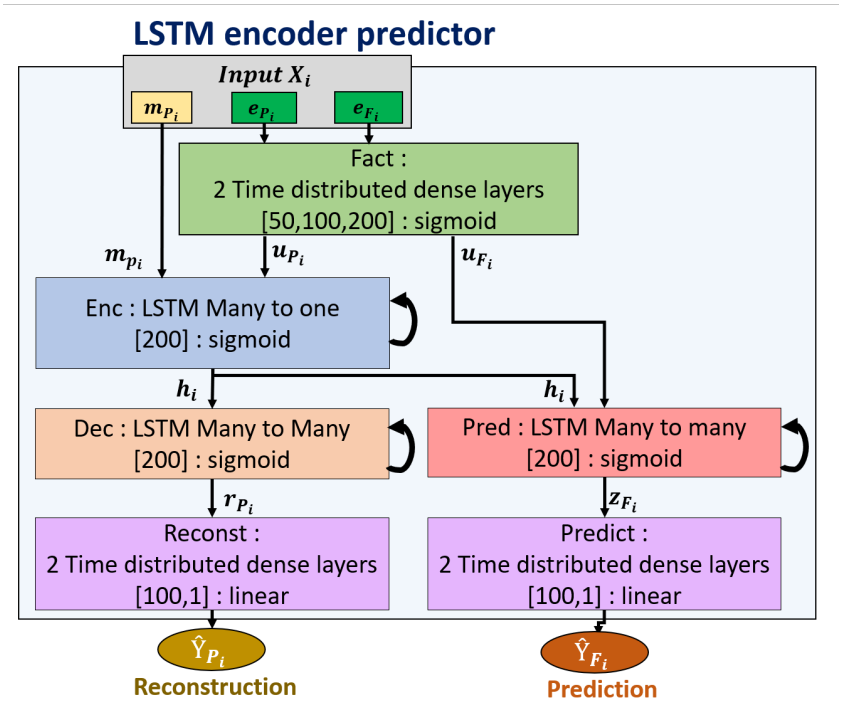

Figure 3.7.: General architecture of the LSTM encoder-predictor network

The arrangement of the different components of the LSTM are detailed in Figure 3.8. A more detailed view of the architecture is provided in Appendix A.3. The sub-parts of the proposed architecture are described as follows.

$$LSTM_{EP}(x_i) = LSTM_{EP}(m_{P_i}, e_{P_i}, e_{F_i}) = (\widehat{y}_{P_i}, \widehat{y}_{F_i}) \tag{3.13}$$

$Fact$: A context factory is dedicated to synthesizing contextual features (e_t) as contextual representations (u_t). It is a preprocessing multilayer perceptron applied on each observation to regularize contextual representations.

$$Fact(e_{P_i}, e_{F_i}) = \bigoplus_{t \in (P_i \cup F_i)} Fact(e_t) = \bigoplus_{t \in (P_i \cup F_i)} u_t = (u_{P_i}, u_{F_i}) \tag{3.14}$$

Enc: A many-to-one LSTM 'encoder' is dedicated to capturing a past latent dynamic (h_i) from the past measures m_{P_i} and the past contextual representations (u_{P_i}).

$$Enc(m_{P_i}, u_{P_i}) = h_i \tag{3.15}$$

Dec: A many-to-many LSTM 'decoder' recurrently decodes latent reconstruction states r_{P_i} of past observations from latent dynamics of the past horizon (h_i). Each latent reconstruction state is then interpreted by **'Reconst'**, a linear reconstruction layers that infers the realization of past observations. From **'Reconst'** outputs we get $\widehat{y}_{P_i}$. These outputs are used as an intermediate objective during the training phase to facilitate the capture of past latent dynamics.

$$Dec(h_i) = r_{P_i} \tag{3.16}$$

$$Reconst(r_{P_i}) = \bigoplus_{t \in P_i} Reconst(r_t) = \widehat{y}_{P_i} \tag{3.17}$$

Enc and Dec form an encoder-decoder structure that synthesizes the dynamics of past observations from their contextual and measurement features.

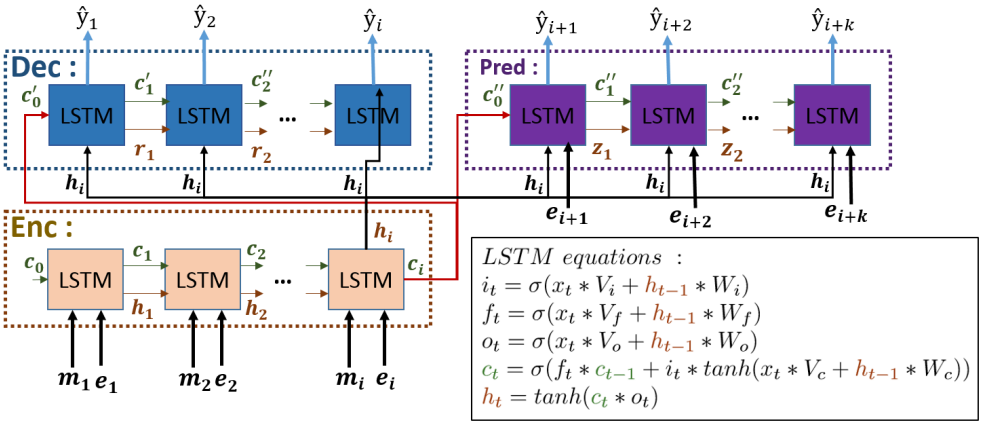

Figure 3.8.: Details of the layout of the LSTMs

$Pred$: A many-to-many LSTM 'predictor' infers latent prediction states (z_{F_i}) of future observations from their contextual representations (u_{F_i}) considering the latent dynamics of the past horizon (h_i). Each latent prediction state is then interpreted by **'Predict'**, a linear prediction layer that infers the realization of future observations. From **'Predict'** outputs we get $\widehat{y}_{F_i}$ which corresponds to the multi-step prediction aim.

$$Pred(h_i, u_{F_i}) = z_{F_i} \tag{3.18}$$

$$Predict(z_{F_i}) = \bigoplus_{t \in F_i} Predict(z_t) = \widehat{y}_{F_i} \tag{3.19}$$

Note that since the model is designed to address variability in the time step, this makes it straightforward to remove observations from the dataset due to missing data. Moreover, once the LSTM encoder-predictor is trained, predictions can be performed on missing data in the future horizon if we are able to reconstruct contextual information.

Optimization

A deep neural network is trained through end-to-end gradient back-propagation by minimizing the following loss function:

$$\begin{gathered} \mathcal{L}(\theta) = \alpha_p * \sum_{t \in P_i} ||y_t - \widehat{y}_t||^2 + \alpha_f * \sum_{t \in F_i} ||y_t - \widehat{y}_t||^2, \\ With\ \theta = (\theta_{Fact}, \theta_{Enc}, \theta_{Dec}, \theta_{Pred}, \theta_{Reconst}, \theta_{Predict}). \end{gathered} \tag{3.20}$$

The first term measures the ability of the model to reconstruct the past observations from the latent past dynamics. It is an intermediate objective that facilitates the learning of the past dynamics. The second term measures the prediction ability of the model. Hyper-Parameters α_p and α_f are the weights of the reconstruction and prediction objectives.

For the learning phase, we perform mini-batch optimization thanks to a *Nadam* optimizer [Sut+13]. Two gradients (prediction and reconstruction) are propagated from their output layers (Predict and Reconst) to the upstream layers towards the context factory through LSTM layers. The encoder-predictor is implemented based on the *TensorFlow* [Aba+16] environment and *Keras* [Cho+15] as a library and high-level neural network API. The parameters were chosen empirically after several experiments based on model performance and learning convergence.

Training is empirically conducted on a batch of size 128 on several thousand iterations, which takes a few hours on a standard GPU card depending on the dataset and time depth. Further work on the choice of parameters is required to improve the convergence.Additional information on the training (Appendix A.2) and architecture of the neural network (A.3), as well as a more complete representation are presented in the appendix.

3.4.4 Modeling in the case of regular time series

The model was designed to manage the structural variability linked to the time series's temporal irregularity and heterogeneity. However, the combination of a recurrent neural

network with an encoder-decoder architecture can also be used to capture the influence of a dynamic context on more regular series. The regular temporal structure and the homogeneity of the time series make it possible to simplify the architecture of the LSTM-EP. The model no longer needs to retrieve information related to the characteristics and the irregular temporal sampling of the time series. Decoder and predictor can be merged into one LSTM layer that simultaneously reconstructs the current load at each time step and infers the evolution of the variable of interest over time through the recursive process.

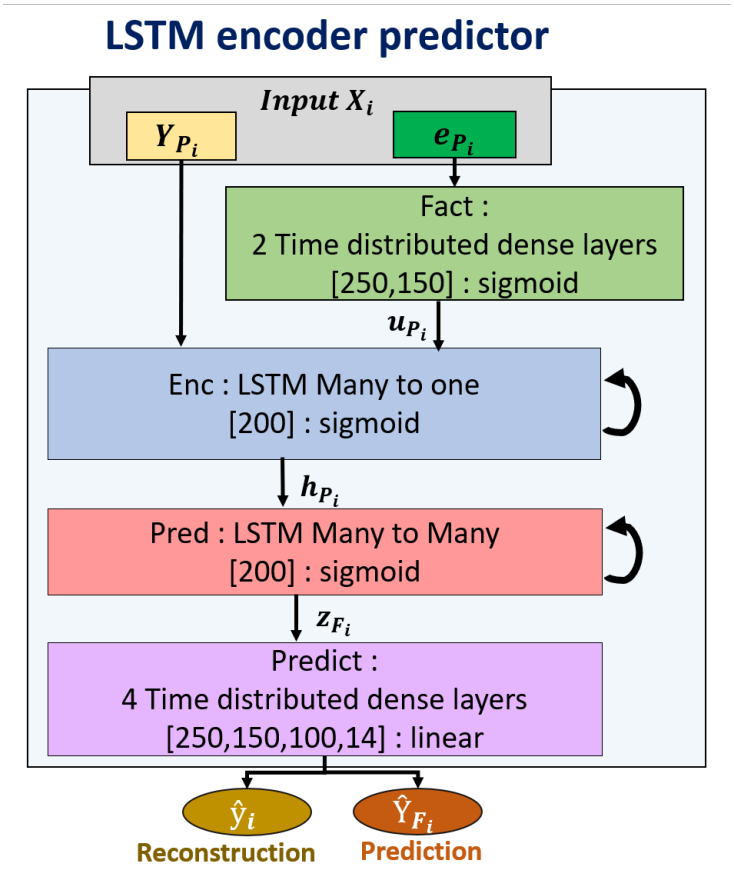

Figure 3.9.: LSTM-EP architecture with the layer size for the real data

The simplified architecture is illustrated in Figure 3.9. First, the long-term features are synthesized through a multi-layer perceptron neural network. Subsequently, a pair of encoder/predictor LSTM layers attempts to capture the contextual influence and infer the multivariate time series's short-term dynamics. Finally, another multi-layer perceptron attempts to interpret the prediction embedding Z^p to produce a prediction $\widehat{y_t}$. This model takes as input the contextual attribute x_t and the past horizon value $y_{P_t} = [y_{t-p}, \ldots, y_t)$ and forecasts a future horizon $[y_t, y_{t+f})]$. Such a model reconstructs the time step t and then infers the temporal evolution on a future horizon $[t+1, t+f]$. Dropout layers are placed in almost every layer to avoid overfitting and to allow variational dropout. Additional information on the training and architecture, as well as a more complete representation are presented in Appendix A.5.

3.5 Train ridership forecasting experiments

3.5.1 Data description

This use case focuses on a dataset collected from a French railway line that serves approximately fifty stations located in the northern area of suburban Paris. The railway line carries approximately 250,000 passengers daily. The dataset covers a period of 18 months from January 2015 to June 2016 on 40 stations for daytime exploitation from 5 am to 2 am of the next day. It includes both timetable information and count data of passengers boarding and alighting at each station collected by radar sensors on trains (2000000 records covering 86% of trains). These heterogeneous sources of data that have been enriched with calendar information enable us to reconstruct the passenger load on each train departing from a station.

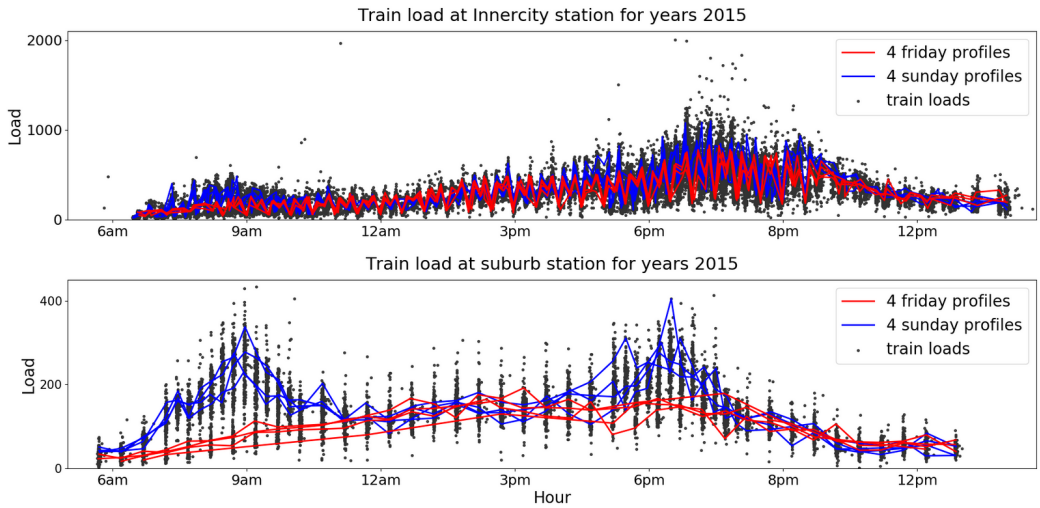

Figure 3.10.: Train loads in year 2015 per hour on suburban and inner-city stations

The main goal of this study concerns the forecasting of univariate train load series for each station. To have an idea of the time series to be predicted, Figure 3.10 shows an example of weekday and weekend daily train passenger load profiles collected from two stations. The suburban station accounts for 22000 train stops with a particularly low train frequency and few train routes that serve this station. Conversely, the inner-city station accounts for 84000 train stops with a high train frequency and multiple train routes that serve the station. Figure 3.10 provides insights into the forecasting problem to be solved and highlights the particularities of our dataset, namely:

- A variable sampling period due to the train timetables and railway operation. Each station has its own train frequency evolution.
- A specific temporal behaviour of each time series, which was found to be linked to the spatial location of public transport stations and geographical aspects of the city (population & employment densities, leisure and so forth).

- Train load series are impacted by calendar factors such as the type of day (weekday or weekend), holiday, public holiday and so on.
- Train load series are also impacted by train characteristics that are closely linked to their services (multi-destination line, various train services).

3.5.2 Description of feature sets

The work focuses on forecasting the passenger load in trains, from a calendar, transportation plan, short-term load, and delay information. As mentioned in Section 3.3.2, these contextual and short-term attributes provide valuable insight to predict the train load series. This section describes the content and processing of the features used to feed the machine learning models in charge of modeling the series related to the train load dynamics.

Calendar representation

Although we do not know all of the factors related to human activities that feed the transportation demand at each station, it possible to model transportation demand approximately by observing the periodic trends and patterns associated with the different calendar factors. As mentioned in Section 3.5, a more relevant representation through cyclic encoding can better retain continuity properties and avoid a bulky one-hot representation. In the following list, we detail the calendar attributes that feed the models. We distinguish the raw features (one-hot) from the processed cyclic features ($^+$).

- $Hour$: quarter-hourly one-hot representation (One-Hot 96-dimension)
- $Hour^+$: minute of the day (1440 possible values) encoded by the cosine and sine of (2×4) periods of 15min, 60min, 1/4day, 1/2day (Cyclic 8-dimension)
- Day: one-hot encoding type of day (One-Hot 7-dimension)
- $Holiday$: marker of Christmas day, School and Public holiday (Binary 3-dimension)
- $Season$: one-hot encoding of a month (One-Hot 12-dimension)
- $Season^+$: day of a year (365 possible values) encoded by the cosine and sine of (2×4) periods of 15days, 1month, 3months and 6months (Cyclic 8-dimension)

Transport service representation

The transport schedule is also a structuring item of information for the train load series. Section 2.3 illustrates the difficulties linked to the variability of the series and the lack of reliable information related to the whole train service code (high percentage of missing data and numerous inconsistencies). A one-hot representation based on these mission codes would be both too bulky and extremely poor because it would lose all the proximity information between missions. To build attributes synthesizing information on the train

service, we first reconstructed the entire train service by data-mining. We were, therefore, able to construct a binary vector of station stops for each train. We see more than 50 different services by analyzing the recurring patterns, but a large part of these services varies very rarely. The various branches of the lines structure the service patterns. It is then possible to synthesize the sparse and redundant information in these binary vectors using a dimension reduction technique, such as principal component analysis (PCA). Here, we used a principal component analysis to perform a dimension reduction from a binary vector of 50 dimensions to a subspace of 8 dimensions that explain 97% of the variance.

- $Service$: missions of the trains encoded under a vector of binary values where each value indicates whether the train is serving the station (50-dimension)
- $Service^{+}$: the eight first principal components of the PCA reducing the set of binary service vectors (8-dimension)

Long term attribute sets

By combining the calendar information and the transport timetable information, we obtain a training set of long term attributes on which the long-term forecasting model can then be trained. Processing long term features provides models with relevant information to better model the long term dynamic. We propose to observe the contribution of both temporal representations based either on one-hot features or features extracted with cyclic transformations and both train service representations based on basic attributes or processed attributes. The composition of the long term feature sets is detailed in Table 3.7.

- $LT_1 : (Hour, Day, Holiday, Season)$: Calendar representation using one-hot attributes (114-dimension).
- $LT_1^{+} : (Hour^{+}, Day, Holiday, Season^{+})$: Calendar representation using cyclic attributes (26-dimension).
- $LT_2 : (Hour, Day, Holiday, Season, Service)$: Calendar and train service representation using basic features (164-dimension).
- $LT_2^{+} : (Hour^{+}, Day, Holiday, Season^{+}, Service^{+})$: Calendar and train service representation using processed features (164-dimension).

Table 3.7.: Sets of long term features

Features:	Day	$Holiday$	$Hour$	$Hour^{+}$	$Season$	$Season^{+}$	$Service$	$Service^{+}$	Size
Set Dimension	7	3	96	8	12	8	50	8	
LT_1	✓	✓	✓		✓				114
LT_1^{+}	✓	✓		✓		✓			26
LT_2	✓	✓	✓		✓		✓		164
LT_2^{+}	✓	✓		✓		✓		✓	34

Short term attributes

Moreover, we also consider short-term attributes to infer the short dynamic of our train load series. Train delay and counts of passengers boarding and alighting are measured at each train passage. The load in the train is then deduced from these counts at each stop. It is possible to use this past information by considering the delay of the train at the last stop, or the load, the alighting or boarding of previous trains passing through the station to estimate the load of the next train. For passenger load, this means using a lagged window of past values of the variable of interest to estimate these future values. Two parameters will be important: The depth p of the past horizon and the lag l of observation. In an ideal real time setting, the observation lag is 0, the information of the preceding train is directly observed:

- $Delay$: difference in minutes between the real and the theoretical schedule of the train at the last station (1-dimension)
- $Lag(l, p)$: short term measures (load, board, alight) for each train from the 'p' past horizon with 'l' observation lag at the considered station (i.e. from the '$p - l$' th last passage to the 'l' th last passage) (3*p-dimension)
  - $load$: number of passengers on the train at departure from the station
  - $boarding$: number of passengers who boarded at the station
  - $alighting$: number of passengers who alighted at the station

Sets of short term attributes

The short term measures provide information to estimate the short-term dynamics of the series. The temporal depth p and the observation lag l are two factors that will impact the quality of the estimate. In the presence of a structuring contextual dynamic, it may be interesting to combine this short-term information with long-term attributes providing information on the contextual structure to thus facilitate the modeling of series dynamics. We are going to build slightly different sets of attributes to observe the contribution and impact of these attributes. The composition of the long term feature sets is detailed in Table 3.8.

- $ST_1 : (Delay, Lag(l = 0, p = 5))$ short-term measures on the last 5 trains (16-dimension)
- $ST_2 : (LT_2, Delay, Lag(l = 0, p = 1))$ combination of long term representation LT_2^+ and short-term measures of the last train (40-dimension)
- $ST_3 : (LT_2^+, Delay, Lag(l = 3, p = 5))$ combination of long term representation LT_2^+ and short-term measures with past horizon composed of the third to eighth last trains (60-dimension)
- $ST_4 : (LT_2^+, Delay, Lag(l = 0, p = 5))$ combination of long term representation LT_2^+ and short-term measures on the last 5 trains (60-dimension)

Table 3.8.: Short term feature sets

Set	Features Dimension	LT_2^+ 34	Delay 1	Lag 3*p		Size	Comment
ST_1			✓	l=0	p=5	16	only short-term attributes
ST_2		✓	✓	l=0	p=1	40	with small past horizon
ST_3		✓	✓	l=3	p=5	60	with observation lag
ST_4		✓	✓	l=0	p=5	60	real time setting

3.5.3 Forecasting Models

We will compare the performance of several baselines and machine learning models introduced in Section 3.2. Two baselines were used as references for long and short term approaches:

Last Value (LV) consists in predicting the next value of a series by the last observed values. It is a relatively efficient Baseline on mobility data whose performance can be used as a minimum benchmark for a short-term approach.

Contextual Average (CA) consists in using average values of a set sharing the same context. On our train load, basic context is defined as the same day type, for a time slice, on seasonal periods. It is also a relatively efficient baseline on mobility data whose performance can be used as a minimum benchmark for the long-term model.

We will also use two standard machine learning approaches to forecast. Both approaches are attributes-based and can perform long or short term predictions based on the type of attributes provided.

Gradient Boosting (XGB) : A regressor model using a succession of decision trees that iteratively improving the model. We can therefore build several models depending on learning attribute sets. The two main models are:

- **XGB-LT**: Model trained on long-term features (LT_2^+ attribute set)
- **XGB-ST**: Model trained on both long and short-term features (ST_4 attribute set).

Random forest (RF): A regressor model based on an ensemble of independent decision trees. RF is based on attributes. We can therefore build several models:

- **RF-LT**: Model trained on long-term features (LT_2^+ attribute set).
- **RF-ST**: Model trained on both long and short-term features (ST_4 attribute set).

Finally, prediction model based on recurrent network approaches: **LSTM**: Basic recurrent network using the ST_4 attribute set. **LSTM-EP**: The proposed RNN network using the ST_4 attribute set.

The parameters of the XGB and RF models were selected though a random search procedure using Sequential 3-fold cross validation (Appendix A.2). The learning procedures of the neural network models are also detailed in Appendix A.2. We evaluated the performance of the models on each time step by root mean square error (RMSE) and weighted absolute percentage error (WAPE) measures:

$$RMSE : \sqrt{\sum_{t}(y_t - \widehat{y}_t)^2} \qquad WAPE : \frac{\sum_t ||y_t - \widehat{y}_t||}{\bar{y}} \tag{3.21}$$

WAPE is a derivative of the MAE that can be interpreted as the percentage of the overall error compared to the average value of the actual observation.

3.5.4 Preliminary experiments on feature contributions

Preliminary experiences were conducted in order to have an overview of the contribution of the different attribute sets to the forecasting task and the induced performances of the proposed machine learning models (XGB, RF) according to these attributes. Two metrics were used for the performance assessment: the mean square error (RMSE) and the weighted absolute percentage error (WAPE) on the train and test sets.

Evaluation was conducted on two groups composed respectively of four inner-city and four suburban stations. The proximity of the target stations to the downtown urban area influences the train service and the stations' ridership profiles. Experiments consisted in evaluating, for both groups of stations, the contribution of the attributes through the mean prediction performance of the ML prediction models (RF and XGB) according to the provided attributes.

Overall, the results of the two experiments reported in Tables 3.9 and Table 3.9 show that the performances of the XGB and RF models are relatively similar for the same set of attributes. We can also see that inner city stations have a higher RSME error which is explained by the higher average load in the city. Conversely, the MAPE error is higher on suburban stations which can be explained by the lower signal-to-noise ratio due to the low ridership.

Table 3.9 provides the prediction performance based on the long term attributes. By firstly looking at the contribution of features processing, a significant gain is observed between raw and processed calendar attributes ($LT_1 < LT_1^+$). Machine learning based models with canonical calendar attributes perform less well than models with processed calendar attributes. We observe the same behavior on train service attributes ($LT_2 < LT_2^+$). These gains are explained by more compact and meaningful attributes that facilitate the inference.

Table 3.9.: Mean of forecast performance on Inner-city and Suburban stations according to attribute sets

Stations		Inner-city				Suburban			
	Metric	MAE		MSE		MAE		MSE	
Set*	Model	Train	Test	Train	Test	Train	Test	Train	Test
-	CA**	13.81	14.29	87.69	91.01	25.76	25.83	46.30	47.94
LT_1	RF	26.12	25.91	132.14	132.02	48.58	47.87	92.04	91.98
LT_1^+	RF	**11.45**	**14.32**	**72.07**	**90.05**	**12.55**	**18.30**	**27.73**	**44.32**
LT_1	XGB	26.07	25.91	131.95	131.99	48.44	47.84	91.94	91.96
LT_1^+	XGB	**11.35**	**14.49**	**69.91**	**91.40**	**14.00**	**18.70**	**32.59**	**44.03**
LT_2	RF	13.68	14.04	84.64	88.01	16.60	16.64	38.98	39.91
LT_2^+	RF	**8.87**	**11.18**	**59.86**	**74.91**	**10.68**	**13.73**	**24.26**	**32.73**
LT_2	XGB	13.40	13.79	83.99	87.35	16.37	16.52	38.63	39.85
LT_2^+	XGB	**8.35**	**11.30**	**55.26**	**76.17**	**10.56**	**13.99**	**24.78**	**32.96**

*Notation of feature sets are defined in Table 3.7 **CA= Calendar average

Table 3.10 provides the prediction performance based on the long and short term attribute sets. We observe that calendar based models perform as well as contextual average baseline models. The contribution of the train service attributes significantly increases the forecasting accuracy ($CA = LT_1^+ < LT_2^+$). For the short term approach, real-time features alone do not really allow good predictions due to the high variability of train services ($ST_1 < CA$). On the other hand, the combination of contextual information and short-term information improves predictions. For short term measurements, the factors of temporal depth and measurement delay impact the quality of the prediction (ST_2 & $ST_3 < ST_4$).

