

Anomaly prediction in mobile networks : A data driven approach for machine learning algorithm selection

Imed Hadj-Kacem

Orange Labs, France

imed.hadjkacem@orange.com

Sana Ben Jemaa

Orange Labs, France

sana.benjemaa@orange.com

Sylvain Allio

Orange Labs, France

sylvain.allio@orange.com

Yosra Ben Slimen

Orange Labs, France

yosra1.benslimen@orange.com

Abstract—In this paper, we propose a model for proactive anomaly detection in mobile networks. We show that when Key Performance Indicators (KPIs) are highly correlated, the linear regression gives a good accuracy in anomaly detection for a short prediction horizon. When the prediction horizon is far, KPIs become weakly correlated. We propose to transform discrete measurements into functional data and apply a functional data regression method to perform the prediction. We compare proposed models using KPI measurements obtained from a real Long Term Evolution (LTE) network. We show that an improvement in prediction performance can be obtained by using functional data analysis.

Index Terms—Anomaly detection, Radio access network, functional regression

I. INTRODUCTION

Flexible and automatic management is key to facilitate the exploitation and optimization of networks. Since 4G (4th generation) networks, Self-Organizing Networks (SONs) concept have been adopted to automate the management of mobile networks [1]. Today SON is already deployed in several 4G networks and it has proven its ability to enhance the network operational efficiency in addition to the OPEX reduction. Designing efficient SON functions relies on building knowledge from the plethora of network data that can be collected from different data sources. In this paper, we focus on Key Performance Indicators (KPIs) and aim at designing efficient anomaly prediction. KPIs are widely used by engineers for radio network management because of their ease of accessibility and interpretation. KPIs are collected by several elements of the network at discrete time instants. An anomaly describes an unexpected behavior of the network such as highly unexpected value of the call drop rate, or unexpectedly low throughput, etc.

In [2], anomaly detection in mobile network was based on the temporal analysis of the variables (KPIs) by considering them as time series. However, time series have some limitations especially when the dataset have missing or noisy values. Therefore, authors in [3] proposed to transform time series into continuous functions using the functional data analysis framework. The prediction is then performed using a functional regression model [3]. Although this model gives good performance in terms of classification errors, its accuracy

and complexity can be improved if a Principal Component Analysis (PCA) is added. Besides, the proposed model performs anomaly detection in medical and industrial contexts [3]. In the context of mobile networks, we aim at designing a model that predicts that an anomaly will occur in the future, namely proactive anomaly detection or anomaly prediction in order to diagnose and correct the problem before it happens.

In [4], KPIs that are measured at discrete time points are transformed into a functional form using a smoothing technique. Functional principal component analysis (FPCA) is used to reduce the size of the functional data space and capture the maximum of correlation between the variables. Then, a Machine Learning (ML) based classification is performed to implicitly detect complex relationships between variables. The resulting trained model is able to distinguish between normal and abnormal behaviors of the network. Results show that the proposed model can effectively predict future anomalies in the network [4]. Similarly, the authors in [5] proposed some approaches based on machine learning for anomaly detection in mobile networks, namely random forest (RF) and support vector machine (SVM). However, some critics can be expressed about the use of machine learning-based algorithms: complexity of the training phase, choice of the most adapted model and difficulties in interpreting the results.

In this work, we consider the problem of anomaly prediction in radio mobile networks. We consider a dataset of KPI measurements collected on a real network in France. KPIs are often correlated since they describe the behavior of a same network. By exploiting the correlation between the different KPIs, we aim at producing a model that predicts the temporal variation of a given KPI (that we name the forecasted KPI) based on the observation of other KPIs (or variables) considered as predictors. We aim at detecting an abnormal behavior of the forecasted KPI in the future, namely that it exceeds a given threshold. The purpose of the paper is to compare the performance of different approaches in terms of computational complexity and prediction accuracy to give a guideline for choosing the most appropriate technique with respect to the correlation level between the predictors and the forecasted KPI:

- The basic approach consists in considering discrete measurements of KPIs. For this case, a linear regression

model and Distributed-Lag Regression (DLR) are proposed.

