Anomaly detection mechanisms to find social events using cellular traffic data

Rosario G. Garroppo
Dipartimento di Ingegneria dell'Informazione
University of Pisa
Pisa, Italy
Email: r.garroppo@iet.unipi.it

Saverio Niccolini
Network Research Division
NEC Laboratories Europe
Heidelberg, Germany
Email: Saverio.Niccolini@neclab.eu

***Abstract*—The design of new tools to detect on-the-fly traffic anomaly without scalability problems is a key point to exploit the cellular system for monitoring social activities. To this goal, the paper proposes two methods based on the wavelet analysis of the cumulative cellular traffic. The utilisation of the wavelets permits to easily filter "normal" traffic anomalies such as the periodic trends present in the cellular traffic. The two presented approaches, denoted as Spatial Analysis (SA) and Time Analysis (TA), differ on how they consider the spatial information of the traffic data. We examine the performance of the considered algorithms using cellular traffic data acquired from one the most important Italian Mobile Network Operator in the city of Milan throughout December 2013.**

The results highlight the weak points of TA and some important features of SA. Both approaches overcome the performance of one reference algorithm present in literature. The strategy used in the SA emerges as the most suitable for exploiting the spatial correlation when we aim at the detection of the traffic anomaly focused on the localisation of social events.

***Index Terms*—Discrete Stationary Wavelets Transform; Anomaly Detection; Cellular traffic; Social Events; Spatial Analysis; Time Analysis.**

I. Introduction

Identifying real-world phenomena through the analysis of network traffic has recently attracted the attention of many researchers. This topic is one of the most promising in the context of cellular networks. In these systems, in addition to the traffic patterns of users, we have rough information on their geographic location. Cellular traffic data can easily be extended and augmented with external information, such as the location of nearby events, etc. Some mobile network operators have launched many research challenges and made available their traffic data to explore the potentialities that the analysis of cellular network data offers [1][2][3]. Based on these data, researchers have carried out a lot of studies considering different problems. For example, in [4] the focus is on the paths of user mobility, while in [5] the authors define a method to determine the levels of poverty in different geographical areas. Other works have proposed methods to link the traffic patterns to external real-world events and observations, see the survey [6] for details.

In this work, we study how the aggregate cellular traffic data can be used to identify social events. We pose this as an online anomaly detection problem and to solve it we propose scalable algorithms, which can run in the operational network. Our mechanisms do not require detailed information on traffic such as user Id, per packet or per flow information as in [7][8]. Instead, they use only cumulative information on the observed traffic at different Base Stations (BSs). The algorithms assume the continuous monitoring of the amount of traffic in non-overlapping Time Slots (TSs) and for each BS. We use wavelet decompositions to analyse the traffic over different timescales. This approach permits to easily filter periodic trends in the observed signals and to highlight the variation of traffic patterns in the time. The basic idea of the proposed methods is the exploiting of the spatial correlation of these traffic variations. The two presented approaches, i.e. Time Analysis (TA) and Spatial Analysis (SA), differ on how they consider the spatial information. In particular, TA is based on the application of hard thresholding de–noising techniques, which consider separately the traffic behaviour for each BS. At a particular TS, TA confirms a detected alarm in a BS only if in the neighbouring area more than a selected number of other BSs have triggered alarms. This strategy permits to take into account the spatial information of the traffic. On the contrary, at a particular TS, SA evaluates if to trigger or not an alarm by means of the analysis of the traffic variations over the whole network.

The rest of the paper is organised as follows. The next Section discussed the related work and summarises the paper contribution. Section III presents the background on wavelets theory, and in particular on the Discrete Wavelets Transform (DWT) and discrete Stationary Wavelets Transform (SWT). Section IV describes the proposed approaches, while Section V presents the dataset and some preliminary tests. Section VI discusses the simulation analysis carried out with the actual data. The analysis compares the performance of the proposed approaches and of one reference algorithm present in literature. Section VII shows the detection performance of the proposed solutions carried out adding random artificial anomalies to the original data. Section VIII draws the concluding remarks.

II. Related Work

In the last years, the problem of detecting traffic anomalies in cellular systems has attracted the attention of a lot of researchers. The results are a set of methods, which differ in the required traffic information, the kind of detected anomalies, the used tools, the online or offline operation, etc. As an example, in [7] the authors present a graph–based anomaly detection to find anomalies that can aide in the security of users, their phones, their personal information, and the companies that provide them services. In [8], the authors address the problem of automatic network traffic anomaly detection and classification using Machine Learning (ML) based techniques.

Different works have considered the design of methods to run on-the-fly in commercial networks. For example, in [9] the authors propose a method for identifying deviations in timeseries distribution based on a statistical change detection algorithm. The study is based on a large dataset from an operational 3G mobile network. The proposed method is able to cope with the marked non-stationarity and daily/weekly seasonality that characterise the traffic mix in a large public network. In [10], the authors present a method based on the change point detection applied to two sets of features extracted from DNS data. The symptomatic features are defined such that their abrupt change directly relates to the presence of abnormal and potentially harmful events, while diagnostic features shall provide contextual details of the anomalies, pointing to their root causes. These methods consider per-user traffic information. This approach requires a large amount of data to be processed, with a higher complexity of the monitoring platform. On the contrary, our methods are based on the knowledge of the aggregated traffic in each BS. This assumption implies a lower amount of data to consider in the detection algorithm, leading to a lower complexity of the monitoring system.