Table 3.10.: Mean of forecast performance on Inner-city and Suburban stations by set of attributes

Stations		Inner-city				Suburban			
	Metric	WAPE		RMSE		WAPE		RMSE	
Set*	Model	Train	Test	Train	Test	Train	Test	Train	Test
-	CA**	13.81	14.29	87.69	91.01	25.76	25.83	46.30	47.94
-	LV	39.60	39.22	190.01	189.27	83.35	81.52	161.59	160.61
LT_1^+	RF	11.45	14.32	72.07	90.05	12.55	18.30	27.73	44.32
LT_1^+	XGB	11.35	14.49	69.91	91.40	14.00	18.70	32.59	44.03
LT_2^+	RF	8.87	11.18	59.86	74.91	10.68	13.73	24.26	32.73
LT_2^+	XGB	8.35	11.30	55.26	76.17	10.56	13.99	24.78	32.96
ST_1	RF	8.29	16.55	48.52	99.02	12.02	21.74	26.54	49.44
ST_1	XGB	7.26	16.39	42.90	97.96	13.91	22.08	26.89	49.19
ST_2	RF	6.51	10.88	44.12	72.94	9.22	13.34	19.87	31.98
ST_2	XGB	5.58	11.00	35.76	73.00	10.09	13.44	21.1	31.46
ST_3	RF	5.26	10.55	35.04	70.39	6.06	12.82	15.23	31.95
ST_3	XGB	4.24	10.55	25.00	69.43	8.38	13.13	17.09	31.46
ST_4	RF	5.19	10.41	34.40	69.64	**5.96**	**12.73**	**14.97**	31.31
ST_4	XGB	**4.96**	**10.40**	**30.55**	**68.87**	8.04	12.98	15.13	**30.82**

*Notation of feature sets are defined in Table 3.7 and 3.8

Feature importance is a measure of the contribution of attributes to the models based on decision trees. It is based on the number and position of constraints for each feature. This measure has limitations described in [Str+07] that restrict its quantitative interpretation but still offers some qualitative feedback. On Figure 3.11, we can observe that the same contribution trends appear for the two types of models. We can also see significant differences between inner-city and suburban stations. For long-term features, train 'Mission' feature makes an important contribution, as it contains exclusive information about train services and shared temporal information through rush-hour or week-end train services. The contribution of the calendar information may appear small, but it remains additional information that may help to enhance the quality of predictions.

The contribution of past short term attributes informs us that:

- For the suburban stations which have homogeneous train services, the most recent trains are more important. In addition, as trains are lightly loaded at suburban station, passenger boarding flows seem to be of significant importance.
- On the other hand, inner-city stations have more heterogeneous train services. We observe that the 4th last train (T-4) makes a higher contribution. Given the irregularity of the transportation plan (frequency, mission, etc.), the model gives more weight to the fourth last train which seems to have, on average, the most similar mission to the train to be predicted. However, the most recent trains can also contain relevant information on the short term dynamics of the train load.

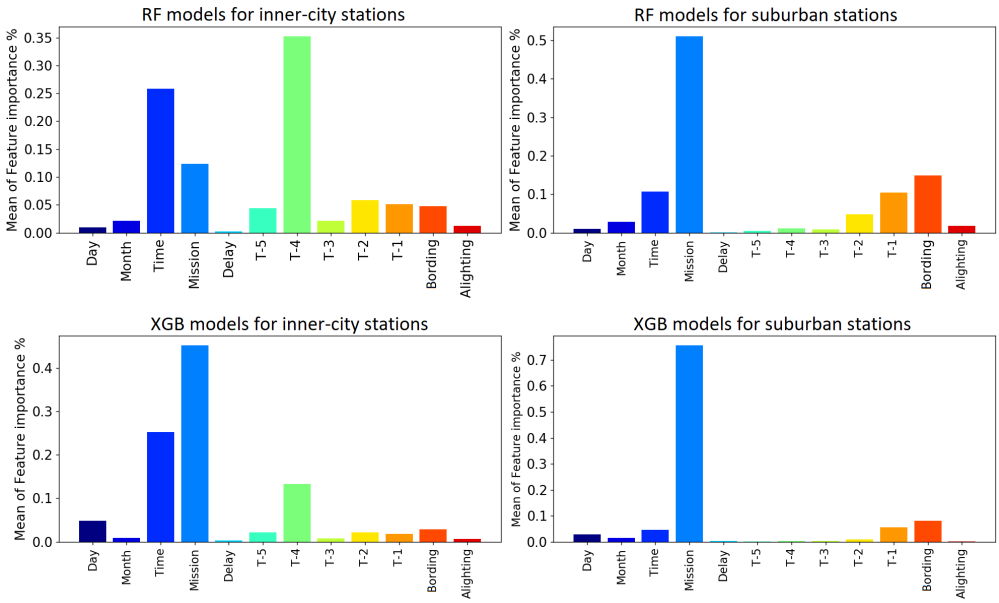

Figure 3.11.: Means of feature importance for RF-ST and XGB-ST models

For other short-term features, past boarding and alighting attributes seem to make a significant contribution to inferring the local station dynamics. The 'Train delay' attributes does not appear to be relevant, but it contains relevant information for a very small portion of the trains which are delayed, which may bias its apparent contribution.

These observations highlight the importance of the short-term features and especially those that encode the train service. Furthermore, it shows the influence of the transportation and traffic plan on the performances of the developed models.

3.5.5 Results of forecasting experiments

For the experiments on advanced models, we focused on a intercity station and a suburban station. The evaluation of the forecasting models was conducted by making comparisons based on performance metrics. These metrics are expressed in terms of RMSE and WAPE and are given for both the training and the test phases for each station. We compared the 5 models defined in Section 3.5.3, namely, the basic last value (LV), the contextual average (CA),the long term gradient boosting (XGB-LT), the short term gradient boosting (XGB-ST), the standard LSTM (LSTM) and the proposed architecture (LSTM-EP). The errors obtained for both the training and test sets are given in Table 3.11.

Table 3.11.: Model performance on the two studied stations

Model	Suburban				Inner-city			
	WAPE		RMSE		WAPE		RMSE	
	Train	Test	Train	Test	Train	Test	Train	Test
LV	17.9	24.1	35.8	47.2	41.9	46.9	186.7	205.0
CA	13.7	19.0	28.7	40.0	14.2	18.5	73.1	96.5
XGB-LT	8.4	18.8	17.2	38.9	8.3	13.4	44.75	76.0
XGB-ST	**7.5**	16.8	**15.1**	35.7	**8.2**	12.7	**43.5**	73.0
LSTM	11.5	16.2	24.3	34.0	8.9	13.7	51.5	75.3
LSTM-EP	10.7	**16.0**	22.1	**33.8**	10.9	**12.9**	57.7	**72.4**

The results show that advanced models (XGB, LSTM) outperform the LV and CA models. This performance improvement can mainly be explained by the fact that the XGB and LSTM models have better generalisation abilities and are able to fit more complex models than LV and CA, which simply predict by forwarding the last observed value or averaging historical data. The basic LSTM model performs less well in the inner-city station compared to the suburban station. This can be explained by its difficulties in dealing with the heterogeneity in terms of train services with that kind of station. Overall, LSTM-EP leads to the best results since it is better able to capture the underlying dynamics of the temporal irregularity related to the heterogeneity of train services by means of its encoder-decoder component.

Looking at the prediction error of the LSTM-EP according to the load to predict (Figure 3.12), we observe that errors increase with the load. The model tends to slightly overestimate weakly loaded trains and greatly underestimate the highly loaded trains. Heavily loaded trains are rare and present contextual information similar to many less loaded trains, which explains the difficulty of the model in predicting large loads. To

remedy this problem, it appears necessary to provision features to distinguish these trains. One could imagine indicators related to the disturbance of the network and the known presence of events near the station.

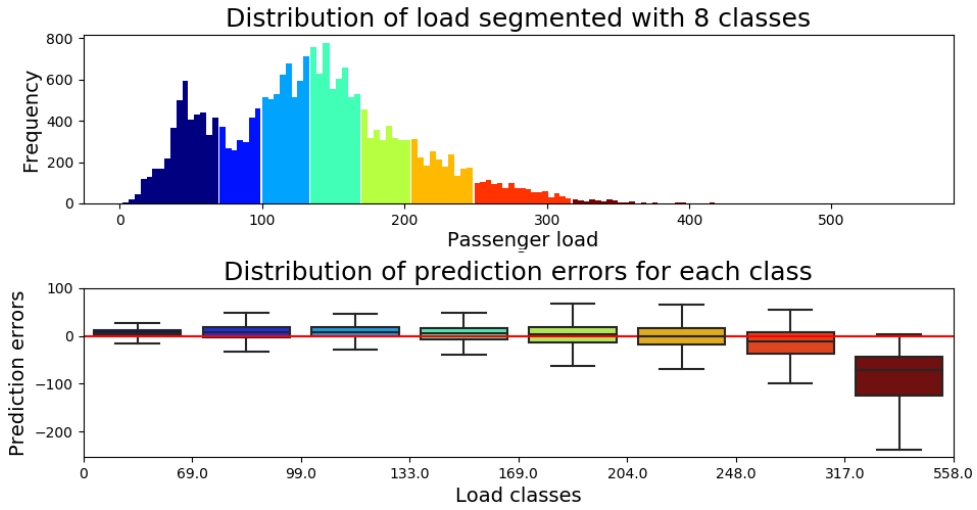

Figure 3.12.: Prediction errors depend on load class for suburban station

As shown in Figure 3.13, the model makes errors of the same order of magnitude for weekdays and weekends with different difficulties. The variance of the error over a time slot is correlated to the magnitude of the load. On weekdays, we observe larger errors in the morning and afternoon peak hours linked to the strong variance and high load. The model makes average errors in the middle of the day and low errors in the morning and evening. On weekends, except in the morning, a relatively similar error variance is observed with a maximum at noon and in the middle of the afternoon. We also observe that the model has more trouble predicting weekend evenings than weekday evenings.

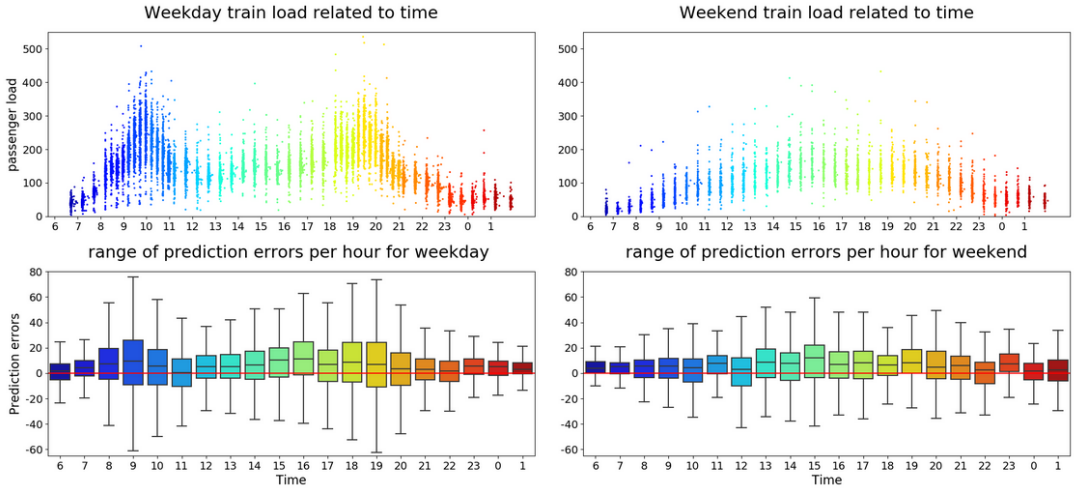

Figure 3.13.: Prediction errors depend on hour class for suburban station

When we examine the performance of the models on multi-step temporal prediction, the LSTM-EP outperforms XGB and the basic LSTM for the next 6 time steps (Table 3.12 and Table 3.13). These time steps correspond to the train passages through the station and

range between 14 and 182 minutes for the suburban station, whereas it ranges between 2 and 61 minutes for the the inner-city where the train passages are more frequent.

Table 3.12.: RMSE test score of the suburban station for the multi-step forecasting models

Model	t+1	t+2	t+3	t+4	t+5	t+6
Time interval*	14-32	29-62	44-92	59-122	75-152	90-182
XGB-LT	38.9	38.9	38.9	38.9	38.9	38.9
XGB-ST	35.7	36.6	36.7	36.7	37.6	38.1
LSTM	34.0	34.4	34.8	35.5	36.3	36.9
LSTM-EP	**33.8**	**34.0**	**34.1**	**34.4**	**34.7**	**34.9**

*The 5th and 95th percentiles of the time interval in the passage of trains at the time T and T+n

Table 3.13.: RMSE test score on the inner-city station for the multi-step forecasting

Model	t+1	t+2	t+3	t+4	t+5	t+6
Time interval*	2-13	5-23	9-31	12-43	15-53	18-61
XGB-LT	76.0	76.0	76.0	76.0	76.0	76.0
XGB-ST	73.0	72.8	73.3	73.8	73.4	73.5
LSTM	75.3	75.4	80.2	83.9	90.5	92.9
LSTM-EP	**72.4**	**72.1**	**72.1**	**72.2**	**72.6**	**72.8**

*The 5th and 95th percentiles of the time interval in the passage of trains at the time T and T+n

Note that these performances were obtained with a single model that simultaneously predicts the load at all the time steps for both LSTM models. The XGB-LT is time-step invariant since it only considers long term features. For XGB-ST, we have as many models as the number of time steps considered. The performance of short-term models is slightly degraded when we move forward in time, excepted for the basic LSTM in the case of the inner-city station, where we can notice a strong degradation of its performance over time steps due to the heterogeneity of train services.The LSTM-EP shows very competitive results and better robustness compared to other the models for both the suburban and inner-city stations for all steps considered. This can be explained by a better understanding of contextual factors through a latent representation that helps to capture the underlying dynamics of train service at the station.

3.5.6 Representation learning exploration

In this section, we explore the latent spaces provided by our neural network to observe the influence captured from the different underlying structures. We analyze the space of contextual representations (u_t) and predictive state representation (z_t) learned by the LSTM-EP model on the suburban station data.

Latent spaces correspond to abstract syntheses of the input features, which are increasingly relevant (over the layers) and better and better arranged (over the layers) with respect

to the target variable (here, Load values). The aim is to distort the initial space of the features to extract a manifold. This manifold associates an inferred load value to each location of the space (symbolizing a state with given input features).

We can obtain a projection of our learning and testing elements on a latent space by recovering the embedding of the related layer during prediction. To facilitate the visual analysis, we apply a dimension reduction process. Dimensional reduction is performed by preserving pairwise distances between points with the help of principal component analysis (PCA) to reduce the embedding of related layer sizes (200 or 300) to 50 dimensions, followed by a T-distributed stochastic neighbor embedding (t-SNE) to reduce the dimensions from 50 to 2 or 3. Each reduced embedding corresponds to the coordinates (in the reduced latent space) of a element. We thus obtain a scatterplot which illustrates the structures of the elements. To allow a visual interpretation of this abstract space, it is possible to draw color plots of the elements according to the values of one attribute.

Figure 3.14 shows the scatter plot of the dimensional reduction of the contextual representation (u_t). It depicts train projection in a space structured by the calendar information. For each sub-plot, the color depends on the influence of a given feature (a: day, b: hour, c: month, d: train load magnitude). The contextual representation (u_t) seems to have a particular arrangement that combines calendar information that characterizing each train. Each point of the obtained structure reflects a train passage whose characteristics are depicted by the color gradient of the 4 color plots.

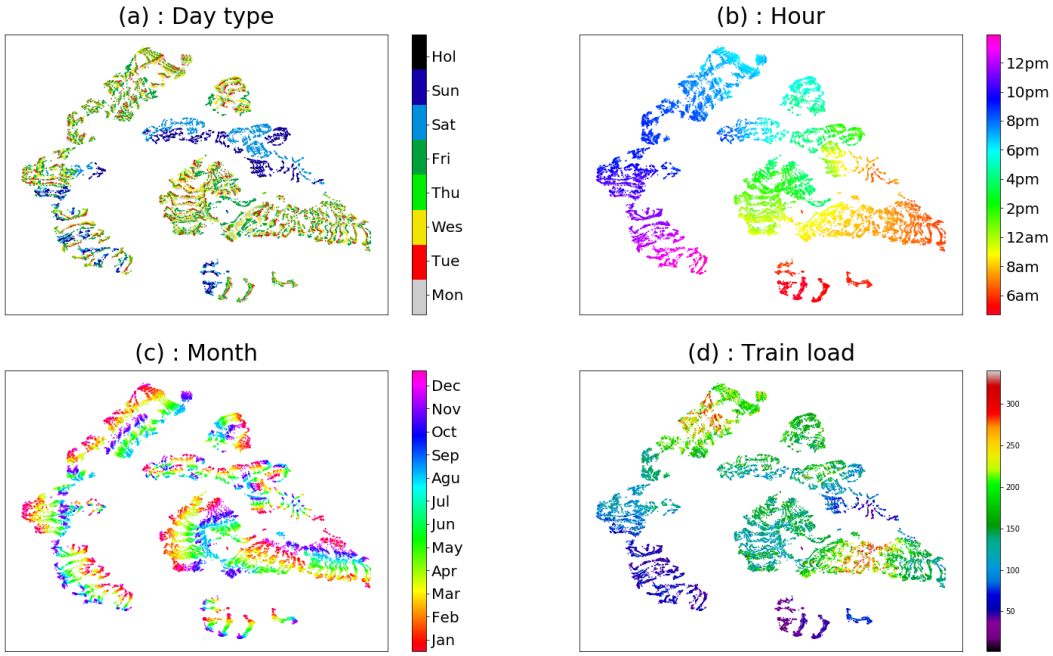

Figure 3.14.: Latent representation (u_t) according to contextual features with dimension reduction

The first plot (Figure 3.14.a) shows that the space distinguishes weekday and weekend behaviors. The second plot (Figure 3.14.b) shows that the hourly information is the main structure of the data: each area of the space will be dedicated to a time slot. Moreover, we observe the continuity of this time structure (rather a diffuse gradient than distinct groups). The third color plot (Figure 3.14.c)) indicates that seasonal information is of a secondary influence. It appears in this space as a local influence that indicates a smaller but still significant influence. This influence also appears to be continuous. Finally, the last color plot (Figure 3.14.d) illustrates the variable of interest that we are trying to infer from the calendar information.

By cross-referencing the color plots, the influence of calendar structures on the load can be highlighted. For example, we can distinguish two red zones (high loads) on plot (d) corresponding respectively to: The weekday trains between 8am and 10am (bottom right) and the weekday trains between 6pm and 8pm outside the summer vacation period (top left). The two rush hour periods are indicated. These representations give a rough illustration of how structural information is captured by the neural network. However, they are not intended to be used explicitly. On the other hand, it is possible to use the vector representation (Embedding) to extract rich information on the topology of the data.

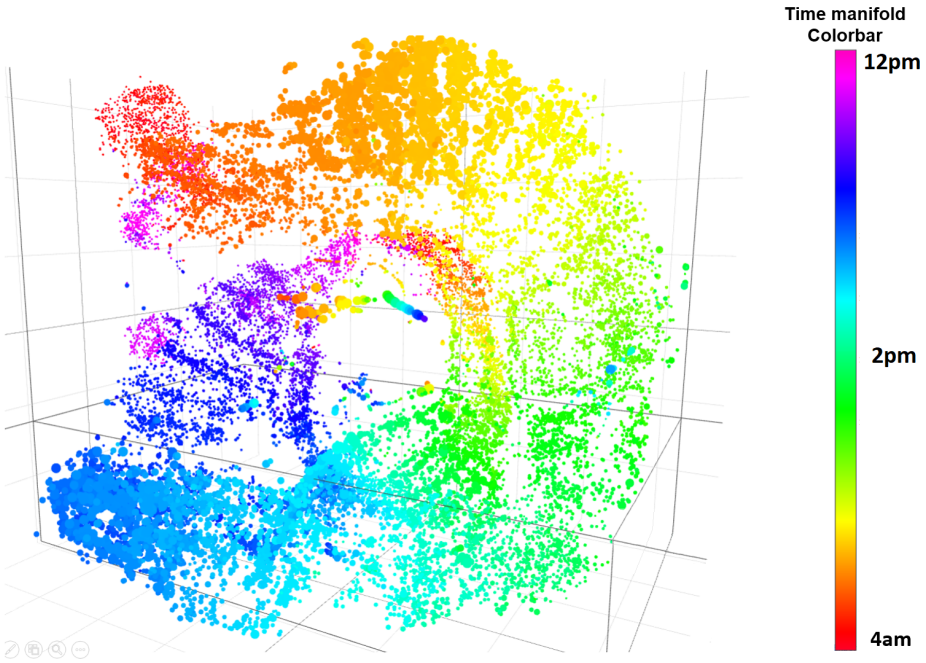

Figure 3.15.: 3D reduction of Predictive latent space (z_t) of suburban station with time coloration

In the same way, Figure 3.15 depicts part of the information captured by the network by an 3dimensional reduction of the predictive latent space (z_t). A more complex structure is observed due to a higher level of abstraction that synthesizes short and long term attributes. All the elements of the point cloud seem to form a coherent manifold that

represents the normal behavior of the data. This manifold takes the form of a kind of continuous coiled spiral which is highly structured by temporal structure as can be seen in the color plot on Figure 3.15.

Figure 3.16 depicts the same manifold with another coloration linked to a basic anomaly score (based on the XGB-ST residue), showing the anomaly distribution in the latent space. In the center of the image, a set of elements far from the 'manifold' have a high anomaly score. After posterior analysis, these surrounded elements (which belong to the test set) could be linked to the the sports event which seems to have had an impact on the train load. The network seems to have capture a trace of an abnormal impact through the inference of the dynamics of short term attributes.

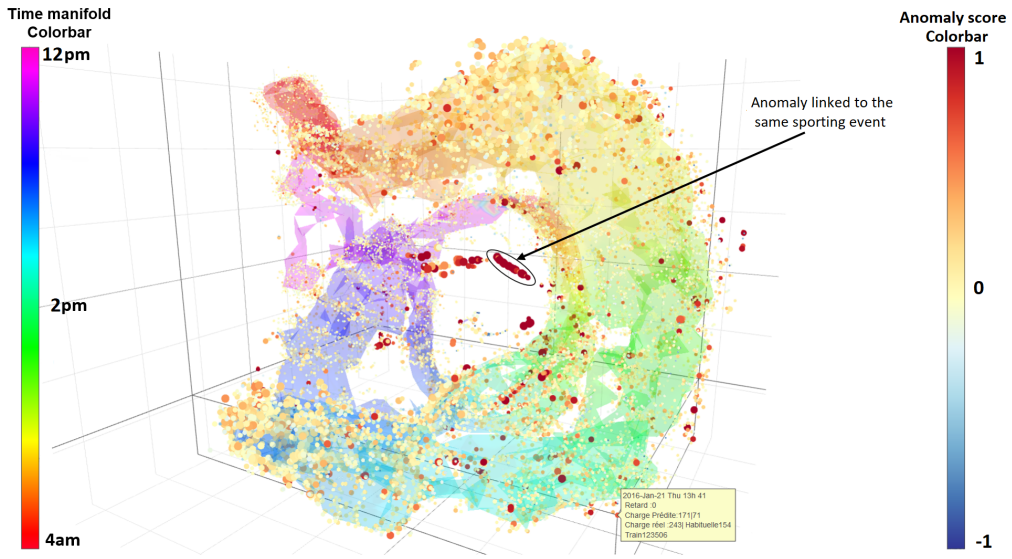

Figure 3.16.: 3D reduction of Predictive latent space (z_t) of suburban station with anomaly score coloration

These examples illustrate the ability of neural networks to construct rich abstract representations that synthesize information from the input attributes. A perspective for future work would be to explore how to use the manifold as a 'normality reference' in order to identify atypical elements that deviate from it.

In this chapter, we have formalized the concept of time series prediction based on structured data. The objective was to capture the cross influences of a set of known and latent factors on a time series. This capture is carried out through contextual attributes that synthesize underlying structures (temporal, seasonal, linked to Heterogeneous properties), and short-term attributes that support the inference of short-term dynamics. The application of this formalism requires understanding and analysis of the data to extract and refine relevant attributes able to synthesize the underlying structures.

In addition to machine learning-based forecasting models, we have proposed a deep learning model called LSTM-EP designed to take full advantage of the sequential data with underlying structure. It is based on a recurrent encoder-decoder architecture combined with a learning representation of contextual factors. This network aims at learning a contextual representation from contextual characteristics, capturing latent past dynamics through recurrent layers, and synthesizing both contextual and dynamic influences through the encoder-decoder architecture. The synthesized information is then used to predict future dynamics using the predictive layer and thus achieve a prediction of future elements.

We have applied this formalism on train load forecasting on both single and multi-step time horizons. The single step aims to predict the load for the next train at the station, whereas the goal of multi-step forecasting is to predict ahead of time the load for the forthcoming train passages. The forecasting problem is particularly challenging due to the high variability in the time series and the irregular time steps of the series to be predicted. This requires building contextual attributes that provide synthetic representations of temporal, calendar, and transportation plan influences. Moreover, part of the latent influences can also be inferred through the analysis of the short term dynamics of the last historical values.

This chapter has illustrated the performance of standard machine learning models (Random forest or Tree Boosting), and advanced neural networks (LSTM and LSTM-EP) to the sequence of the future train loads from processed attributes providing long and short term information. The results show on the one hand the importance of extracting and refining contextual attributes to make good predictions using standard learning machine models. It also shows the potential of the LSTM encoder-predictor to address short-term prediction on sequential data with an underlying structure. We evaluated the performance of the proposed model on two real datasets related to suburban and inner-city stations for single and multi-step forecasting horizons of the train load. On both configurations, the LSTM-EP outperforms the LSTM, XGB and baseline models by maintaining robustness in the quality of the forecasts throughout the time horizon.

Finally, the exploration of the latent spaces of the advanced LSTM-EP model opens up many perspectives. To realize its prediction, the LSTM-EP model learn a data representation corresponding to the synthesis of the information provided in the network inputs. The analysis of this representation can be used to extract synthetic feedback on the nature and behavior of the data. Future research should explore of the learned representation. In particular, it would be interesting to investigate the ability of the predictive latent space to characterize abnormal situations, such as disturbances and traffic anomalies.

4 Anomaly detection in a dynamic context

4.1 Introduction

This chapter addresses the issue of anomaly detection in multivariate time series evolving in a dynamic context. The applicative motivation behind this academic research is the evaluation of the impact of disturbances due to events or incidents in transit station ridership. The main specificity lies in the inherent variability of the series considered whose dynamic context is structured by a set of influencing factors (calendar and spatial in particular) which makes the modeling of "normal" ridership difficult. The proposed method is based on residuals calculated from the prediction model developed in the previous chapter, which are used to estimate the contextual bias and variances.These two quantities are then used to calculate a test statistic that is robust to contextual variations. Two questions have to be addressed: How to characterize and model the dynamic context? How to quantify statistical anomalies in time series that consider the dynamic context?

This chapter is structured as follows: Section 1 introduces the application framework and issues associated with anomaly detection. Section 2 reviews the state of the art techniques in anomaly detection applied to multivariate time series in dynamic contexts. Then, in section 3, we formalize the problem of detecting contextual anomalies in time series in the presence of dynamic contexts. In this framework, we detail the methodology and the bias-variances estimations proposed to tackle this problem. Section 4 is dedicated to the evaluation of the proposed methodology on a synthetic data set. Section 5 shows the application of the approach on real-world data related to human mobility in the transit network of the city of Montreal thanks to the availability of a database of incidents and events provided by the transport operator. Lastly, section 6 concludes the chapter.