- The second approach aims at capturing and exploiting the correlation between the predictors and the forecasted KPI even if it is low. We first represent the data by functions (functional data) using a smoothing technique. Then, we apply a PCA analysis on the functional data to obtain the main modes of variation of the different KPIs. Then we use a classification model based on logistic regression. We compare the performance of this simple classification model to a machine learning based classification approach [4].

The reminder of this paper is organized as follows. Section II develops two regression models based on time series then on functional data. Section III describes the different KPIs considered in the study and calculates their correlation coefficients and Section IV compares the anomaly prediction performance of the different approaches proposed in the study using measurements from a real Long Term Evolution (LTE) network. Section V concludes the paper and gives some perspectives on further studies.

II. ANOMALY PREDICTION USING DISCRETE DATA

KPIs are derived from counters collected from different elements of the network. They are computed periodically with different temporal granularities. When the value of the KPI becomes higher (or lower) than a predefined threshold, the network presents an anomaly which requires a diagnosis or triggers an action on the network parameters for example. KPIs can be more or less correlated. For instance, when the *Average Active User per cell* KPI increases, the *Drop Call Rate* increases as well. Then, anomalies may be correlated. Our aim is to predict the anomaly for a given KPI using the values of the other KPIs by exploiting the correlation between them. Since KPIs are generally considered as time series or multivariate vectors [2], the simple method to predict one of them given the others is the regression approach. The regression method creates a relationship between the forecasted variable (the KPI to be predicted) and p other KPIs (or variables) which will be used as predictors. We consider two cases, the case where the prediction is performed using only one measurement per KPI and the case where KPIs are considered as time series since they are recorded on regular time intervals.

A. Linear regression for discrete dataset

Our aim is to propose a model to predict whether or not an abnormal behavior of the network will occur in a prediction horizon. In this study, KPIs are computed from the LTE Operation Subsystem each 15 minutes. The prediction horizon is the parameter to be set. The expression of the regression that links the variable of interest (to be predicted) Z in a prediction horizon h with p prediction variables $\mathbf{X} = (X_1, \dots, X_p)^T$ is given by

$$z_{i+h} = \delta + \sum_{j=1}^p \kappa_j x_{j,i} + \varepsilon_{i+h}, \quad (1)$$

where i corresponds to the observation time index, $x_{j,i}$ is the i^{th} observation (actual observation) of the j^{th} KPI, ε_{i+h} refers to the error of the model and $\theta = (\delta, \kappa_1, \dots, \kappa_p)^T$ is the vector of the parameters of the model. Note that when $h = 0$, it is a detection problem, otherwise it is a prediction problem. For another time instant i' , the prediction of Z in a prediction horizon h is

$$\hat{z}_{i'+h} = \hat{\delta} + \sum_{j=1}^p \hat{\kappa}_j x_{j,i'}, \quad (2)$$

where $\hat{\theta} = (\hat{\delta}, \hat{\kappa}_1, \dots, \hat{\kappa}_p)^T$ is the estimated vector of the parameters using a least square estimator at the training phase. Then, the anomaly of the considered KPI is detected by comparing $\hat{z}_{i'+h}$ to a threshold. In this paper, we consider some KPIs for which the behavior of the network is normal when they are lower than a predefined threshold. Otherwise, if the KPI is higher than the threshold, an anomaly is detected and an alarm must be triggered to resolve the associated problem. Thus,

- If $\hat{z}_{i'+h} \geq \text{Threshold}$, then $Y = 1$ (Anomaly)
- If $\hat{z}_{i'+h} < \text{Threshold}$, then $Y = 0$ (No Anomaly)

The prediction of Z in a prediction horizon h given the other observations is better when it is strongly correlated with the p variables of $\mathbf{X}$. When the prediction horizon is null the correlation between the forecasted variable and the predictors is maximum. When the horizon of prediction increases, this correlation decreases. Thus, the linear regression is only valid when the variables are strongly correlated and for a near prediction horizon. If not, the regression method will give an erroneous prediction of anomalies. In this case, it is better to consider a series of observations per KPI and perform the prediction based on the trend of variation of these measures, thus fully exploiting the correlation between the variables.