Traffic anomaly techniques are based on different approaches, as shown in [11][12][13]. For example, we can mention the machine learning approach proposed in [14][15], the combination of filtering and statistical methods discussed in [16], the technique based on principal subspace tracking suggested in [17], the traffic feature distributions used in [18][19], the utilisation of big data analytics presented in [20], the method based on the variation in the entropy associated to the network traffic [21], and the kernel recursive least squares used in [22].

We have chosen to exploit the properties of wavelet decomposition for designing our scheme. In [23], the authors suggested the wavelet transform for the modelling and the synthetic generation of multifractal traffic. More recently, this tool has represented the base for the design of some online traffic anomaly detection methods. In [24], the authors proposed a tool exploiting the wavelet packets in order to detect network traffic attacks in real time. The detection mechanism of the tool considers the iterated cumulative sums of squares (ICSS) algorithm and the Schwarz information criterion (SIC) algorithm for the identification of multiple variance change points in sequence data. These algorithms are integrated with another approach aimed at detecting sharp jumps and cusps in the data. However, the proposed method does not consider spatial information. In [25], the authors present a technique for traffic anomaly detection based on the analysis of the correlation of destination IP addresses in outgoing traffic at an egress router. The address correlation data are transformed through discrete wavelet transform for effective detection of anomalies through statistical analysis. The technique requires the IP address information, which is expensive to acquire.

In [26], the authors propose a signal analysis technique for detecting network traffic anomalies. They analyse the applications of general wavelet filters to the traffic data representing the byte and packet counts, over five minute sampling intervals, from wide-area routing links. They show that wavelets provide a powerful tool for isolating characteristics of signals via a combined time-scale representation. With respect to [26], we add the spatial information associated to the BSs in order to locate the alarm and then the related social event.

Focused on the anomaly detection for DoS attack prevention or for network malfunctioning identification, other works proposed wavelet theory to improve well-known techniques, e.g. CUSUM in [27], or to design new algorithms that jointly use other techniques, e.g., PCA in [28]. This paper differs from these works because the anomaly detection is aimed at the triggering of alarms related to social events. Furthermore, we exploit the available spatial information.

A. Cellular traffic data and real-world phenomena

A lot of works have used cellular traffic data to relate network information and real-world phenomena. Most of them are based on the data made available in the framework of the Data For Development (D4D) challenges [1][2].

In [6], the authors provide an extensive review of results on the analysis of mobile phone datasets, characterised by detailed information. A large amount of data in these datasets permit to carry out research on social networks, mobility, geography, urban planning, help towards development, and security. Detailed information available in the D4D datasets permits to study the relation between cellular traffic and social environment and events. For example, in [4] the authors describe some strategies to identify the cities, the city population, the strength of social ties among cities, the urban mobility in the largest city of Abidjan, the residential districts and the work areas. In [29], the authors have monitored six million users of a mobile network in Wuxi (China) from October 24, 2013, to March 24, 2014 for recovering individuals' commute routes. The dataset includes a huge amount of information, but it is not publicly available.

In this work, we use the Open Data of Telecom Italia (ODTI) [3], which contains cumulative information on the traffic acquired in different BSs. Unlike D4D datasets, ODTI does not provide data on individual users. ODTI dataset does not permit to carry out analysis aimed at finding anomalies for aiding in the security of users or at evaluating the mobility pattern. However, ODTI dataset is complete and contains all the measured data for each BS in each TS of the observation period. In [30] the authors present early experiments in predicting land use and demographics considering jointly heterogeneous open data and an ODTI [3] dataset. These

ODTI data provide phone activity information over time and also over space (due to the positioning of transceiver towers, i.e. BSs), which is a strong indicator of the presence of people and of the mobility in urban environments. They demonstrate that an approach leveraging diverse datasets can be effective in aiding smarter urban planning efforts. In [31], the authors demonstrate that ODTI data can be used to infer information about the behaviour of foreign people in Milan, using simple statistical tools.

Recent studies aim at the estimation of the road traffic starting from mobility data acquired from cellular networks. As an example, the authors of [32][33][34] present systems able to detect incidents and predict travel times on main traffic roads. These systems are based on the monitoring of the mobility paths of cellular network devices or on the analysis of signalling traffic in the core network. The proposed approaches require per-user traffic information.

The contribution of this paper significantly differs from these previous works because we aim at defining an on-the-fly and scalable algorithm that permits us to isolate and characterise patterns that substantially deviate from the "normal" behaviour of the network. The idea is to regard the traffic load of the different BSs in a large cellular network as the measurements from a large (spatially) distributed sensor network. As such this work addresses the problem of spatio–temporal anomaly detection in a smarter city context with large numbers of deployed sensors.

B. Paper contribution

The main contribution of the paper can be summarised as follows.

- We define two different approaches based on SWT for detecting cellular traffic anomalies related to social events. The utilisation of the wavelets allows to easily filter the periodic trends present in the cellular traffic and to have scalable algorithms that can be deployed in a real operational context. The key feature of the proposed algorithms is the utilisation of the spatial information available in the cellular traffic data. Furthermore, they are based on the analysis of cumulative traffic per each BS and do not require detailed traffic information.
- We provide an extensive analysis of the proposed approaches using actual traffic data both with and without artificial anomalies. The simulation results provide a guideline for the design of algorithms able to exploit spatial traffic correlation for the detection of traffic anomalies related to localised social events. The proposed SA approach shows two key advantages. First, it is able to detect traffic anomalies in some critical scenarios, such as the "Christmas Day". In this day, we have a noticeable traffic increment over the whole network, which hides "abnormal" traffic growth related to localised social events. Second, SA is able to detect the anomaly also in time periods when the traffic volume is negligible w.r.t. the average. The simulation results point out that the strategy used in SA is the most suitable for detecting localised traffic anomaly exploiting the spatial correlation of cellular traffic.