4.1.1 Application context and objectives

Numerous punctual hazards punctuate the operation of transportation networks, disrupting the demand (events) and the transportation supply (incidents). Managing the network in a disrupted situation is an extremely complex challenge that often consists in blindly navigating through a blurred situation trying to make the least bad decisions.

Analyzing and understanding the impact of a hazard on the transportation network would provide valuable information that could facilitate crisis management. The anomaly detection field offers data analysis tools providing valuable information for crisis management.

- To extract a posteriori rich knowledge in order to characterize the impact of the various hazards already observed.
- To calculate current and forecast indicators in order to estimate the evolution of the state of the transportation network in a disrupted situation.
- To detect abnormal signals to anticipate a critical situation.

From a regulation point of view, the analysis of past disturbance situations will enable us to extract knowledge that will allow us to better analyze future situations and predict more appropriate regulation scenarios. The extraction of indicators and abnormal signals can make it possible to better estimate the state of the transportation network, inferring evolution due to a hazard, and anticipate the emergence of critical situations. This will then facilitate the regulation decisions by offering a clearer view of their consequences and the state of the situation.

The application area concerns human mobility data and, in particular, smart card data about the ridership in transit networks. We have to deal with data aggregated in the form of a multivariate time series evolving in a dynamic context. This dynamic context results from the continuous interaction over time of a set of latent or observed influential factors that can be calendar (hour, day, season,...) or spatial, related to the geographical properties of the network (incoming /outgoing transit population, employment area, land use....). Therefore, we consider a theoretical framework that analyzes the series with additional information on some 'observed' influential factors through contextual attributes. Within this framework, it is possible to capture the main time series dynamics linked to observed influences through contextual attributes and recent history. However, as shown in Figure 4.1, complex variability remains, which is produced by the cross-interactions of hidden influential factors.

The work aims to capture both the dynamics of the series and a significant part of the variability in the contextual attributes by approximating the statistical means and variances related to the dynamic context. Such 'contextual' means and variances are valuable information that allows a better perception of anomalies occurring among the standard data noises. It can be used to measure the deviation from the normal behavior related to a specific context to detect elements that have 'contextual' abnormal values, i.e., statistical extrema. Formalized in this way, two main questions have to be addressed, which are as follows: How should the dynamic context be characterized and captured? How can the statistical abnormalities in the time series data be quantified by considering the dynamic context?

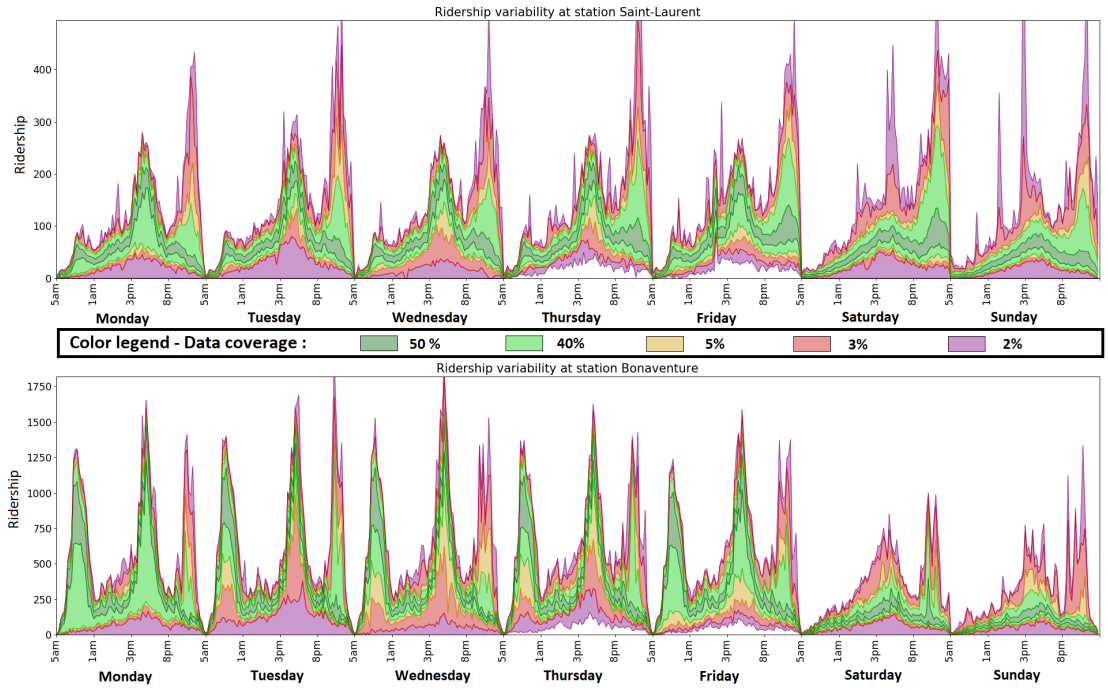

Figure 4.1.: Two examples of ridership variability at two metro stations
Weekly time series from three years of smart card data of Montreal City metro (2015-2017)

Whereas most studies perform agnostic anomaly detection on a series, our work concerns contextual anomaly detection on a series with a dynamic context using additional information in the form of contextual attributes. Based on our previous research on short-term prediction models of train load series [Pas+19b] detailed in Chapter 3, the proposed approach is positioned within the paradigm of contextual anomaly detection based on prediction residuals.

4.1.2 Anomaly detection issues

Anomaly detection consists in distinguishing some atypical elements within a dataset by their non-standard behaviors. The anomaly detection field was partly structured by a Chandola's review of the literature [CBK09] which summarized the issues and problems of anomaly detection at the end of the 2000s. This issue concerns many application domains including mobility [ZL17], cybersecurity [YLC17], industry [Hun+18], medicine [Sch+17], and fraud [AMZ16], among many others.

By definition, anomalies are rare and poorly known which means that little information is available about them. The main issue is often related to the definition of the 'norm' which appears to be a delicate question on complex data (Structured, non-stationary, with high variability or many contextual influences). Another problem is related to the high variability of anomaly behaviors which may have various causes. Even in some applications, malicious intention can try to hide anomalies or make them evolve. Another

critical issue concerns the confidence of the detection system and the management of borderline cases.

Anomaly detection also depends on the nature of the data, especially on the different structures that organize elements. The data can be represented through combinations of several data structures: mixed attribute list, sequence of tokens or numerical values, Spatial Matrix, Interaction Graph, ... The construction of the most relevant representation will enable the maximum amount of information to be extracted and synthesized. It therefore largely conditions the viability of anomaly detection by facilitating the capture of normal behavior or the expression of atypical characteristics.

Likewise, there are several types of anomalies (see Figure 4.2), each of which is related to different angles of attack. The nature of the anomaly will depend on the application.

- **Punctual Anomalies:** Elements that can be distinguished from the data set directly by these attributes.
- **Contextual Anomalies:** Elements that can be distinguished from a neighborhood (Temporal, Spatial, Contextual or other) by these attributes. The definition of the relevant neighborhood can be a complex task in the presence of multiple influences.
- **Collective Anomalies:** Groups of elements whose interactions form an irregular pattern with respect to the norm. The definition of a group of elements within structured data (Sequences, Matrices, Graphs) can be a complex task.

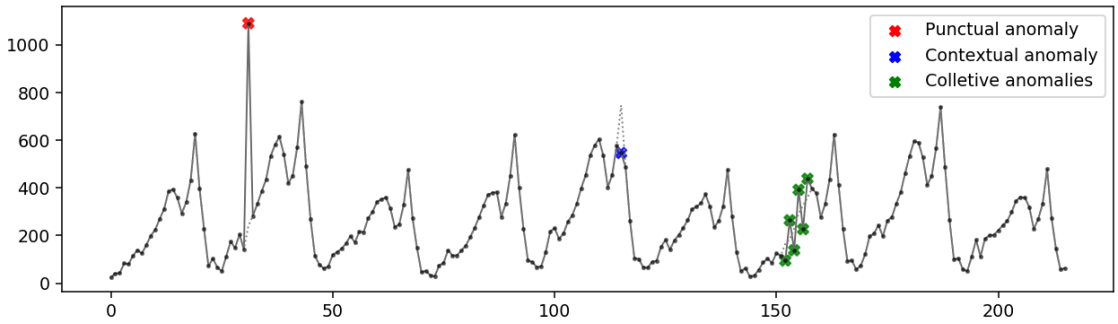

Figure 4.2.: Example of the different types of anomalies

Finally, there are several anomaly detection frameworks that will depend on the information available and more specifically on the presence of a label identifying the anomalies.

- **The supervised framework** where anomaly markers are available. The objective is then to learn the specificities of the abnormal elements in order to be able to recognize them. The task is then very similar to that of an unbalanced classification problem.
- **The semi-supervised framework** where only part of the anomaly markers are available (either a small proportion, or a data set without anomalies). As a de-

graded variant of the supervised framework, this approach often tries to adapt the classification techniques to learn the normal behaviour and then discriminate atypical elements (One-class classification).

- **The unsupervised framework** where no information on anomalies is available. The objective is then to build types of anomaly scores capable of discriminating atypical elements by comparing the different elements.

The domain of anomaly detection is extremely vast. Several anomaly detection approaches have been developed for a large variety of application domains. The availability of qualified and labeled databases that guarantee robust detection is often lacking in several application areas, and performing this task is tedious and costly. An unsupervised learning framework is often considered. Moreover, depending on the nature of the data available, the anomaly detection task can present many specific methodological challenges. Our work focuses on a specific question: **The importance of considering the contextual variance in contextual anomaly detection on time series structured data with an unsupervised detection framework.**

Our main investigation focuses on the difficult task of capturing the influence of a dynamic context on a series through descriptive contextual attributes. The influences of several of the factors that make up the dynamic context are complex. Indeed, they are non-linear, dynamic (changing over time), intermix (interacting with each other), and only partially observed through contextual attributes. Our contribution involves combining a short term forecasting model based on contextual attributes and short term dynamics with a contextual variance estimation also based on contextual attributes. The goal is to approximate the influence of the underlying dynamic context on the series through time-varying statistical measures, such as 'contextual' means and variances. These statistics are exploited in the form of prediction residuals normalized by a contextual variance to form a context-normalized anomaly score. Different forecasting and context-estimation methods are conducted and detailed in the proposed approach. The estimation methods can be basic methods or based on advanced methods, such as machine learning techniques for time-series prediction models.

The issue addressed is assimilable to a statistical sampling applied to multivariate time series elements. In our work, we aim to build 'continuous' homogeneous contextual subsamples based on contextual attributes to extract normal behaviors and the statistical dispersion of the time series. It is from these characteristics that it is possible to construct a contextually normalized anomaly score.

The main contributions can be summarized as follows:

- An anomaly detection formalism for a multivariate time series in dynamic contexts with contextual attributes is developed.
- Contextual and variability estimations of the time series data are performed.

- Computation of robust contextual anomaly scores based on the normalization of the forecasting residuals by the dynamic contextual estimation (bias and variance) is performed.
- The methodology is applied to both synthetic and real data. The Montreal transportation authority has provided us with three years of data on station ridership, as well as information about events and incidents. This allows us to evaluate and highlight our anomaly scores regarding these events and incidents.

4.2 Time series anomaly detection literature

This section addresses unsupervised contextual anomaly detection in a multivariate time series. Contextual anomaly detection is a complex issue due to the definition of normality, the sparsity and non-redundancy of abnormal observations, the lack of data labeling, and the consideration of dynamics that structure data evolution. Moreover, the choice of an anomaly detection approach is often application-oriented. The anomaly detection field was summarized at the end of the 2000s by Chandola in his thesis research (([CBK09], [Cha09]). More recent surveys are dedicated to more specific subjects with, for instance, the analysis of work on real-time Big Data [Hab+19] or different approaches based on a deep learning framework [CC19] and applied to road transportation networks. Other more complex objectives are emerging, notably related to the interpretability of detection systems [Cho+16] and the characterization and impact of anomaly forecasting [Cao+17].

4.2.1 Paradigms of anomaly detection in time series

There are three main paradigms to perform anomaly detection:

- Proximity based clustering methods consist in explicitly exploiting data topology by forming groups of similar elements that express normal behaviours.
- Distance-based isolation methods consist in implicitly exploiting data topology through neighbor distance considerations in order to identify atypical elements which are far from their closest neighbors.
- Prediction residuals based methods consist in explicitly exploiting data topology by analysing residuals of a prediction / reconstruction model. Here, the aim is to learn a simplification of the generative model of the data, and to detect observations that deviate from the learned model.

The majority of anomaly detection approaches are based on one of the three paradigms and use diverse mathematical tools, distance choices and data representations in order to perform the same essential steps:

1. build a rich representation of the studied objects, integrating contextual, temporal and spatial information.
2. identify nominal behaviors in the data that may fluctuate in time and space (Explicitly or implicitly).
3. Evaluate for each element, the distance from the most suitable reference behavior.
4. Compute an anomaly score to discriminate normal and abnormal elements on the basis of an estimated boundary threshold.

Table 4.1 lists various studies on anomaly detection.

Table 4.1.: Synthesis of some Anomaly detection studies

Works	Data	Model	Representation
Proximity clustering approach			
[HC14]	Sensor data	Parallel K-mean Clustering	Gaussian synthesis
[Bri+19]	Subway ridership	constrained hierarchical clustering	ridership profile
[He+19]	Mobility flow	Gaussian mixture clustering	Mobility Embedding
[Liu+17]	Image series	EM-clustering	Pixel group
[BBC18]	Sensor data	Spectral clustering + DTW	Window features
[LBE17]	Traffic data	Latent dirichlet allocation clustering	Window features
[Wit+13]	Mobile data	Kmean + Hidden Markow Model	Sequences profile
Distance isolation approach			
[Che+09]	Satellite survey	Neighborhood Graph + RandomWalk	Sub-sequences
[DF13]	Streaming data	Isolation forest on Window features	Window features
[Yeh+16]	Text & Various data	Matrix Profile	Sub-sequences
[YKR08]	Various data	Discord pattern detection	Sub-sequences
[Nak+20]	Various data	Discord pattern detection	Sub-sequences
[Fer+19]	Human activity	Pattern mining + Isolation forest	Mixed-type series
[Smo+20]	Mobility data	Infinite Gaussian mixture GAN	Sub-sequences
Forecasting model approach			
[RT+14]	Road Trafic speed	Exponential moving average	Sub sequences
[Ton+18]	Subway ridership	Non-Negative Factorisation (NMF)	Factorisation
[KTA14]	Traffic data	Latent Dirichlet allocation (LDA)	Window features
[Mal+16]	Various sensor	LSTM encoder decoder	Series batch
[Guo+18]	Various data	GRU variational autoencoder	Series batch
[Mun+18]	Various data	CNN Forecast model	Series batch
[Hun+18]	Spatial sensor	LSTM Forecast model	Series batch

Proximity-based clustering methods

Proximity-based clustering methods aim to extract several clusters that synthesize the behaviors of all the elements of the series. Each cluster contains relatively homogeneous elements. It is then possible to calculate the degree of abnormality of an element through the distance from these cluster neighbors.

In [HC14], the authors computed the difference from average contextual profiles on massive data. These profiles were obtained by a parallelizable 'K-means' clustering, which was designed to handle massive data.

The authors of [Liu+17] tackled the task of anomaly detection on an image series through EM-based clustering. Clustering was performed on pixel group representations based on

the combination of pixel group information and its difference from spatial or temporal neighborhoods. Clustering was then refined by merging similar clusters and excluding outliers.

The authors of [BBC18] proposed an approach based on iterative optimization that performed spectral clustering using the dynamic time warping (DTW) distance with an actualized weight to estimate the contribution of anomalous observations. It was an iterative optimization procedure based on the contribution of the series to the 'order' set. They proposed different weightings based on entropy measurement, a ridge penalty or a local approach.

[LBE17] also proposed a probabilistic latent dirichlet allocation (LDA) model based on a reduced centered speed categorization of road portions from the AVL data. The role of the probabilistic model was to group similar behaviors to extract a measure of uncertainty called perplexity, close to entropy, which was used to capture abnormal behaviors using a dynamic threshold based on the median.

[Wit+13] performed an event detection by extract sequence profiles from time series of mobile activity densities by a discrete quantification using K-means clustering. These sequences were then manipulated by a hidden Markov model in charge of detecting abnormal behaviors signalling an unusual event.

In the context of mobility, [Bri+19] proposed an atypical event detection from public transport ridership. The work proposed to extract clusters based on daily ridership profiles from a constrained hierarchical clustering according to calendar contexts. It then constructed an anomaly score based on the difference between the measured ridership and a reference profile obtained through the mean to the most relevant cluster. This difference was then normalized by the interquartile variance of the same cluster to obtain an anomaly score that considered contextual variability.

The work by [He+19] also dealt with anomaly detection in mobility flows. They proposed to construct a representation of the state of flow from an aggregation of flow graphs combined with a dimension reduction method. Then, they applied a probabilistic clustering based on a Gaussian mixture model associated with a statistical test to detect anomalies.

Distance-based isolation methods

Distance-based isolation methods are based on a similarity measure of an element or a set of elements with others. In contrast to the previous paradigm, these methods aim to isolate the atypical elements that are often presented as regular sub-sequences on the basis of a similarity measure of an element or a set of elements with others. There are many forms of similarities using various mathematical tools.

For example, [Che+09] computed a similarity graph between the sub-sequences of the multivariate series with a Radial Basis Function (RBF) kernel. Then, they obtained a similarity score by using a random walk on the neighborhood graph. This similarity score was used to isolate atypical sub-sequences.

Isolation forest

Some methods propose applying the isolation forest concept [LTZ08; LTZ12] to discriminate elements based on their features by storing them in binary trees with an entropy measure. Isolation forest consists in building a binary tree assigning each observation of a data-set to a dedicated leaf, according to binary constraints on its attributes. Abnormal observations with atypical information (high entropy) are easy to isolate and are placed close to the root. Normal observations that require more constraints to be isolated, are then associated to the deep leaves. Thus, an anomaly score can be defined for each element. It is based on the depth of the assigned leaf which directly informs us on the similarity of an observation with the data set.

In [DF13], Isolation forest was used to perform anomaly detection on sub-sequences with a vector representation. The authors also proposed to add a mechanism to further detect distribution changes in the data based on monitoring the percentage of anomaly detection over a short time horizon. If this percentage is too high, model actualisation is performed by learning a new Isolation forest on recent data.

The authors of [Fer+19] proposed an anomaly detection approach that considered contextual elements. The approach proceeded in two steps: first, a catalog of frequent patterns is created. Each sub-sequence was thus characterized by its contextual elements and a weighted combination of the catalog elements. An Isolation forest was performed on these representations to detect abnormal sub-sequences by considering contextual information.

Nearest neighbor distance

It is also possible to isolate elements by analyzing their neighbor proximity using a relevant distance. Work has focused on comparing subsequences according to their nearest neighbor distances to detect atypical elements. The choice of distance is crucial. Different types of distance are used: Euclidean, Z-normalised Euclidean or dynamic time warping (DTW) Distance.

The authors of [Dim+17] proposed a clustering based on the k-nearest neighbor (KNN) approach using the dynamic time warping (DTW) distance between the subsequences of the time series. These sets were filtered to remove 'outliers' and then used as a reference. The anomaly score of an element then corresponded to its average distance from the most relevant set of filtered elements of the cluster.

Matrix profile approaches directly exploit distances between the sub-sequences of a time series to perform pattern mining and anomaly detection. These approaches are designed to detect agnostic anomalies through pattern mining on massive time series data without additional information. The Matrix profiles representation is a synthesis of the calculations of distances between sub-sequences. It provides a 'normality' score for each element of a series based on its nearest neighbor proximity.

Some work, such as time series matrix profiles, concentrated on the exhaustive calculation of distances. These approaches require solving an issue related to the high cost of calculating a large number of distances n^2 on a large dataset. Several strategies have been designed to reduce the computational cost, such as [Yeh+16] who proposed the rapid computation of an exact Matrix profile using Fourier transformation. They also proposed an algorithm 'Scalable Time series Anytime Matrix Profile Incrementally' able to approximate in a incremental way and distance parallelized computation.

Other approaches avoid computing all the sub-sequence distances. For example, using a discord sequence approach, [YKR08] focused rather on the detection of atypical candidate sub-sequences called 'discords' using a threshold on the nearest neighbor distance. The detection was performed over the series by identifying 'discord candidates' which were updated by browsing the series allowing detection at a low computational cost. Recent work sought to deepen various aspects of the approach, for example by automating the threshold and the choice of the size of the sub-sequences considered, which proved to be a critical parameter [Nak+20].

Generative adversarial networks

The GAN anomaly detection approach consists in projecting the data into a latent Z space in order to capture an approximation of the distribution of the normal data, often in the form of a multivariate Z law, in a large dimensional space. Detected Anomalies moving away from the center of the Z distribution, or have large reconstruction errors.

[Zen+18] was one of the first applications of GAN to anomaly detection. The anomaly score was a combination of the reconstruction error and the discriminator loss that measures how well the element follows the Z law. In more recent work, [Smo+20] proposed to apply a bi-GAN that approximates an infinite Gaussian mixture to perform anomaly detection on mobility data. Then, the authors used a Mahalanobis distance on the Z space to detect anomalies.

Methods based on the residuals of the forecast

The last paradigm is explicitly based on the prediction/reconstruction of the series. A model aims to capture 'normal behavior' by learning an approximation of the data's generative function. Such a model is assimilated to a bottleneck (encoder-decoder)

approach, which captures the maximum 'normal' information from the data to extract the normal behavior of the series. Forecasting the model's residuals is then used to detect observations that deviate from the 'normal' model. We detail below the paradigm standards upon which our approach is based.

The underlying idea consists in using the residuals of a prediction or reconstruction model to highlight the anomalies occurring in a time series. Forecasting models capture frequently recurring patterns and overlook less predictable phenomena, such as anomalies. A significant residual can be considered a sign of a deviation from normal behavior.

The related studies in this domain have applied several types of predictive models, such as the autoregressive integrated moving average (ARIMA) model [AP14], parametric models [RT+14], probabilistic models such as hidden Markov models (HMMs)[LPJ17], multiple linear regression models [Sal+14], machine learning models, such as support vector machines (SVMs) [KK13], random forests [Has+14], or the Matrix-decomposition reconstruction technique [KKK16; Ton+18].

[RT+14] proposed to perform traffic accident analysis using real time speed measurement thanks to a forecasting model based on a parametric exponential moving average. The differences in the speed properties of the portions of upstream and downstream roads were used as anomaly markers thanks to a parametric marker.

The authors of [Ton+18] performed anomaly detection on ticketing data of subway stations using spectral theory. They built hour and daily reference profiles using a 'non-negative matrix factor' decomposition allowing it to decompose the trends into atomic components, thus providing a synthetic reference profile. The anomaly score was given by the difference between the current state and the synthesis.

[KTA14] proposed a probabilistic model based on Bayesian estimators using the average speed of road portions. It was a probabilistic LDA (latent dirichlet allocation) model based on a hidden traffic state (fluid, congested, saturated) that analyzes the difference between the short and long term estimation to detect anomalies.

Recent research has investigated recurrent neural networks by exploiting predictive residuals to detect anomalies in time series. The authors in [Mal+15] were the first to use predictive residuals from recurrent models for anomaly detection. The residuals were assumed to follow a Gaussian distribution with a non-zero mean. Anomaly detection was performed by using a threshold on the anomaly score based on the likelihood that these residuals will occur.

More recently, the authors extended their proposal by using an LSTM encoder-decoder as a predictive model [Mal+16]. Several datasets (ECGs, industrial sensors, and spatial

sensors) were used to evaluate the proposed multivariate anomaly detection methodology, and the detection threshold was optimized using a sample of anomalies.

Moreover, in [Mun+18], the authors proposed applying a convolutional neural network (CNN) for prediction. These models required fewer data while having a better generalization capacity than LSTM networks, which are more appropriate to capture the temporal dynamics of a time series.

The authors of [Guo+18] proposed a detection approach based on the residual reconstruction of a recurrent gated recurrent unit (GRU) network, such as a "Gaussian-based mixture" approximation. The architecture used was a GRU-based recurrent encoder/decoder structure associated with a variational layer playing the role of the Gaussian mixture. The reconstruction residuals of the model were then analyzed to detect anomalies.

In the same vein, the authors of [Hun+18] proposed using a classic approach based on residuals from an LSTM predictive model for anomaly detection in space probe sensors. The model achieved a univariate prediction by using attributes over a past time horizon. The anomaly detection was then based on smoothed residuals using an exponential filter with a dynamic threshold. This threshold was estimated from the mean and variance of the prediction errors over the past time horizon, which led to more robust detection against false positives.

Another aspect of anomaly detection is related to prediction models that are able to provide a confidence interval associated with the prediction. In this framework, a prediction interval containing a prediction with a confidence of X% is inferred. One can then consider the observation to be abnormal if it does not belong to this interval with the predefined confidence. The authors of [Mei06] proposed learning prediction quantiles based on the extraction of the distribution of predictions. In [Car19], the authors proposed performing learning on the prediction residuals using a second model (either linear or not) that captured the variance of the error to determine a confidence interval associated with each prediction.

In the context of neural networks, a comparison with Bayesian theories introduced within the variational paradigm, formalized by [KW14], offers relevant alternatives. Recently, the authors of [GG16] proposed approximating Bayesian behavior in a deterministic network by conserving the dropout in the prediction phase. The dropout induces a form of random draw by randomly forcing some neurons to have zero weight, which disturbs the prediction. Relying on this principle, [ZL17] proposed the use of LSTMs to obtain a prediction with a confidence interval. This technique exploited the variational dropout by performing several runs of predictions to provide a prediction interval.

4.2.2 Positioning and contribution

Mobility time series data in public transportation have several specificities, including the fact that they evolve in a dynamic context for which certain influential factors are known and observed [Toq+18]. One of the challenges lies in considering the contextual variability in the data for the detection and characterization of the anomaly's impact on a multivariate series.

Proximity-based approaches often use a discrete representation of the context that does not fit well with a complex dynamic context that evolves continuously. Approaches based on isolation by distance notions will have a strong impact on the relationship between normal data and anomalies and the topology of the anomalies. Finding a relevant distance that considers the dynamic context is a complex issue. Moreover, the slight redundancy (link to public transportation issues) can make the isolation task harder. Likewise, forecast residual approaches have some weaknesses handling dynamic contexts because of their extreme dependence on the quality of the forecast. Sadly, in a multivariate non-stationary time series, the prediction task can be a real challenge. The issues of prediction reliability and the existence of biases in the models are therefore critical. Furthermore, most of the work in the literature is not designed to use additional context information, which is valuable for capturing contextual variability, which can easily disrupt the anomaly detection process.