B. Distributed-lag regression

In the following, anomaly prediction is performed using a set of measurements collected during a time window. The chosen size of window is one day with a step size of 15 minutes which gives $m = 96$ values of KPI per day. We consider the distributed-lag regression model where a regression method is used to predict the forecasted variable based on the past values (lagged) of the p predictors [7] in prediction horizon h . The DLR model is given by

$$z_{i+h} = \delta + \sum_{j=1}^p \sum_{l=0}^{m-1} \kappa_{j,l} x_{j,i-l} + \varepsilon_{i+h}, \quad (3)$$

where $\theta = (\delta, \kappa_{1,0}, \dots, \kappa_{1,m-1}, \dots, \kappa_{p,0}, \dots, \kappa_{p,m-1})^T$ is the vector of the parameters of the model. This means that z_i is detected given the actual and the $(m-1)$ last observations of the p predictors when $h = 0$. When $h > 0$, future anomaly in horizon prediction h is predicted using m previous values.

As the case of the linear regression, the predicted value of the considered KPI in a horizon prediction h at time instant i' is

$$\hat{z}_{i'+h} = \hat{\delta} + \sum_{j=1}^p \sum_{l=0}^{m-1} \hat{\kappa}_{j,l} x_{j,i'-l}, \quad (4)$$

where $\hat{\theta} = (\hat{\delta}, \hat{\kappa}_{1,0}, \dots, \hat{\kappa}_{1,m-1}, \dots, \hat{\kappa}_{p,0}, \dots, \hat{\kappa}_{p,m-1})^T$ is the estimated vector of the parameters using a least square estimator at the training phase. Then, the anomaly is prevented by comparing $\hat{z}_{i'+h}$ to the threshold.

Thus, DLR method captures more correlation between the variable of interest and the predictors since several measurements per KPI are considered. However, the measurements within each KPI can be noisy. The accumulation of the corresponding noise terms in the DLR model leads to the degradation in the prediction performance.

III. ANOMALY PREDICTION USING FUNCTIONAL DATA

To cope with the aforementioned issue related to noisy measurements, in [4], the authors consider the KPIs as functional data in order to find the latent relationship between the variables and reduce the amount of noise. A smoothing technique builds a functional form from the discrete observations of the KPIs. A principal component analysis shows the linear combination between the variables that expresses their main modes of variation. However, machine learning algorithms are used instead of regression, which leads to high computational complexity [4]. In the following, we use regression method for functional data to predict anomaly for a given KPI aiming at reducing the computational complexity.

We assume that KPIs are collected at regular time steps. Functional data $\mathbf{X} = (X_1(t), \dots, X_p(t))^T$, $t \in [0, T]$ created from the p KPIs, is the observation of a stochastic process during a continuous time interval T . Each observation at time instant i is a set of p functions. The functions of each KPI are obtained based on the collected samples of the corresponding KPI during the time window T . For the KPI of interest Z , a label $Y \in \{0, 1\}$ is created by comparing its values to the threshold, where $Y = 1$ indicates that there is an anomaly in the prediction horizon h and $Y = 0$ indicates that the future behavior of the network is normal at the same time horizon.

A. Anomaly prediction model

For anomaly prediction, we consider the following model given in Figure 1. The KPI observations are first collected from the network. These observations are noisy and may contain missing values. Hence, we propose as in [4] to transform them into functional data using a smoothing technique in order to reduce the impact of noise. So that, each observation $\mathbf{x}_i$ is described by a set of p functions or curves. The functions are functional estimates that describe the daily evolution of all KPIs. The second step is to perform a functional principal component analysis. This allows optimizing the representation of the functional data by computing their principal components. In this work, all the observations (training and test) are projected on the same smoothing basis and FPCA space.

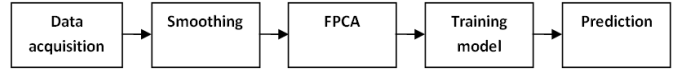


Fig. 1. Prediction model

The third step consists in using a proportion of the labeled dataset (training observations) to develop the training model. Unlike the previous study of [4] where machine learning-based classification algorithms are considered to create the training model, we propose here to use the logistic regression method for functional data. Our aim is to simplify the prediction model by considering the regression which presents a much lower complexity compared to machine learning classification methods (such as the random forest), especially when the number of observations and/or variables is great. We compare both methods in terms of prediction performance and computation time. At the inference phase, the new observation, after being transformed with smoothing and FPCA, is used by the training model to prevent a possible future anomaly in the prediction horizon.

