Neural Computing and Applications (2022) 34:1483–1507
https://doi.org/10.1007/s00521-021-06455-z

ORIGINAL ARTICLE

Contextual anomaly detection on time series: a case study of metro ridership analysis

Kevin Pasini[1,2] · Mostepha Khouadjia[1] · Allou Samé[2] · Martin Trépanier[3] · Latifa Oukhellou[2]

Received: 25 February 2021 / Accepted: 26 August 2021 / Published online: 9 September 2021

Abstract

The increase in the amount of data collected in the transport domain can greatly benefit mobility studies and create high value-added mobility information for passengers, data analysts, and transport operators. This work concerns the detection of the impact of disturbances on a transport network. It aims, from smart card data analysis, to finely quantify the impacts of known disturbances on the transportation network usage and to reveal unexplained statistical anomalies that may be related to unknown disturbances. The mobility data studied take the form of a multivariate time series evolving in a dynamic environment with additional contextual attributes. The research mainly focuses on contextual anomaly detection using machine learning models. Our main goal is to build a robust anomaly score to highlight statistical anomalies (contextual extremums), considering the variability within the time series induced by the dynamic context. The robust anomaly score is built from normalized forecasting residuals. The normalization of the residuals is carried out using the estimated contextual variance. Indeed, there are complex dynamics on both the mean and the variance in the ridership time series induced by the flexible transportation schedule, the variability in transport demand, and contextual factors such as the station location and the calendar information. Therefore, they should be considered by the anomaly detection approach to obtain a reliable anomaly score. We investigate several prediction models (including an LSTM encoder–decoder of the recurrent neural network deep learning family) and several variance estimators obtained through dedicated models or extracted from prediction models. The proposed approaches are evaluated on synthetic data and real data from the smart card riderships of the Quebec Metro network. It includes a basis of events and disturbances that have impacted the transport network. The experiments show the relevance of variance normalization on prediction residuals to build a robust anomaly score under a dynamic context.

Keywords Contextual anomaly detection · Forecasting · Machine learning · Multivariate time series · Recurrent neural network

1 Introduction

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 are often missing 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 specific methodological challenges.

In our case, 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

Git-lab of experiments on synthetic data: https://gitlab.com/Haroke/contextual-anomaly-detection.

✉ Latifa Oukhellou
latifa.oukhellou@univ-eiffel.fr

[1] Institut de Recherche Technologique IRT-SystemX, Paris-Saclay, France

[2] Université Gustave Eiffel, Cosys-Grettia, Champs-sur-Marne, France

[3] Polytechnique Montréal, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport (CIRRELT), Montreal, Canada

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 could be calendar (hour, day, season, etc.) or spatial related due to the geographical properties of the network (incoming /outgoing transit population, employment area, land use, etc.). Therefore, we consider a theoretical framework that analyzes the series with additional information on some "observed" influencing 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.

This work addresses the problem of contextual anomaly detection on ridership time series evolving in a dynamic context. Figure 1 illustrates the issue that can be summarized as follows: how to accurately quantify the abnormal character of an observation y_t of a time series while considering its contextual variability.

In this figure, the mean (green curve) and the variance (shown by the yellow envelope) depend on the dynamic context, which implies that they are structured by a set of latent factors related to a calendar or other dynamic contexts. The main idea investigated in the paper is to exploit the relationship between residuals and variances to provide an abnormality marker of an observation y_t. Due to the dynamic context, this task requires estimating the contextual means and variances, which is a difficult task. The proposed approach is based on three standard forecasting concepts, which we refer to in the rest of the article as follows:

- "Contextual average" performs local averages for data subsets that share a similar context (i.e., with similar input attributes).
- "Contextual influence capture" means building a model that infers a target variable from input attributes (i.e., explanatory variables) related to calendar information in our case study.
- "Short-term dynamics capture" corresponds to an autoregressive model estimating a target value from its past observations.

Therefore, this work addresses the tasks of estimating contextual means $\hat{y}_t$ and variances σ_t that require capturing the influence of many known and hidden factors. Such contextual statistics are valuable information that allow a better distinction between anomalies and usual data noise. Thus, these statistics are used to build context-normalized deviations of observations. 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?

Different methodologies can be applied to perform anomaly detection in time series [1]. There are a variety of approaches to either capture normal behavior to extract a measure of deviation or to isolate atypical elements using various similarity/distance measures (Euclidean, strings, dynamic time warping, pairwise, entropy, etc.) and tools (clustering, distance profiles, pattern matching, forecasting). We can list some recent works carried out using clustering techniques [2, 3], pairwise subsequences [4, 5], isolation forest [6, 7], matrix decomposition [8], or prediction residuals [9, 10].

Whereas most of the extant work performs agnostic anomaly detection on a series, our work concerns contextual anomaly detection on a series with a dynamic context partly available through contextual attributes. Our previous research focused on short-term prediction models applied to train ridership series [11]. More precisely, the work aims to capture contextual influences and short-term dynamics on structured data with dynamic context using a neural network LSTM encoder–decoder. Following this previous work, the proposed approach is positioned on the paradigm of contextual anomaly detection based on prediction residuals.

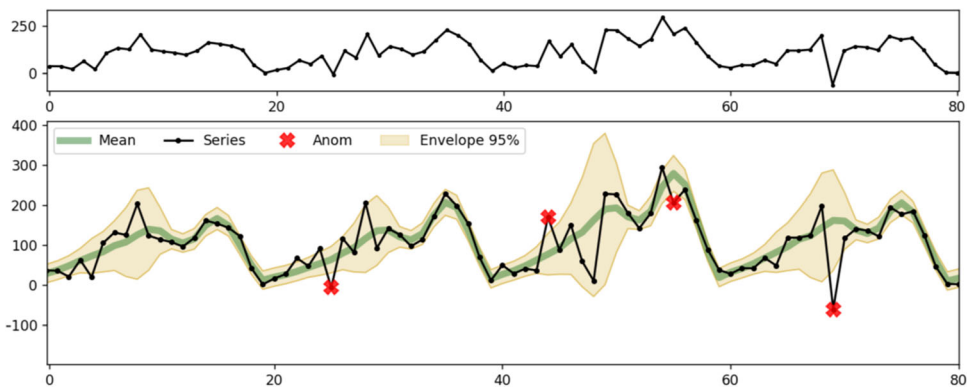

Fig. 1 Raw time series (top) and time series with the mean, the variance, and the anomalies (bottom)

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 feed the dynamic context are complex. Indeed, they are nonlinear, dynamic (changing over time), intermix (interacting with each other), and only partially observed through contextual attributes. The goal is to approximate the influence of the underlying dynamic context on the series through time-varying statistical measures. Our approach starts by forecasting a target variable (ridership in the station) $\boldsymbol{y}_t$ by $\hat{\boldsymbol{y}}_t$ using a dynamic prediction model. We then assume that the forecasting residuals ($\boldsymbol{r}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$) follow a Gaussian distribution $\mathcal{N}(\boldsymbol{B}_t, \boldsymbol{\sigma}_t)$ whose parameters evolve over time according to the dynamic context. Our contribution lies in the dynamic estimation of residual mean B_t and variance σ_t using contextual attributes. These quantities $(\boldsymbol{r}_t, \boldsymbol{B}_t, \boldsymbol{\sigma}_t)$ are then used to build a context-robust anomaly score based on prediction residuals normalized by contextual variance. 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 addressed issue 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.

Our 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 residual 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 attendance, as well as information about events and incidents. This allows us to evaluate and highlight our anomaly scores regarding these events and incidents.

This paper is structured as follows: Section 2 reviews the state-of-the-art techniques in anomaly detection applied to multivariate time series in dynamic contexts. Then, in Sect. 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 proposed to tackle this problem. Section 4 is dedicated to the evaluation of the proposed methodology on synthetic and 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, we define an experimental protocol that highlights the impacts of incidents and events on transit network ridership data.

2 Related work

2.1 Time series anomaly detection literature

This paper concerns the field of contextual anomaly detection, which aims to qualify the abnormal behavior in a time series. This is a rather complex issue to solve due to the definition of normality, the sparsity and non-redundancy of abnormal observations, the cost of data labeling, and the consideration of the structure of the studied data (sequences, spatial structures, graphs, text, or images). Moreover, the choice of an anomaly detection approach is often application-oriented. A wide variety of application areas are involved, including mobility [12], cybersecurity [13], industry [14], medicine [15], and fraud [16]. Other more complex objectives are emerging, notably related to the interpretability of detection systems [17] and the characterization and impact of anomaly forecasting [18]. The anomaly detection field is summarized by a review of the literature ([1, 19]) that structures and details the various issues by distinguishing between several forms of anomalies, as follows:

- Anomaly points that deviate individually from the static distribution of the dataset.
- Contextual anomalies corresponding to an observation that is considered abnormal with respect to a set of observations sharing common properties, for example, belonging to the same temporal or spatial neighborhood.
- Collective anomalies made up of several observations that form an atypical pattern with respect to the rest of the data. Examples include a set of points forming an irregular pattern within a time series and a community of nodes in a graph with abnormal properties.

More recent surveys are dedicated to more specific subjects with, for instance, the analysis of work on real-time big data [20] or different approaches based on a deep learning framework [21] and applied to road transportation networks.