Our goal is to build a robust anomaly score to highlight statistical anomalies (contextual extrema) within the normal contextual variability. This anomaly score would enable us to better detect and characterize the anomalies in the observed multivariate time series by facilitating the analysis of their impact and refining the evaluation of their severity. Based on our work forecasting the passenger load for commuter trains [Pas+19b], we propose a methodology to compute an anomaly score that considers the dynamic context of multivariate time series. Following [Mal+15], we assume that the prediction residuals follow a Gaussian distribution $\mathcal{N}(B_t, \sigma_t)$. Our contribution lies in considering the contextual modeling of the mean B_t and the variance σ_t linked to the dynamic context. The anomaly score is therefore based on a contextual normalization of the forecasting residuals over the underlying context. It expresses, in a statistical sense, the deviation from normality.

We apply the proposed approach to synthetic and real data collected from the Quebec transportation network. The goal is to analyze the anomaly scores provided by our method and cross-reference this analysis with datasets of events and incidents provided by the transport operator. The main investigation that we carry out here concerns the following question: How can one qualify with finesse the deviation from the normal values of ridership in a disturbed situation?

4.3 Formalization of the proposed detection approach

In a heterogeneous dataset, estimations of contextual mean and variance are essential to identify "contextual normalities" which can allow us to qualify abnormal values with regard to the context. Our goal is then to infer the impact of the dynamic context on the time series (y_t) by obtaining better knowledge of the contextual mean (prediction task) and contextual variability of our data in order to build a robust context-normalized anomaly score. Given the formalism described in Section 3.3.1, the proposed detection strategy is based on two main steps:

Extraction of prediction residuals:
First, the forecasting steps aim to extract the dynamic contextual mean structured by the dynamic context by using a forecasting model. We will therefore use the different types of prediction models detailed in the 'Prediction' chapter to extract the dynamic mean from the signal and thus obtain residuals that can indicate the presence of anomalies.

Dynamic modeling of the residuals:
Second, dynamic residual modeling aims to consider the influence of the dynamic context on the variability and noise of the data. The estimation of the mean and variance of the prediction residuals can be used to refine the residuals to obtain a better anomaly score. It is this problem that is at the heart of the following work.

4.3.1 Multivariate time series structured by a dynamic context

The work follows the formalism described in Section 3.3.1. In this framework, we study a time series (multivariate) structured by a dynamic context. This dynamic context is the result of the mixing of a set of influential factors which are partly observed through a contextual state vector ($\mathbf{c}$) and partly latent ($\boldsymbol{\ell}$).

The time series (y_t) can be decomposed as a signal $\boldsymbol{M_t}$ and a noise $\boldsymbol{\epsilon_t}$ which are both structured by the dynamic context. Moreover, the signal $\boldsymbol{M_t}$ can be split into several components linked to specific sets of known and latent factors.

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{M}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{M}_t &= \boldsymbol{f}^c(\mathbf{c}_t) + \boldsymbol{f}^d(\mathbf{c}_t, \boldsymbol{y}_t^p) + \boldsymbol{f}^a(\mathbf{c}_t, \boldsymbol{y}_p, \boldsymbol{a}_t) \\ \boldsymbol{\epsilon}_t &\sim \mathcal{N}(B_t(\mathbf{c}_t, \boldsymbol{\ell}_t), \sigma_t(\mathbf{c}_t, \boldsymbol{\ell}_t)) \end{aligned} \tag{4.1}$$

- $\boldsymbol{y}_t^p = (\boldsymbol{y}_{t-1}, .., \boldsymbol{y}_{t-p})$ is the previous temporal horizon.
- $\boldsymbol{f}^c$ is the long-term contextual component linked to the known influential factors (contextual attributes).

- $\boldsymbol{f}^d$ is the short-term dynamic component resulting from the mixture between some of the known and latent factors. We want to infer this component through the short-term dynamics induced during the past temporal horizon.
- $\boldsymbol{f}^a$ is the abnormal component linked to anomalies that significantly impact the dynamics of the series over a short range. $\boldsymbol{a}_t$ is a characteristic series of anomalies that encodes the presence of anomalies at a time step t for each dimension.
- $\boldsymbol{\epsilon}_t$ is the unexplained variability in the components f^c, f^d, f^a of $\boldsymbol{M}_t$. This variability is structured by known and latent influential factors $(\boldsymbol{\ell}, \mathbf{c})$ and can be represented as noise with a dynamic mean $\boldsymbol{B}_t$ and variance $\boldsymbol{\sigma}_t$. The use of a non-zero mean $\boldsymbol{B}_t$ makes it possible to consider any bias in the prediction model $\boldsymbol{M}_t$.

Based on the work of the previous chapter, we will compare several prediction approaches (Contextual, Dynamic) based on several types of models (Random forest, Artificial Neural Networks) in order to construct the most suitable residuals for the construction of a robust anomaly score.

4.3.2 Prediction residual and anomaly score

Prediction residuals are given by the difference between predictions $\widehat{\boldsymbol{y}}_t$ and observations $\boldsymbol{y}_t$. These differences are due to errors in the contextual and dynamic impact capture, variability in the data, and noise. In the decomposition framework (Eq. 4.1), we can decompose the prediction residuals as follows :

$$\begin{aligned} \boldsymbol{r}_t &= \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t \\ &= (\boldsymbol{f}_t^c + \boldsymbol{f}_t^d + \boldsymbol{f}_t^a + \boldsymbol{\varepsilon}_t) - (\widehat{\boldsymbol{f}}_t^c + \widehat{\boldsymbol{f}}_t^d + \widehat{\boldsymbol{f}}_t^a) \\ &= \boldsymbol{e}_t^c + \boldsymbol{e}_t^d + \boldsymbol{e}_t^a + \boldsymbol{\varepsilon}_t \end{aligned} \tag{4.2}$$

- The error $\boldsymbol{e}^c$ is related to the capture of the contextual impact $\mathbf{c}$ (Bias).
- The error $\boldsymbol{e}^d$ is related to the capture of the nominal dynamics $\mathcal{D}$ (Bias).
- The error $\boldsymbol{e}^a$ is related to anomalies a (Anomalies).
- The noise $\boldsymbol{\varepsilon}$ is related to the unexplained variability in the data (Variance).

The usual anomaly detection approach using the absolute prediction residual implicitly assumes that the residue is independent of the context and overwrites the strongly unexplained variability (Equation 4.3), which means that:

$$||\boldsymbol{e}_{t_1}^a|| > ||\boldsymbol{e}_{t_2}^d|| + ||\boldsymbol{e}_{t_2}^c|| + ||\boldsymbol{\varepsilon}_{t_2}||. \tag{4.3}$$

Both assertions are questionable in a series structured by a dynamic context in which the variability and errors are context-dependent. A contextual normalization of these

residuals can provide a robust anomaly score s_t that allows us to disentangle anomalies ($\boldsymbol{e}_t^{\mathrm{a}}$) from normal variance $\mathcal{E}_t$.

The proposed treatments are applied dimension by dimension $(*)$ to simplify the application (co-variance will be used later), but using a multivariate method can also be relevant depending on the application. The treatment begins by reducing the bias related to errors induced by the context and nominal dynamics ($e^c + e^d$) and then by estimating the variance related to the unexplained variability $\mathcal{E}$. This contextual anomaly score s_t is intended to statistically evaluate abnormalities in a series related to the dynamic context. We propose estimating the contextual bias $\widehat{B}_t \approx e_t^c + e_t^d$ and contextual standard deviation $\widehat{\sigma}_t = \sqrt{\mathcal{E}_t}$ with learning algorithms.

For each dimension of the time series, we define the anomaly score s_t as follows (Equation 4.4):

$$s_t = \frac{r_t - \widehat{B}_t}{2\widehat{\sigma}_t} = \frac{(e_t^c + e_t^d - \widehat{B}_t) + e_t^a + \varepsilon_t}{2\widehat{\sigma}_t} = \frac{e_t^{\widehat{B}} + e_t^a + \varepsilon_t}{2\widehat{\sigma}_t}\ (*) \tag{4.4}$$

* Applied dimension by dimension.

For each element y_t of the time series (y_t), the anomaly score quantifies what percentage of the extrema the value belongs to in relation to its context. It is also possible to quantify the probability of detection, which depends on the ratio between the magnitude of centered error and the contextual variance under certain Gaussian residual assumptions.

Contextual anomaly score

The multivariate residual series $\boldsymbol{r}_t$ (Eq. (4.2)) expresses the difference between $\boldsymbol{y}_t$ and the nominal prediction $\widehat{\boldsymbol{y}}_t$ for each time step t in each dimension d. The proposed approach to the anomaly score is based on these residues normalized by an estimation of the contextual variance. As it performs anomaly detection on multivariate series, the approach provides both **local** and **global** scores. The **local score** provides the context-normalized anomaly score for each dimension of the series. The **global score** synthesizes the local scores from all dimensions through the Mahalanobis distance.

The following process is used to build the score:

Step 1. Contextual mean estimation $F^m(\boldsymbol{c}, \boldsymbol{y}_P) = \widehat{\boldsymbol{y}}$

The process starts by estimating the contextual mean due to contextual attributes and short-term features. It is a multivariate forecasting task that can be performed by several models. From experience, we use the following four main types of models: the categorical mean, long and short-term random forests and the long short-term memory encoder predictor (LSTM-EP) model.

Algorithm 1 Contextual anomaly score

$Score_{anom}(y \in \mathbb{R}^{T\times D}, \boldsymbol{c}, \beta = 95, \beta_{agg} = 95, q = 1)$

Step 1. Contextual mean estimation: $\begin{cases} F^m(\boldsymbol{c_t}, \boldsymbol{y_{P_t}}) = \widehat{\boldsymbol{y_t}} \in \mathbb{R}^{T\times D} \\ \boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{y}} \in \mathbb{R}^{T\times D} \end{cases}$

Step 2. Bias-variance estimation: $\begin{cases} F^B(\boldsymbol{c_t}) = \widehat{\boldsymbol{B_t}} \in \mathbb{R}^{T\times D} \\ F^\sigma(\boldsymbol{c_t}) = \widehat{\boldsymbol{\sigma_t}} \in \mathcal{R}^{T\times D} \end{cases}$

Step 3. Reduction bias variance by dimension: $\boldsymbol{s} = [\boldsymbol{s}^1, \ldots, \boldsymbol{s}^D]$ with $\boldsymbol{s}^d = \frac{\boldsymbol{r}^d - \widehat{\boldsymbol{B}}^d}{(\widehat{\boldsymbol{\sigma}}^d)^q}$

Step 4. Spatial aggregation (Mahalanobis distance): $\boldsymbol{s}_{agg} = D_M(\boldsymbol{s})$

Step 5. Threshold normalization $\boldsymbol{s} = Norm_\beta(\boldsymbol{s})$ and $\boldsymbol{s}_{agg} = Norm_{\beta_{agg}}(\boldsymbol{s}_{agg})$ with $Norm_\beta$ a threshold normalization linked to β

Output: $\boldsymbol{s} \in \mathcal{R}^{T\times D}, \boldsymbol{s}_{agg} \in \mathcal{R}^T$

Step 2. Bias and variance estimators $F^B(r) = \widehat{\boldsymbol{B}}$ *and* $F^\sigma(\boldsymbol{r}) = \widehat{\boldsymbol{\sigma}}$

Then, an estimation of the contextual bias $\widehat{\boldsymbol{B}_t}$ and variance $\widehat{\boldsymbol{\sigma}}_t$ is performed with the contextual attributes **c**. It can be performed in univariate or multivariate ways. These estimators give us the contextual bias and variance series with values for each dimension d at each observation t.

We propose three ways to perform the bias-variance estimation, whose details are given in the Section 4.3.3:

- By naive methods that provide the classic scores:
  - **AE:** the absolute error score $\widehat{B} = 0$* and $\widehat{\sigma} = 1$.
  - **RE:** the relative error score $\widehat{B} = 0$* and $\widehat{\sigma} = y$.

(* We could take a bias equal to the average of the errors but with machine learning forecasting models this average is already assumed to be almost zero by the minimization of the loss function.)

- Learning from the residual with a bias-variance estimation model.
  - **EMP (empirical estimation):** By using prior knowledge to build homogeneous samples from our data, the bias and variance of each "local" subsample can be extracted. Thus by associating to each element the value of bias and variance of the sub-sample to which it belongs, we obtain our contextual bias and variance estimations according to the prior sampling.
  - **ML (machine learning estimation:)** A combination of two ML models using contextual attributes can replace the prior sampling to perform bias-variance estimation from the forecasting residue. First, a “bias-estimation” model aims to predict the forecasting residue from the context. Then, a second “variance-estimation” model learns to predict the square of the centered residue (the residue minus the previously estimated bias). In practice, we use random forest regressors that explicitly perform a type of sampling like the ML model.

- Extracting a type of variance directly from the forecasting models.
  - **RF (Random forest extraction):** For the random forest model, the authors of [Mei06] showed that we often exploit the valuable information about the distribution learned from the random forest. We consider only the mean of the 'subsample'. Based on this assumption, we propose to extract the variance based on the learned subsample of our random forest forecasting model.
  - **DEEP (Deep neural network extraction):** For a deep neural network, the variational dropout [GG16] can be used to approximate a Bayesian behavior on a deterministic network by adding random draws. Like the authors of [ZL17], we propose to use the variational dropout to estimate the variance of our LSTM-EP model by virtual sampling. Instead of performing explicit sampling, it is simulated using the random draws of the variational paradigm.

Step 3. Bias and variance reduction $s = \frac{r - \widehat{B}}{\widehat{\sigma}^q}$

The contextual bias is removed, and the contextual variance is reduced from the residual series r dimension by dimension to stretch the residue to a standard normal distribution. The context-normalized scores qualify the contextual abnormality of each time step t, ensuring that a high score is linked to an extremum data according to the context. We introduce a hyperparameter q, which allows us to modulate the importance of variance normalization. A low value of q ($q < 1$) makes the score more sensitive to a context with a high variance, whereas a high value of q ($q > 1$) makes the score more sensitive to a context with a low variance.

Step 4. Spatial aggregation $\boldsymbol{s}_{agg} = \sqrt{(\boldsymbol{s} - \bar{\boldsymbol{s}})^\intercal \, \Sigma_{\boldsymbol{s}}^{-1} \, (\boldsymbol{s} - \bar{\boldsymbol{s}})}$

Spatial aggregation aims to synthesize across dimensions to obtain a single synthesis score. The Mahalanobis distance will better take into account the variances and covariances of the different spatial dimensions. The use of one-dimensional statistics is not always appropriate because part of the information related to the phenomena of cross interactions between dimensions remains inaccessible. Working with multivariate statistics allows us to better consider atypical co-occurrences in multivariate observations.

Step 5. Formatting normalization

To perform anomaly detection, a threshold normalization is performed to discriminate the detected anomaly ($s > 1$) according to a prior anomaly ratio assumption. This normalization can be performed independently or at the same time on each dimension, depending on the hypothesis of a homogeneous anomaly distribution across dimensions. It corresponds to a division by a constant that is chosen either in an explicit way (by percentiles of the real score) or based on implicit criteria, such as $N * \sigma$ or entropy criteria.

Several mathematical processes can also be applied to format the anomaly score. For example, temporal convolution induces a time-local consideration in the anomaly score linked to the convolution filter. A squared score can enhance the dispersion of anomaly

scores and facilitate visualization. In Table 4.2, the normalization types used to construct the anomaly scores are listed.

Table 4.2.: List of different types of normalization

Normalization	Bias $\widehat{B}$	Variance $\widehat{\sigma}$
Absolute error (N-AE)	0	1
Relative error (N-RE)	0	y
Empirical variance (N-EMP)	B_{emp}	σ_{emp}
Variance ML (N-ML)	B_{ml}	σ_{ml}
Variance RF* (N-RF)	0	σ_{rf}
Variance DEEP* (N-DP)	0	σ_{deep}
Variance EXACT** (N-EX)	0	σ

*The N-RF/N-DP variance extractions are restricted to the RF and DEEP forecasting models

**N-EX is an artificial model that gives the variance used during data generation.

4.3.3 Bias-variance estimation approaches

The prediction and bias-variance estimation tasks are equivalent to the mean and variance estimation on well-formed contextual subsampling. Machine learning models can be useful for building such subsamples by considering the relationship between the contextual attributes and the predicted or residual values. The extraction of the contextual means, bias, and variances of the data are essential to define the contextual "normality" by considering the contextual variability through the model's confidence. It is necessary to build a robust contextual anomaly score that allows us to quantify and detect the statistical contextual anomalies present in our transportation time series data.

First, we propose two ways to learn and estimate the bias-variance based on the prediction residues produced by the forecasting models.

1. EMP: Empirical estimation on a prior sampling

The estimation model is based on prior knowledge. We segment the contextual attribute space **c** into prior subspaces (subsampling) defined by a set of constraints ($\boldsymbol{V}^{inf}, \boldsymbol{V}^{sup}$) given by expert knowledge. The bias $\widehat{B}$ and variance $\widehat{\sigma}$ estimators are summarized in three steps, as follows:

1. Extract from each prior E_k the subsampling bias and variance.
2. Associate each time step t to its subsampling E_k.
3. Return the bias $\widehat{B}_t$ and variance $\widehat{\sigma}_t$ for each time step t.

$$\{E_k : t \in E_k \, | \boldsymbol{V}_k^{inf} < \mathbf{c}(t) < \boldsymbol{V}_k^{sup}\}$$
$$\widehat{B}_{E_k} = \sum_{t \in E_k} \frac{r_t}{\#E_k} = \widehat{r}_{E_k} \qquad \widehat{\sigma}_{E_k} = \sqrt{\sum_{t \in E_k} \frac{(r_t - \widehat{B}_{E_k})^2}{\#E_k}} \tag{4.5}$$

2. ML: Machine learning-based estimation

The estimation model can be learned by a machine learning algorithm. We train two prediction models to learn the bias and variance of the residues of the predictions from the contextual attributes. The two models are similar in terms of estimating a type of mean (absolute for the bias and quadratic-centered for the variance) on a learned contextual subsample.

$$Bias \ : \ \theta = \underset{\theta}{\operatorname{argmin}} \sum_t |M_\theta^{\widehat{B}}(x_t) - r_t| \qquad \widehat{B}(t) = M_\theta^{\widehat{B}}(x_t) = \widehat{r}_t$$
$$Variance \ : \ \theta = \underset{\theta}{\operatorname{argmin}} \sqrt{\sum_t |M_\theta^{\widehat{\sigma}}(x_t) - (r_t - \widehat{B}(t))^2|} \qquad \widehat{\sigma}(t) = \sqrt{M_\theta^{\widehat{\sigma}}(x_t)} \tag{4.6}$$

Second, we propose directly extracting an estimation of the bias and variance from a forecasting model. We propose exploring the extraction for a random forest and a deep neural network. Often, extracting the estimated bias from the model itself will lead to a result of zero since the model has been optimized to minimize this bias.

3. RF: Random forest extraction

In [Mei06], the authors show that we often exploit valuable information about the distribution learned from a random forest by considering only the mean of the subsamples. From this assumption, we propose extracting the variance based on a learned subsampling of our random forest forecasting model.

Let M be a random forest composed of $(T^1, .., T^n)$ binary trees. Each tree T^k is composed of a set of leaves L^k. Values j_i are assigned to each leaf during the learning phase according to their attribute modalities x_i. We define a tree walk operator $F^k(x_t)$ that takes attributes x_t and returns for the associated leaf L_i^k, the set of assigned values.

$$M(x_t) = \frac{1}{n} * \sum_{k \in [1,n]} \Big(\sum_{j \in F^k(x_t)} \frac{j}{\#F^k(x_t)} \Big) = \widehat{y}_t \tag{4.7}$$

The prediction of an element by an RF model is similar to the weighted mean of a subsample formed by elements sharing a leaf. The weighting depends on the shared leaf number and shared element tree number. Shared leaf elements can be considered contextual neighbors on the basis of their attributes. Then, we can extract the bias (equal to 0) and variance from this contextual subsampling.

$$\widehat{B}(t) = 0 \qquad \widehat{\sigma}(t) = \sqrt{\frac{1}{n} * \sum_{k \in [1,n]} \Big(\sum_{j \in F^k(x_t)} \frac{(j - \widehat{y}_t)^2}{\#F^k(x_t)} \Big)} \tag{4.8}$$

4. DEEP: Neural network extraction

A second form of extraction is based on variational dropout [GG16], which aims to approximate Bayesian behavior in a deterministic network. A study in [ZL17] applies this technique to an LSTM neural network to extract the confidence in the prediction model. Following the same line of research, we use the variational dropout to estimate the variance from our LSTM encoder-predictor model.

$$\begin{gathered} Let\ M_{\widehat{\theta}}\ \text{be a neural network that infers } y_t \text{ from } x_t \\ \theta = \underset{\theta}{\operatorname{argmin}} \sum_t |M^{\theta}(x_t) - y_t|^2 \qquad M_{\theta}(x_t) = \widehat{y}_t \end{gathered} \tag{4.9}$$

The neural network aims to capture the link between the attributes and prediction targets through an embedding of the attribute space into the prediction space. Successive nonlinear projections in the abstract space Z are used to this end. These abstract spaces give us abstract representations z_t of our elements that capture the topological structure of our data. We can exploit such spaces to perform contextual subsampling by defining a neighborhood in Z space. The contextual subsampling will be based on the contextual information captured by M. The main issue comes from the definition of a neighborhood $\mathcal{B}(z_t)$ in Z space.

$$\begin{gathered} \mathcal{B}(z_t) : \{k \mid z_k \in [z_t \pm \epsilon]\}\ with\ z, \epsilon \in \mathcal{R}^{\#Z} \\ \widehat{B}(t) = \sum_{k \in \mathcal{B}(z_t)} \frac{|\widehat{y_k} - y_t|}{\#\mathcal{B}(z_t)} = \widehat{r}_t \qquad \widehat{\sigma}(t) = \sqrt{\sum_{k \in \mathcal{B}(z_t)} \frac{((\widehat{y_k} - y_t) - \widehat{B}(t))^2}{\#\mathcal{B}(z_t)}} \end{gathered} \tag{4.10}$$

This issue can be avoided with a variational neural network M_{θ}^{var} based on an explicit (variational layer) or implicit (variational dropout) random drawing by generating a virtual sampling that self-defines the neighborhood in Z space.

$$\begin{gathered} \theta = \underset{\theta}{\operatorname{argmin}} \sum_t |(M_{\theta}^{var}(x_t) - y_t)|^2 \\ \sum_1^m \frac{M_{\theta}^{var}(x_t)}{m} = \widehat{y}_t \end{gathered} \tag{4.11}$$

The stochastic projections of the model M_{θ}^{var} transform the latent representations z_t into a collection of probabilistic points. We can access the probabilistic clouds of predictions for an element by making many predictions. This gives us a virtual contextual subsampling from which we can estimate the mean and variance.

$$\widehat{\mathcal{B}} = 0 \qquad \widehat{\sigma}(t) = \sqrt{\sum_m \frac{(M_{\theta}^{var}(x_t) - \widehat{y}_t)^2}{m}} \tag{4.12}$$

4.4 Experiments on a synthetic data set

4.4.1 Evaluation setting

Data generation Contextual anomaly detection assessment requires a complete knowledge of anomalies. To establish the foundations for our assessment protocol, we first experiment with the methodological framework on synthetic data generated with contextual anomalies. Once our framework is well defined, we apply it to the time series data related to the transportation domain.

This use case is a toy example that aims to illustrate the interest of variance estimation for anomaly detection. The purpose is not to illustrate the detection performance, which depends on the magnitude of the anomalies. The aim is to show the detection gain due to the addition of the contextual variance. For better comprehension and rendering of the results, we use periodic influences. As the approaches are attribute-based, they can capture more complex influences as long as relevant contextual attributes are available.

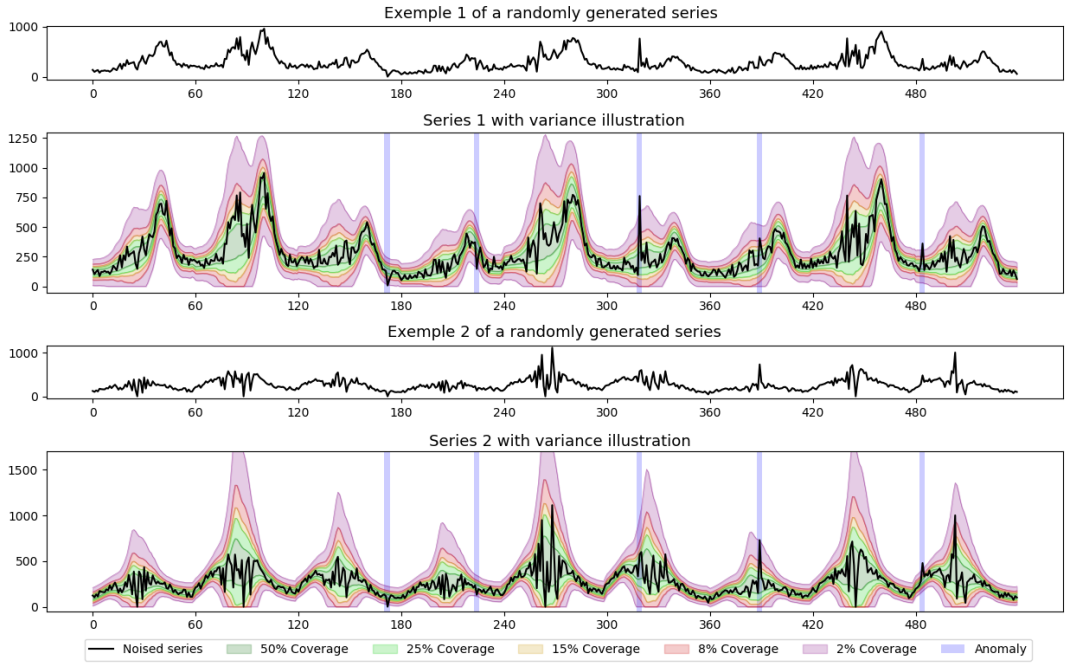

Figure 4.3.: Example of time series generated on 9 "hypothetical days"

We develop a generation process for multivariate time series with a dynamic context and anomalies. Figure 4.3 shows the example of two signals generated by this process. For our synthetic application, we generate a time series of 8000 time steps in three dimensions (3D). The generated data look like ridership time series with a short periodic pattern of 20 time steps ("Daily pattern"), a medium periodic pattern of (20×3) timesteps (a type of "weekly trend"), and a long periodic trend of (20×100) timesteps (a "seasonal trend").

The data generation process (Equation 4.13) follows several steps. (i) First, we generate a regular pattern corresponding to daily trends, and we combine the regular pattern with a combination of sinusoidal trends representing weekly and yearly influences. (ii) Then, we add a dynamic component built by multiplying the contextual component with a random daily magnitude coefficient. (iii) We introduce variability based on multiplicative Gaussian noise modulated by other regular patterns corresponding to contextual variability. (iv) Additive noise is introduced to disturb the series. (v) Finally, anomalies are generated randomly and applied through a predefined impact by considering both types of noise. For more details, the generation process is presented in Appendix A.1.

$$
\begin{aligned}
f^c &= Periodic\ pattern * Trends\ (i) \\
f^d &= f^c * Daily\ magnitude\ (ii) \\
\mathcal{E} &= (f^c + f^d) * (Noise_{mult} * Daily\ variance\ pattern) + Noise_{add}\ (iii + iv) \\
f^a &= (f^c + f^d + \mathcal{E}) * (\alpha_{mult} * Anom) + \alpha_{add} * Anom\ (iv) \\
y &= f^c + f^d + f^a + \mathcal{E}
\end{aligned}
\tag{4.13}
$$

Our data generation process is designed to create a context linked to a virtual hour with a specific magnitude and partially correlated variance through different generations. Each anomaly is linked to a context depending on its temporal position. Table 4.3 summarizes the setting of each anomaly context. The contextual anomaly detection performance of the combination of the forecasting models with the bias-variance estimations is evaluated on our synthetic multivariate time series.