B. Logistic regression for functional data

Let $\mathbf{x}_i$ be a realization of the random process $\mathbf{X}$. The optimal Bayes classifier is given by [8]

$$\hat{y}(\mathbf{x}_i) = \arg \max_{a \in \{0,1\}} P(Y = a | \mathbf{X} = \mathbf{x}_i). \quad (5)$$

By defining $\pi(\mathbf{x}_i) = P(Y = 1 | \mathbf{X} = \mathbf{x}_i)$, the expectation of Y given $\mathbf{x}_i$, the classifier of (5) can be expressed by

$$\begin{cases} \hat{y}(\mathbf{x}_i) = 1 & \text{if } \pi(\mathbf{x}_i) \geq 0.5, \\ \hat{y}(\mathbf{x}_i) = 0 & \text{if } \pi(\mathbf{x}_i) < 0.5. \end{cases}$$

Let $\lambda(\mathbf{x}_i) = \ln \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}$, then the optimal classifier becomes equivalent to

$$\begin{cases} \hat{y}(\mathbf{x}_i) = 1 & \text{if } \lambda(\mathbf{x}_i) \geq 0, \\ \hat{y}(\mathbf{x}_i) = 0 & \text{if } \lambda(\mathbf{x}_i) < 0. \end{cases} \quad (6)$$

Hence, the prediction can be performed by comparing the logit $\lambda(\mathbf{x}_i)$ to 0.

When a logistic regression model is considered, the expectation of Y given $\mathbf{x}_i$ can be modeled as [9]

$$\pi(\mathbf{x}_i) = \frac{\exp [\alpha + \int_T \gamma^T(t) \mathbf{x}_i(t) dt]}{1 + \exp [\alpha + \int_T \gamma^T(t) \mathbf{x}_i(t) dt]}, \quad (7)$$

where α is a real parameter and $\gamma(t)$ is a p multivariate function. Then,

$$\lambda(\mathbf{x}_i) = \alpha + \int_T \gamma^T(t) \mathbf{x}_i(t) dt. \quad (8)$$

The collected data corresponding to different KPIs are observed at discrete time points whereas functional data $\mathbf{x}_i(t)$ belong to an infinite dimensional space. The most widely used method to find the functional form from discrete measurements is to assume that the data belong to a finite-dimensional space. This space is represented by a basis of

functions. Several basis of functions exist in the literature (see [10] for more details). Hence, the j^{th} curve of the observation $\mathbf{x}_i(t)$ can be represented on the basis of function as $x_{i,j}(t) = \sum_{l=1}^M c_{i,j,l} \varphi_l(t)$, $1 \leq j \leq p$, where $\varphi(t) = (\varphi_1(t), \dots, \varphi_M(t))^T$ is the basis of M functions and $c_{i,j,l}$, $1 \leq l \leq M$ are the basis coefficients that can be fit by least squares. In this work, the B-splines functions are considered [10]. Let $\Phi(t)$ be the $(p \times Mp)$ matrix having the q^{th} row $(\mathbf{0}_{1 \times (q-1)M}, \varphi^T(t), \mathbf{0}_{1 \times (p-q)M})$ and $\mathbf{c}_i$ be the $(Mp \times 1)$ vector given by $\mathbf{c}_i = (c_{i,1,1}, \dots, c_{i,1,M}, \dots, c_{i,p,1}, \dots, c_{i,p,M})^T$ [3], then we have $\mathbf{x}_i(t) = \Phi(t)\mathbf{c}_i$. Thus,

$$\lambda(\mathbf{x}_i) = \alpha + \int_T \boldsymbol{\eta}^T(t) \mathbf{c}_i dt = \alpha + \sum_{k=1}^{Mp} c_{i,k} \int_T \eta_k(t) dt, \quad (9)$$

where $\boldsymbol{\eta}^T(t) = \boldsymbol{\gamma}^T(t)\Phi(t)$ and $\eta_k(t)$ is the k^{th} element of $\boldsymbol{\eta}(t)$.

If we define $\beta_k(t) = \int_T \eta_k(t) dt$, for $1 \leq k \leq Mp$, we obtain

$$\lambda(\mathbf{x}_i) = \alpha + \mathbf{c}_i^T \boldsymbol{\beta}, \quad (10)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{Mp})^T$.