This paper addresses unsupervised contextual anomaly detection in a multivariate time series. In this framework,

the issues addressed pertain to the time series representation, the similarity and deviation distances, and the paradigm used. Naturally, a series takes the form of a continuous value sequence, but it is possible to modify or enrich its representation to facilitate the handling, for example, by considering the set of regular subsequences, by discretizing the continuous values to obtain a sequence of tokens, by breaking the structure to work with a set of independent elements based on windowed attributes, or by enriching a series or an ensemble set with contextual information. Likewise, element comparison requires the use of an appropriate similarity measure or distance. The most used standard measure is often the Euclidean distance, but many works are interested in the use of other measures with interesting properties related to sequence handling (string metrics, dynamic time warping), variance consideration (Mahalanobis distance), or measuring the similarity of an element or a set with a set (entropy, Wasserstein distance). We can distinguish three main paradigms, as follows:

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 [2], the authors perform outlier detection on power-smith sensor data by computing the difference from average contextual daily profiles on massive data. These daily profiles are obtained by parallelizable "K-means" clustering, which is designed to handle massive data. The authors in [22] propose constructing a similarity graph using a radial basis function (RBF) kernel between the subsequences of the multivariate series. The similarity is expressed as a random walk on the neighborhood graph. The proposed approach obtains an average AUC performance (area under the curve) of 0.9 on the anomaly detection benchmark related to several time series datasets. The authors of [23] propose achieving traffic anomaly detection by clustering based on the K-nearest neighbor (KNN) approach using the dynamic time warping (DTW) distance between subsequences of the time series. The performances depend on anomaly durations with an F1-score greater than 0.6 for anomalies over several hours. In [3], the authors propose an approach based on iterative optimization to perform spectral clustering using the dynamic time warping distance with an actualized weight to estimate the contribution of anomalous observations. The proposed approach outperforms several competitors (one-class SVM, local outlier factors) with an average AUC of 0.81 on the time series classification benchmark of 30 datasets adapted to the anomaly detection framework.

Distance-based isolation methods

In contrast to the previous paradigm, these methods aim to isolate the atypical elements that are often presented as regular subsequences on the basis of a similarity measure of an element or a set of elements with others. Some methods propose applying the Isolation forest concept [24, 25] to discriminate elements based on their features by storing them in binary trees with an entropy measure. The authors in [6] apply isolation forests to time series window attributes to compute an anomaly score. The approach leads to an average AUC of 0.89 on four datasets (2 cyberattack, one forest density, and one shuttle sensor). More recently, [7] perform an isolation forest on features extracted from time series processing. The approach outperforms competitors such as matrix profiles or local outlier factors. The authors consider datasets of different time series types (univariate, multivariate, mixed type) and domains (water/electrical/human activities/temperatures sensor). They obtained an average AUC of 0.811 (univariate), 0.911 (multivariate), and 0.85 (mixed types) by type of time series. Other works consider subsequence comparisons to detect atypical elements mainly on massive raw time series without contextual information as big data sensor applications, from their 1-nearest neighbor distance. Strategies have been built to reduce the computational cost, such as computing the matrix profile with a fast Fourier technique [4] to avoid calculating all the subsequence distances with the discord-based approach [5, 26].

Methods based on prediction/reconstruction residuals

The last paradigm is based explicitly 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.

2.2 Methods based on prediction residuals

The underlying idea consists of using the residuals of a prediction or reconstruction model to highlight the anomalies occurring in a time series. Forecasting models capture frequent recurrent 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 works in this domain have applied several types of predictive models, such as the autoregressive integrated moving average (ARIMA) model [27], probabilistic models such as hidden Markov Models (HMM)[28], multiple linear regression models [29], machine learning models, such as support vector machines (SVMs) [30], random forests [31], or the matrix-decomposition reconstruction technique [8, 32].

The recent research has investigated recurrent neural networks by exploiting predictive residuals to detect anomalies in time series. The authors in [9] were the first to use predictive residuals from recurrent models for anomaly detection. The residuals are assumed to follow a Gaussian distribution with a nonzero mean. Anomaly detection is performed by using a threshold on the anomaly score based on the likelihood that these residuals will occur.More recently, the authors have extended their proposal by using an LSTM encoder–decoder as a predictive model [33]. The last model is evaluated on various sensor datasets (space shuttle, industrial engine/power, or vital signal) with average (precision, recall) scores of (0.96,0.18). Moreover, in [34], the authors propose applying a convolutional neural network (CNN) to obtain root squared residuals used as anomaly scores. The proposed approach equals or outperforms 15 other methods on 10 different time series datasets related to road traffic, network utilization, online advertisement, net traffic, space shuttle, and health. The authors of [10] propose 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 a Gaussian mixture. The proposed approach outperformed several competitors (EM-GMM and several GRU architectures), providing between 5.7% and 7.2% improvements in accuracy and F1 score on room sensor (light, humidity, temperature) and Yahoo anomaly detection datasets.

Similarly, the authors of [14] propose using a classic approach based on residuals from an LSTM predictive model for anomaly detection in space probe sensors. The model achieves a univariate prediction by using attributes over a past time horizon. The anomaly detection is then based on smoothed residuals using an exponential filter with a dynamic threshold. This threshold is estimated from the mean and variance in the prediction errors over the past time horizon, which led to more robust detection against false positives. Their best approach achieves a precision/recall score of (0.875/0.80) on two space shuttle sensor datasets.

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 [35] propose learning prediction quantiles based on the extraction of the distribution of predictions. In [36], the authors propose performing learning on the prediction residuals using a second model (either linear or not) that captures 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 [37] offers relevant alternatives. Recently, the authors of [38] proposed approximating Bayesian behavior in a deterministic network by conserving the dropout in the prediction phase. The dropout induces a form of a random draw by randomly forcing some neurons to have zero weight, which disturbs the prediction. Relying on this principle, [12] propose the use of LSTMs to obtain a prediction with a confidence interval. This technique exploits the variational dropout by performing several runs of predictions to provide a prediction interval.

2.3 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 influencing factors are known and observed [39]. One of the challenges lies in taking into account the contextual variability in the data for the detection and characterization of the anomaly's impact on a multivariate series. First, most of the literature does not have access to additional contextual information, which is valuable information for considering the contextual variability that can easily disrupt the anomaly detection process. Then, there is some limitation in each of the three paradigms to face issues of the dynamic context:

- Proximity-based approaches often use a discrete representation of the context that does not fit well with a complex dynamic context that evolves continuously.
- Distance isolation approaches may have difficulties incorporating additional contextual information in sequence comparison step unless performing heavy preprocessing to extract contextual influences (Feremans et al. [7]), which could be insufficient to capture dynamic context influence.
- Prediction residuals approaches seem adequate to deal with the modeling of contextual influence, but forecasting under a complex dynamic context (even more in a multivariate time series framework) is still a challenging task. The reliability and bias of the forecasting

models have a crucial impact on the relevance of a residue-based anomaly detection approach.

By setting our work in the paradigm of "prediction residuals," our positioning consists of giving particular attention to modeling the dynamic influence of the context on the means and variances of the target variable. Our goal is to build a robust anomaly score to highlight statistical anomalies (contextual extremum) 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 previous work, which proposes an encoder-predictor LSTM model dedicated to passenger ridership prediction for public transport trains [11], we propose a methodology to compute an anomaly score that considers the dynamic context of multivariate time series. The forecasting model focuses on capturing contextual influences and short-term dynamics to achieve the prediction. Following [9], we assume that the prediction residuals follow a Gaussian distribution $\mathcal{N}(B, \sigma)$. 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 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?

3 Formalization of the proposed detection approach

The major specificity of our work is the processing of time series structured by a dynamic context, i.e., a set of contextual factors that evolve in time. Our work focuses on dynamic modeling of contextual factors influencing both the means and variances of a multivariate time series. An explicit estimation of contextual means $\hat{\boldsymbol{y}}_t$) and variances $\boldsymbol{\sigma_t}$ is performed for this purpose. In a heterogeneous dataset, contextual mean and variance estimations are essential to define "contextual normality." This estimation allows us to qualify abnormal values through a context-normalized anomaly score.

3.1 Multivariate time series structured by a dynamic context

In this paper, we focus on a multivariate time series structured by a dynamic context, which can be denoted as follows:

$$y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T) \quad \text{with} \quad \boldsymbol{y}_t \in \mathbb{R}^d. \tag{1}$$

3.1.1 Dynamic context

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 define $\mathbf{c}_t$ as the contextual vector as follows:

$$\forall\, t = 1, \ldots, T \;\; \mathbf{c}_t = (c_{i,t})_{i=1,\ldots,m} \quad \text{with} \quad c_{i,t} \in E^i \tag{2}$$

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 by obtaining better knowledge of the contextual mean (prediction task) and contextual variability (variance analysis task) of our data to define a "contextual normality." This "contextual normality" is then used as a reference to quantify the contextual abnormality of each time step.

3.1.2 Proposed decomposition of the times series

We suppose that the time series y 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.

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{m}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{m}_t &= \boldsymbol{f}^c(\mathbf{c}_t) + \boldsymbol{f}^d(\mathbf{c}_t, \boldsymbol{y}_t^p) + \boldsymbol{f}^a(\mathbf{c}_t, \boldsymbol{y}_p, \boldsymbol{a}_t) \\ \boldsymbol{\epsilon}_t &\sim \mathcal{N}(B_t(\mathbf{c}_t, \ell_t), \sigma_t(\mathbf{c}_t, \ell_t)) \end{aligned} \tag{3}$$

where

- $\boldsymbol{y}_t^p = (\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-p})$ 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.
– $\boldsymbol{\varepsilon}_t$ 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 $(\ell, \mathbf{c})$ and can be represented as noise with a dynamic mean $\boldsymbol{B}_t$ and variance $\boldsymbol{\sigma}_t$. The use of a nonzero mean $\boldsymbol{B}_t$ makes it possible to take into account any bias in the prediction model $\boldsymbol{m}_t$.

Given this model, the proposed detection strategy is then based on two main steps, which are detailed below, namely the forecasting and the dynamic modeling of the residuals.

3.1.3 Attribute-based forecasting

A multivariate forecasting model F aims to predict a multivariate series y through a set of attributes X_t: $F(X_t) = \hat{\boldsymbol{y}}_t \approx \boldsymbol{y}_t$. 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.