Table 4.3.: The different contexts for synthetic data generation with anomalies

Name	Hour	Magnitude
C1	1-5	Very Low
C2	6-10	Very Strong
C3	11-15	Strong
C4	16-20	Low

Forecasting models

Two types of attributes can be distinguished. The first is long-term contextual attributes, which are known influential factors such as calendar factors, including the hour, type of day, and season. Forecasting models using long term attributes perform as a type of seasonal decomposition using cyclic attributes. Short-term dynamic attributes summarize the past dynamics of the time series. Forecasting models use both 18 long-term attributes and 12 short-term attributes that can be likened to an auto-regressive model with exogenous variables.

Long-term contextual attributes (LT): Linked to known influential factors, such as the hour, day type and season attributes, encoded by a sinusoidal transformation (Appendix A.3.1), yielding (6 * 3) contextual features.

Short-term dynamic attributes (ST): The latest historical values on a horizon [t-4,t] of each spatial dimension (3), which are used to capture the dynamic component.

Overall, we use 18 long-term attributes and 12 short-term attributes.

We compare the performances of six forecasting models, as follows:

- **The last value (LV) model**, based on the last observed value (t-1).
- **The categorical (CAT) model**, based on a categorical mean (linear regression where regressors are indicator functions) computed on long-term attributes (hour, day type and seasonality).
- **A long-term random forest (RF-LT)** ensemble of decision trees using only long-term attributes.
- **A short-term random forest (RF-ST)** ensemble of decision trees using long- and short-term attributes.
- **An encoder-predictor LSTM (LSTM-EP)**, which is better able to capture the dynamics of the time series and to achieve multi-step forecasting. More details are provided in appendix (A.5) and in the article [Pas+19b].
- **A "virtual model" called EXACT**, which obtains the synthetic data distribution without the variability or anomaly components. It corresponds to the best feasible forecasting model.

The two models LV and EXACT are used to provide the minimum and maximum forecasting performances.

Evaluation criteria

To achieve a robust evaluation, we generate five datasets with their own regular and variance patterns. We inject 1.5% anomalies. The anomalies are generated to be "mostly detectable", which means that some anomalies can be difficult to differentiate from standard noise. For each dataset, we train all forecasting models and bias-variance estimation methods.

The forecasting performance is measured through the root-mean-square error (RMSE) (see Table 4.4) computed on the training and test sets for the six forecasting models. Here, we distinguish between the abnormal subsamples composed of time steps impacted

by anomalies and the normal subsamples composed of the remaining time steps without anomalies.

The anomaly detection performance is measured through the sensitivity (% detected anomalies). We calculate the global sensitivity (Table 4.7) and the specific context and magnitude sensitivity (Table 4.8). As the number of detections is fixed by the prior anomaly ratio, the specificity is redundant with respect to the sensitivity. However, we observe the influence of the prior anomaly ratio on the detection performance through the receiver operating characteristic (ROC) curve metrics (Figure 4.6).

4.4.2 Results on the synthetic data

Forecasting results

The forecasting results are presented in Table 4.4. We observe classic forecasting trends in terms of the performance. The long-term models (CAT and RF-LT) provide similar results, and the short-term models (RF-ST and LSTM-EP) improve the prediction performance due to their ability to capture the dynamic component of the time series. The LSTM-EP seems to slightly improve the prediction compared to the RF-ST model. Regardless of the model prediction, the forecasting error is higher for abnormal samples, supporting our idea that the prediction residuals can highlight anomalies in the time series. Despite the measures, we observe overfitting, in particular on the abnormal subsamples, which can be explained by the strong corrective gradient.

Table 4.4.: One-step-ahead forecasting performance (RMSE) on the synthetic dataset

Sample	Normal		Abnormal	
Model	Train	Test	Train	Test
LV	87.4±6.8	87.2±8.9	253.3±23	270.0±42
CAT	67.4±3.3	70.0±4.8	242.2±23	265.2±43
RF-LT	64.3±3.1	70.8±4.8	233.1±23	265.6±45
RF-ST	**44.6±2.8**	61.8±4.9	**189.5±21**	**264.8±46**
LSTM-EP	51.3±3.1	**59.6±4.9**	206.5±12	265.9±43
EXACT	51.6±4.1	51.1±4.3	248.3±21	264.0±45

Variance learning result

Variance learning is a critical task that is complex to evaluate on heterogeneous datasets because variance can only be indirectly observed. Fortunately, work on generated data makes it possible to recover the variance introduced during the generation. As our estimators estimate a relative variance based on different techniques, it must therefore be normalized by their means before comparison.

Table 4.5.: Performance of variance estimation (Mean of 5 experiments)

Variance Estimation	Forecast Models	Metrics MAE	MSE
S-AE	-	0.547	0.786
S-RE	-	0.527	0.798
S-EMP	CAT	0.273	0.447
	RF-LT	0.272	0.445
	RF-ST	0.271	0.479
	LSTM	0.253	0.432
	EXACT	0.275	0.479
S-ML	CAT	0.230	0.404
	RF-LT	0.231	0.402
	RF-ST	0.215	0.416
	LSTM	0.199	0.374
	EXACT	**0.195**	0.392
S-RF	RF-LT	0.231	**0.372**
	RF-ST	0.233	0.391
S-DEEP	LSTM	0.495	0.745
*S-EXACT	-	**0.000**	**0.000**

* Theoretical variance used as reference.

Table 4.5 provides the performance of variance estimation. Absolute and quadratic mean error metrics (MAE and MSE) are used to evaluate the difference between the theoretical variance recovered and the different estimated variances. By taking the average of MAE and MSE performances over several experiments, a more relevant evaluation is performed.

The 'pseudo-estimators' S-AE and S-RE (which are implicitly used when studies do not consider variance) do not really estimate the variance. They serve as a baseline. The real estimators 'S-EMP", 'S-ML' and 'S-RF' show better performances. The empirical S-EMP estimator estimates the contextual variance on the basis of a predefined context which guarantees a good performance as long as the priority is relevant. On the other hand, the S-ML and S-RF estimators directly capture the variance by exploiting contextual attributes. They perform better than the estimator based on a learned context. These results corroborate the interest of variance capture by ML learning models using contextual attributes.

The poor performance of the 'DEEP' estimator seems to indicate that the variation dropout is insufficient to capture the variance. Several explanations can be invoked (Network too deep, variational dropout too low, Interference related to short term dynamics). More in-depth work is needed to explore the relevance of this technique.

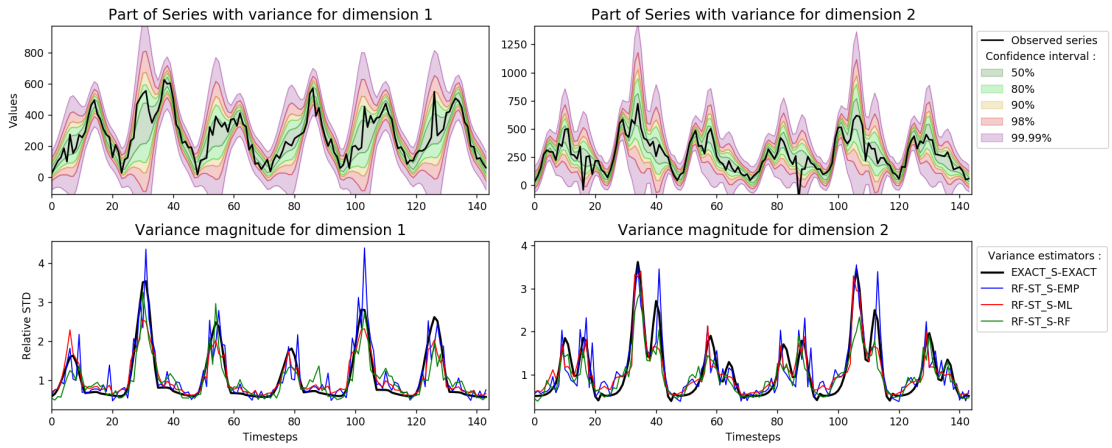

Figure 4.4.: Variance estimations illustration

These variance captures are performed by multivariate models that must capture variance on several dimensions. Figure 4.4 illustrates two of the dimensions of a portion of the generated data; the variances are estimated by the different methods (including the recovery of the theoretical variance).

Finally, Figure 4.5 illustrates the impact of the parameter q that defines the power coefficient of the normalization for an anomaly score based on the RF-ST+S-ML models. It shows that the parameter q modulates the distribution of anomaly scores as a function of the magnitude of contextual variance.

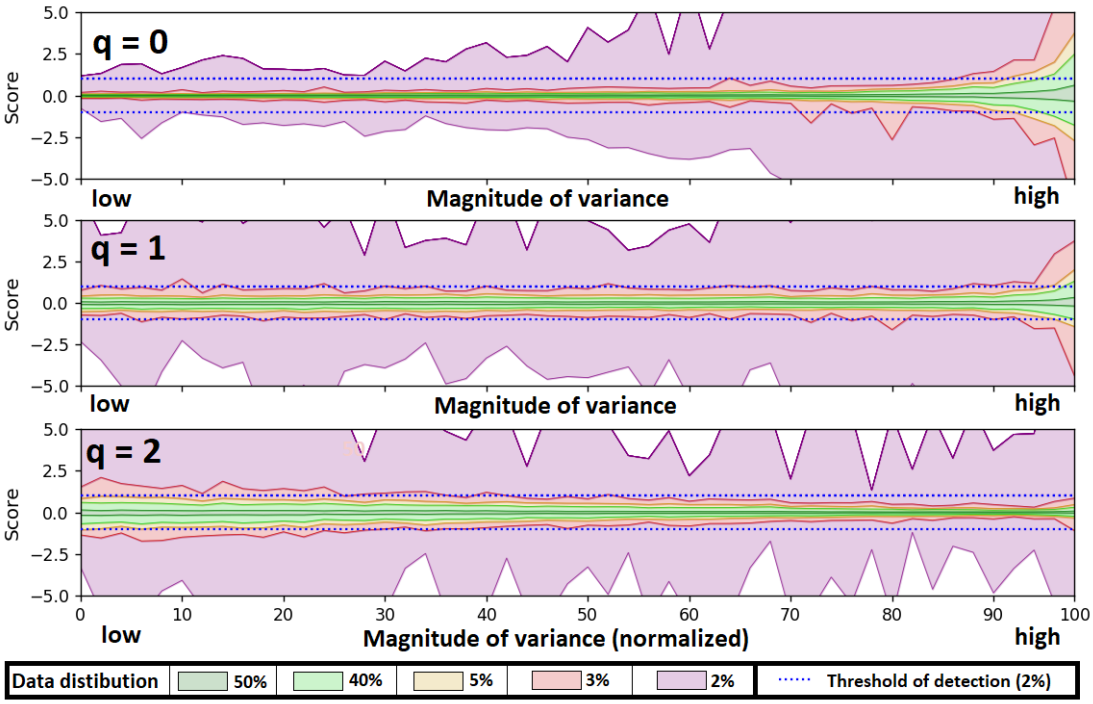

Figure 4.5.: Distribution of anomaly score (2%) as a function of variance magnitude

As announced, we can observe that the parameter q will modify the sensitivity of detection linked to the contextual variance amplitude. On the generated data, 2% anomalies were injected homogeneously regardless of the magnitude of the contextual variance. Thus a perfect anomaly score should have a constant anomaly score distribution that does not vary with contextual variance. This means on Figure 4.5 that the violet/red boundary should be aligned with the detection threshold.

For $q = 0$, equivalent to not considerate variance (it is the same as S-AE), we observed that elements linked to a high variance are more easily detected. Much more than 2% of elements exceed the threshold line for the most extreme contextual variances, which implies false detection. Conversely, the purple/red border is below the threshold for low contextual variances, which implies undetected anomalies.

For $q = 1$, which assumes a homogeneous anomaly distribution, we see that the distribution of the score is more homogeneous and invariant of the amplitude of the context. Nevertheless, we note a slight over-detection on extreme variances. This can be explained by the bias of prediction models which have difficulties predicting in the presence of a high variance, and the difficulty to estimating extreme variance.

For $q = 2$ we observe the inverse of $q = 0$: a sensitivity of detection on the elements with low variances to the detriment of the elements with high variances.

The hypothesis of a homogeneous and context-invariant anomaly score distribution is not necessarily always relevant. It depends on the application, the data and the nature of the anomalies. However, when knowledge of the nature of this distribution is lacking, it is possible to adapt the q parameter by looking for the optimal value that maximizes a labeled anomaly detection number.

Anomaly detection results

The univariate results obtained with all prediction model combinations and bias-variance estimation approaches are presented in Table 4.6. In the columns of the table, one can observe an improvement in the detection performance directly linked to the performance of the prediction models. In fact, improving the forecasting performance leads to forecasting residues that are more closely correlated with the anomalies and results in a better detection capability. Analyzing the different anomaly scores according to each row of the table allows us to conclude that the approaches based on the variance estimation outperform the other approaches, and that the obtained results are close to those provided by the virtual model EXACT, which exploits the true variance. To summarize, for the synthetic dataset, the ranking in ascending order of the detection rate can be given as follows: N-AE $>$ N-EMP $=$ N-DP $>$ N-ML $=$ N-RF $>$ N-EX.

Table 4.6.: Performance of all combinations (sensitivity with a 2% detection ratio)

Norm / Model	N-AE	N-RE	N-EMP	N-ML	N-RF	N-DP	N-EX*
CAT	40	53	69	**75**	-	-	74
RF-LT	42	53	69	74	**76**	-	74
RF-ST	51	54	75	79	**80**	-	85
LSTM-EP	56	54	77	**82**	-	75	**86**
EXACT*	60	56	85	**90**	-	-	95

* Virtual models not available for the real data.

An in-depth analysis of the anomaly scores, according to the magnitude and the context of the generated anomaly, is provided in Table 4.7.

Table 4.7.: Detailed performance of the LSTM-EP-based scores (2% detection ratio)

Magnitude	Minor anomaly				Major anomaly				Total
Context	C1	C2	C3	C4	C1	C2	C3	C4	-
Norm									
N-AE	30	57	58	28	52	80	83	61	56
N-RE	42	49	44	47	65	60	60	63	54
N-EMP	54	72	71	68	83	88	88	92	77
N-ML	**64**	**76**	**74**	**70**	**89**	**93**	**92**	**95**	**82**
N-DP	62	65	62	66	87	88	82	89	75
N-EX*	76	77	74	79	97	94	94	98	86

* Virtual models not available for the real data.

The forecasting is achieved here by the LSTM-EP model. These results show the weakness of the naive scores; i.e., low-magnitude contexts (C1 & C2) induce weakness in the N-AE score, while high-magnitude contexts (C1 & C4) induce weakness in the N-RE score. Conversely, the context-normalized anomaly scores (N-EMP, N-ML, N-RF, and N-DP) show better context robustness and better detection rates for both minor and major anomalies. As expected, minor anomalies are more difficult to detect.

Table 4.8.: Local/global score performance (sensitivity with a 2% detection ratio)

Norm	N-AE		N-RE		N-EMP		N-ML		N-RF		N-DP		N-EX*	
Type	Loc	Glo	Loc	Glo	Loc	Glo	Loc	Glob	Loc	Glo	Loc	Glo	Loc	Glo
Model														
CAT	40	40	53	57	69	80	75	**89**	-	-	-	-	74	86
RF-LT	42	43	53	57	69	82	74	89	76	**90**	-	-	74	87
RF-ST	51	53	54	58	75	88	79	94	80	**96**	-	-	85	96
LSTM-EP	56	62	54	58	77	89	82	**95**	-	-	75	90	86	**97**
EXACT*	60	59	56	58	85	93	90	**98**	-	-	-	-	95	100

* Virtual optimal model not available for the real data.

Table 4.8 compares the **local** and **global** score performance (as defined in Section 4.3.3). Spatial aggregation improves all the detection rates. The improvement is particularly significant for context-normalized anomaly scores. Indeed, an anomaly can be masked by the variance in one dimension, whereas it can be detectable in the other two dimensions.

The analysis of the detection results can also be achieved through ROC curves. The area under the ROC curve can be used to quantify the effectiveness of a detection approach. The detection threshold varies in the range of 0% to 15%. As shown in Figure 4.6, the ROC curves based on the naive scores are located below the other context-normalized score curves. Furthermore, the ROC curves obtained with aggregation improve the anomaly detection compared to the univariate ROC curves.

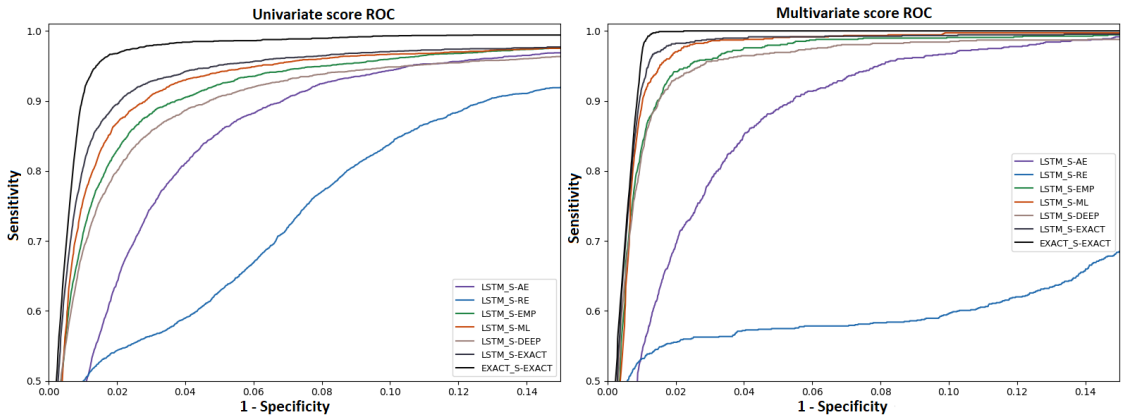

Figure 4.6.: ROC curves for the local and global scores based on the LSTM-EP residue

4.4.3 Conclusion of synthetic experiments

The experiments carried out on synthetic data show the relevance of using the prediction residues jointly with a bias-variance estimation for contextual anomaly detection. Several conclusions may be drawn from these experiments, as follows:

- The scores based on the short-term prediction residue give better results in terms of anomaly detection.
- The bias-variance estimation makes the anomaly detection more robust with regard to context.
- The variance can be estimated by learning on forecasting residues or by extraction in a random forest prediction model or a neural network forecasting model.
- The multidimensional aggregation of univariate scores significantly improves the detection performance.

4.5 Experiments on a real smart card ticketing dataset

The Montreal Transit Corporation (STM) provided us with smart card ticketing data from the logs of the automatic fare collection (AFC) recorded at 50 metro stations in the city. In addition, we obtained a disturbance database that lists the special events and incidents over the studied period, namely, from January 2015 to December 2017. We aimed to apply our detection anomaly approaches and to confront the statistical anomaly scores with the disturbance database. This would enable us to characterize the impacts of the different disturbances on the metro ridership in Montreal.

4.5.1 Data description

Smart card ticketing time series

For each station of the Montreal subway, smart card tap-in logs are aggregated with a temporal step of 15 minutes starting at 5am each day until 1am the following day.

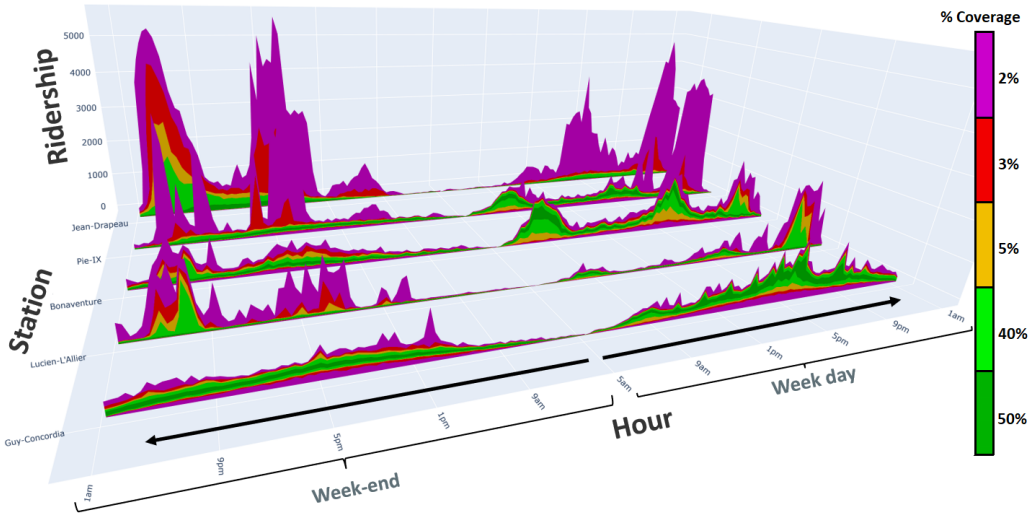

Figure 4.7.: Profiles of the Montreal metro ridership for 6 stations

In this study, we focus on the fourteen stations located in downtown Montreal. The data consist of a multivariate time series of 14 dimensions and 87860 timesteps, corresponding to 1096 days with 80 daily time steps of 15 minutes. Figure 4.7 shows an example of the tap-in log time series collected at six metro stations. A differentiation is made between weekday and weekend profiles. The figure highlights the impact of the context (station, hour, and day type) on metro ridership, which makes the forecasting and anomaly detection tasks quite complicated.

Disturbance data

The disturbance data contain some events and incidents that occurred within the studied period and that might impact the stations' ridership. Based on the tap-in smart card logs, the disturbances are characterized by the date, start and end times, impacted station, and class of the underlying disturbance. We can distinguish between minor and major incidents/events. The 2076 **minor incidents**, which have an average duration of 35 minutes, are divided into the following seven categories: technical failure (482), minor failure (184), door failure (112), track operation (96), works (42) and miscellaneous (43). The 960 **major incidents**, which have an average duration of 45 minutes, may be attributable to a variety of causes, including malignancy (299), accidents (281), intrusion (202), and fire (75). The **event data** include 1772 events with 10 event categories, including exhibition (414), hockey match (385), festival (365), concert (178), sport (174), show (172), tennis (37), football (30) and other (144). We note the highly variable durations of the events, which range from a few hours for soccer events to an entire day for exhibitions.

Note that the operator disturbance database is a rich information source, but it does not constitute a reliable and full dataset of anomalies. Since it is an incomplete source, it is not a ground-truth reference for all the events, incidents and other phenomena that can impact the station ridership. Consequently, the goal is not to detect all disturbance database elements but rather to evaluate which and how disturbances impact the smart card activity through the prediction residues based anomaly scores. An additional objective is to map the unexplained detection with the disturbance dataset to either evaluate their impacts or investigate the causes afterward.

4.5.2 Forecasting results

The goal is to forecast the ridership for each of the fourteen stations at time step t+1. Depending on the forecasting model, long and short-term attributes can be used. The **short-term dynamic attributes (ST)** are considered to capture the dynamic components. They consist of a 70-dimensional vector of the historical time series collected in the past temporal horizon $[t-5, t]$ for the 14 stations. The **long-term contextual attributes (LT)** are linked to the well-known influential factors :

- The time schedule (8 dimensions), given by the time step position in a day (80 possible values) encoded by a sine transformation (see Appendix A.3.1) at four frequencies (1/2, 1/4, 1/8, 1/24) related to the hour pattern.
- The day type (7 dimensions), encoded by a one-hot vector.
- The seasonality (8 dimensions), encoded by the day of the year (365 possible values) as a sine transformation (see Appendix A.3.1) at four frequencies (1/2, 1/4, 1/8, 1/12) related to the seasonality pattern.

- The year (3 dimensions), encoded by a one-hot vector.
- Holidays (7 dimensions), including July, August, Winter, Christmas, New Year, and other holidays, encoded by a one-hot vector.

Table 4.9.: Forecast performance (RMSE metrics) on differentiated samples

Sampling	Normal		Minor-incidents		Major-incidents		Events		All	
Data %	69.5%		5.6%		2.6%		24.1%		100%	
Model	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
LV	75.35	75.54	90.92	90.92	101.96	100.36	101.77	103.75	83.13	84.31
CAT	56.13	58.76	71.24	75.62	71.18	105.75	104.78	120.44	70.55	79.53
RF-LT	35.84	53.14	46.43	70.20	57.51	97.42	68.08	109.93	45.84	72.39
RF-ST	28.74	**35.66**	35.68	48.33	43.41	75.00	47.99	76.42	34.51	49.81
LSTM-EP	31.01	**35.80**	40.30	**47.30**	50.25	**68.92**	48.59	**71.07**	36.25	**47.73**

In Table 4.9, we evaluate the forecasting performance of the five forecasting models. The evaluation is conducted by splitting the training and test sets based on the following different sample types: normal, with minor incidents, with major incidents, and with events.

Compared to the synthetic data, similar conclusions can be drawn for the prediction of real data. The machine learning models exhibit a better forecasting performance compared to the categorical model. Short-term attributes improve the forecast, particularly for the LSTM-EP model, which slightly improves the prediction performance for both normal and disturbed contextual samples. However, this does not lead to a noticeable gain when the prediction residues are used for anomaly detection. We note an overfitting between the training and the test performance. The data collected in 2015 and 2016 are used for training, while the test performance is evaluated on data collected in 2017. Learning on sliding windows can reduce this problem.

4.5.3 Anomaly detection results

Methodology

First, it is essential to emphasize that these results can only be interpreted as trends because the operator disturbance dataset is incomplete and only partially reliable (see 5.1.2). Moreover, the link between a disturbance and a ridership anomaly is not straightforward. Therefore, the usual evaluation conducted in anomaly detection is not possible. To this end, we aim to qualitatively evaluate the relevance of the anomaly score on the basis of the overlap with the disturbance dataset. A comparison of the different approaches will be performed based on the ratio of the detected disturbances with regard to their associated score. For the relevance comparison, each approach must detect the same number of statistical anomalies (based on a prior anomaly ratio).

This work aims to show that for a predefined threshold (based on a prior anomaly ratio), it is possible to refine the detection of contextual anomalies based on the prediction residues by taking the variance into account. Considering the fourth variance estimation method detailed in Section 4.3.3 and the prediction models based on the categorical model (CAT), the short-term random forest model (RF-ST) and the LSTM-EP neural network, we evaluate nine combinations of a prediction model and variance estimator. The results are shown in Tables 4.10 and 4.11. For each combination, we analyze at two different scales through the **local** (per station) and **global** (Network scale) scores (introduced in Section 4.3.2).