III. Background on the discrete wavelet transform (DWT)

In this section, we recall some basic concepts on wavelets analysis useful to understand important aspects of the proposed methods. Details on the wavelets theory can be found in [35], while Table I summarises the notation used in the paper.

Symbol	Meaning
$S_i(n)$	The timeseries of the measured traffic at TS n, $n = 1,...,N$, in the BS i – The subscript i is not reported when the BS identifier is not important for understanding the text
$j = 1,\cdots,J$	Set of layers, $j = 1$ is the lowest
$AC_j(A,f)$	Approximation Coefficient (AC) of index $f = 1,...,N$ of signal A at layer j – When we refer to the set of all ACs at layer j, we omit f
$DC_j(A,f)$	Detail Coefficient (DC) of index $f = 1,...,N$ of signal A at layer j – When we refer to the set of all DCs at layer j, we omit f
H	Linear filter used to obtain AC_j
L	Linear filter used to obtain DC_j
dbK	Daubechies wavelet of order K
$Th^{\hat{i}}_{Hj}$	High threshold at layer j for the $DC_j(S_{\hat{i}})$ in the TA approach – $\hat{i}$ is the considered BS
$Th^{\hat{i}}_{Lj}$	Low threshold at layer j for the $DC_j(S_{\hat{i}})$ in the TA approach – $\hat{i}$ is the considered BS
$Th_{Hj}(\hat{t})$	High threshold at layer j for the $DC_j(S_i, f = \hat{t})$ in the SA approach – $\hat{t}$ is the considered TS
$Th_{Lj}(\hat{t})$	Low threshold at layer j for the $DC_j(S_i, f = \hat{t})$ in the SA approach – $\hat{t}$ is the considered TS

TABLE I
Notation used in the paper

The first step of a wavelet analysis is the choice/design of the wavelet function, which depends on the considered application. This function allows to obtain the filters H and L. The choice should be based on a careful balance between time localisation and frequency localisation. The time localisation characteristic is strictly related to the length of the filters. This feature has an impact on the ability to easily distinguish between strong short duration change in traffic volume, with respect to a milder change of longer duration. In fact, short filters lead to localise significant traffic volume changes in the time.

The frequency localisation can be measured by means of the number of vanishing moments of the filter H. By definition, a filter H has r vanishing moments when the first r derivatives of its Fourier transform evaluated at 0 are null, i.e. $\hat{H}(0) = \hat{H}^{'}(0) = \hat{H}^{''}(0) = ... = \hat{H}^{r}(0) = 0$, where $\hat{H}$ is the Fourier series of H. To have a wavelet transform with a high number of vanishing moments, longer filters are needed. Filters with a low number (usually one or two) of vanishing moments may lead to increase the number of false positive alarms because large wavelet coefficients may appear at times when no significant event is occurring.

In our study, we analysed two common wavelet families characterised by being orthogonal and compactly supported (see [35] for details). These families are the Daubechies and the symlets. The orthogonality provides the properties deriving from the projection of a signal over an orthogonal basis. Among these, the most significant is the maintaining of the energy features. Both families have the important properties of a small support, a well-defined number of vanishing moments and a known impulse response for both H and L.

The dbK has a support width equal to $2K-1$, a number of vanishing moments equal to K, and the filter length equal to $2K$. The dbK family is asymmetric, whereas the symlet family is near symmetric. This property of the symlet family implies that H and L are near linear-phase filters. In our study, we do not reconstruct the signal from the coefficients, but we use directly DC_j to detect alarms. Thus, the linear-phase property of the symlet family is not significant. Based on this analysis, we carried out some preliminary tests with our dataset. The results of this analysis suggested selecting the $db4$ wavelets for our study.

A. *Discrete Stationary Wavelets Transform (SWT)*

The DWT is time variant because the alignments between features in the signal and features of basis elements can generate differences in the coefficients. Signals exhibiting similar features, but with slightly different alignment in time or frequency might generate fewer of the artefacts we are interested in. To correct unfortunate mis-alignments between signal and basis features, one approach is to shift signals. However, it is hard to establish a priori the best shift that leads to the alignment. In a signal with several discontinuities, these may interfere with each other: the best shift for one discontinuity in the signal may also be the worst shift for another discontinuity. Different studies have considered this problem and have proposed solutions for overcoming it. These solutions are based on the same basic idea: to adopt a frame expansion that is essentially an undecimated discrete wavelet transform. One of the first implementation, known as *algorithme á trous* [36], is based on the up–sampling of the filters response. This result can be obtained inserting zeros between non-zero filter taps, instead of decimating the filter output. The approach allows to overcome the alignment problems of DWT. Given the signal $S(n)$, for each layer j the SWT generates $AC_j(S)$ and $DC_j(S)$, each one having the same number of samples as $S(n)$.

IV. The proposed approaches

For each BS, a timeseries $S(n)$ represents the measured traffic in each TS n. $S(n)$ can be divided into two components, i.e. $S(n) = T(n) + W(n)$. $T(n)$ indicates the traffic composed by its "normal" pattern and its anomalies. The "normal" pattern includes the trends commonly observable in traffic data, such as night/day, working/no working day, etc.. $W(n)$ is the noise we assume to be white. The BSs located at different geographic points have dissimilar traffic intensity. The related timeseries consequently have a different energy. In order to have a meaningful comparison among the traffic of the different BSs, all timeseries should have the same energy. Thus, we apply an energy normalisation to the acquired timeseries before the SWT analysis.