By considering that N_{tr} observations are dedicated to the training phase, the estimates of parameters $(\alpha, \boldsymbol{\beta})$ in the sense of the least squares method are those that minimize the quantity

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg \min_{(\alpha, \boldsymbol{\beta})} \sum_{i=1}^{N_{tr}} (\tilde{y}_i - \lambda(\mathbf{x}_i))^2, \quad (11)$$

where $\tilde{y}_i = 2y_i - 1$ (i.e. $\tilde{y}_i \in \{-1, 1\}$) is introduced to take into account that $\lambda(\mathbf{x}_i)$ is positive when an anomaly is recorded and negative otherwise (see equation (6)) and bring the values of $\tilde{y}_i$ and $\lambda(\mathbf{x}_i)$ closer. So that parameters estimation becomes more accurate.

Let $\mathbf{H}$ be the matrix of dimensions $N_{tr} \times (Mp + 1)$ having the i^{th} row $(1, \mathbf{c}_i^T)$, then the least square estimator is given by

$$(\hat{\alpha}, \hat{\mathbf{g}}) = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \tilde{\mathbf{y}}, \quad (12)$$

where $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_{N_{tr}})^T$.

The predicted value of $\lambda(\mathbf{x}_{i'})$ at time instant i' is then

$$\hat{\lambda}(\mathbf{x}_{i'}) = \hat{\alpha} + \mathbf{c}_{i'}^T \hat{\boldsymbol{\beta}}. \quad (13)$$

In order to improve the estimation of the model parameters, we model the variables of prediction by their principle components that have maximum variance. Let $\mathbf{C}$ be the matrix of dimensions $(N \times Mp)$ of the basis coefficients of the all N collected observations having the i^{th} row $\mathbf{c}_i^T$, then we create the matrix $\mathbf{W} = \mathbf{C}\mathbf{V}$, whose columns are the principal components of $\mathbf{C}$, with $\mathbf{V}$ being the matrix whose columns are the eigenvectors of the covariance matrix of $\mathbf{C}$. If we denote by $\boldsymbol{\lambda} = (\lambda(\mathbf{x}_1), \dots, \lambda(\mathbf{x}_N))^T$, equation (10) can be rewritten as

$$\boldsymbol{\lambda} = \alpha \mathbf{1} + \mathbf{W} \boldsymbol{\beta}_{PCA}, \quad (14)$$

where $\mathbf{1} = (1, \dots, 1)^T$ is a $N \times 1$ vector of 1 and $\boldsymbol{\beta}_{PCA} = \mathbf{V}^T \boldsymbol{\beta}$. Parameters $(\alpha, \boldsymbol{\beta}_{PCA})$ are then estimated by the least

square estimation method as in (12) based on the first N_{tr} rows of $\mathbf{W}$. The predicted value of the logit at time instant i' is then

$$\hat{\lambda}(\mathbf{x}_{i'}) = \hat{\alpha} + \mathbf{w}_{i'}^T \hat{\boldsymbol{\beta}}_{PCA}, \quad (15)$$

where $\mathbf{w}_{i'}^T$ is the i'^{th} row of $\mathbf{W}$.

IV. EVALUATION USING LTE KPIs

In this section, we apply the anomaly prediction models on measurements collected from a real LTE network in France. Our aim is to compare the different approaches and to measure the contribution of functional data pre-processing to the anomaly prediction performance.

A. Dataset description

In this study, KPIs are computed from the LTE Operation Subsystem each 15 minutes. The observation window is one day with a step size of 15 minutes which gives 96 values of each KPI. The data set consists of historical data collected on a period of 68 days which results in a total number of observations of 6573 per KPI. We consider the four KPIs given in Table I. The motivation behind this choice is that these KPIs are typically used for detecting and analyzing congestions in an LTE network. Congestion on a cell occurs if the value of one of the four KPIs exceeds a certain threshold.