Dynamic prediction: $F^d(\mathbf{c}_t, \boldsymbol{y}_t^p) \approx f_t^c + f_t^d$ is intended to make a short-term prediction by inferring contextual influence from the attributes (contextual average) and short-term dynamics from the past values of the target variable (autoregressive approach).

Dynamic prediction with anomalies: $F^a(\mathbf{c}_t, \boldsymbol{y}_t^p, \mathrm{a}_t) \approx \boldsymbol{y}_t$ is intended to make a short-term prediction in an abnormal context related to known anomalies.

Learning is always performed on the real series y, which generates some bias that may or may not be negligible. We also make the following assumptions:

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.

In this work, we focus on dynamic prediction with standard machine learning models, such as random forest or recurrent neural network forecasting. Contextual prediction is used only to compare the forecasting performance.

3.2 Prediction residual and anomaly score

Prediction residuals are given by the difference between predictions $\hat{\boldsymbol{y}}_t$ 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. 3), we can decompose the prediction residuals as follows:

$$\begin{aligned} \boldsymbol{r}_t &= \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t \\ &= (\boldsymbol{f}_t^c + \boldsymbol{f}_t^d + \boldsymbol{f}_t^a + \boldsymbol{\varepsilon}_t) - (\hat{\boldsymbol{f}}_t^c + \hat{\boldsymbol{f}}_t^d + \hat{\boldsymbol{f}}_t^a) \\ &= \boldsymbol{e}_t^c + \boldsymbol{e}_t^d + \boldsymbol{e}_t^a + \boldsymbol{\varepsilon}_t \end{aligned} \tag{4}$$

- 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 (Eq. 5), which means that:

$$||\boldsymbol{e}^a|| > ||\boldsymbol{e}^d|| + ||\boldsymbol{e}^c|| + ||\boldsymbol{\varepsilon}||. \tag{5}$$

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 $(\boldsymbol{e}^{\mathrm{a}}(t))$ from normal variance ε_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 ε. 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 $\hat{B}_t \approx e_t^c + e_t^d$ and contextual variance $\hat{\sigma}_t = \sqrt{\varepsilon_t}$ with learning algorithms.

For each dimension of the time series, we define the anomaly score s_t by the following:

$$s_t = \frac{r_t - \hat{B}_t}{2\hat{\sigma}_t} = \frac{(e_t^c + e_t^d - \hat{B}_t) + e_t^a + \varepsilon_t}{2\hat{\sigma}_t} = \frac{e_t^{\hat{B}} + e_t^a + \varepsilon_t}{2\hat{\sigma}_t} \quad (*) \tag{6}$$

* Applied dimension by dimension.

For each element y_t of the time series y, 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.

3.2.1 Contextual anomaly score

The multivariate residual series $\boldsymbol{r}_t$ (Eq. (4)) expresses the difference between reality $\boldsymbol{y}_t$ and the nominal prediction $\hat{\boldsymbol{y}}_t$ 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:

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: $F^m(\boldsymbol{c_t}, \boldsymbol{y_t^p}) = \hat{\boldsymbol{y}}_{\boldsymbol{t}} \in \mathbb{R}^{T\times D},\ \boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} \in \mathbb{R}^{T\times D}$

Step 2. Bias-variance estimation: $F^B(\boldsymbol{c_t}) = \hat{\boldsymbol{B}}_{\boldsymbol{t}} \in \mathbb{R}^{T\times D},\ F^\sigma(\boldsymbol{c_t}) = \hat{\boldsymbol{\sigma}}_{\boldsymbol{t}} \in \mathcal{R}^{T\times D}$

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

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 1. Contextual mean estimation $F^m(\boldsymbol{c}, \boldsymbol{y}^p) = \hat{\boldsymbol{y}}$

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. We use the following four main types of models: the categorical mean, long- and a short-term random forests, and the long short-term memory encoder–predictor (LSTM-EP) model.

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

Then, an estimation of the contextual bias $\hat{\boldsymbol{B}}_t$ and variance $\hat{\boldsymbol{\sigma}}_t$ is performed with the contextual attributes $\mathbf{c}$. 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. The details are given in part A.1 of Appendix.

By naïve methods that provide the classic scores:

- **AE:** The absolute error score $\hat{B} = 0$ and $\hat{\sigma} = 1$.
- **RE:** The relative error score $\hat{B} = 0$ and $\hat{\sigma} = y_t$.

By learning from the residual with a bias-variance estimation model.

- **EMP (empirical estimation):** Using a classic method, prior knowledge can be used to build homogeneous samples from our data. Then, the bias and variance are extracted from the residual values of each subsample as local contextual estimations. The combination of local estimations gives the contextual bias and variances.
- **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 as the ML model.

By extracting a type of variance directly from the forecasting models.

- **RF (Random forest extraction):** For the random forest model, the authors of [35] have shown 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 extracting 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 [38] can be used to approximate the Bayesian behavior of a deterministic network. Like the authors of [12], we propose using the

variational dropout to estimate the variance of our LSTM-EP model by virtual sampling.

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

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 the contextual extremum data. 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 $\boldsymbol{s}_{agg} = \sqrt{(\boldsymbol{s}-\bar{\boldsymbol{s}})^{\intercal}\ \boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1}\ (\boldsymbol{s}-\bar{\boldsymbol{s}})}$

Spatial aggregation aims to synthesize across dimensions to obtain a unidimensional synthesis score. We choose the Mahalanobis distance due to its covariance consideration.

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 1, the normalization types used to construct the anomaly scores are listed.

3.2.2 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 taking into account 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.

Table 1 List of different types of normalization

Normalization	Bias $\hat{B}$	Variance $\hat{\sigma}$
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 Results of experiments on a synthetic dataset

4.1 Evaluation setting

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

4.1.1 Data generation

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 conspicuity 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.

We develop a generation process for multivariate time series with a dynamic context and anomalies. For our synthetic application, we generate a time series of 8000 time steps in three dimensions (3D). The generated data include look like ridership time series with a short periodic

pattern of 20 time steps ("Daily pattern"), a medium periodic pattern of (20 × 3) time steps (a type of "weekly trend"), and a long periodic trend of (20 × 100) time steps (a "seasonal trend").

The data generation process (Eq. 7) follows several steps.

$$\begin{aligned} f^c &= Periodic\ pattern * Trends\ (i) \\ f^d &= f^c * Daily\ magnitude\ (ii) \\ \varepsilon &= (f^c + f^d) * (Noise_{mult} * Daily\ variance\ pattern) \\ &\quad + Noise_{add}\ (iii + iv) \\ f^a &= (f^c + f^d + \varepsilon) * (\alpha_{mult} * Anom) + \alpha_{add} * Anom\ (iv) \\ y &= f^c + f^d + f^a + \varepsilon \end{aligned} \tag{7}$$

(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 taking into account both types of noise. For more details, the generation processes are available in the Git[1]

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. An illustration of two synthetic time series is provided in Fig. 2. Each anomaly is linked to a context depending on its temporal position. 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.

4.1.2 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 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 multistep forecasting. More details are provided in Appendix (A.2.3) and in the article [11].
- **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.

4.1.3 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 2) 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 sensibility (% detected anomalies). We

[1] Git-lab of experiments on synthetic data: https://gitlab.com/Haroke/contextual-anomaly-detection.

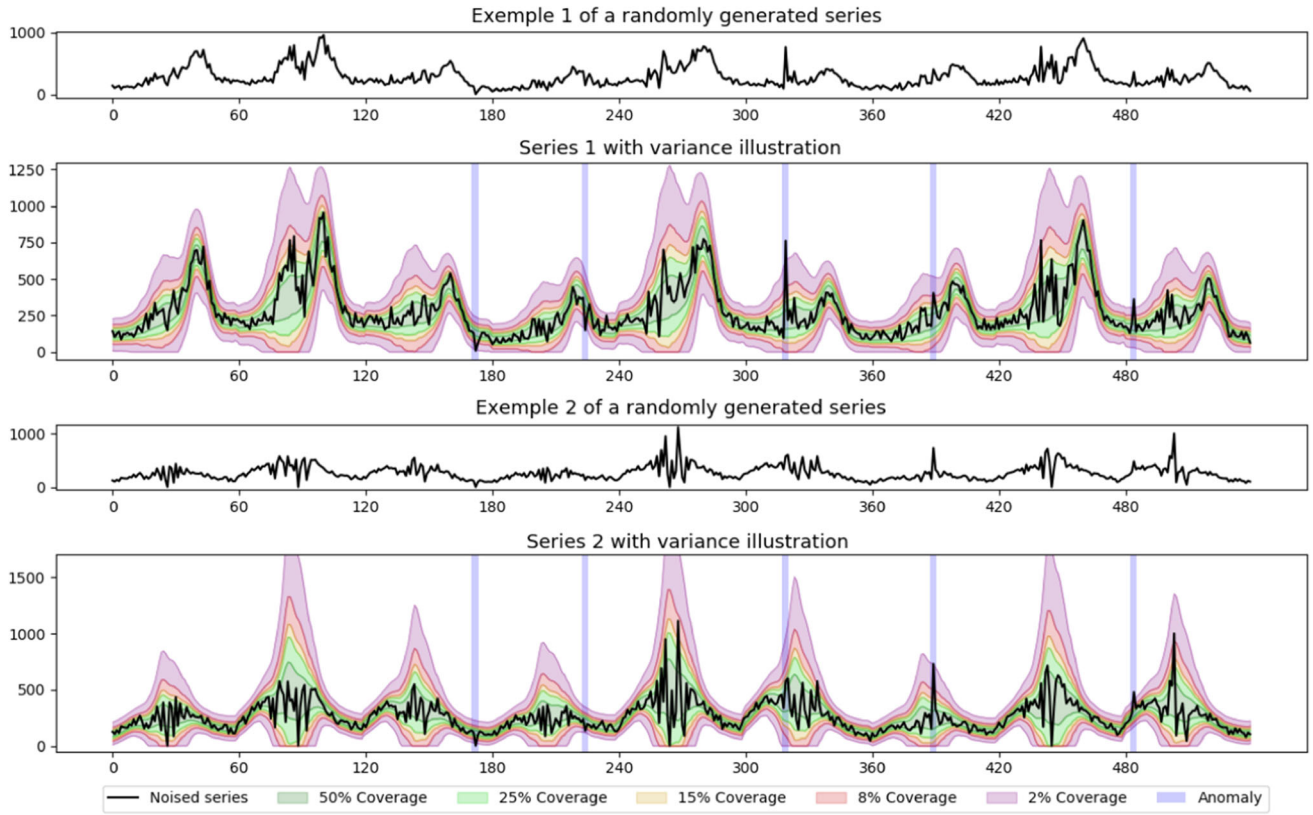

Fig. 2 Example of curves generated on 9 "hypothetical days"

calculate the global sensibility (Table 4) and the specific context and magnitude sensibility (Table 5). 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 (see Fig. 3).