The wavelet transform is additive. Consequently, the coefficients of $S(n)$ are the sum of those obtained from $T(n)$ and $W(n)$. In other words, the SWT of $S(n)$ will generate

$$AC_j(S,f) = AC_j(T,f) + AC_j(W,f) \tag{1}$$

$$DC_j(S,f) = DC_j(W,f) + DC_j(W,f) \tag{2}$$

At each layer j, $AC_j(T)$ describes the "normal" trends of $T(n)$ given that these coefficients are obtained from a low pass filter. $DC_j(T)$ contains information on the deviation from the normal behaviour of $T(n)$. Both $AC_j(T)$ and $DC_j(T)$ are noisy coefficients given the presence of $W(n)$. We observe that in the case of white noise, the related SWT coefficients on all the scales are white noise with decreasing variance for increasing layer, see [37] for details. In many cases, $DC_j(T)$ coefficients are smooth enough, except in some intervals localised for example at the neighbour of transitory phenomena. This structure makes $DC_j(T)$ sparse. The traffic anomalies with respect to the normal trends are then well represented by $DC_j(T)$. De-noising $DC_j(S)$, we can obtain an estimate of $DC_j(T)$ useful for localising transitory phenomena in the time. The de-noising is obtained using threshold-based approaches, where large $DC_j(S,f)$ coefficients are retained assuming that the localised energy of anomalous events leads to having $DC_j(W,f)$ values negligible w.r.t. $DC_j(T,f)$.

A. *Time Analysis approach*

Different de-noising strategies are available, see [38] for details. In our solution, we refer to hard thresholding, which exactly preserves the coefficients above the threshold. We choose this solution because it can keep the height of the peaks.

In the de-noising process, we need to choose if the threshold is independent or not from the layer j. In general, if the noise is stationary and correlated, the wavelet transform has a de-correlating effect. In other words, the $var[DC_j(W,f)]$ is independent of f and can be assumed equal to σ_j^2. Hence, the obvious solution is to estimate a threshold for each layer j.

To select the threshold, the simple way is to set it to q times σ, with q that can be set to 3 or 5, after some tests. However, some strategies having a theoretical basis can lead to defining automatic algorithms. These strategies can be divided in fixed and data-dependent. The first one sets the threshold in advance of data observation, such as to set the threshold to $q\sigma_j$ or to the Universal Threshold (UT),

$$UT = \sigma\sqrt{2\log(N)} \tag{3}$$

where N is the number of samples. The second one sets the threshold after the data observation, as for example the False Discovery Rate (FDR) thresholding, presented in [38].

In the fixed approaches, when we do not have a model for σ, the estimation of this parameter requires the analysis of the data. In our approach, the estimate of σ at each layer is obtained assuming that the noise coefficients on all the scales are white noise. The SWT does not imply that the variance of $DC_j(W)$ is the same for all j. Assuming $DC_j(W)$ as a sequence of i.i.d. random variables with Gaussian distribution, the σ_j estimator has the form

$$\widehat{\sigma_j} = \frac{median(|DC_j(W)|)}{0.6745} \tag{4}$$

This equation is obtained taking into account that for l i.i.d. Gaussian random variables $X_1, X_2, \cdots, X_l$ with null mean and standard deviation σ, we have

$$E(median(|X_i|, 1 \leqslant i \leqslant l)) \approx 0.6745\sigma \tag{5}$$

Established the thresholds $Th^{\hat{i}}_{Hj}/Th^{\hat{i}}_{Lj}$, the simplest algorithm triggers an alarm at the time $\hat{t}$ in the BS $\hat{i}$ when there is at least a layer j where the $DC_j(S_{\hat{i}}, \hat{f} = \hat{t})$ is higher/lower than $Th^{\hat{i}}_{Hj}/Th^{\hat{i}}_{Lj}$. This simple solution does not exploit the available spatial information and leads to a high number of False Alarms (FAs). To account for the spatial information, to the previous algorithm we add the strategy described in the following.

Let $AL(n,i)$ be the number of BSs triggering the alarm at TS n in an observation area around BS i.

Let TBA be the total number of BSs in the observation area around i.

If $\frac{AL(n,i)}{TBA} > P_{al}$ then the alarm of BS i at TS n is confirmed, else it is assumed to be FA.

The strategy requires two parameters: the size of the observation area around i, and the threshold P_{al}.

B. Spatial Analysis approach

Using the notation of Table I, the spatial approach is based on the distribution analysis of the $DC_j(S_i, f)$ over the space. In other words, fixed the index $\hat{f}$, and the layer $\hat{j}$, we analyse the distribution of $DC_{\hat{j}}(S_i, \hat{f})$ over the BS identifiers i. Given that we normalised the energy of all timeseries, in the "normal" conditions, $DC_j(S_i, f)$ should be mainly impacted by the noise. Indeed, the $DC_j(T_i, f)$ contains information on the variation of the T_i at layer j. In the case of "normal" conditions these variations tend to be null, as j increases, given that the information on the "normal" pattern of the traffic is summarised by the $AC_j(T_i, f)$. The weight of each $DC_j(S_i, f)$ is mainly due to $DC_j(W_i, f)$, except for the set of f where the variations of T_i lead to have negligible values of $DC_j(W_i, f)$ w.r.t. $DC_j(T_i, f)$.