TABLE I
LIST OF KPIs USED FOR CONGESTION PREDICTION

KPI	Description
Average Latency	Average latency inside the cell (without distinction of the QCI)
Average Active User	Average active users within the cell
DL Traffic Volume	Traffic volume in downlink
DL Load	Utilization rate of PRB (Physical Resource Block) in downlink

As mobile networks are continuously monitored and optimized, the number of congestions is rather small, which makes the class 1 (presence of anomaly) minority compared to class 0 (no anomaly). In this study, the imbalance proportion is almost 1/10. The imbalance between classes can lead to errors in the model during the training phase [11]. Hence, some observations of the majority class are ignored in this work to ensure balance in the dataset. Then, 90% of the labeled dataset are used for training and 10% are used for test. We consider the k -fold cross-validation technique with $k = 10$ [10]. This technique divides randomly the original dataset into k subsamples of equal size of which only one is retained as test data, and the remaining $k - 1$ subsamples are used as training data. The process is then repeated k times, so that each of the k subsamples is used exactly once as validation data. Then, all observations are used for both training and test.

In the following, we give the performance of anomaly prediction in the two cases considered in this paper: the case of discrete dataset for which we use the regression technique

Predicted Class \ Actual Class	Anomaly=0	Anomaly=1
Anomaly=0	True Negative (TN)	False Positive (FP)
Anomaly=1	False Negative (FN)	True Positive (TP)

Fig. 2. Confusion Matrix

given in (2) and the case of functional data for which a pre-processing of data is performed (smoothing and PCA). The results are then compared with those obtained by a technique based on machine learning as in [4]. In order to measure the quality of the classification methods, we built the confusion matrix (CM) which calculates the number of the true and false predictions for both classes [6]. As shown in Figure 2, the columns of the confusion matrix represent the number of occurrences of a predicted class, while the rows represent the number of occurrences of an actual class. Four performance indicators that evaluate the classification method can be deduced from the CM:

- Correct Classification Rate or Accuracy ($A = \frac{TP+TN}{TP+TN+FP+FN}$): defines the proportion of true predictions.
- Recall ($R = \frac{TP}{TP+FN}$): is defined by the ratio of the detected anomalies to the total number of anomalies in the labeled Dataset. A low value of R means that the classification method does not predict anomalies well.
- Precision ($P = \frac{TP}{TP+FP}$): measures the reliability of a prediction of anomalies. When P is low, the number of false alarms (false alert of the presence of an anomaly) is high.
- F-measure ($F = 2 \frac{P \cdot R}{P+R}$): is the weighted harmonic mean of the precision and the recall. A high value of F-measure means that both precision and recall are high.

B. Impact of correlation on anomalies prediction

Now, we study the performance of the predictor based on linear regression and using discrete data in terms of congestion prediction in mobile network. In fact, because of the random distribution of the users over cells and the variety of sessions, some cells are more loaded than others. When the load is unbalanced between cells, it can cause congestion and the victim cell can no longer serve the users located in its coverage area. This results in a degradation of the quality of service perceived by the user. Thus, prediction of the future congestion is essential for a mobile operator to guarantee a better quality to its customers. For the network engineers, the congestion is produced if one of the KPIs described in Table I shows degradation. The data is then labeled based on this condition. Figure 3 shows the performance metrics of the prediction of congestion based on linear regression using discrete dataset. The prediction horizon varies between 15 minutes and 4 hours. We note that when the prediction horizon increases, the performance of the model degrades because the correlation between the future congestion and the actual values of the KPIs falls especially when the temporary

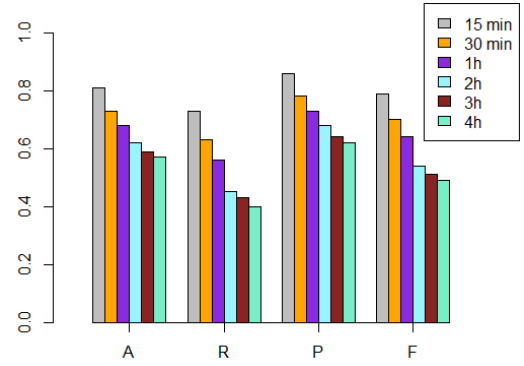


Fig. 3. Performance of regression based on discrete dataset in terms of prediction of congestion

shift is important. This is illustrated in Table II where we calculate the correlation coefficient between the KPIs (current values) and the congestion in the prediction horizon (future values). We note that for a horizon greater than one hour, the correlation becomes very weak. As a result, the model is not able to reliably prevent future congestions 1 hour in advance (or more) since $R \simeq 55\%$. Even worse, when the prediction horizon is equal to 2 hours, the model detects only 45% of the anomalies and 32% of them are false alarms (as $P = 68\%$).