4.2 Results on the synthetic data

4.2.1 Forecasting results

The forecasting results are presented in Table 2. 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.

4.2.2 Anomaly detection results

The univariate results obtained with all prediction model combinations and bias-variance estimation approaches are presented in Table 3. In the column 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 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 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 following 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.

An in-depth analysis of the anomaly scores according to "anomaly strength" and "magnitude of the context" is provided in Table 4. To facilitate the analysis of the results,

Table 2 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

we sampled the data in four contexts (C1,C2,C3,C4) according to their magnitude (average value). The forecasting is achieved here by the LSTM-EP model. These results show the weakness of the naïve scores; low-magnitude contexts (C1 & C4) induce weakness in the N-AE score, while high-magnitude contexts (C2 & C3) induce weakness in the N-RE score. Conversely, the context-normalized anomaly scores (N-EMP, N-ML, N-RF, and N-DP) show better contextual robustness and better detection rates for both minor and major anomalies. As expected, minor anomalies are more difficult to detect.

Table 5 compares the **local** and **global** score performance (as defined in Sect. 3.2.2). Spatial aggregation improves all the detection rates. The improvement is particularly significant for context-normalized anomaly scores. Indeed, an anomaly can be covered 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.

As shown in Fig. 3 which make focus between range of [0,15] of anomaly ratio, the ROC curves based on the naïve 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.

4.3 Conclusion

The experiments carried out on synthetic data show the relevance of using the prediction residues jointly with a

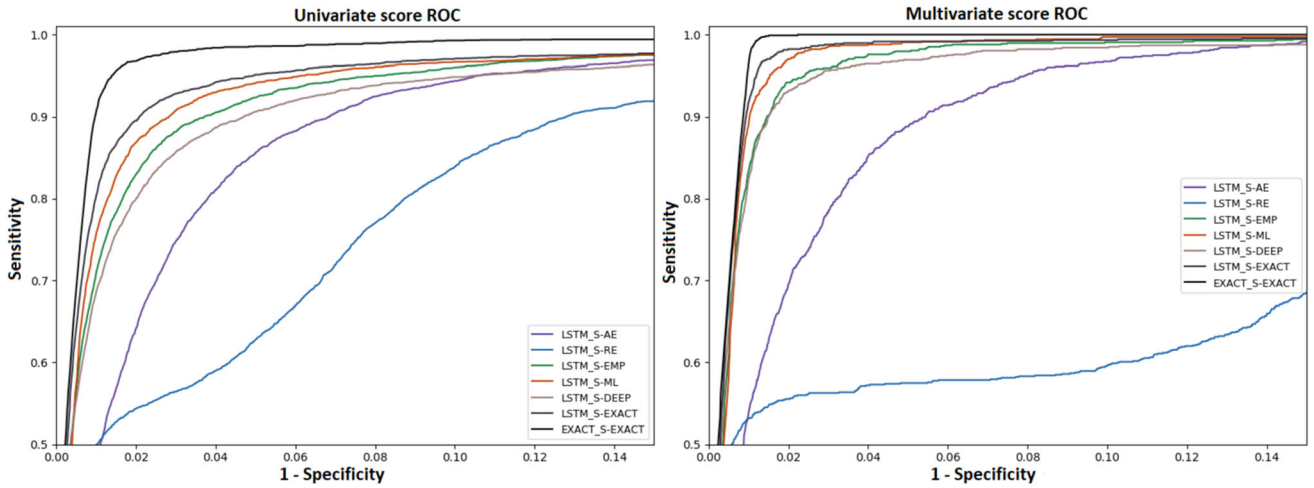

Fig. 3 ROC curves for the local and global scores based on the LSTM-EP residue

Table 3 Performance of all combinations (sensibility 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.

Table 4 Performance of the LSTM-EP-based approaches (sensibility with a 2% detection ratio)

Strength	Minor anomaly				Major anomaly				Total
Context	C1	C2	C3	C4	C1	C2	C3	C4	–
Magnitude norm	Low	High	High	Low	Low	High	High	Low	–
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 real data

Table 5 Local/global score performance (sensibility with a 2% detection ratio)

Norm	N-AE		N-RE		N-EMP		N-ML		N-RF		N-DP		N-EX*	
Type Model	Loc	Glo	Loc	Glo	Loc	Glo	Loc	Glo	Loc	Glo	Loc	Glo	Loc	Glo
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

bias-variance estimation for contextual anomaly detection. Several conclusions may be drawn from these experiments, as follows:

1. The scores based on the short-term prediction residue give better results in terms of anomaly detection related to the gain between long- and short-term predictions. In our synthetic experiment, we note a 15% forecasting gain between long- and short-term approaches (see Table 2). We still observe an anomaly detection gain between 5% and 10% (see Table 3, the difference between lines CAT/RF-LT and RF-ST/LSTM-EP/EXACT*).
2. The bias-variance estimation makes anomaly detection more robust concerning context, which improves the detection performance. In our synthetic experiment, anomalies have two specificities to be suitable for variance consideration: (i) anomalies are incorporated homogeneously (the same chance for all time steps), and (ii) anomaly magnitudes are proportional to their application contexts. This frame allows us to observe up to 30% detection gain (see Table 3, differences between columns N-AE/N-RE and N-EMP/N-ML/N-DP/N-EX).
3. We can estimate the variance by learning on forecasting residues or by extracting it from random forest or neural network forecasting models.
4. The multidimensional aggregation of univariate scores significantly improves detection performance. In our synthetic experiment, we observe up to 15% detection gain on global scores based on Mahalanobis aggregation of local scores. Table 5 shows the difference between columns Loc (local score) and Glo (global score).

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. By doing so, we could characterize the impacts of the different disturbances on the metro ridership in Montreal.

5.1 Data description

5.1.1 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 5:00 am each day until 1:00 am the next day. In this study, we focus on the fourteen stations located at the center of Montreal. The data consist of a multivariate time series of 14 dimensions and 87860 time steps, corresponding to 1096 days with 80 daily time steps of 15 minutes. Figure 4 illustrates the contextual envelopes (confidence intervals) of the target variable (ridership values) at two stations. We observe a high variability of the ridership time series that depends on the station and its dynamic context. This complex variability is fed by a set of factors of known and unknown influences that estimate the contextual means and variances.

5.1.2 Disturbance data

The disturbance data contain some events and incidents that occurred within the studied period and that might impact the station's 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 2, 076 **minor incidents**, which have an average duration of 35 minutes, are distributed into the following seven categories: technical failure (482), staff issues (117), minor failure (184), door failure (1112), 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 (399), accidents (281), intrusion (205), and fire (75). The **event data** include 1772 events with 10 event categories, including exhibition (414), hockey match (178), festival (385), concert (173), sport (174), show (172), tennis (37), football (30), soccer (65), 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 partial source, it is not a ground-truth reference for all the events, incidents and other phenomena that can impact the station ridership and some disturbance. Lastly, some disturbances may have no impact on the studied variable (entrance 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's residues based anomaly scores.
- to investigate afterward the unexplained detection in order to enrich the disturbance database.