The proposed strategy assumes that the $DC_{\hat{j}}(S_i, \hat{f})$ distribution over i is Gaussian with mean, $\mu_{\hat{j}}(\hat{f})$, and standard deviation, $\sigma_{\hat{j}}(\hat{f})$. These two parameters can be estimated on-the-fly. The high threshold and the low threshold are respectively

$$Th_{H\hat{j}}(\hat{f}) = \mu_{\hat{j}}(\hat{f}) + Q\sigma_{\hat{j}}(\hat{f}) \tag{6}$$

$$Th_{L\hat{j}}(\hat{f}) = \mu_{\hat{j}}(\hat{f}) - Q\sigma_{\hat{j}}(\hat{f}) \tag{7}$$

The parameter Q is the α–percentile of the normalised Gaussian distribution, i.e. with zero mean and unitary standard deviation. At each TS $\hat{t}$, BS $\hat{i}$ and layer $\hat{j}$, an alarm is triggered if assuming $\hat{f} = \hat{t}$ the $DC_{\hat{j}}(S_{\hat{i}}, \hat{f})$ is higher than $Th_{H\hat{j}}(\hat{f})$ or lower than $Th_{L\hat{j}}(\hat{f})$.

V. Dataset and Preliminary Tests

We have carried out some preliminary tests using ODTI dataset. This dataset belongs to "Open Big Data" initiative of Telecom Italia (TI). The data have been obtained by the processing of the Call Detail Records (CDRs) acquired in TI cellular network over the city of Milan during December 2013. The dataset presents normalised measurements of the Received/Sent SMS and Incoming/Outgoing Calls. Internet traffic data are also available, but these do not have the same normalisation parameter used in the SMS and call activity traffic. For sake of clarity, we do not take into account Internet traffic data because they confirm the results obtained analysing SMS and Calls traffic. The ODTI dataset reports the activity measurements obtained by aggregating CDRs in TS of 10 min. Furthermore, the CDRs processing has spatially aggregated the data using a virtual grid. Each spatial slot of this grid is a square of side 250 m. A matrix composed by $100 * 100$ non-overlapped spatial slots covers the whole Milan area, and represents the so-called Milan GRID [3]. In the remainder of the paper, we assume that each spatial slot of the Milan GRID is covered by a BS and will be referred to as BS.

A. Preliminary tests on the wavelet approach

As described in Section III, the ACs take into account the "normal" pattern, while the DCs summarise the changes. To highlight this feature, figures 1 and 2 show the traffic observed in the BS 5735, the ACs and the DCs, evaluated at layer 1 and 3 respectively. The figures clearly point out that at each layer the ACs replicate the "normal" trend of traffic, while the DCs summarise the changes at each TS.

We concentrate the analysis of these changes at each layer on the determination of their statistical features. In particular, we first consider the autocorrelation aspects. Referring to the BS 5735, the figure 3 shows the autocorrelation function of the traffic data, the AC_1 and the DC_1. Furthermore, the plots report the approximate upper and lower 95%-confidence bounds estimated assuming that each considered dataset is a Gaussian white noise. The results of the figure show the similarity of the traffic sample and the AC_1 in terms of the autocorrelation function, while the DC_1 has a small autocorrelation. The correlation is limited to few lags, although it is not a white noise.

The second analysis considers the distribution of the DCs. The results are summarised in figure 4, which plots the histogram of the observed DCs at layer 1-4. The shown histograms suggest that the Gaussian assumption can be accepted.

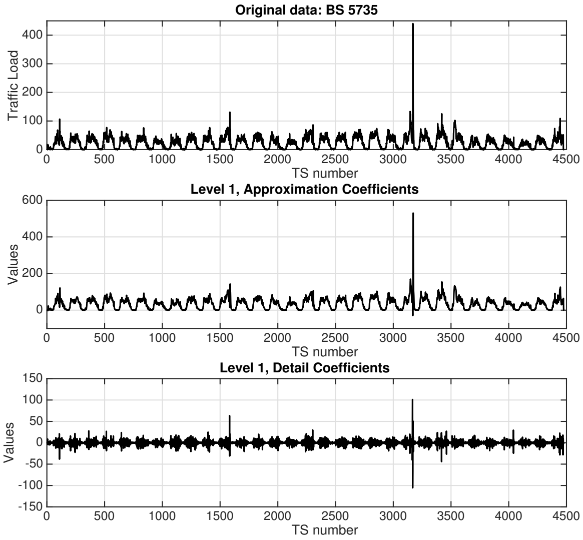

Fig. 1. Traffic data, AC_1 and DC_1 - BS 5735

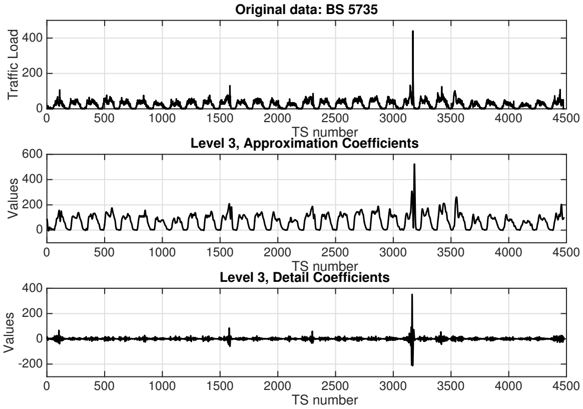

Fig. 2. Traffic data, AC_3 and DC_3 - BS 5735

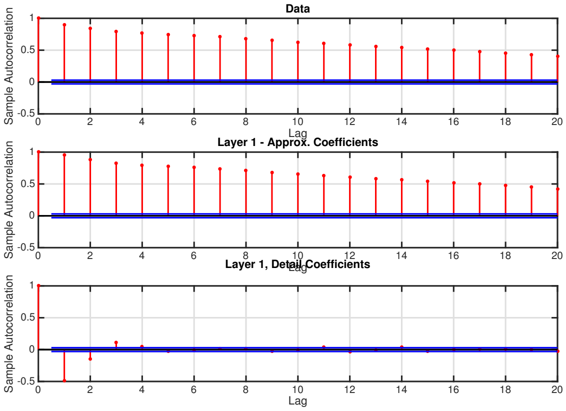

Fig. 3. Autocorrelation function of Traffic data, approximation and detail coefficient at layer 1 - BS=5735