The cross-referencing between the anomaly scores and the disturbance dataset requires additional processing. (i) First, statistical anomalies must be extracted from the anomaly scores. A statistical anomaly is defined by a time interval (a start and end time) in which the score values are higher than a threshold. The local and global thresholds are defined according to prior knowledge of the anomaly percentage in the dataset (2.5% for the local score and 5% for the global score). Then, for each operational disturbance, an anomaly impact is detected if at least one statistical anomaly arises in the temporal and spatial perimeter (for local anomaly detection).

Among the parameters of our anomaly score process, the parameter q (introduced in Section 4.3.2, Step 3) manages the homogeneity of detection with regard to the contextual variance. The assumption of the contextual anomaly distribution is directly linked to the anomalous behavior of the data. In our application, events are heterogeneously distributed (they often impact specific contexts, such as Friday evenings, for example). Conversely, the incidents overall seem to be more homogeneously well distributed (they occur more equally in each context). Therefore, the parameter q is manually tuned for each combination through a trade-off between event and incident detection in the range [0,1], which slightly promotes detection in high-variance contexts.

Results

We will therefore confront disturbances coming from a labeled operator base and anomalies coming from the detection models. However, as mentioned in Section 4.5.1, disturbances are events, minor incidents, or major incidents that "**could significantly impact** ‘ the ridership at the station. Some of these disturbances will not have a significant impact because the link between the disturbances provided by the transport operator and the ridership time series is not straightforward. Hence, the quantitative results must be interpreted with caution.

Usual **true positives** become **detected disturbances**. This means that disturbances can be associated with spatially and temporally consistent anomalies that reflect their impacts. Conversely, **false negatives** become **undetected disturbances**, which means that no

significant impact on the ridership time series was found by the detection models. On the other hand, **false positives** correspond to **unexplained anomalies**. This means that a significant statistical impact was found by using the detection approaches, but no known disturbances can explain it. These detection incidents cannot be qualified as false detection incidents because of the incompleteness of the disturbance base. Finally, **True Negatives** correspond to **normal periods** with no known disturbances or statistical anomalies detected by the detection models.

An exhaustive validation would require more human expertise and further investigations to enrich the dataset disturbances with meaningful labels that can help in the evaluation step. Moreover, trends and markers that can be inferred from the results can also contribute to the labeling task.

Nevertheless, we will carefully analyze some quantitative results next. Table 4.10 shows the results of the matching that we carried out between the anomalies detected by our models and the operator's declarative disturbances.

Table 4.10.: Disturbance and local anomaly explanation for a prior local anomaly ratio of 2.5%

		% Detected disturbances			% Anomalies explained by disturbances			
Anomaly score		Incident-min	Incident-maj	Event	Explained	Unexplained	Number	Average
Residue	Norm	N=2076	N=960	N=1772	Anomaly	Anomaly	Total*	Duration
CAT	N-AE	10.8	15.1	74.6	37.3	62.6	7163	4.54
RF-ST	N-AE	20.3	28.7	89.9	35.1	64.9	16351	1.98
	N-EMP	21.5	33.4	91.2	34.2	65.8	16316	1.99
	N-ML	**24.2**	**34.6**	90.4	30.7	69.3	20371	1.57
	N-RF	23.8	34.5	**91.7**	30.2	69.8	19861	1.64
LSTM-EP	N-AE	20.2	27.4	89.2	34.3	65.8	16737	1.93
	N-EMP	21.6	30.9	89.8	34.2	65.8	16653	1.96
	N-ML	21.6	32.8	**89.9**	31.3	68.7	19937	1.61
	N-DP	**24.3**	**33.6**	89.4	29.4	70.6	19114	1.70

*Local anomaly number after temporal aggregation

*All approaches have the same fixed ratio of 2.5% of anomalous time steps

These disturbances are focused only on the local level, i.e., at the station scale. Each row corresponds to a combination of the residue of either of the prediction models (CAT, RF-ST, LSTM) with one of the proposed normalizations (N-AE, N-EMP, N-ML, ...). The first part of the table provides **the percentages of detected disturbances** for the three sub-categories (Minor incident, Major incident, Events). This corresponds to the disturbance ratio covered by anomalies. The second part of the table provides the ratio of anomalies explained by disturbances. It also contains information on the number of anomalies after temporal aggregation and the average duration.

If we examine the disturbance detection-ratio according to the class of disturbance (Event, Minor incident, Major incident), the results show that it is easier for all approaches to detect the 'event-class' disturbances that often have a direct influence through ridership increase more than the 'incidents-class' disturbance whose influence may be more complex.

In the same vein, 'major incident-class' disturbances are more easily detectable than 'minor incident-class' disturbances due to their often more significant impacts.

Concerning the comparison of the different approaches linked to forecasting models, there is a very significant gain in detection between the long-term (CAT) and short-term (RF/LSTM) approaches, which can be explained by a better modeling of normal dynamic behavior. The performances based on LSTM residuals appear to be worse than those based on RF residuals. This can be explained by the better prediction performance of the LSTM model in disturbed situations, which counter-intuitively will reduce the anomaly signal of the residuals.

Concerning the type of normalization, we also observe a slight but significant improvement of the method without contextual normalization (N-AE) in comparison with approaches such as (N-EMP/N-ML/N-RF/N-DP). The gain is more measurable on the detection of the impact of major incidents. The gains can be explained by the contribution of contextual normalization, which will reduce the importance of the magnitude by considering the variance of each context. The combination that seems to provide the best performance is the S-RF normalized RF prediction couple. However, a formal decision cannot be made without a more quantitative evaluation that requires a more complete perturbation dataset.

Table 4.11.: Disturbance and global anomaly explanation for a prior global anomaly ratio of 5%

		% Detected disturbances			% Anomalies explained by disturbances			
Anomaly score		Incident-min	Incident-maj	Event	Explained	Unexplained	Number	Average
Residue	Norm	N=2076	N=960	N=1772	Anomaly	Anomaly	Total*	Duration
CAT	N-AE	10.8	16.2	62.9	69.6	30.4	1306	3.54
RF-ST	N-AE	13.5	20.6	75.4	69.4	30.6	2127	2.10
	N-EMP	12.6	20.3	71.4	71.2	28.8	1989	2.35
	N-ML	15.0	24.1	76.6	64.8	36.2	2529	1.84
	N-RF	**16.4**	**24.2**	**78.0**	64.1	36.9	2654	1.78
LSTM-EP	N-AE	13.5	19.6	**72.1**	69.0	30.7	2026	2.31
	N-EMP	13.6	20.7	71.3	70.2	29.8	1981	2.36
	N-ML	**15.7**	**22.2**	71.2	66.0	34.0	2309	2.01
	N-DP	15.3	20.8	66.5	61.5	38.5	2197	2.15

*Global anomaly number after temporal aggregation

*All approaches have the same fixed ratio of 5% of anomalous time steps

Table 4.11 contains the information resulting from the confrontation of the disturbances with anomalies linked with the **global anomaly score** (introduced in Section 4.3.2) of the proposed detection models. In contrast to the local score, which analyzes the anomalies at the scale of the stations, the global score analyzes at the scale of the network through a spatial synthesis carried out by the Mahalanobis distance. The local and the global detection do not have the same granularity nor the same fixed number of detections since they do not detect the same type of 'anomaly".

There is a pattern of results similar to that of the local score, with greater ease of event detection followed by major and minor incidents. The approaches based on short-term

predictions show a significant gain (RF-ST > LSTM-EP > CAT) and the best detection seems to be provided by approaches with contextual normalization (RF-ST+N-ML and RF-ST+N-RF > RF-ST+N-AE). Nevertheless, we notice that some approaches (RF-ST+N-EMP and LSTM-EP+N-EMP combined with contextual normalizations) seem to misbehave when combined with Mahalanobis spatial aggregation.

Even if the comparison between local and global scores does not really make sense, the lower explained disturbance ratio of the global scale than that of the local scale can be explained by the fact that global detection should detect a smaller number of anomalies that should, however, have a high impact on the network. In the same way, global detection has an higher ratio of anomalies with explanations. Indeed anomalies with a severe impact on the network are often explained by known disturbances. Conversely small magnitude and highly localized anomalies are less often linked to known causes.

4.5.4 In-depth analysis of the results

In-depth analysis of the spatio-temporal impact of various disturbances is a non-trivial and time-consuming task that requires large investigations with human expertise and validation. Automatic analysis tools save time and provide valuable help in focusing attention on relevant elements. The following section aims to provide insights regarding the interpretation of the results within the real-world application context.

Confidence intervals of the predictions

The aim here is to provide a detailed analysis of the results obtained from the real dataset. One of the first issues is related to the confidence interval of the prediction given by the bias and the variance. These parameters can teach us some valuable information about the contextual variability in the data.

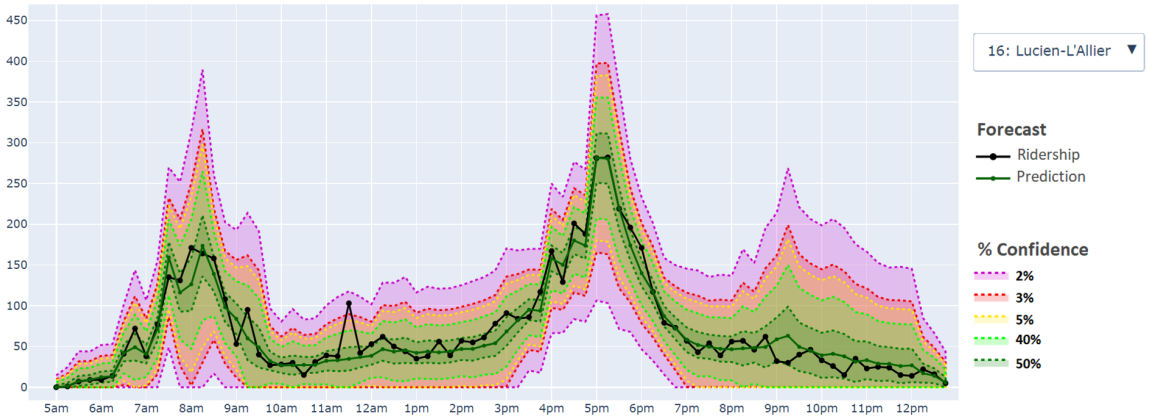

Figure 4.8.: Ridership prediction confidence at Lucien l'Allier station on Monday, February 27

Figure 4.8 shows the observed and predicted transport ridership at Lucien l'Allier station on February 27. The confidence intervals are also shown in this figure. We observe three periods of high variability. Both the morning and afternoon periods are expected since they are linked to rush hours. The evening period can be explained by numerous events taking place close to this station in the evening. The impact of these events is not considered by the forecasting models and induces high variability in the prediction. Adding event features would allow forecasting models to refine the evening forecasts and thus reduce the evening variability. We also observe an abnormal peak of ridership at 11:30 AM. This unusual peak greatly exceeds the contextual envelopes, allowing us to qualify the abnormality.

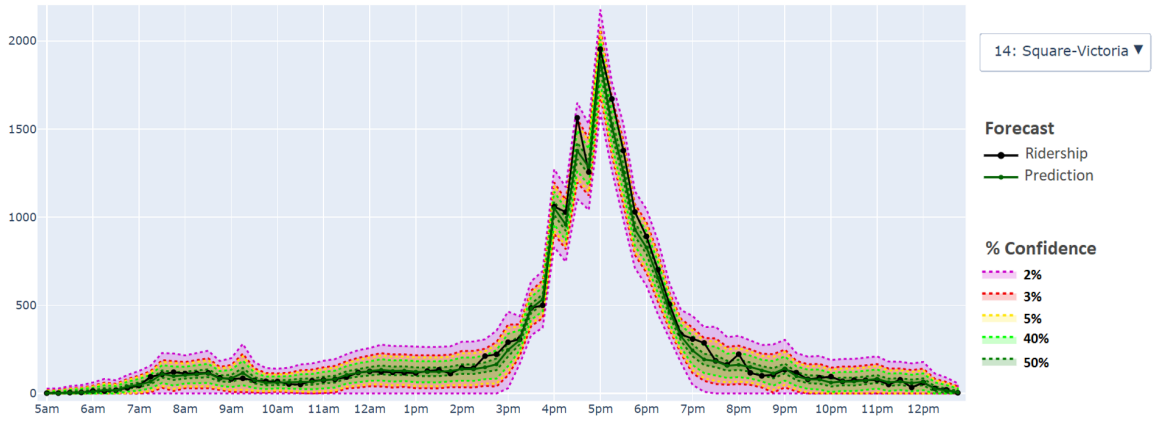

Figure 4.9.: Ridership prediction confidence at the Square Victoria station, Monday February 27

Considering another metro station (Square Victoria station) on the same day, ridership series are presented in Figure 4.9. Here, we observe a daily profile that is different from that of the Lucien l'Allier station. In particular, the metro ridership exhibits lower variability.

Analysis of two particular days

Monday, 27 February 2017

On Monday, February 27, a severe accident induced a partial traffic stop on a metro line (the green line at Pie-IX station) between 7:30 AM and 9:30 AM. The commuters used a transport hub station (Berri-UQAM station) to move to other metro lines. Figure 4.10 shows 2 anomaly scores based on naive normalization (N-AE) and contextual normalization (N-RF) computed for the fourteen metro stations for Monday, 27 February.

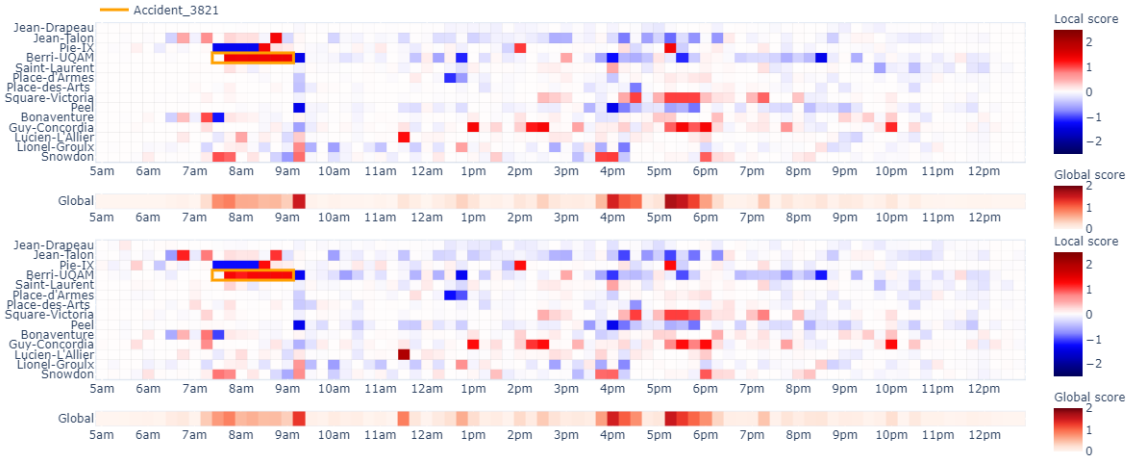

Figure 4.10.: Two anomaly scores (N-AE, N-RF) computed on Monday, 27 February, 2017

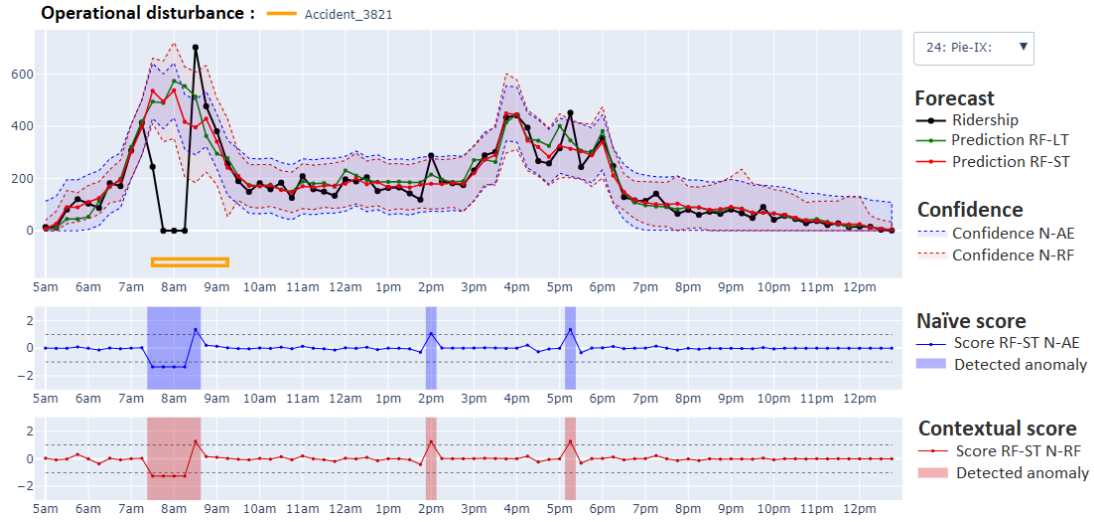

Figure 4.11.: Ridership and anomaly scores at Pie IX station on Monday, 27 February, 2017

For both anomaly scores, a short-term random forest (RF-ST) is used to achieve forecasting. We observe that the two scores seem almost identical. We detect two abnormal temporal periods. The first one occurs between 7:30 AM and 9:30 AM, with high negative scores for Pie-IX station (under-crowded) linked to the traffic stop and high positive scores for BERRI-UQAM station (overcrowded) linked to the commuter shift. The second period, between 4:00 PM and 6:00 PM, involves several stations but cannot be explained by our disturbance dataset. Figure 4.11 shows ridership and anomaly scores at the Pie X station for that day.

Wednesday, 5 August, 2015

On Wednesday, 5 August 2015, there were three major disturbances (one incident and two events), as follows:

- Between 5:30 AM and 7:30 AM, a traffic stop on the green line affected several stations on the perimeter of our study (Saint-Laurent, Places des Arts, Square-Victoria, and Guy Concordia).
- Between 7:00 PM and 11:00 PM, Philharmonic concerts occurred as part of a festival near the Pie-IX station. This event attracted approximately 45,000 people.
- Between 9:00 PM and 11:00 PM, a soccer match involving a popular team took place, with an attendance of 20,000 people. This event also occurred near the Pie-IX station.

This example illustrates the usefulness of contextual normalization to refine the precision of the anomaly score on small contexts. On the matrix anomaly scores shown in Figure 4.12, the naive score (N-AE) and the context-normalized score (N-RF) show significant differences in the morning. The disturbance does not affect the two scores in the same way: it is invisible for the naive score, whereas it seems to have a high magnitude for the context-normalized anomaly score. On the other hand, for the two evening disturbances, all scores show an overcrowded metro ridership linked to the events occurring near the Pie-IX station and other stations located around the area of the events.

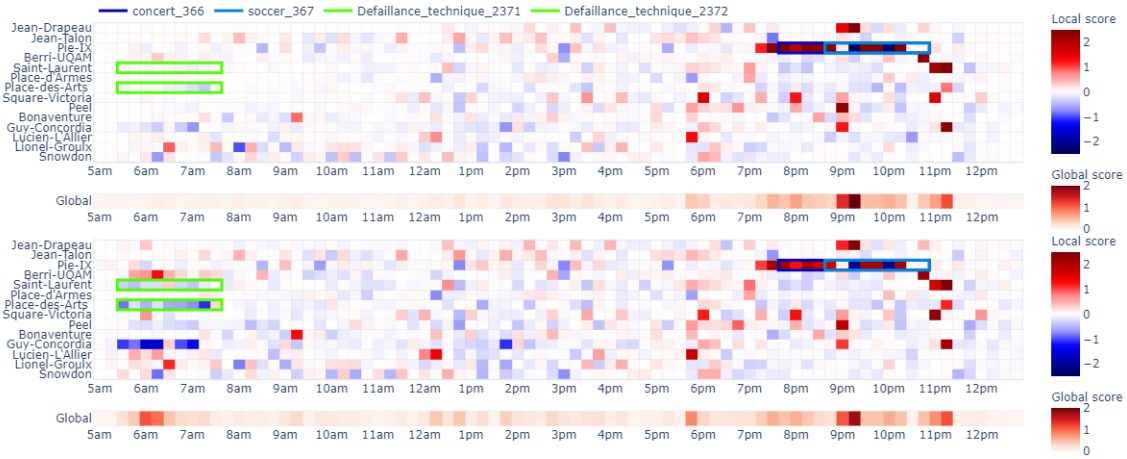

Figure 4.12.: Two anomaly scores (N-AE, NS-RF) computed on Wednesday, 5 August, 2015

If we examine Figure 4.13, which presents the ridership entry logs collected and predicted at Berri-Uqam station, one can observe that the prediction confidence for the context-normalized score is lower than that of the naive score. This confidence in the forecasting allows the context-normalized score to detect the overcrowding situation induced by the traffic interruption. In contrast, the naive score tends to focus on high impact (or magnitude) anomalies. We also notice that both scores detect an overcrowding situation

due to the end of the events. Figure 4.13 shows ridership and anomaly scores at the Pie X station for that day.

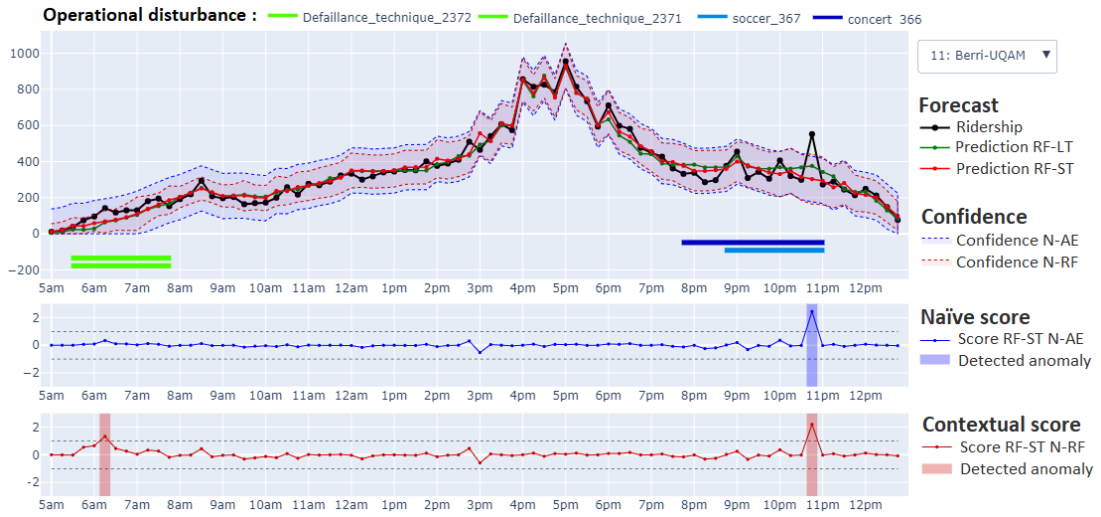

Figure 4.13.: Ridership and anomaly scores at Pie IX station on Wednesday, 5 August, 2015

4.6 Conclusion

In this chapter, we have proposed a general methodology to detect contextual anomalies on multivariate time series with a dynamic context. The influence of the dynamic context on these series can be summarized by a contextual mean (as reference behaviour) and a contextual variance (as variability measure). Our approach consists in estimating both contextual mean and variance using contextual and short-term features through different techniques. This contextual characterization can be useful in qualifying the prediction confidence, including both the strength and the weakness of the forecast, or in defining the contextual normality of the data.

This work used the short-term prediction models previously studied. It focused on the analysis of prediction residuals by proposing a model for estimating bias and variances. The combination of prediction and variance estimation models was then used to produce a context-normalized anomaly score. These scores allow the definition of the statistical abnormality of each time step with respect to its context.

The proposed method was evaluated on a synthetic dataset. The detection performance shows the effectiveness of the anomaly detection approaches involving context normalization. We also applied the proposed methodology to a real smart card dataset collected from a metro network. The statistical anomalies were compared with a disturbance dataset provided by the transport operator that contains minor incidents, major incidents, and events. The results show that a contextual normalized anomaly score can improve

the detection of disturbances and reveal incidents that are not listed in the disturbance dataset. This reinforces our idea that such an analysis tool can provide transport operators and urban stakeholders with knowledge and insights on the temporal and spatial impacts of disturbances.

Although this has been done empirically, the choice of both the parameter q and the detection threshold remains an open question. These two parameters are directly related to modeling hypotheses about data distribution :

- q for the nature of the distribution (homogeneous or heterogeneous) of anomalies as a function of variance.
- The detection threshold for the percentage of anomalies present in the data.

It is possible to associate default values based on knowledge of data distribution such as for example the case of a homogeneous data set with 5% anomalies by taking a $q = 1$ and a Threshold $= 5\%$. It is also possible to tune these parameters empirically with a representative anomaly based on a training set.

Although the detection approach is illustrated with a posteriori analyses, it is designed to be applicable in real time in a rather naive way. The models are trained on past data. The contextual prediction, variance estimation, and anomaly score calculation bricks provide their output values as observations are reported. These indicators can support operators in charge of regulation. It would be possible to deepen certain problems related to Updating or Learning models in real time. In the same way, one could try to anticipate the anomaly score over a future horizon by comparing the forecasting values of a nominal model and a model designed to make predictions in disturbed situations.

This work is a preliminary step to address more advanced issues. Rather than detection, this research has focused more on on the extraction of a context-invariant abnormal signal in time series with a dynamic context. Additional work on the analysis of anomalous patterns will address one of the questions that originally motivated the thesis: How do the incidents or events that occur in a transportation network, whether or not they are significant, impact the passenger flows at mass transit stations?. Several direct perspectives are envisaged on different aspects of the work such as Anomaly characterisation, Short-term error, deep learning approaches (See Section 5.3).

5

Conclusion & Perspectives

This thesis has focused on the analysis of data collected in public transport networks. Through the valorization of real data (train loads, station ridership), it has addressed two main issues:

- How to forecast the short-term evolution of trainload.
- How to detect the impact of a disruption on station ridership

These problems are related to major academic issues concerning the evolution of temporal data in dynamic contexts, namely short-term prediction and anomaly detection.

Today, urban mobility is a central challenge related to crucial societal issues (ecology, quality of life, land use planning). Public transportation systems allow many people to be transported at a low economic and ecological cost, making them one of the main levers of sustainable mobility policies. The spread of metropolitan areas, the increasing complexity of transportation networks, and the growing demand for mobility make these transportation networks increasingly complex. Malfunctions in these networks have very significant economic and social consequences.

The emergence of connected infrastructures equipped with sensors allows the development of near-real-time capture of numerous data sources in the transport domain. These data can provide valuable information describing many facets of the transportation systems' state, from the demand (e.g., station ridership, congestion) to the supply (e.g., delays, changes in transportation schedules, incidents). The present work involved formalizing the analysis of time series structured through a dynamic context resulting from the entanglement of a set of known and unknown influential factors. The capture of the 'Dynamic context", particularly adapted to mobility data, is was carried out through contextual attributes that synthesize underlying structures (temporal, seasonal, or linked to heterogeneous properties) and short-term attributes that support the inference of short-term dynamics. The application of this formalism required understanding and analyzing the data to extract and refine relevant features to synthesize underlying structures. Then, dedicated models for forecasting and anomaly detection were built to optimally exploit the feature information.