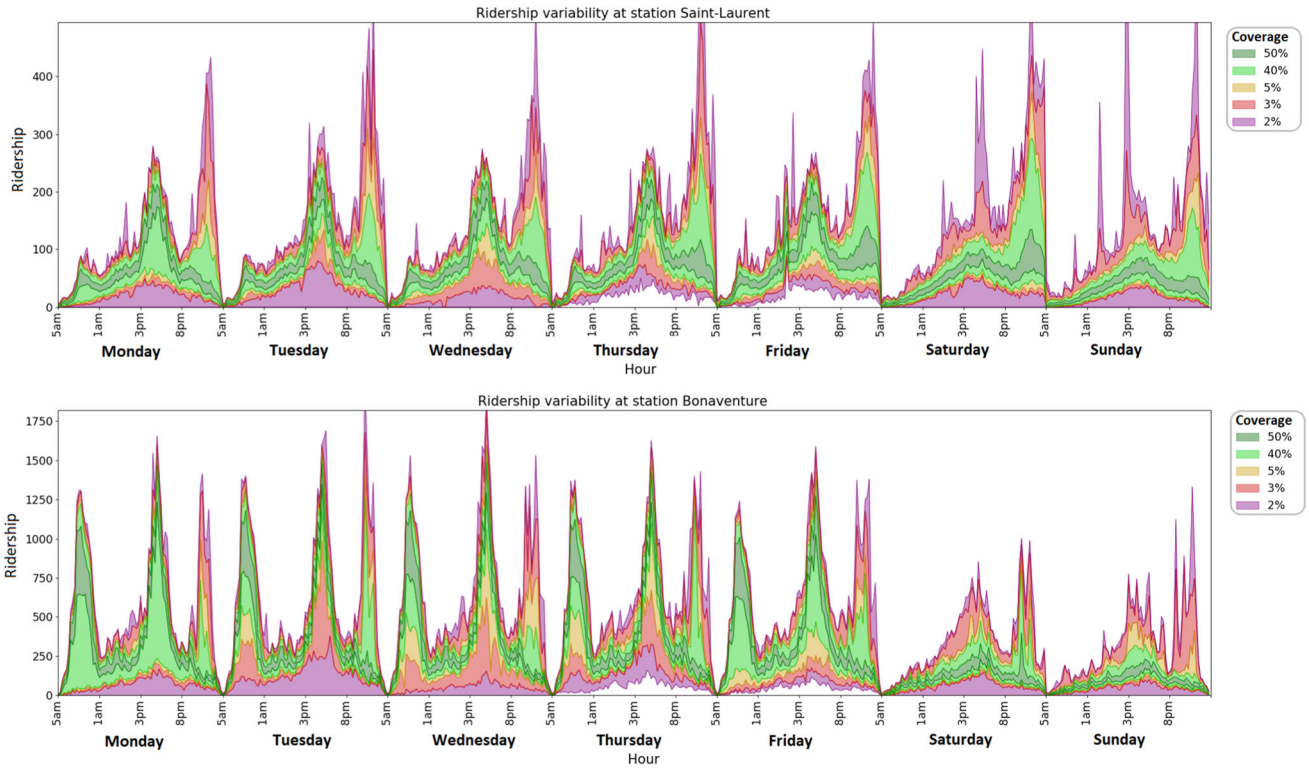

Fig. 4 Ridership contextual envelopes (confidence intervals) extracted from three years of smart card data (2015-2017) for two stations of Montreal City metro

Future work will be mandatory to perform a more robust evaluation with more metrics and additional disturbance information data sources.

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-term and short-term attributes can be used. The **short-term dynamic attributes (ST)** are considered to capture the dynamic components. They consist of five ridership values of the past temporal horizon [t - 4, t] for each of the 14 stations, which makes a numerical vector of size 70. The **long-term contextual attributes (LT)** are linked to well-known influencing factors, such as the following:

- The time schedule (8 dimensions), given by the time step position in a day (80 possible values) encoded by a sine and cosine 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 and cosine 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's, and other holidays, encoded by a one-hot vector.

In Table 6, 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 have 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.

5.3 Anomaly detection results

5.3.1 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 Sect. 5.1.2). Moreover, the link between a disturbance and a ridership anomaly is not straightforward. Therefore, an usual evaluation 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 Sect. 3.2.2 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 7 and 8. For each combination, we analyze at two different scales through the

Table 6 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**

Table 7 Disturbance and local anomaly explanation for a prior local anomaly ratio of 2.5%

Anomaly score		% Detected disturbances			% Anomalies explained by disturbances			
Residue	Norm	Incident-min N=2076	Incident-maj N=960	Event N=1772	Explained Anomaly	Unexplained Anomaly	Number Total*	Average 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

Table 8 Disturbance and global anomaly explanation for a prior global anomaly ratio of 5%

Anomaly score		% Detected disturbances			% Anomalies explained by disturbances			
Residue	Norm	Incident-min *N*=2076	Incident-maj *N*=960	Event *N*=1772	Explained Anomaly	Unexplained Anomaly	Number Total*	Average 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

local (per station) and **global** (Network scale) scores (introduced in Sect. 3.2.1).

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 that is a temporal and spatial fit (for local anomaly detection).

Among the parameters of our anomaly score process, the parameter q (introduced in Sect. 3.2.1, 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.

5.3.2 Results

We will therefore confront disturbances coming from a labeled operator base and anomalies coming from the detection models. However, as mentioned in Sect. 5.1.2, disturbances are events, minor incidents, or major incidents that **"could significantly impact"** the ridership at 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 put into perspective.

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 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 instants** with no known disturbances or statistical anomalies detected by the detection models.

An exhaustive validation will 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 indicators that can be inferred from the results can also contribute to the labeling task.

Nevertheless, we will carefully analyze some quantitative results next. Table 7 shows the results of the matching that we carried out between the anomalies raised by our models and the operator's declarative disturbances. These disturbances are focused only on the local level, i.e., at the station scale. Each line corresponds to a combination of the residue of either of the prediction models (CAT, RF-ST, LSTM) with one of the proposed normalization (N-AE, N-EMP, N-ML, etc.). The first part of the table provides **the percentages of detected disturbance** for the three subcategories (minor incident, major incident, events). This corresponds to the disturbance ratio covered by anomalies. The second part of table provides the explained 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 taking into account 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 8 contains the information resulting from the confrontation of the disturbances with anomalies linked with the **global anomaly score** (introduced in Sect. 3.2.1) of 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 $LSTM - EP$ 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 at network scale are often explained by known disturbances. Conversely small magnitude and highly localized anomalies are less often linked to known causes.

5.4 In-depth analysis of the results

In-depth analysis of the spatiotemporal 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.

5.4.1 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 5 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 taken into account 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.

Considering another metro station (Square Victoria station) on the same day, the collected and predicted ridership curves are presented in Fig. 6. 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.

5.4.2 Analysis of two particular days

(1) 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 7 shows 2 anomaly scores based on naïve normalization (N-AE) and contextual normalization (N-RF) computed for the fourteen metro stations for Monday, 27 February. 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 (undercrowded) 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.

This clearly shows the benefits of crossing the statistical anomaly score with the transport operator disturbance

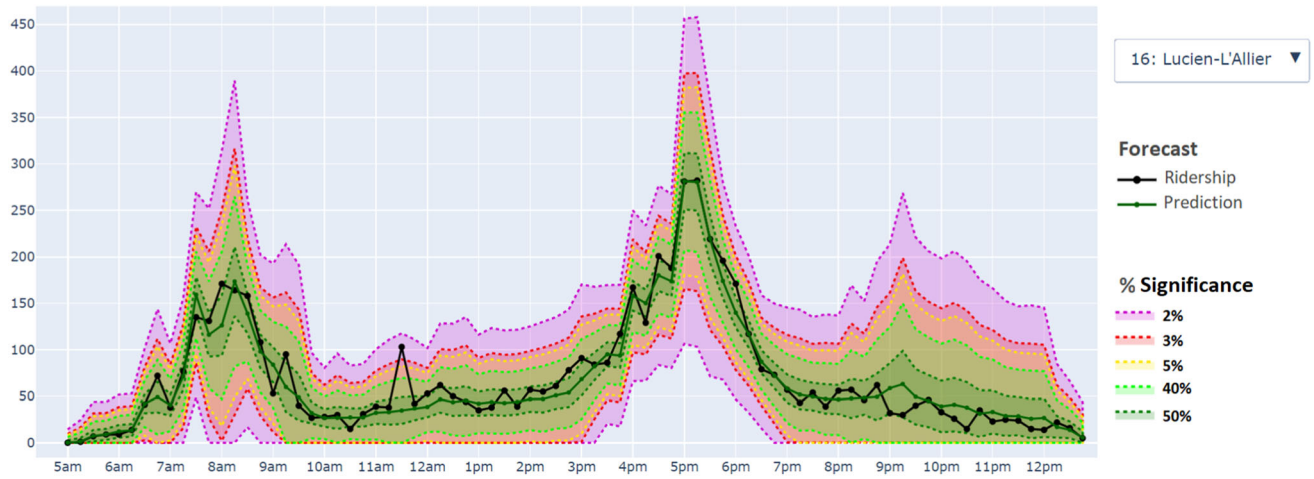

Fig. 5 Ridership prediction confidence at Lucien l'Allier station on Monday, February 27

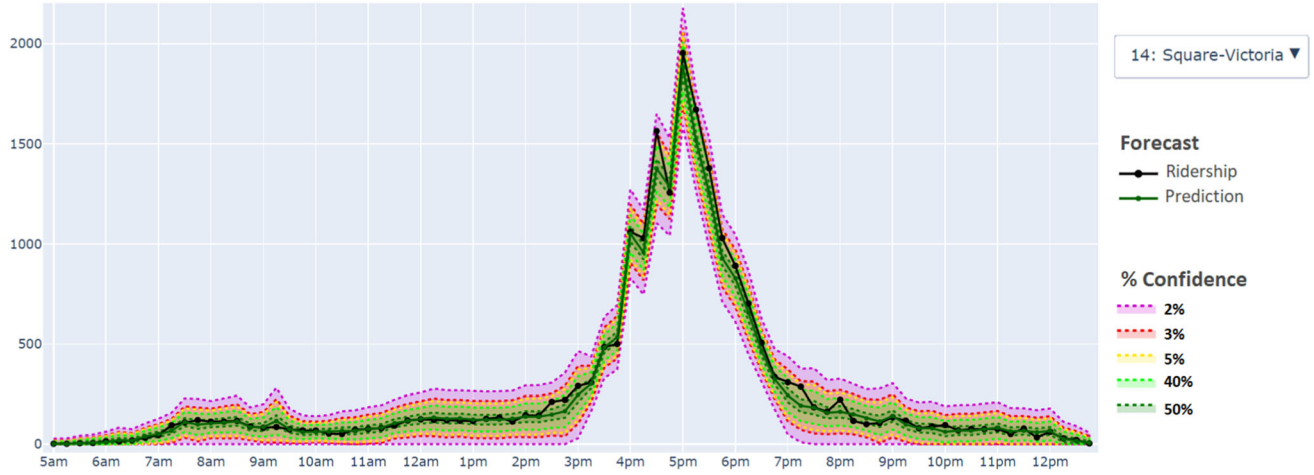

Fig. 6 Ridership prediction confidence at the Square Victoria station, Monday February 27