This assumption is confirmed by the quantile–quantile (QQ) plot displayed in figure 5 that, for sake of simplicity, reports only the QQ-plot of DC_1. The figure shows the QQ curve obtained by comparing the Gaussian distribution with the obtained DC_1. We set mean and standard deviation of the Gaussian distribution according to the compared dataset. To immediately visualise the quality of the fitting results, in the figure we draw the “Best Fitting” curve, which corresponds to the bisector of the cartesian plane. The QQ curve is close to the “Best Fitting” curve, indicating that the DC_1 can be accurately modelled by the Gaussian distribution. The two curves deviate for very high and very low quantile values. This deviation can be related to “abnormal” values of the DC_1 due to traffic anomalies with respect to the “normal” pattern.

VI. Performance analysis using actual data

For the two presented approaches, we have defined and analysed different methods. In the case of the TA approach, we have mainly analysed two algorithms: one based on the setting of the thresholds using the UT procedure presented in subsection IV-A and another one based on the α–percentile of the Gaussian model used for the DC_j. In this case, the study has considered different settings of α. The preliminary analysis carried out using the UT procedure has highlighted a large number of alarms associated with traffic patterns with negligible changes with respect to the “normal” behaviour. This outcome is due to the low value of the thresholds given by the UT procedure. Given this conclusion, we have focused our study on the second algorithm. In this case, we set the thresholds in a similar way as described in the SA approach in subsection IV-B. In particular, for each layer j we calculate the thresholds $Th^{\hat{i}}_{Hj}$ and $Th^{\hat{i}}_{Lj}$ considering the estimated mean and standard variation of $DC_j(S_{\hat{i}})$. Obviously, $Th^{\hat{i}}_{Hj} = -Th^{\hat{i}}_{Lj}$ when the estimated mean of $DC_j(S_{\hat{i}})$ is zero. For sake of clarity, we point out that in this TA algorithm we examine the statistical features of $DC_j(S_{\hat{i}})$ for the selected BS $\hat{i}$ to calculate the thresholds. In the SA, the thresholds are calculated studying the $DC_j(S_i, \hat{f})$ of all signals S_i (i.e. of all BSs) at a fixed $\hat{f} = \hat{t}$. We denote this algorithm as Gaussian Thresholding (GT). The GT algorithm assumes that we have an alarm in the BS $\hat{i}$ at the TS $\hat{t}$ if there is at least a layer j where the $DC_j(S_{\hat{i}}, \hat{f} = \hat{t})$ is higher/lower than $Th^{\hat{i}}_{Hj}/Th^{\hat{i}}_{Lj}$.

A first simulation analysis indicates that the setting $\alpha = 99.99\%$ achieves a good tradeoff between the FA and the Missed Detection (MD). Given the lack of the grand truth for the whole observation period, we inspect the relation between traffic pattern and alarms to assess the FA and the MD. We assume to have an FA when the traffic variation near the alarm is negligible w.r.t. the “normal” traffic pattern. For the evaluation of the MD we investigate the traffic pattern in some specific BSs and time periods, where we have information on social events. For example, we select the BSs near the San Siro Stadium during football matches.

We add to the GT, the procedure of spatial filtering described in subsection IV-A, setting the size of the observation area equal to a square of side 2750 m. We set the threshold