5.1 Forecasting

The first part of the work focused on short-term forecasting by considering highly structured data on passenger load in commuter trains. The models were built on data collected on a commuter train line in the Paris region. The problem is tricky because this is a new data source that imposes particular constraints: temporal irregularity due to non-regular temporal sampling and contextual dynamics induced by numerous cross-influential factors, including transportation schedule variability.

The first task consisted in building and comparing different prediction models (baseline models, machine learning models including neural networks), focusing on the development and learning of models and the extraction/refinement of attributes. The idea was to extract and synthesize from raw data a relevant representation that capable of capturing the influence of underlying structures (calendar, spatial location, transport plan, events). We also propose a 'deep learning' Model named 'LSTM encoder-predictor". This model is dedicated to multi-step forecasting on irregular and heterogeneous time series with contextual dynamics. The model is based on the learning of an abstract synthesis of contextual influences (representation learning) combined with a RNN encoder-decoder architecture designed to capture the temporal dynamics present in the data.

The results show the importance of contextual representation and dynamic inference to perform short-term prediction of train loads. The 'Deep learning' approach slightly outperforms standard learning machine models and shows more stability for multi-step prediction. Neural networks can be particularly relevant to exploit complex attributes (for example, related to incidents). They could also be suitable tools to capture spatial dynamics that have not been studied in this work and produce a synthetic abstract embedding state used in analysis and decision-making algorithms.

5.2 Detection anomaly

The second part of the thesis focused on analyzing the impact of contextual anomalies on multivariate regular time series structured by a dynamic context. The Study capitalizes on previous work through the detection paradigm based on the residuals of a prediction model. It focuses on the influence of the dynamic context on the contextual variance and proposes a variance estimation to achieve robust contextual anomaly detection. This work performs a complete prediction task by focusing on contextual variance, which can be associated with prediction confidence.

The core of this work aimed to capture the short-term dynamics and disentangle the influence of numerous factors that compose the dynamic context in order to infer normal

data behavior and estimate contextual variability. Contextual variability can provide valuable information that makes it possible to discriminate the standard elements of a high variance context from abnormal extrema by avoiding amplitude-bias detection.

The detection approach is based on a context-robust anomaly score obtained from a prediction residual normalized by an estimated contextual variance. The proposed approach consists in combining a prediction model with a contextual variance estimation method and a normalization treatment. Several prediction models (based on previous work) were compared with several variance estimation methods.

The proposed methodology was applied on a synthetic data set, including anomalies. This provides an opportunity to illustrate and quantitatively evaluate the interest of considering variance. The advanced variance estimation (S-ML, S-RF) approaches showed their ability to learn the contextual variance. The deep variance estimation (S-DEEP) provided weak results and requires further investigation. Results show that the detection performances are linked to the quality of the forecasting models. Moreover, variance normalization based on the advanced estimation of contextual variance improves detection quality for a given forecasting model. The combination of high-performance prediction models (RF-ST, LSTM-EP) and advanced variance estimation (S-ML, S-RF) provides the best performances.

The proposed methodology was applied to another real data set containing two years of Montreal metro station ridership with a complementary incident database. The quantitative evaluation is complex because of the incompleteness of anomaly information. However, the approaches proposed here proved able to detect many known incidents and highlight unknown anomalies that significantly impact ridership data. The variance normalization refines the anomaly score by making it contextually robust. This makes it possible to detect a finer anomalies that could go unnoticed due to their minor amplitudes. The variance normalized anomaly score acquires interesting statistical properties to evaluate the deviation to the contextual normality for each observations.

5.3 Perspectives

This work, therefore, opens up many new perspectives related to different issues:

Data sources

We focused our analysis on data sources related to ridership. However, it will be essential to consider and confront several data sources (Train loads, station ridership, transportation schedule). Analyzing them through models that can either be dedicated to each data

source or be cross-source should extract valuable knowledge about the transportation system's behavior.

Prediction task

Our work related to prediction models has focused on the capture of temporal dynamics. Nevertheless, capturing spatial dynamics is an important issue that could be improved by using spatial synthesis mechanisms such as convolutional neural network structures (CNN) or attention mechanisms. Besides, the problem of graph-structured data, a growing academic theme, could be particularly relevant to consider for the scalability (graph Convolution, multi-spatial scales) of data analysis related to the transportation network.

Variance estimation

The proposed work related to contextual variance estimation is mainly exploratory. It aims at illustrating the benefit of considering the variance for contextual anomaly detection. The proposed estimation approaches mainly aim to extract the contextual variance by neglecting the short-term dynamics. However, it is possible to analyze the short-term dynamics of the prediction errors through 'autoregressive' variance estimation models.

Similarly, the deep approach based on the 'Variational Dropout' requires in-depth analysis to handle some issues (drop out qualification, network depth, contribution limitation of short-term attributes). We can build prediction models that incorporate variance estimation, such as probabilistic models and variational neural networks.

Model confidence

Prediction and variance estimation should be simultaneously considered to obtain a prediction model and provide a margin of error. We can use this error margin to estimate a confidence interval related to the data's normal behavior (statistical test) in the context of the detection of contextual anomalies. We can also use this error margin in an exploratory analysis to identify contexts with high variances and look for complementary information to explain these phenomena. Furthermore, these error margins will be useful to evaluate the reliability of the models, or even to identify weaknesses linked to specific contexts.

Anomaly characterisation

One of the main motivations of the thesis was related to the characterization of the impact of anomalies. The thesis aimed to produce context-normalized anomaly scores that filter the nominal contextual and dynamic influences in order to capture the abnormal

components. This abnormal signal can then be studied by considering different pattern-matching/clustering techniques to extract regular anomaly patterns. It would then be possible to characterize and cluster the impacts of known incidents and anomalies based on these anomaly patterns.

Anomaly detection at the variance scale

In our work, we have noticed that considering variance affects how anomalies are perceived. At first sight, abnormal signals of negligible magnitude may turn out to be extremely significant after being adjusted according to the variance of the associated context. In further research, it may be interesting to analyze anomaly scores produced with different variance considerations (as a kind of analytical filter) to characterize the nature of the observed anomalies better. In this way, we can identify extreme anomalies detected for all magnitudes of variances and characterize finer anomalies associated with specific ranges of magnitudes of variances. For example, in some applications (cybersecurity, medicine), detecting fine anomalies related to the context of very low variances could enable the identification of precursor signals preceding specific crises.

Learning with anomaly weighting

The contextual anomaly score could be incorporated in the learning to identify atypical elements and reduce their importance. Using an iterative learning mechanism with updated learning weights will specialize a model on the normal data. These nominal models can then be used as reference behavior for anomaly detection.

Deep Model perspectives

The idea that motivated our anomaly detection work was to design a stacked multi-objective LSTM-EP forecasting model for time series analysis. Firstly, the network aims to perform a contextual prediction, then a short-term forecast, and finally an anomaly prediction impact. Future research could be conducted on the weighting of learning examples based on the constructed anomaly score in order to guide the learning of the different layers - for example, by limiting the contribution of anomalies by reducing their learning weights for learning layers dedicated to nominal prediction, or conversely by increasing the weights of abnormal elements to specialize learning on atypical elements. This kind of network might better discriminate abnormalities and provide predictions that identify the influence of the main components (context, dynamic, anomalies) of time series.

Representation learning

Deep learning approaches may seem very cumbersome to deploy in view of the gain in performance they can achieve. However, as illustrated in Section 3.5.6, these models achieve a synthesis on a relevant latent embedding. This 'representation learning' may offer a rich synthesis of several aspects of the transit system state that can be exploited for visualization or decision support on complex tasks. This would facilitate the analysis and the management of large-scale transportation networks.

Closing remark

This thesis is a first step towards analyzing structured temporal data by better capturing the various influences, extracting knowledge, and learning synthetic representation, particularly in smart card data analysis. These analysis tools are the basis of 'artificial intelligence systems', which will gradually become essential tools in managing complex systems such as public transport.

Appendix

A

A.1 Synthetic Data

Contextual anomaly detection assessment requires complete knowledge of anomalies and their impact on the time series handled. To establish the foundations for our assessment protocol, we first experimented with the methodological framework on synthetic generated data with contextual anomalies. Once our framework was well defined, we applied it to time series data related to the transportation domain.

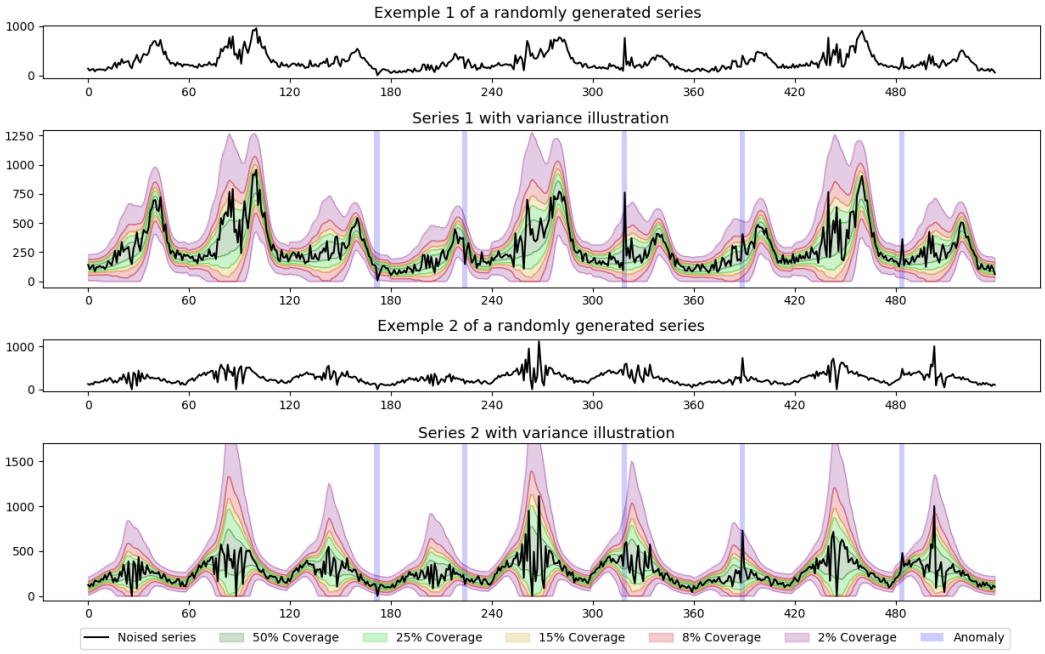

Figure A.1.: Example of curves generated on 12 “hypothetical days”

We developed a generation process for multivariate time series with a dynamic context and anomalies. Figure A.1 illustrate the behaviour of two generated data examples. For our synthetic application, we generated a time series of 8000 time steps in three dimensions (3D). The generated data look like a ridership time series with a short periodic pattern of 20 time steps (“Daily pattern”), a medium periodic pattern of (20×3) timesteps (a kind of “weekly trend”), and a long periodic trend of (20×100) timesteps (a “seasonal trend”).

The data generation process (Equation A.1) follows several steps: (i) First, we generate a regular pattern corresponding to daily trends, and we combine the regular pattern with a

combination of sinusoidal trends representing weekly and yearly influences. (ii) Then, we add a dynamic component built by multiplying the contextual component with a random daily magnitude coefficient. (iii) We introduce variability based on multiplicative Gaussian noise modulated by other regular patterns corresponding to contextual variability. (iv) Additive noise is brought in to disturb the series. (v) Finally, anomalies are generated randomly and applied through a predefined impact by considering both types of noise.

$$
\begin{aligned}
f^c &= Periodic\ pattern * Trends\ (i) \\
f^d &= f^c * Daily\ magnitude\ (ii) \\
\mathcal{E} &= (f^c + f^d) * (Noise_{mult} * Daily\ variance\ pattern) + Noise_{add}\ (iii + iv) \\
f^a &= (f^c + f^d + \mathcal{E}) * (\alpha_{mult} * Anom) + \alpha_{add} * Anom\ (iv) \\
y &= f^c + f^d + f^a + \mathcal{E}
\end{aligned}
\tag{A.1}
$$

The generation of the daily pattern is based on a mixture of three Gaussians with means, variances and magnitudes chosen randomly in a range defined by previously set parameters. We recover the histogram of the length dim_t in addition to a constant. Then, we stack dim_n times the pattern to obtain a regular series of length $dim_t * dim_n$. This process is carried out for means and variances over each dimension.

$$
\mathcal{P} = \sum_{i=1}^{3} \mathcal{N}(m \pm m_i, \sigma \pm s_i, N \pm n_i)
$$

$$
M_{day} = 0.02 + hist(PDF(\mathcal{P}^a), dim_t) \qquad M^{\sigma}_{day} = 0.5 + hist(PDF(\mathcal{P}^b), dim_t)
$$

The contextual patterns are obtained by adding two cosine harmonics with frequencies of (3 days x dim_t hours) for the weekly variability (a week of 3 days) and of (dim_n x dim_t) for the seasonal variability (a year of dim_n days). Then, a spatiotemporal convolution $Conv_c$ is applied to cross structures between dimensions.

$$
seas = 1 + \alpha_{seas} * cos(\frac{2 * \pi * (dim_n * dim_t)}{dim_n * 0.25}) + \alpha_{seas} * -sin(\frac{2 * \pi * (dim_n * dim_t)}{dim_t * 3})
$$

The daily dynamic magnitudes are simulated by multiplying each daily time step by the same random values given by the normal distribution $\mathcal{N}(1, \theta)$. Therefore, we have to make a random selection of a dimension $dim_g * dim_n$ on $\mathcal{N}(1, \theta)$. Another spatiotemporal convolution $Conv_d$ is applied here.

The additive and multiplicative noises are generated by a random drawing dim_g of length $dim_n * dim_t$ following a normal law $\mathcal{N}(0, \theta)$. The multiplicative noise is multiplied by

the variance pattern (specific to each dimension) to form the contextual variance. This simulates the unpredictable variability according to the forecasting model.

Anomalies are injected a posteriori with a random temporal position, a sign, and a magnitude (weak or strong). The anomaly impact is applied to the series. α_{a} defines the anomaly magnitudes.

The boundaries $[env_t^{bot}, env_t^{up}]$ form the confidence interval link to the normal data ratio (η). An abnormal time step (impact of the injected anomaly) has a probability of ($\alpha_{\mathrm{a}} * \eta\%$) of being outside the confidence interval.

Agorithm 2 summarizes the generation process and its different stages.

Algorithm 2 Data generation process

Gen($dim_n, dim_h, dim_d, \theta_{add}, \theta_{mult}, \theta_{add}, N_{\mathrm{a}}, \alpha_{\mathrm{a}}$):
Init :
$M_{day}, M_{day}^{\sigma}$= Gen_M() *Mean and variance, daily pattern*
$T_{\mathbf{c}}$ = Gen_T_context() *Contextual trends*
$Conv_{\mathbf{c}}, Conv_{\mathcal{D}}$= Gen_M() *Spatio-temporal convolution pattern*
$Ampl_{day}$= Gen_ampl_day(θ_{day}) *Daily dynamic* $\mathcal{N}(0, \theta)$
$\epsilon_{mult}, \epsilon_{add}$= Gen_Bruit($\theta_{mult}, \theta_{add}$) *Mult and added noise* $\mathcal{N}(0, \theta)$
I_{a} = Gen_I_anom(N_{a}) *Generated anomaly impacts*
$\phi_{mult} = percentile(mult, \frac{N_{\mathrm{a}}}{2})$ $\quad \phi_{add} = percentile(add, \frac{N_{\mathrm{a}}}{2})$

$$y^c = M_{day} * T_{\mathbf{c}} * Conv_{\mathbf{c}}$$
$$y^d = y^{\mathbf{c}} * Ampl_{day} * Conv_{\mathcal{D}}$$
$$\mathcal{E} = (y^c + y^d) * (\epsilon_{mult} * M_{day}^{\sigma}) + \epsilon_{add}$$
$$y^a = (y^c + y^d + \mathcal{E}) * (\phi_{mult} * M_{day}^{\sigma} * I_{\mathrm{a}} * \alpha_{\mathrm{a}}) + \phi_{add} * I_{\mathrm{a}} * \alpha_{\mathrm{a}}$$

$$y = y^c + y^d + y^a + \mathcal{E} \quad .$$

$Env^{mid} = y^c + y^d$
$Env^{up} = (y^c + y^d) * (1 + M_{day}^{\sigma} * \phi_{mult}) + \phi_{add}$
$Env^{bot} = (y^c + y^d) * (1 - M_{day}^{\sigma} * \phi_{mult}) - \phi_{add}$
Return($y, Env^{up}, Env^{mid}, Env^{bot}$)

A.2 Training of prediction models

Machine learning algorithms aim to approximate a decision function (Classification, Regression) whose parameters are chosen by an optimization loop of a cost function to be minimized. The objective is to build the most efficient decision function on a training sample, but also one that will be the most generalizable on unobserved data. This induces a trade-off between complexity and generalization. Their high approximation

capacity (thanks to the non-linear combination) raises the problem of ‘over-fitting". Bad parametrisation and training can lead to low generalization capacities. Some techniques related to parameter tuning and learning procedure aim to reduce ‘over-fitting’ in order to ensure generalization ability.

In addition, real data collected by industrial experiments have some quality concerns that can deteriorate the learning process of algorithms such as neural network, inducing convergence and inference issues. In addition, as we work on phenomena that evolve in time, the data distribution can change slightly over time, inducing another difficulty for generalisation.In the following sections, we will detail the learning process behind our machine learning experiments.

Ensemble models training procedure

The developments was conducted in Python. It uses some classical machine learning packages such as ‘Scikitlearn’ [Ped+11] and which contain implementations of standard machine learning models and implementation procedures.

Parameter choices were based on the Greedsearch (or Randomsearch) procedure which consists in comparing the performance of models using all (or a random part) of parameter combinations related to parameter grid values. Tables A.1 and A.2 provide the parameter grids evaluated for the random forest (RF) and gradient boosting (XBG) models. For the gradient boosting approach, an Earlystopping mechanism is also used to limit overfitting by interrupting learning in the event of test performance stagnation.

Table A.1.: Parameter grid of Random Forest

Parameter	Grid parameter values	Type of parameter
n estimators	[100,200,250,300,350]	Tree maximum number
max depth	[5,7,9,11,15]	Tree maximum depth
min samples split	[2,5,10,20]	Tree growth parameter
min impurity decrease	[0.05,0.1,0.2,0.4]	Tree growth parameter
min samples leaf	[2,5,10,20]	Leaf set Representativeness
max features	[0.7,0.8,0.9,0.95]	Features bagging meta-parameter
max samples	[0.7,0.8,0.9,0.95]	Sample bagging parameter

Table A.2.: Parameter grid of Grading Boosting on decision Tree*

Parameter	Grid parameter values	Type of parameter
n estimators	[75,150,200]	Tree maximum number
max depth	[7,10,15,20]	Tree maximum depth
learning rate	[0.1,0.05,0.01]	Weight of learning gradient
min child weight	[1,2,10,20]	Leaf set Representativeness
gamma	[0.0001,0.00001]	Tree growth parameter
subsample	[0.8,0.9,1.0]	Sample bagging parameter
colsample	[0.8,0.9 1.0]	Features bagging parameter

Finally, to guarantee the relevant complexity/generalization trade-off, tuning performance evaluations are always carried out using a sequential k=5 cross-validation procedure dedicated to handling sequential data on training sets. It consists in splitting a dataset into K temporally ordered sub-samples, evaluating the performances of models having learned on the first n samples (training base) using the N+1 sample as a test base, and aggregating performances for n=1 to k. The tuned models are then trained from scratch on the training set, and then evaluated on a test set that has never been seen by the models.

Deep learning training procedure

The neural networks and learning procedures were developed in Python using to the Tensorflow [Mar+15] environment combined with the high neural network API Keras [Cho+15]. The learning of neural networks was performed on a 24-gigabyte GPU card. The learning procedure takes a few hours, depending on the size of the data-set. To apply the deep neural network approach efficiently to industrial data collected on transportation infrastructure, certain choices related to the type of architecture (Encoder-Decoder Structure), the model parameters and the learning procedure were made.

Hyperparametrization was done manually through numerous successive experiments. The sigmoid activation functions led to better performance and faster convergence compared to the tangent functions and the Rectified Linear Unit (ReLU). The number of neurons per layer and the number of layers were chosen as a compromise between learning convergence and generalization performance. A strong dropout (10%) was added on all layers of the network to reduce over-fitting.

Learning was performed in an end-to-end way using continuous gradient propagation through the whole network. We performed mini-batch optimization of size 128 using an adaptive gradient (Nadam-optimizer [Sut+13]). Encoder-predictor models combine prediction and reconstruction objectives. The weighting of the objectives promotes reconstruction at the beginning of learning and prediction at the end of learning to accelerate convergence. We performed several iterations of learning loops that include learning rate reduction and an early stopping procedure based on stagnant test performance to fight against learning convergence issues.

A.3 Detailed architecture of deep models

LSTM encoder-predictor for irregular series

Table A.3.: Notations and variables of the irregular LSTM-EP

Notation	
t	A time step $t \in [1, T]$
$y_1, ..., y_T$	(y_t) Realization series
$S_1, ..., S_t$	(S_t) Observation sequences
$e_1, ..., e_T$	(e_t) Sequence of feature contextual vectors
$m_1, ..., m_T$	(m_t) Sequence of feature measure vectors
	Windows
W_i	$[i-k, i+k']$: Window associated to the ith observation
P_i	$[i-k, i[$: Past horizon of window W_i
F_i	$[i, i+k']$: Prediction horizon of window W_i
x_i	$(m_{P_i}, e_{P_i}, e_{F_i})$ Input features from the window W_i
	Latent space
$u_1, ..., u_T$	(u_t) Contextual representation
$h_1, ..., h_T$	(h_t) Latent past dynamic
$r_1, ..., r_T$	(r_t) Latent reconstruction state
$z_1, ..., z_T$	(z_t) Latent prediction state

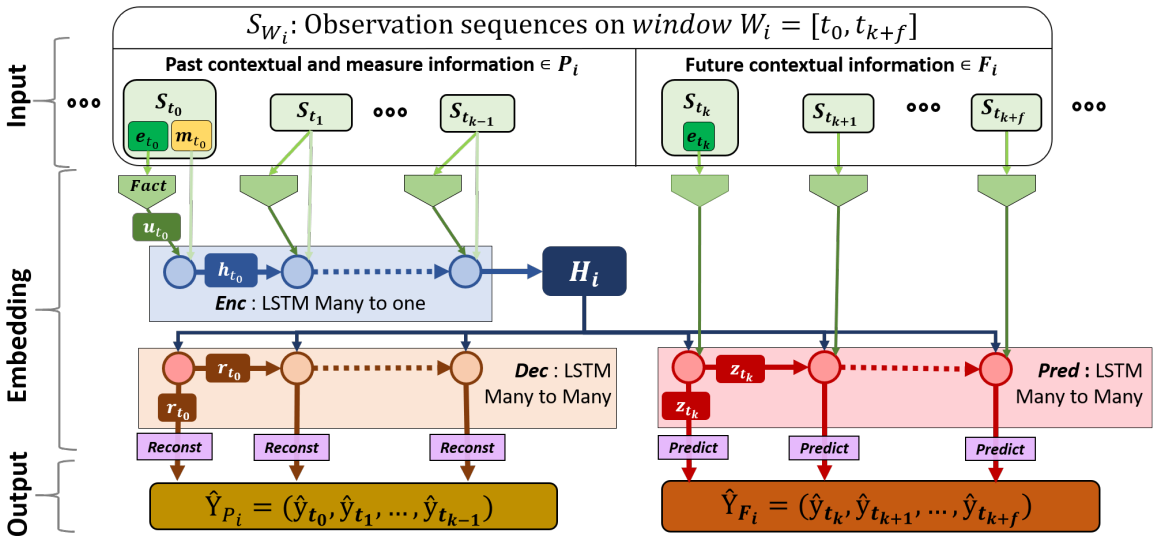

Figure A.2.: Irregular LSTM-encoder predictor with underlying structured data

Table A.4.: Layer of irregular LSTM-EP

Sub-part	layer Type	layer size	Activation function
Fact : MLP factoring contextual features	3 denses	[50,150,200]	Sigmoid
Enc : Recurrent encoder of past observation	1 LSTM	[200]	Sigmoid
Dec : Recurrent decoder of past observation	1 LSTM	[200]	Sigmoid
Pred : Recurrent predictor of future observation	1 LSTM	[200]	Sigmoid
Reconst : MLP to reconstruct past realizations	2 dense	[100,1]	Linear
Predict : MLP to predict future realizations	2 dense	[100,1]	Linear

LSTM encoder-predictor for regular series

Table A.5.: Notations and variables of the regular LSTM-EP

Notation	
t	A time step $t \in [1, T]$
$y_1, ..., y_T$	(y_t) Realization series
$x_1, ..., x_T$	(x_t) Sequence of feature contextual vectors
	Latent space (see subsection 3.1)
$u_1, ..., u_T$	(u_t) Contextual representation
$h_1, ..., h_T$	(h_t) Latent past dynamic
$z_1, ..., z_T$	(z_t) Latent prediction state

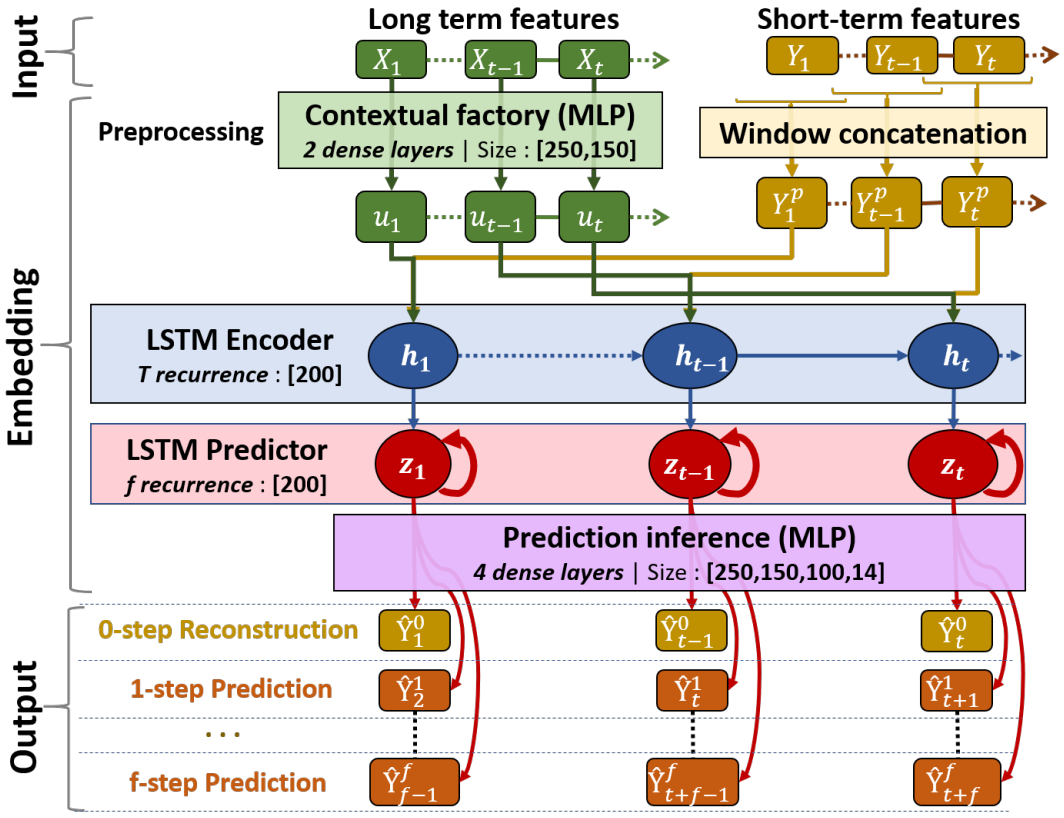

Figure A.3.: Regular LSTM-encoder predictor with underlying structured data

Table A.6.: Layer of regular LSTM-EP

Sub-part	layer Type	layer size	Activation function
Contextual factory	2 dense	[250,150]	Sigmoid
LSTM Encoder (T recurrence)	1 LSTM	[200]	Sigmoid
LSTM Predictor (f recurrence)	1 LSTM	[200]	Sigmoid
Prediction inference	4 dense	[250,150,100,14]	Linear

A.4 List of notations, equations, tables and figures

List of Notations :

Notation	Description
General : Section 3.3.1, A.3	
t	A time step $t \in [1, T]$
$(y_t) : y_1, ..., y_T$	Time series
$(x_t) : x_1, ..., x_T$	Characteristic attributes of time series observations
$(\widehat{y_t}) : \widehat{y}_1, ..., \widehat{y}_T$	Estimation (forecasting) of the time series
$\mathbf{c}_t$, $\boldsymbol{\ell}_t$, a_t	Contextual, latent and abnormal influential factors
M_t	Temporal average of the time series (Variability explained by the attributes)
f^c, f^d, f^a	Virtual components of M_t (c : contextual, d : dynamic, a : abnormal)
$\boldsymbol{\epsilon}_t \sim \mathcal{N}(B_t, \sigma_t)$	Temporal noise of the time series (unexplained variability)
$\boldsymbol{B}_t$, $\boldsymbol{\sigma}_t$	Temporal bias and variance of between noise
Forecasting : Section 3.4, A.4	
$\boldsymbol{S}_t$	Element sequence
$\boldsymbol{W}_t$, $\boldsymbol{P}_t$, $\boldsymbol{F}_t$	Temporal windows
$\boldsymbol{e}_t$, $\boldsymbol{m}_t$	Contextual (Planned) and dynamic (measured) attributes
$\boldsymbol{u}_t$, $\boldsymbol{h}_t$, $\boldsymbol{r}_t$, $\boldsymbol{z}_t$	(u_t) Latent spaces/representations
$Fact$, $Reconst$, $Predict$	Part of LSTM-EP neural network (Multilayers perceptron)
Enc, Dec, $Pred$	Part of LSTM-EP neural network (LSTM layers)
Anomaly detection : Section 4.3.1	
$\boldsymbol{r}_t$	Forecasting residue
$\boldsymbol{s}_t, \boldsymbol{s}_t^{agg}$	Local and global anomaly scores
$\boldsymbol{e}^c$, $\boldsymbol{e}^d$, $\boldsymbol{e}^a$	Error components (c : contextual, d : dynamic, a : abnormal)
$\boldsymbol{\varepsilon_t}$	Variance residue

List of Equations

List of Tables

List of Figures

Bibliography

[Aba+16] Martı n Abadi et al. “Tensorflow: A system for large-scale machine learning”. In: *OSDI Symposium on Operating Systems Design and Implementation* (2016), pp. 265–283 (cit. on p. 67).