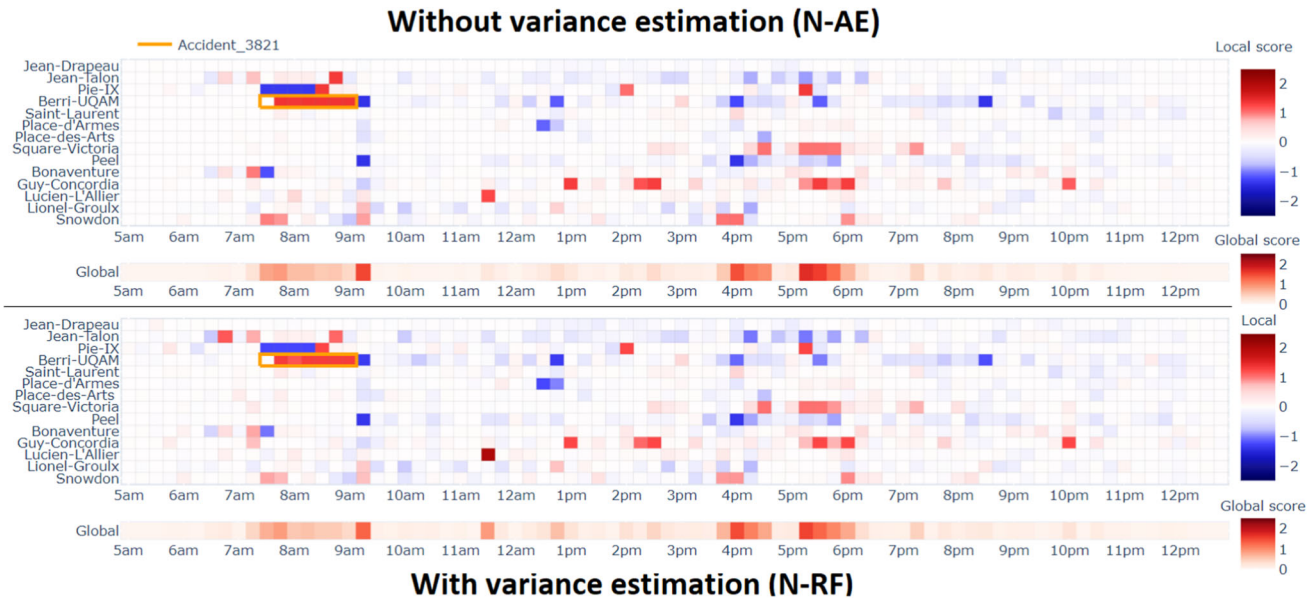

Fig. 7 Two anomaly scores (N-AE, N-RF) computed on Monday, 27 February, 2017

dataset. This crossing can be helpful in completing the labeling of the disturbance data.

If we examine the ridership details at Pie-IX station in Fig. 8, we observe a distinct period of traffic stop in the morning, followed by a ridership peak at 8:45 AM linked to the traffic recovery. Both the naïve and the context-normalized scores greatly exceed the normality envelopes. Here, we also observe two unexplained anomalies with high anomaly scores, i.e., at 2:00 PM and 5:30 PM.

(2) 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 Fig. 9, the naïve score (N-AE) and the context-normalized score (N-RF) show significant differences in the morning.

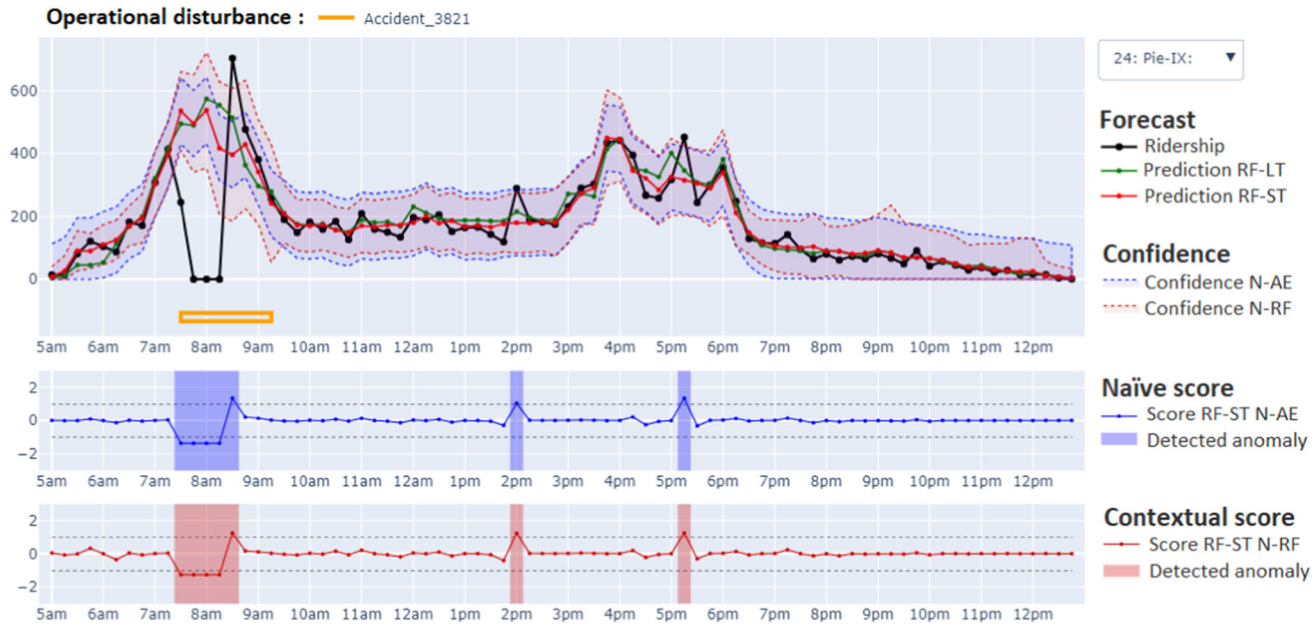

Fig. 8 Ridership and anomaly scores at Pie IX station on Monday, 27 February, 2017

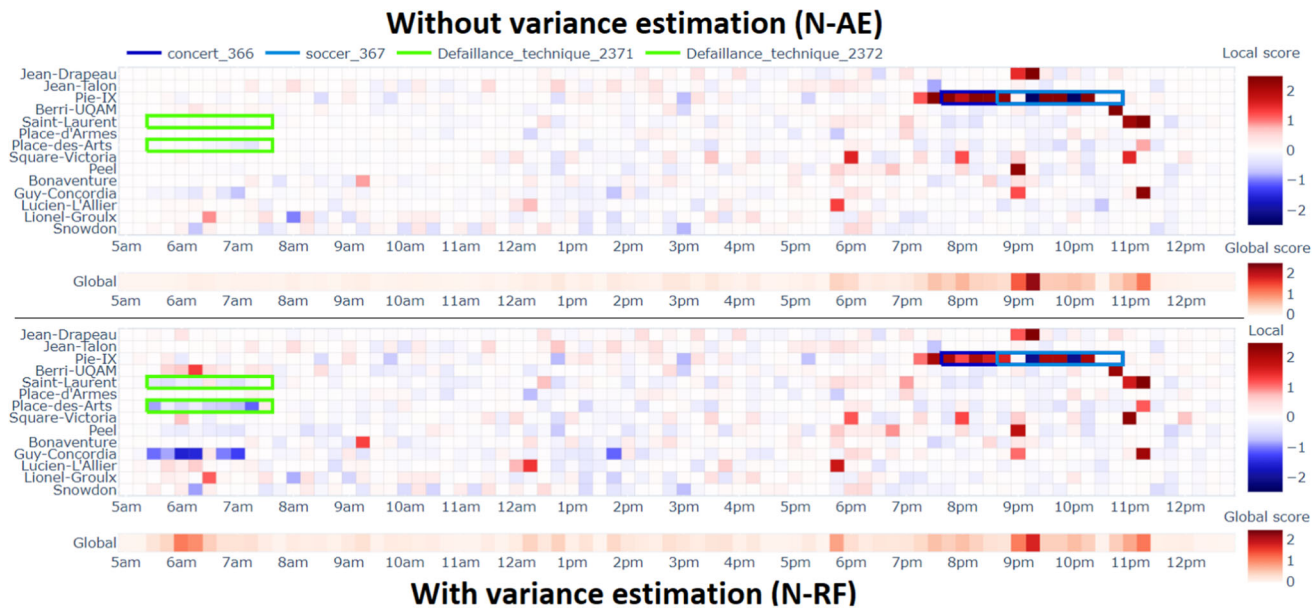

Fig. 9 Two anomaly scores (N-AE, NS-RF) computed on Wednesday, 5 August, 2015

The disturbance does not affect the two scores in the same way: it is invisible for the naïve 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.

If we examine Fig. 10, 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 naïve score. This confidence in the forecasting allows the context-normalized score to detect the overcrowding situation induced by the traffic cut. In contrast, the naïve 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.