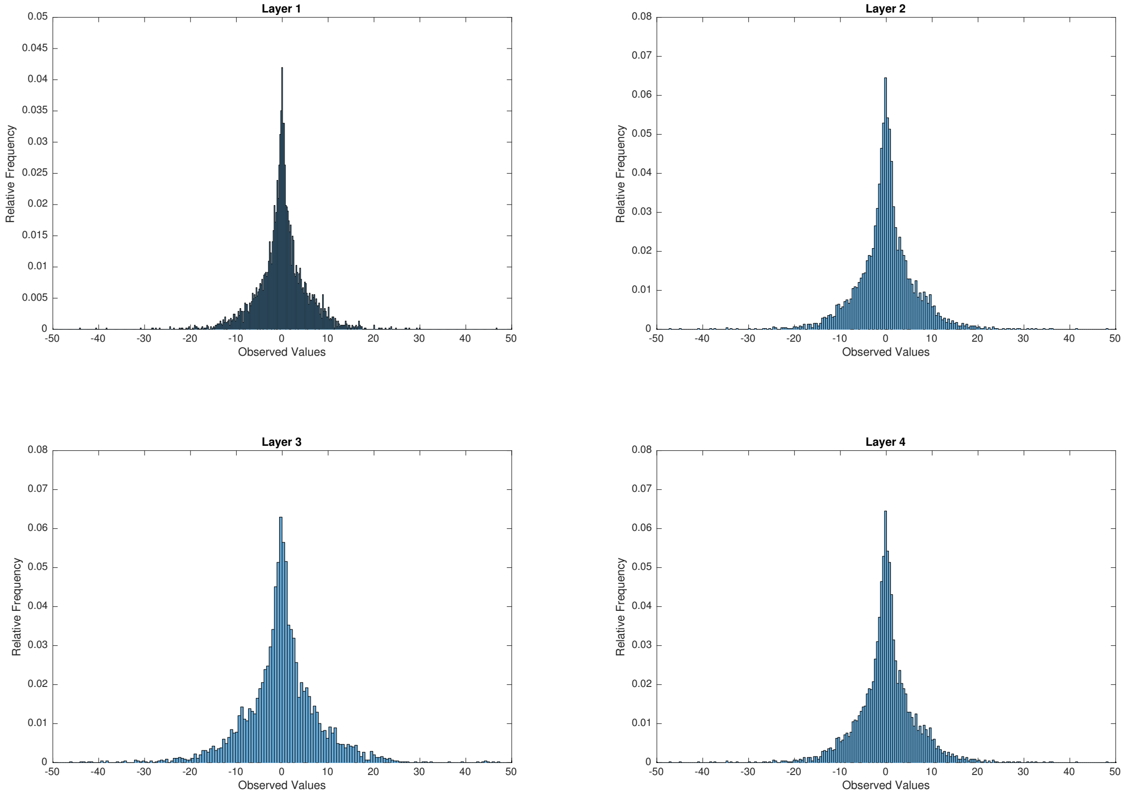

Fig. 4. Histogram of the detail coefficients at layer 1–4 - BS=5735

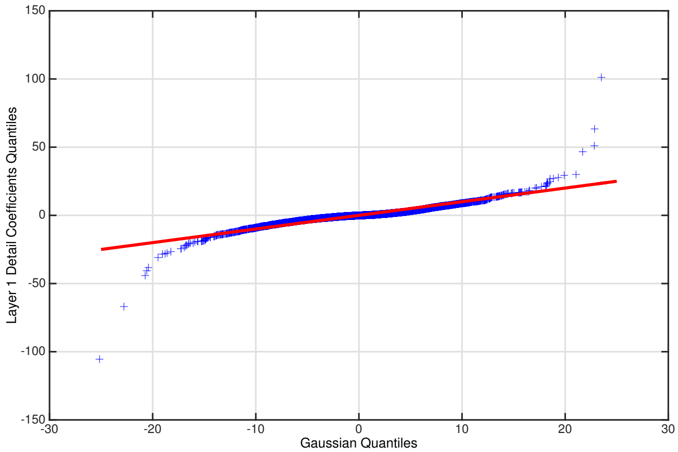

Fig. 5. QQ curve of DC_1 - BS=5735

P_{al} to 0.25. We denote this version of GT as GT with Spatial Filtering (GTSF).

As regarding SA, the basic algorithm detects an alarm for the BS $\hat{i}$ in the TS $\hat{t}$ if there is at least a layer j where $DC_j(S_{\hat{i}},\hat{t})$ is higher/lower than $Th_{H_j}(\hat{t})/Th_{L_j}(\hat{t})$. We denote the algorithm as Spatial Analysis with Gaussian model (SAG). During the preliminary tests, we experimentally observed that some detected alarms are characterised by the exceeding of the thresholds at only one layer (usually higher than layer 5). In many cases, this kind of events is often associated with the presence of a high traffic peak in a particular TS that generates large DCs at the highest considered layer. These large DCs are observable in a relatively long time interval around the TS where the peak is localised. They generate FAs in this interval. This observation suggests the definition of a new SAG algorithm version aimed at reducing the described phenomenon. With respect to the SAG, the new algorithm triggers an alarm for the BS $\hat{i}$ in the TS $\hat{t}$ when the estimated thresholds are exceeded in more than one layer. We denote this solution as Spatial Analysis with Gaussian model and with Confirmation (SAGC).

We choose to compare the performance of the proposed methods with the Kernel-based Online Anomaly Detection (KOAD) algorithm presented in [22]. The KOAD is an online, sequential, anomaly detection algorithm, which is suitable for use with multivariate data. The algorithm is based on the kernel version of the recursive least squares method. It assumes no model for network traffic or anomalies, and constructs and adapts a dictionary of features that approximately spans the subspace of normal behaviour. The results shown in [22] indicate detection performance similar to other compared approaches, with KOAD offering lower computational complexity and faster time-to-detection. To adapt the KOAD to our scenario, we assume that the multivariate data are represented by the timeseries related to the traffic observed in a set of BSs. We consider this set as composed by all BSs within the square of side 2750 m having in the centre the reference BS. This assumption is connected with the GTSF method. In this manner, we establish the multivariate data necessary to detect traffic anomaly over the time and simultaneously we have information on the location of the detected events. We use the code available in [39] for obtaining the KOAD performance. We set a training time equal to 400 TSs. During this period, we do not collect the alarms triggered by the KOAD algorithm.

For the performance analysis, we have selected some particular areas of Milan and some days of December 2013. We did the selection checking the availability of information on the social events. The selected areas and days are the following:

- The area of the San Siro Stadium, during the six main football matches of December 2013;
- December 16, the day of the strike of the Milan public transportation system;
- The area of the Politecnico of Milan, during all Sundays of December 2013;
- Selected BSs through Christmas Day.

Figures summarising the results report the network traffic activity, labelled with `Data`, during the selected days. The figures depict either the observation of the same day for different BSs or of different days for a selected BS.

In the first case, the abscissa labels report the BS Ids. The vertical double-dashed lines represent the border between the data belonging to BSs in adjacent positions in the plot. As an example, referring to Figure 6, the first vertical double-dashed line on the left divides the traffic data acquired at the end of the day (i.e. up to the TS 23.50 p.m.-00.00 a.m.) in the BS 5734 from those acquired at the beginning of the same day (i.e. starting from TS 00.00 a.m.-00.10 a.m.) in the BS 5736.