TABLE II
PEARSON CORRELATION COEFFICIENT BETWEEN CONGESTION FOR A GIVEN PREDICTION HORIZON AND THE OTHER KPIs

Prediction horizon	Average Latency	Average Active User	DL Traffic Volume	DL Load
0 min	0.62	0.51	0.62	0.71
15 min	0.41	0.43	0.46	0.52
1 hour	0.23	0.31	0.30	0.32
2 hours	0.14	0.21	0.22	0.21

C. Improved prediction by using functional dataset

Since the performance of the predictor based on discrete data is bad when the variables are weakly correlated or when the prediction horizon is far, we study in this section the impact of the use of functional data on the prediction of anomalies. We consider that $m = 96$ measurements per KPI are used for DLR prediction while each KPI observation (96 measurements) is represented by $M = 25$ coefficients after passing through the smoothing block. Figure 4 gives the experimental results in terms of accuracy and F-measure of four prediction models: linear regression (LiReg) and DLR based on discrete dataset, logistic regression (LoReg) using functional dataset and the random forest algorithm also using functional dataset. We choose RF as ML based method because it presents the best performances for this use case compared to the other ML based techniques. The horizon of prediction varies between 0 (detection) and 4h. Prediction of congestions in the future allows to perform proactive load balancing. From Figure 4, several interpretations can be identified:

- When the prediction horizon is short (less than 15 minutes), the classical linear regression offers the same

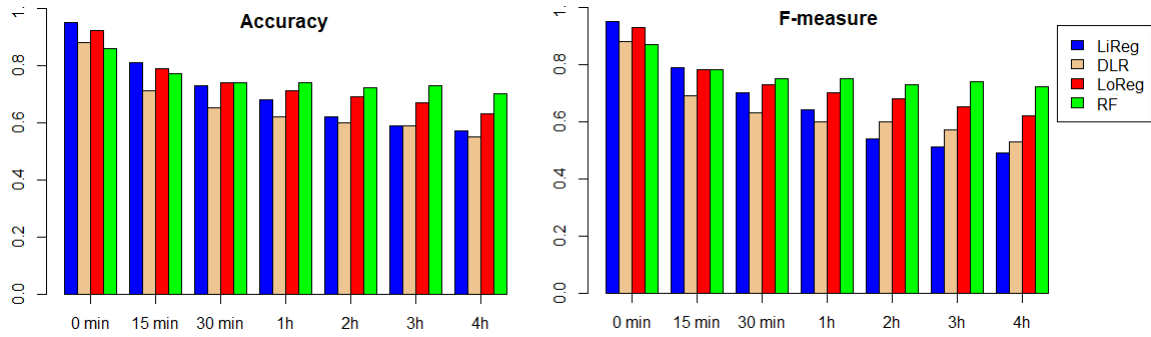


Fig. 4. Experimental results of the considered methods in terms of F-measure and accuracy for congestion prediction

performance as other robust methods that performs data processing before setting up the prediction model (LoReg and RF). This is explained by the strong correlation between future anomalies and current KPI measurements when the prediction horizon is small.

- DLR performance is lower compared to linear regression when the prediction horizon is short. Indeed, the use of several previous measurements per KPI will weaken the correlation between the forecast variable and the predictors for a close horizon. Thus, it is useless to consider the trend of KPIs in this case.
- When the prediction horizon is far, the correlation between future congestion and KPIs is weakened. It is therefore interesting to predict anomalies based on the trend. This justifies the improved performance of the DLR technique compared to the linear regression.
- However, it is clear that DLR suffers from the too noisy measurements of KPIs in finding that the transition to the functional data greatly improves the performance of the prediction. When the prediction horizon is equal to 30 minutes, we have an accuracy of 74% for LoReg but 65% for DLR and an F-measure of 73% for LoReg but 63% for DLR. Indeed, the values of KPIs collected by the network correspond to an average of 15 minutes and this average does not reflect the fluctuation of the KPI which could be fast because of the random behavior of the users which would lead to measurement errors.
- The two prediction methods based on functional data have the potential to effectively prevent future congestions even 4 hours in advance, unlike the classical regression that uses discrete KPI measurements that fails to predict the anomaly 1 hour in advance. This is due to the functional data processing based on the smoothing and the PCA which manages to extract a maximum of correlation between the variables and reduce the amount of noise associated with the data which improve the performances of the models.
- We note that regression of functional data offers performances close to the RF method. For example, when the prediction horizon is 1 hour, the values of accuracy and F-measure are respectively 71% and 70% for the logistic regression model and those of RF method are respectively 74% and 75%.