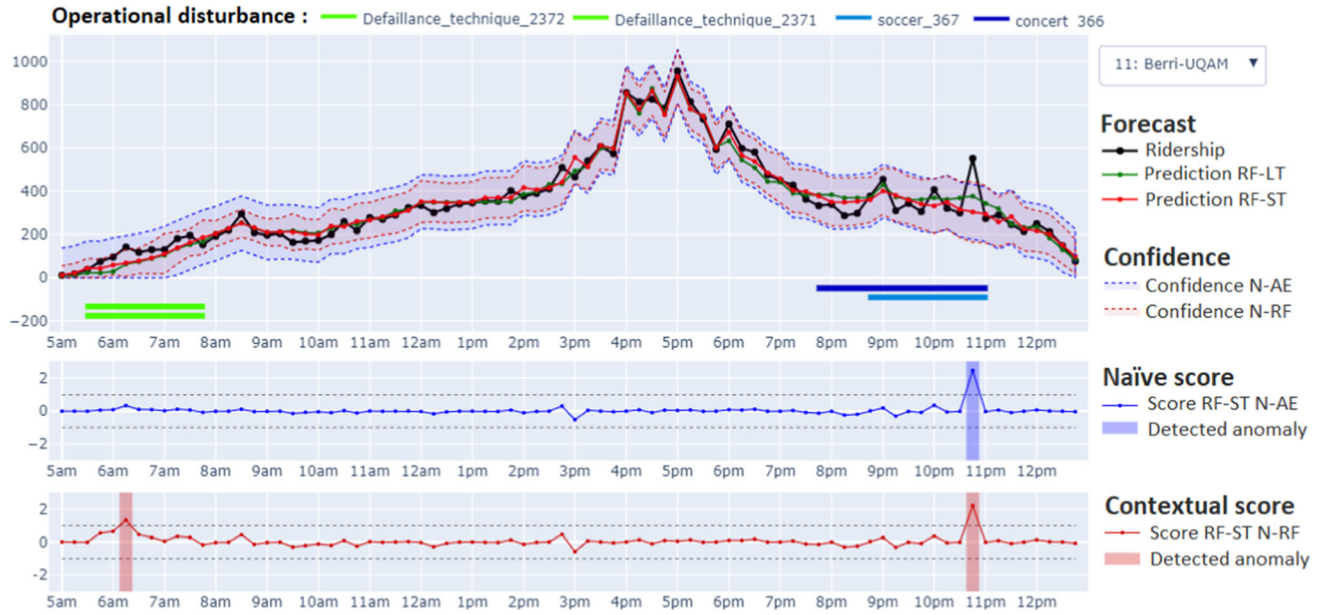

Fig. 10 Ridership and anomaly scores at Pie IX station on Wednesday, 5 August, 2015

6 Conclusion and perspectives

In this paper, we propose a general methodology to capture the influence of a dynamic context which structures a multivariate time series as well as to detect anomalies in the time series through a context-normalized anomaly score. Two parameters can synthesize the influence of the dynamic context, namely the contextual mean and the contextual variance (or variability). Based on the prediction model residues, we propose estimating the contextual mean using contextual and short-term features and the contextual variance through an empirical estimation or a machine learning model. 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.

In this paper, we focus on anomaly detection in a multivariate time series with contextual attributes. In particular, we propose different context-normalized anomaly scores that we compare to naïve anomaly scores. The main strength of the context-normalized anomaly scores lies in their ability to define the statistical abnormality of each time step with respect to its context. Such anomaly scores are context-robust, allowing them to detect anomalies more precisely in a dynamic context that is often neglected by standard anomaly detection systems.

The proposed method is evaluated on a synthetic dataset. The detection performance shows the effectiveness of the anomaly detection approaches involving context normalization. We also apply the proposed methodology to a real smart card dataset collected from a metro network. The statistical anomalies are compared with a disturbance dataset provided by the transport operator. The disturbances include minor incidents, major incidents, and events (cultural, sporting, etc.). The results show that the detection score is closely related to the seriousness of the incident. They also highlight the fact that the impact of a disturbance on the transport ridership is not homogeneous. Events impact human mobility in the transit network, while the impact of an incident is not straightforward. The analysis of anomaly scores can 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 disturbance impacts.

Further work is mandatory to deepen the quantitative evaluation of the results, requiring an active collection of complementary data sources informed on the reasons for the disturbances and confrontation between the detected anomalies and explained disturbances. This work will also make it possible to consider using other more advanced metrics (selectivity & sensitivity by perturbation type, distinction between false positives, and posterior explained anomalies) that will lead to a better understanding of the impact of disturbances.

Plus, this work could be extended in other several ways. We have used the anomaly score to perform an anomaly detection based on a dynamic threshold. It would be relevant to perform a pattern analysis of the context-normalized anomaly score that could detect and characterize different types of anomaly impacts. In addition, the extant works focus mainly on capturing contextual variance. It

would also be relevant to capture the variance that can be related to the dynamic component by considering autoregressive variance estimation models. The construction of an unsupervised anomaly score could be enhanced through an iterative learning phase aiming to focus learning on normal data, i.e., having a low anomaly score by adjusting the learning weights. In the same way, further research can be conducted to modify the architecture of the deep forecasting model to perform multiple learning procedures to infer several prediction targets (contextual, dynamic, anomaly-based) that can allow the consideration of anomalies with low signals but that are significant from the view of an operator.

Appendix

Bias-variance Estimators

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 $\mathbf{c}$ into prior subspaces (subsamplings) defined by a set of constraints (V^{inf}, V^{sup}) given by expert knowledge. The bias $\hat{B}$ and variance $\hat{\sigma}$ 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 $\hat{B}_t$ and variance $\hat{\sigma}_t$ for each time step t.

$$\{E_k : t \in E_k \mid V_k^{inf} > \mathbf{c}(t) > V_k^{sup}\}$$

$$\hat{B}_{E_k} = \sum_{t \in E_k} \frac{r_t}{\#E_k} = \hat{r}_{E_k} \qquad \hat{\sigma}_{E_k} = \sqrt{\sum_{t \in E_k} \frac{(r_t - \hat{B}_{E_k})^2}{\#E_k}}$$

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.

$$Bias \ : \ \theta = \text{argmin}_\theta \sum_t |M_\theta^{\hat{B}}(X_t) - r_t| \hat{B}(t) = M_\theta^{\hat{B}}(X_t) = \hat{r}_t$$

$$Variance \ : \ \theta = \text{argmin}_\theta \sqrt{\sum_t |M_\theta^{\hat{\sigma}}(X_t) - (r_t - \hat{B}(t))^2|}$$

$$\hat{\sigma}(t) = \sqrt{M_\theta^{\hat{\sigma}}(X_t)}$$

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.

- **RF: Random forest extraction**

  In [35], 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.

$$M(X_t) = \frac{1}{n} * \sum_{k \in [1,n]} \left(\sum_{j \in F^k(X_t)} \frac{j}{\#F^k(X_t)} \right) = \hat{y_t}$$

  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.

$$\hat{B}(t) = 0 \quad \hat{\sigma}(t) = \sqrt{\frac{1}{n} * \sum_{k \in [1,n]} \left(\sum_{j \in F^k(X_t)} \frac{(j - \hat{y_t})^2}{\#F^k(X_t)} \right)}$$

- **DEEP: Neural network extraction**

  A second form of extraction is based on variational dropout [38], which aims to approximate Bayesian behavior in a deterministic network. A study in [12] 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.

  Let $M_{\hat{\theta}}$ be a neural network that infers y_t from X_t.

$$\theta = \text{argmin}_\theta \sum |M^\theta(X) - y|^2 \quad M_\theta(x_t) = \hat{y_t}$$

  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.

$$\{\mathcal{B}(z_t) : k \; tq \; z_k \in [z_t \pm \varepsilon]\} \; with \; z, \varepsilon \in \mathcal{R}^{\#Z}$$

$$\hat{B}(t) = \sum_{k \in \mathcal{B}(z_t)} \frac{|\hat{y_k} - y_t|}{\#\mathcal{B}(z_t)} = \hat{r_t}$$

$$\hat{\sigma}(t) = \sqrt{\sum_{k \in \mathcal{B}(z_t)} \frac{((\hat{y_k} - y_t) - \hat{B}(t))^2}{\#\mathcal{B}(z_t)}}$$

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.

$$\theta = \text{argmin}_\theta \sum |(M_\theta^{var}(X) - y)|^2 \quad \sum_m \frac{M_\theta^{var}(x_t)}{m} = \hat{y_t}$$

The stochastic projections of 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.

$$\hat{\mathcal{B}} = 0 \quad \hat{\sigma}(t) = \sqrt{\sum_m \frac{(M_\theta^{var}(x_t) - \hat{y_t})^2}{m}}$$

Forecasting models

Encoding cyclical features

Cyclical encoding aims to encode continuous cyclic attributes by preserving their cyclic structure. Instead of having a large one-hot vector per feature, the sine and cosine encodings project each attribute on a two-dimensional plane. However, some contextual information contains more than one cyclical structure. Using several pairs of sines and cosines with different frequencies can allow us to better express meaningful and compact structures. For instance, we can express several pieces of periodical information (weekly, monthly, and seasonal) by encoding the position of the day in the year with several pairs of sines and cosines with well-chosen frequencies, i.e., 1/53 for a weekly structure, 1/31 for monthly, and 1/1 for a yearly structure.

This technique yields compact and meaningful attributes, in contrast to the bulky and sparse hot encoding.

Random forest training

The forecasting and bias-variance estimation models are optimized through a mean-square error (MSE) optimization loss. We use a standard scikit-learn random forest regressor [40]. The random forest parameters are tuned through a random search using cross-validation combined with early stopping to control the number, size and depth of trees to avoid overfitting.

LSTM EP : architecture and training

In our previous research [11], we proposed an LSTM encoder predictor for ridership forecasting by using both long-term and short-term attributes.

The model was designed to manage the structural variability in the data induced by the transport plan. A simplified version of the model (Fig. 11) is applied in the current work thanks to the regular structure of the data.

First, the long-term features are synthesized through a multilayer perceptron neural network. Then, a pair of encoder–predictor LSTM layers attempt to capture the contextual influence and infer the short-term dynamics of the multivariate time series. Finally, another multilayer perceptron attempts to interpret the prediction embedding Z^p to produce a prediction $\hat{y}$. This model takes as input the contextual attribute X_t and past horizon value $y_t^p = [y_{t-p}, \ldots, y_t)$ and aims to forecast 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. The size is manually chosen through a compromise between three components of size, performance and overfitting.

The encoder predictor model is implemented based on the *TensorFlow* [41] environment with *Keras* [42] as a library and a high-level neural network API. Training is performed through 3 training loops with gradient reduction and early stopping. We use an adaptive gradient (ADAM), and we reduce the batch size between each loop. The first training loop is a type of initialization in which we keep only the reconstruction task in the learning loss. Then, we add multistep forecasting with a higher weight to the t+1 prediction loss.