[AL09] Bovas Abraham and Johannes Ledolter. *Statistical methods for forecasting*. Vol. 234. John Wiley & Sons, 2009 (cit. on p. 51).

[AMZ16] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. “Fraud detection system: A survey”. In: *Journal of Network and Computer Applications* 68 (2016), pp. 90–113 (cit. on p. 87).

[AP14] Hermine N Akouemo and Richard J Povinelli. “Time series outlier detection and imputation”. In: *2014 IEEE PES General Meeting*. IEEE. 2014, pp. 1–5 (cit. on p. 95).

[APK19] Ahmed Amrani, Kevin Pasini, and Mostepha Khouadjia. “Predictive Multimodal Trip Planner: A New Generation of Urban Routing Services”. In: *TRANSITDATA2019 5th International Workshop and Symposium, Paris.* 2019 (cit. on p. 16).

[APK20] Ahmed Amrani, Kévin Pasini, and Mostepha Khouadjia. “Enhance Journey Planner with Predictive Travel Information for Smart City Routing Services”. In: *2020 Forum on Integrated and Sustainable Transportation Systems (FISTS)*. IEEE. 2020, pp. 304–308 (cit. on p. 17).

[Bac+19] Danya Bachir et al. “Inferring dynamic origin-destination flows by transport mode using mobile phone data”. In: *Transportation Research Part C: Emerging Technologies* 101 (2019), pp. 254–275 (cit. on pp. 28, 32).

[BBC18] Seif-Eddine Benkabou, Khalid Benabdeslem, and Bruno Canitia. “Unsupervised outlier detection for time series by entropy and dynamic time warping”. In: *Knowledge and Information Systems* 54.2 (2018), pp. 463–486 (cit. on pp. 91, 92).

[BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828 (cit. on pp. 46, 50, 65).

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152 (cit. on p. 52).

[Bre+84] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984 (cit. on p. 52).

[Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 53).

[Bri+17] Anne-Sarah Briand et al. “Analyzing year-to-year changes in public transport passenger behaviour using smart card data”. In: *Transportation Research Part C: Emerging Technologies* 79 (2017), pp. 274–289 (cit. on pp. 28, 32).

[Bri+19] Anne-Sarah Briand et al. “Detection of atypical events on a public transport network using smart card data”. In: *European Transport Conference 2019Association for European Transport (AET)*. 2019 (cit. on pp. 31, 32, 91, 92).

[Cao+17] Nan Cao et al. “Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data”. In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 23–33 (cit. on p. 90).

[Car19] Léna Carel. “Big data analysis in the field of transportation”. PhD thesis. Université Paris-Saclay, 2019 (cit. on p. 96).

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58 (cit. on pp. 87, 90).

[CC19] Raghavendra Chalapathy and Sanjay Chawla. “Deep learning for anomaly detection: A survey”. In: *arXiv preprint arXiv:1901.03407* (2019) (cit. on p. 90).

[CGR02] Roberto Camagni, Maria Cristina Gibelli, and Paolo Rigamonti. “Urban mobility and urban form: the social and environmental costs of different patterns of urban expansion”. In: *Ecological economics* 40.2 (2002), pp. 199–216 (cit. on p. 8).

[Cha09] Varun Chandola. “Anomaly detection for symbolic sequences and time series data”. PhD thesis. University of Minnesota, 2009 (cit. on p. 90).

[Che+09] Haibin Cheng et al. “Detection and characterization of anomalies in multivariate time series”. In: *Proceedings of the 2009 SIAM international conference on data mining*. SIAM. 2009, pp. 413–424 (cit. on pp. 91, 93).

[Cho+15] Francois Chollet et al. *Keras*. 2015 (cit. on pp. 67, 137).

[Cho+16] Edward Choi et al. “Retain: An interpretable predictive model for healthcare using reverse time attention mechanism”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 3504–3512 (cit. on p. 90).

[Cho+17] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *EMNLP Empirical Methods in Natural Language Processing* (2017), pp. 1724–1734 (cit. on pp. 50, 57, 64).

[CSC12] Irina Ceapa, Chris Smith, and Licia Capra. “Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data”. In: *ACM SIGKDD International Workshop on Urban Computing* (2012), pp. 134–141 (cit. on pp. 30, 32, 47).

[Cui+16] Chunsheng Cui et al. “Fuzzy multivariate narx model for passenger entrance flow prediction in the shanghai subway system”. In: *Journal of Intelligent & Fuzzy Systems* 31.6 (2016), pp. 3047–3054 (cit. on pp. 29, 47, 57).

[DF13] Zhiguo Ding and Minrui Fei. “An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window”. In: *IFAC Proceedings Volumes* 46.20 (2013), pp. 12–17 (cit. on pp. 91, 93).

[Dim+17] Giorgos Dimopoulos et al. “Detecting network performance anomalies with contextual anomaly detection”. In: *2017 IEEE International Workshop on Measurement and Networking (M&N)*. IEEE. 2017, pp. 1–6 (cit. on p. 93).

[Din+16] Chuan Ding et al. “Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees”. In: *Sustainability* 8.11 (2016), p. 1100 (cit. on pp. 29, 32, 47–49, 54, 57).

[Din+17] Chuan Ding et al. “Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility”. In: *IEEE Transactions on Intelligent Transportation Systems* 19.4 (2017), pp. 1054–1064 (cit. on pp. 47, 48, 51).

[Dru+97] Harris Drucker et al. “Support vector regression machines”. In: *Advances in neural information processing systems*. 1997, pp. 155–161 (cit. on p. 52).

[EB20] Oscar Egu and Patrick Bonnel. “How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? An empirical investigation in Lyon”. In: *Transportation Research Part A: Policy and Practice* 138 (2020), pp. 267–282 (cit. on pp. 28, 32).

[El +17] Samy El Tawab et al. “Data analysis of transit systems using low-cost IoT technology”. In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. 2017, pp. 497–502 (cit. on pp. 27, 32).

[Fer+19] Len Feremans et al. “Pattern-based anomaly detection in mixed-type time series”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 240–256 (cit. on pp. 91, 93).

[Fri01] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232 (cit. on p. 53).

[GG16] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*. 2016, pp. 1050–1059 (cit. on pp. 96, 102, 105).

[GL16] Aditya Grover and Jure Leskovec. “node2vec: Scalable feature learning for networks”. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864 (cit. on p. 60).

[GSC99] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. “Learning to forget: Continual prediction with LSTM”. In: *ICANN International Conference on Artificial Neural Networks* (1999) (cit. on p. 48).

[Guo+18] Yifan Guo et al. “Multidimensional time series anomaly detection: A gru-based gaussian mixture variational autoencoder approach”. In: *Asian Conference on Machine Learning*. 2018, pp. 97–112 (cit. on pp. 91, 96).

[Hab+19] Riyaz Ahamed Ariyaluran Habeeb et al. “Real-time big data processing for anomaly detection: A survey”. In: *International Journal of Information Management* 45 (2019), pp. 289–307 (cit. on p. 90).

[Ham20] James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020 (cit. on p. 51).

[Has+14] Md Al Mehedi Hasan et al. “Support vector machine and random forest modeling for intrusion detection system (IDS)”. In: *Journal of Intelligent Learning Systems and Applications* 2014 (2014) (cit. on p. 95).

[HC14] Michael A Hayes and Miriam AM Capretz. “Contextual anomaly detection in big sensor data”. In: *2014 IEEE International Congress on Big Data*. IEEE. 2014, pp. 64–71 (cit. on p. 91).

[He+19] Mingyi He et al. “Pattern and anomaly detection in urban temporal networks”. In: *arXiv preprint arXiv:1912.01960* (2019) (cit. on pp. 31, 32, 91, 92).

[Hey+18] Léonie Heydenrijk-Ottens et al. “Supervised learning: Predicting passenger load in public transport”. In: *CASPT Conference on Advanced Systems in Public Transport and TransitData* (2018) (cit. on pp. 30, 32, 47–49, 54).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 56).

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009 (cit. on p. 51).

[Hun+18] Kyle Hundman et al. “Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 387–395 (cit. on pp. 87, 91, 96).

[Jen19] Erik Jenelius. “Data-driven metro train crowding prediction based on real-time load data”. In: *IEEE Transactions on Intelligent Transportation Systems* 21.6 (2019), pp. 2254–2265 (cit. on pp. 30, 32, 47, 54).

[JFG17] Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. “Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore”. In: *IEEE Transactions on Big Data* 3.2 (2017), pp. 208–219 (cit. on pp. 28, 32).

[Ke+17] Jintao Ke et al. “Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach”. In: *Transportation Research Part C: Emerging Technologies* 85 (2017), pp. 591–608 (cit. on pp. 47, 48, 50, 51, 54, 57).

[KK13] Rolands Kromanis and Prakash Kripakaran. “Support vector regression for anomaly detection from measurement histories”. In: *Advanced Engineering Informatics* 27.4 (2013), pp. 486–495 (cit. on p. 95).

[KKK16] Hiroyuki Kasai, Wolfgang Kellerer, and Martin Kleinsteuber. “Network volume anomaly detection and identification in large-scale networks based on online time-structured traffic tensor tracking”. In: *IEEE Transactions on Network and Service Management* 13.3 (2016), pp. 636–650 (cit. on p. 95).

[Kon+18] Daniel Kondor et al. “Towards matching user mobility traces in large-scale datasets”. In: *IEEE Transactions on Big Data* (2018) (cit. on pp. 27, 32).

[KTA14] Akira Kinoshita, Atsuhiro Takasu, and Jun Adachi. “Traffic Incident Detection Using Probabilistic Topic Model.” In: *EDBT/ICDT Workshops*. Citeseer. 2014, pp. 323–330 (cit. on pp. 91, 95).

[KW14] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*. 2014 (cit. on p. 96).

[LB+95] Yann LeCun, Yoshua Bengio, et al. “Convolutional networks for images, speech, and time series”. In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995 (cit. on p. 60).

[LBE17] Pierre-Antoine Laharotte, Romain Billot, and Nour-Eddin El Faouzi. “Detection of non-recurrent road traffic events based on clustering indicators.” In: *ESANN*. 2017 (cit. on pp. 91, 92).

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444 (cit. on p. 55).

[LC11] Neal Lathia and Licia Capra. “Mining Mobility Data to Minimise Travellers’ Spending on Public Transport”. In: *17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2011), pp. 1181–1189 (cit. on pp. 28, 32).

[Li+17] Yang Li et al. “Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks”. In: *Transportation Research Part C: Emerging Technologies* 77 (2017), pp. 306–328 (cit. on pp. 30, 32).

[Liu+17] Qi Liu et al. “Unsupervised detection of contextual anomaly in remotely sensed data”. In: *Remote Sensing of Environment* 202 (2017), pp. 75–87 (cit. on p. 91).

[LPJ17] Jinbo Li, Witold Pedrycz, and Iqbal Jamal. “Multivariate time series anomaly detection: A framework of Hidden Markov Models”. In: *Applied Soft Computing* 60 (2017), pp. 229–240 (cit. on pp. 47, 49, 54, 57, 95).

[LTZ08] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation forest”. In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE. 2008, pp. 413–422 (cit. on p. 93).

[LTZ12] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation-based anomaly detection”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.1 (2012), pp. 1–39 (cit. on p. 93).

[Lu+17] Shiwei Lu et al. “Understanding the representativeness of mobile phone location data in characterizing human mobility indicators”. In: *ISPRS International Journal of Geo-Information* 6.1 (2017), p. 7 (cit. on pp. 27, 32).

[Ma+17] Xiaolei Ma et al. “Understanding commuting patterns using transit smart card data”. In: *Journal of Transport Geography* 58 (2017), pp. 135–145 (cit. on pp. 29, 32).

[Mal+15] Pankaj Malhotra et al. “Long short term memory networks for anomaly detection in time series”. In: *Proceedings*. Vol. 89. Presses universitaires de Louvain. 2015 (cit. on pp. 95, 97).

[Mal+16] Pankaj Malhotra et al. “LSTM-based encoder-decoder for multi-sensor anomaly detection”. In: *Anomaly Detection Workshop of the 33rd International Conference on Machine Learning (ICML 2016)*. 2016 (cit. on pp. 91, 95).

[Mar+15] Abadi Martin et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015 (cit. on p. 137).

[McN07] Michael G McNally. *The four-step model*. Emerald Group Publishing Limited, 2007 (cit. on p. 46).

[Mei06] Nicolai Meinshausen. “Quantile regression forests”. In: *Journal of Machine Learning Research* 7.Jun (2006), pp. 983–999 (cit. on pp. 96, 102, 104).

[Mik+13] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013) (cit. on p. 60).

[MP12] Marcela A Munizaga and Carolina Palma. “Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile”. In: *Transportation Research Part C: Emerging Technologies* 24 (2012), pp. 9–18 (cit. on pp. 27, 32).

[Mun+18] Mohsin Munir et al. “DeepAnT: A deep learning approach for unsupervised anomaly detection in time series”. In: *IEEE Access* 7 (2018), pp. 1991–2005 (cit. on pp. 91, 96).

[Nak+20] Takaaki Nakamura et al. “MERLIN: Parameter-Free Discovery of Arbitrary Length Anomalies in Massive Time Series Archives”. In: *2020 IEEE 16th international conference on data mining (ICDM)*. Ieee. 2020 (cit. on pp. 91, 94).

[NHG16] Ming Ni, Qing He, and Jing Gao. “Forecasting the subway passenger flow under event occurrences with social media”. In: *IEEE Transactions on Intelligent Transportation Systems* 18.6 (2016), pp. 1623–1632 (cit. on pp. 47, 48, 51, 54).

[Pas+19a] Kevin Pasini et al. “Forecasting passenger load in a transit network using data driven models”. In: *WCRR 2019, 12th World Congress on Railway Research*. 2019 (cit. on pp. 15, 17).

[Pas+19b] Kevin Pasini et al. “LSTM encoder-predictor for short-term train load forecasting”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham. 2019, pp. 535–551 (cit. on pp. 15, 17, 32, 47, 50, 57, 87, 97, 108).

[Pas+19c] Kevin Pasini et al. “Modèle LSTM encodeur-prédicteur pour la prévision court-terme de l’affluence dans les transports collectifs”. In: *CAP 2019, Conférence sur l’Apprentissage Automatique*. 2019 (cit. on pp. 15, 17).

[Pas+19d] Kevin Pasini et al. “Representation Learning of public transport data. Application to event detection”. In: *5th International Workshop and Symposium TransitData 2019*. 2019 (cit. on pp. 16, 17).

[Pas+20a] Kevin Pasini et al. “Contextual anomaly detection on time series: A case study of metro ridership analysis”. 2020 (cit. on pp. 16, 17).

[Pas+20b] Kevin Pasini et al. “Modèle LSTM Encodeur-Prédicteur pour la Prévision Court-Terme de la Charge Passagers dans les Transports en Commun”. Workshop virtuel - Nouvelles méthodes pour l’analyse descriptive et prédictive de données massives et structurées. 2020 (cit. on p. 17).

[Pas+21] Kevin Pasini et al. “Anomaly detection on time series evolving in a dynamic context. Application to smart card data analysis”. To be submitted to ESANN. 2021 (cit. on pp. 16, 17).

[Ped+11] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 136).

[Pou+15] Mickaël Poussevin et al. “Mining ticketing logs for usage characterization with nonnegative matrix factorization”. In: *Big Data Analytics in the Social and Ubiquitous Context*. Springer, 2015, pp. 147–164 (cit. on pp. 29, 32).

[RBG16] Jérémy Roos, Stéphane Bonnevay, and Gérald Gavin. “Short-Term Urban Rail Passenger Flow Forecasting: A Dynamic Bayesian Network Approach”. In: *ICMLA International Conference on Machine Learning and Applications* (2016), pp. 1034–1039 (cit. on pp. 29, 32, 47, 48, 50, 51).

[RT+14] Jamal Raiyn, Tomer Toledo, et al. “Real-time road traffic anomaly detection”. In: *Journal of Transportation Technologies* 4.03 (2014), p. 256 (cit. on pp. 91, 95).

[Sal+14] Osman Salem et al. “Anomaly detection in medical wireless sensor networks using SVM and linear regression models”. In: *International Journal of E-Health and Medical Communications (IJEHMC)* 5.1 (2014), pp. 20–45 (cit. on p. 95).

[Sch+17] Thomas Schlegl et al. “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery”. In: *International conference on information processing in medical imaging*. Springer. 2017, pp. 146–157 (cit. on p. 87).

[Sim+14] Ralph Sims et al. *2014: Transport.* Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 2014 (cit. on p. 8).

[SKO16] Neveen Shlayan, Abdullah Kurkcu, and Kaan Ozbay. "Exploring pedestrian Bluetooth and WiFi detection at public transportation terminals". In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2016, pp. 229–234 (cit. on pp. 27, 32).

[Smo+20] Daniel Smolyak et al. "Coupled IGMM-GANs with Applications to Anomaly Detection in Human Mobility Data". In: *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 6.4 (2020), pp. 1–14 (cit. on pp. 91, 94).

[SMS15] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. "Unsupervised learning of video representations using lstms". In: *International conference on machine learning*. 2015, pp. 843–852 (cit. on pp. 57, 65).

[SSC16] Grzegorz Sierpiński, Marcin Staniek, and Ireneusz Celiński. "Travel behavior profiling using a trip planner". In: *Transportation Research Procedia* 14 (2016), pp. 1743–1752 (cit. on pp. 25, 32).

[Str+07] Carolin Strobl et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1 (2007), p. 25 (cit. on p. 76).

[Sut+13] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning." In: *ICML International Conference on Machine Learning* 28.1139-1147 (2013), p. 5 (cit. on pp. 67, 137).

[TCA05] Martin Trépanier, Robert Chapleau, and Bruno Allard. "Can trip planner log files analysis help in transit service planning?" In: *Journal of Public Transportation* 8.2 (2005), p. 5 (cit. on pp. 25, 32).

[TLW09] Tsung-Hsien Tsai, Chi-Kang Lee, and Chien-Hung Wei. "Neural network based temporal feature models for short-term railway passenger demand forecasting". In: *Expert Systems with Applications* 36.2 (2009), pp. 3728–3736 (cit. on pp. 47, 48).

[Ton+18] Emeric Tonnelier et al. "Anomaly detection in smart card logs and distant evaluation with Twitter: a robust framework". In: *Neurocomputing* 298 (2018), pp. 109–121 (cit. on pp. 31, 32, 91, 95).

[Toq+16] Florian Toqué et al. "Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks". In: *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)*. IEEE. 2016, pp. 1071–1076 (cit. on pp. 27, 32, 45).

[Toq+18] Florian Toqué et al. "Short-Term Multi-Step Ahead Forecasting of Railway Passenger Flows During Special Events With Machine Learning Methods". In: *CASPT 2018, Conference on Advanced Systems in Public Transport and TransitData 2018*. 2018 (cit. on pp. 30, 32, 47, 48, 50, 51, 54, 97).

[TTC07] Martin Trépanier, Nicolas Tranchant, and Robert Chapleau. "Individual trip destination estimation in a transit smart card automated fare collection system". In: *Journal of Intelligent Transportation Systems* 11.1 (2007), pp. 1–14 (cit. on pp. 27, 32).

[Val+] Flore Vallet et al. "Outil de visualisation pour l'analyse du trafic et des flux voyageurs dans un réseau de transport multimodal". In: *48ème Congrès ATEC ITS France* (cit. on p. 16).

[Wit+13] Apichon Witayangkurn et al. “Anomalous event detection on large-scale gps data from mobile phones using hidden markov model and cloud platform”. In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 2013, pp. 1219–1228 (cit. on pp. 91, 92).

[WT16] Yuankai Wu and Huachun Tan. “Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework”. In: *arXiv preprint arXiv:1612.01022* (2016) (cit. on pp. 47, 48, 54, 57).

[Xin+15] SHI Xingjian et al. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting”. In: *NIPS Advances in neural information processing systems* (2015), pp. 802–810 (cit. on pp. 48, 57).

[Yan+19] Xin Yang et al. “Concepts of artificial intelligence for computer-assisted drug discovery”. In: *Chemical reviews* 119.18 (2019), pp. 10520–10594 (cit. on p. 54).

[Yao+18] Huaxiu Yao et al. “Deep multi-view spatial-temporal network for taxi demand prediction”. In: *arXiv preprint arXiv:1802.08714* (2018) (cit. on pp. 47, 49–51, 54, 57).

[Yeh+16] Chin-Chia Michael Yeh et al. “Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets”. In: *2016 IEEE 16th international conference on data mining (ICDM)*. Ieee. 2016, pp. 1317–1322 (cit. on pp. 91, 94).

[YKR08] Dragomir Yankov, Eamonn Keogh, and Umaa Rebbapragada. “Disk aware discord discovery: finding unusual time series in terabyte sized datasets”. In: *Knowledge and Information Systems* 17.2 (2008), pp. 241–262 (cit. on pp. 91, 94).

[YLC17] Yang Yu, Jun Long, and Zhiping Cai. “Network intrusion detection through stacking dilated convolutional autoencoders”. In: *Security and Communication Networks* 2017 (2017) (cit. on p. 87).

[YYZ17] Bing Yu, Haoteng Yin, and Zhanxing Zhu. “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting”. In: *arXiv preprint arXiv:1709.04875* (2017) (cit. on pp. 47, 51, 54, 57).

[Zen+18] Houssam Zenati et al. “Efficient gan-based anomaly detection”. In: *Workshop track - ICLR 2018* (2018) (cit. on p. 94).

[Zha+15] Fang Zhao et al. “Exploratory analysis of a smartphone-based travel survey in Singapore”. In: *Transportation Research Record* 2494.1 (2015), pp. 45–56 (cit. on pp. 28, 32, 57).

[Zha+17] Juanjuan Zhao et al. “Spatio-temporal analysis of passenger travel patterns in massive smart card data”. In: *IEEE Transactions on Intelligent Transportation Systems* 18.11 (2017), pp. 3135–3146 (cit. on pp. 31, 32).

[Zha+19] Xing Zhao et al. “Clustering analysis of ridership patterns at subway stations: A case in Nanjing, China”. In: *Journal of Urban Planning and Development* 145.2 (2019) (cit. on pp. 29, 32).

[Zhe+14] Yu Zheng et al. “Urban computing: concepts, methodologies, and applications”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.3 (2014), pp. 1–55 (cit. on p. 10).

[Zia+17] Ali Ziat et al. “Spatio-temporal neural networks for space-time series forecasting and relations discovery”. In: *IEEE International Conference on Data Mining ICDM* (2017), pp. 705–714 (cit. on pp. 45, 49, 57).

[ZKZ18] Zhan Zhao, Haris N Koutsopoulos, and Jinhua Zhao. “Individual mobility prediction using transit smart card data”. In: *Transportation research part C: emerging technologies* 89 (2018), pp. 19–34 (cit. on p. 28).

[ZL17] Lingxue Zhu and Nikolay Laptev. “Deep and confident prediction for time series at uber”. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2017, pp. 103–110 (cit. on pp. 87, 96, 102, 105).

[Zmu+13] Johanna Zmud et al. *Transport survey methods: Best practice for decision making*. Emerald Group Publishing, 2013 (cit. on pp. 28, 32).

[ZZQ17] Junbo Zhang, Yu Zheng, and Dekang Qi. “Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction.” In: *AAAI Association for the Advancement of Artificial Intelligence* (2017), pp. 1655–1661 (cit. on pp. 45, 47, 49–51, 57).

[0]This thesis was typeset with Latex using the *Clean Thesis* style developed by Ricardo Langner.
`http://cleanthesis.der-ric.de/`