In order to compare the complexity of the different methods, we give their convergence times in Table III. It should be mentioned that the models based on functional data are more complex than the ones based on discrete data since the collected measurements of KPIs must be projected on a function basis with the smoothing method and a matrix calculation is required with the PCA technique. Also, the random forest model is the most complex technique as it performs parallel learning on multiple decision trees randomly constructed and trained on different subsets of data according to the principle of bagging [12].

TABLE III
CONVERGENCE TIME OF THE DIFFERENT CONSIDERED METHODS

	Discrete data	Functional data	Execution time (min)
LiReg	x		0.5
DLR	x		1.3
LoReg		x	8
RF		x	15

V. CONCLUSION

In this paper, we compared classification approaches for anomalies prediction in a real mobile network. We showed that the basic method of linear regression based on discrete measurements gives the best performance-complexity trade-off when the prediction horizon is short. When the prediction horizon is far, we recommended preprocessing data based on smoothing and PCA before using classification techniques. In this study, we considered a limited set of variables to supervise the congestion while the network uses many more KPIs to continuously monitor the quality of service. In future work, we propose to include a variable selection method for the different prediction models [13]. The difficulty in the mobile network context is that the KPIs used as predictors are usually correlated. Therefore, some of them will be insignificant since they provide the same information as others. Several methods for selecting non-correlated predictors have been proposed in the literature (see [14] for example) and can be applied in a future study.

REFERENCES

- [1] 3GPP, "Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements," 3rd Generation Partnership Project (3GPP), TS 32.500, Jul. 2008.

- [2] J. Wu, P. P. C. Lee, Q. Li, L. Pan, and J. Zhang, "CellPAD: Detecting Performance Anomalies in Cellular Networks via Regression Analysis", in *17th International IFIP Networking Conference*, 2018, pp.19–27.
- [3] T. Górecki, M. Krzyśko, and W. Wolyński, "Classification problems based on regression models for multi-dimensional functional data," *Statistics in Transition new series*, vol. 16, No. 1, pp. 97–110, 2015.
- [4] Y. Ben Slimen, S. Allio, and J. Jacques, "Anomaly prediction in Radio Access Networks Using Functional Data Analysis," in *IEEE Global Communications Conference*, 2017, pp. 1–6.
- [5] P. Casas, P. Fiadino, and A. D'Alconzo, "Machine-learning Based Approaches for Anomaly Detection and Classification in Cellular Networks", in *Proc. of IFIP TMA*, 2016.
- [6] A. Gunawardana and G. Shani, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks," *The Journal of Machine Learning Research*, vol. 10, pp. 2935–2962, 2009.
- [7] D. S. Batten and D. L. Thornton, "Polynomial. Distributed Lags and the Estimation of the St. Louis Equation", *Fed. Res. Bank St. Louis Rev.*, vol.65, no.4, pp. 13–25, 1983.
- [8] T. W. Anderson, *An introduction to Multivariate Statistical Analysis*, 3rd Edition, Wiley and Sons, Inc., New Jersey, 2003.
- [9] G. M. James, "Generalized linear models with functional predictors", *Journal of the Royal Statistical Society, Series B*, vol. 64, no. 3, pp. 411–432, 2002.
- [10] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd Edition, Springer, New York, 2009.
- [11] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal Of Artificial Intelligence Research*, vol. 16, pp 321–357, 2002.
- [12] L. Breiman, "Random Forests", *Machine Learning*, Vol. 45, pp. 5–32, 2001.
- [13] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection- A review and recommendations for the practicing statistician," *Biometrical journal*, vol. 60, no. 3, pp. 431–449, 2018.
- [14] D. B. Sharma, H. D. Bondell, and H. H. Zhang, "Consistent Group Identification and Variable Selection in Regression With Correlated Predictors", *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 319–340, 2013.