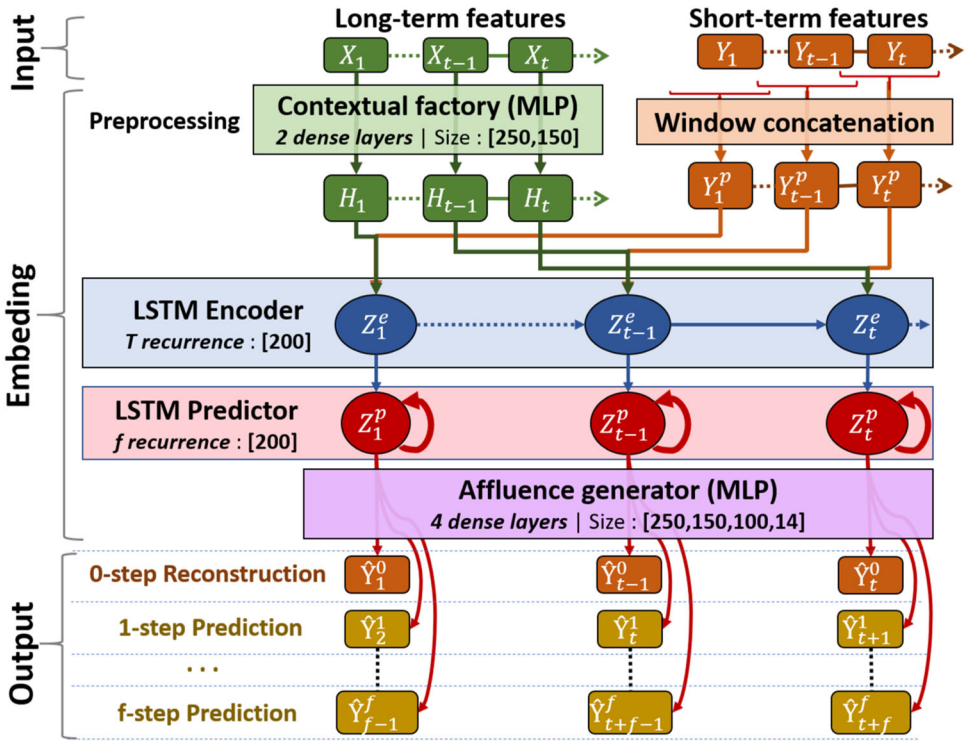

Fig. 11 LSTM-EP architecture with the layer size for the real data

Acknowledgements This research is a part of the IVA Project, which aims to develop machine learning approaches to enhance traveler information. The project is carried out under the leadership of the Technological Research Institute SystemX, with the partnership and support of the transport organization authority Ile-De-France Mobilités (IDFM), SNCF, Université Gustave Eiffel and public funds under the scope of the French Program "ANR - Investissements d'Avenir." The authors also wish to thank the Montreal Transit Corporation (STM) for providing ridership data and the database of events and disturbances.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Kevin PASINI. The first draft of the manuscript was written by Kevin PASINI and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research is a part of the IVA Project, which aims to enhance traveler information. The project is carried out under the leadership of the Technological Research Institute SystemX, with the partnership and support of the transport organization authority Ile-De-France Mobilités (IDFM), the french railway operator NCF, and public funds under the scope of the French Program "Investissements d'Avenir."

Availability of data and material Data provided by the Société de Transport de Montréal (STM) are private and the authors do not have the right to provide it to third parties. However, synthetic data and the generation process can be shared.

Declarations

Conflicts of interest There is no conflict of interest.

Code availability The code related to experiments on real data is the property of the project partners. The part of the code related to experiments on synthetic data can be shared. https://gitlab.com/Haroke/contextual-anomaly-detection.

References

1. Chandola V (2009) Anomaly detection for symbolic sequences and time series data, Ph.D. thesis, University of Minnesota
2. Hayes MA, Capretz MA (2014) Contextual anomaly detection in big sensor data. In: 2014 IEEE International Congress on Big Data, IEEE, pp 64–71
3. Benkabou S-E, Benabdeslem K, Canitia B (2018) Unsupervised outlier detection for time series by entropy and dynamic time warping. Knowl Inf Syst 54:463–486
4. Yeh C-CM, Zhu Y, Ulanova L, Begum N, Ding Y, Dau HA, Silva DF, Mueen A, Keogh E (2016) 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, pp 1317–1322
5. Nakamura T, Imamura M, Mercer R, Keogh E (2020) Merlin: Parameter-free discovery of arbitrary length anomalies in massive time series archives. In: 2020 IEEE 16th international conference on data mining (ICDM), IEEE
6. Ding Z, Fei M (2013) An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proc. Vol. 46:12–17
7. Feremans L, Vercruyssen V, Cule B, Meert W, Goethals B (2019) Pattern-based anomaly detection in mixed-type time series, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp. 240–256
8. Tonnelier E, Baskiotis N, Guigue V, Gallinari P (2018) Anomaly detection in smart card logs and distant evaluation with twitter: a robust framework. Neurocomputing 298:109–121

9. Malhotra P, Vig L, Shroff G, Agarwal P (2015) Long short term memory networks for anomaly detection in time series. In: Proceedings, vol 89, Presses universitaires de Louvain
10. Guo Y, Liao W, Wang Q, Yu L, Ji T, Li P (2018) Multidimensional time series anomaly detection: a gru-based gaussian mixture variational autoencoder approach. In: Asian Conference on Machine Learning, pp 97–112
11. Pasini K, Khouadjia M, Same A, Ganansia F, Oukhellou L (2019) LSTM encoder-predictor for short-term train load forecasting. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp 535–551
12. Zhu L, Laptev N (2017) Deep and confident prediction for time series at uber. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, pp 103–110
13. Yu Y, Long J, Cai Z (2017) Network intrusion detection through stacking dilated convolutional autoencoders. Security and Communication Networks 2017
14. Hundman K, Constantinou V, Laporte C, Colwell I, Soderstrom T (2018) Detecting spacecraft anomalies using LSTMS and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 387–395
15. Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G (2017) Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging, Springer, pp. 146–157
16. Abdallah A, Maarof MA, Zainal A (2016) Fraud detection system: a survey. J Netw Comput Appl 68:90–113
17. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W, Retain, (2016) An interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in Neural Information Processing Systems 3504–3512
18. Cao N, Lin C, Zhu Q, Lin Y-R, Teng X, Wen X (2017) Voila: visual anomaly detection and monitoring with streaming spatiotemporal data. IEEE Trans visual Comput Graph 24:23–33
19. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv (CSUR) 41:1–58
20. Habeeb RAA, Nasaruddin F, Gani A, Hashem IAT, Ahmed E, Imran M (2019) Real-time big data processing for anomaly detection: a survey. Int J Inf Manag 45:289–307
21. Chalapathy R, Chawla S (2019) Deep learning for anomaly detection: a survey, arXiv preprint arXiv:1901.03407
22. Cheng H, Tan P-N, Potter C, Klooster S (2009) Detection and characterization of anomalies in multivariate time series. In: Proceedings of the 2009 SIAM international conference on data mining, SIAM, pp 413–424
23. Dimopoulos G, Barlet-Ros P, Dovrolis C, Leontiadis I (2017) Detecting network performance anomalies with contextual anomaly detection. In: 2017 IEEE international workshop on measurement and networking (M&N), IEEE, pp 1–6
24. Liu FT, Ting KM, Zhou Z-H (2008) Isolation forest. In: 2008 Eighth IEEE international conference on data mining, IEEE, pp 413–422
25. Liu FT, Ting KM, Zhou Z-H (2012) Isolation-based anomaly detection. ACM Trans Knowl Discov Data TKDD 6:1–39
26. Yankov D, Keogh E, Rebbapragada U (2008) Disk aware discord discovery: finding unusual time series in terabyte sized datasets. Knowl Inf Syst 17:241–262
27. Akouemo HN, Povinelli RJ (2014) Time series outlier detection and imputation. In: 2014 IEEE PES General Meeting, IEEE, pp 1–5
28. Li J, Pedrycz W, Jamal I (2017) Multivariate time series anomaly detection: a framework of hidden Markov models. Appl Soft Comput 60:229–240
29. Salem O, Guerassimov A, Mehaoua A, Marcus A, Furht B (2014) Anomaly detection in medical wireless sensor networks using svm and linear regression models. Int J E-Health Med Commun IJEHMC 5:20–45
30. Kromanis R, Kripakaran P (2013) Support vector regression for anomaly detection from measurement histories. Adv Eng Inf 27:486–495
31. Hasan MAM, Nasser M, Pal B (2014) Ahmad S (2014) Support vector machine and random forest modeling for intrusion detection system (ids). J Intell Learn Syst Appl
32. Kasai H, Kellerer W, Kleinsteuber M (2016) Network volume anomaly detection and identification in large-scale networks based on online time-structured traffic tensor tracking. IEEE Trans Netw Serv Manag 13:636–650
33. Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G (2016) Lstm-based encoder-decoder for multi-sensor anomaly detection. In: Anomaly Detection Workshop of the 33rd International Conference on Machine Learning (ICML 2016)
34. Munir M, Siddiqui SA, Dengel A, Ahmed S (2018) Deepant: a deep learning approach for unsupervised anomaly detection in time series. IEEE Access 7:1991–2005
35. Meinshausen N (2006) Quantile regression forests. J Mach Learn Res 7:983–999
36. Carel L (2019) Big data analysis in the field of transportation, Ph.D. thesis, Université Paris-Saclay
37. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: 2nd international conference on learning representations, ICLR 2014, Conference Track Proceedings
38. Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), pp 1050–1059
39. Toqué F, Côme E, Oukhellou L, Trépanier M (2018) Short-term multi-step ahead forecasting of railway passenger flows during special events with machine learning methods
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
41. Martin A et al. (2015) TensorFlow: Large-scale machine learning on heterogeneous systems
42. Chollet F et al. (2015) Keras,