In the second case, the abscissa labels report the observed days. The vertical double-dashed lines represent the border between the traffic data of two different days. At the right of the vertical double-dashed line, we have the first TS of a day, while at the left side the last TS of the previously considered day, adjacent to the plot.

These figures also show the alarms triggered by each considered algorithm. In particular, the points having ordinate equal to 50 indicate the alarms generated by the GT, while those with ordinate equal to 70 are obtained by the GTSF. The alarms obtained with the SAG are depicted by the points with ordinate equal to 90. The points at ordinate 110 represent the alarms generated by the SAGC, those at 150 refer to the KOAD.

A. San Siro Stadium Area

In this area, in December 2013 we had six main events, as reported in Table II. The San Siro Stadium is used by the two big football teams of the city, i.e. Milan and Inter. The relevance of the event can be deduced by the number of spectators, which was high during the Champions League match Milan-Ajax and the derby Inter-Milan.

Day	Time	Event	Event Type	Spectators	MTV
1	15.00	Inter-Sampdoria	Serie A	43706	173
4	20.45	Inter-Trapani	Coppa Italia	12714	149
8	20.45	Inter-Parma	Serie A	33732	120
11	20.45	Milan-Ajax	Champ. League	61744	516
16	20.45	Milan-Roma	Serie A	37987	232
22	20.45	Inter-Milan	Serie A	79311	889

TABLE II
DATA ON EVENTS AT SAN SIRO STADIUM AND RELATED IMPORTANCE - MTV STANDS FOR MAXIMUM TRAFFIC VALUE DURING THE EVENT

Given that the derby is the most important event, we analyse the behaviour of a set of BSs around the San Siro area through December 22. We note a lot of alarms in the city. However, the temporal position of the triggered alarms changes in the different areas of the city. As an example, there are some areas of Milan where a lot of supporters meet to organise their football fans choreography to show at the stadium during the match. Focusing our attention on a time interval close to the start of the match, i.e. 20.45, we notice a lot of alarms near the stadium. For a subset of the BSs near the stadium, the traffic pattern and the detected alarms are shown in figure 6. The figure points out a traffic pattern that depends on the distance from the Stadium. The BS 5738 shows a high traffic peak during the hours of the match, while little traffic appears in the other hours of December 22. On the contrary, other BSs, such as 5734 or 5740, show the presence of a high traffic activity both before and after the hours of the match. The geographic position of these cells justifies this different behaviour. The coverage area of BS 5738 is mainly the stadium, while the others cover the neighbours of the stadium, i.e. the parking area and the ways to arrive/leave at/from the stadium by car, metro, and bus.

The figure 7 shows the curves related to BS 5738 for all days of December considered in the table II. To improve the clarity of the figure, we limit the represented ordinate to 200. The data show similar patterns for all considered days. The differences are the size of the peak, which is equal to the registered maximum traffic value reported in the table, and the temporal position of the peak on December 1. Indeed, the match of this day was the only one of December played at 15.00.

Comparing the behaviour of the different algorithms, we can observe that the GT is not able to detect the traffic changes associated with the less meaningful matches, i.e. those played December 4 and 8. This problem is due to the presence of the high peaks during the two main matches, played December 11 and 22. These peaks increase the estimated variance of the Gaussian model of DCs. At each layer, the high value of the estimated variance implies the increase of the related threshold. This high threshold is not exceeded by the DCs associated with the traffic variations observed during

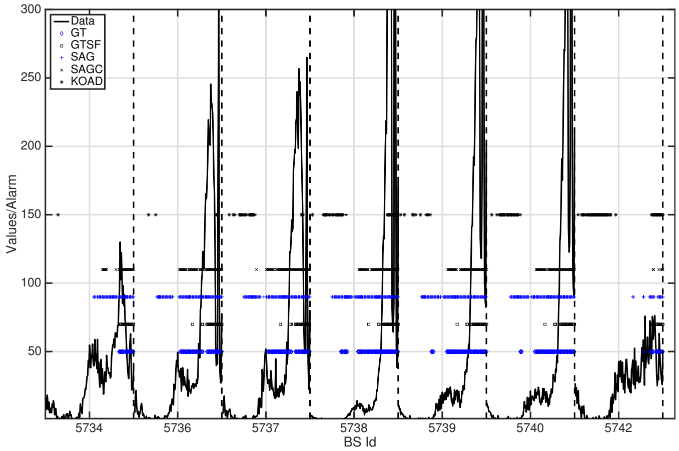

Fig. 6. Set of BSs around San Siro Stadium, through December 22. The day of the Derby

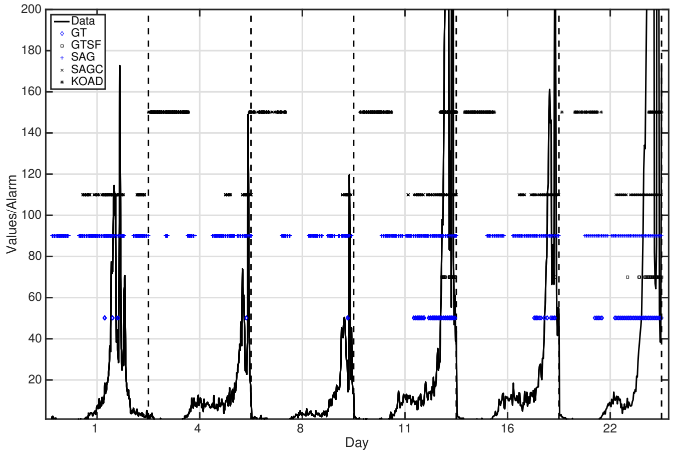

Fig. 7. Observation of all days with football matches at San Siro Stadium, BS 5738

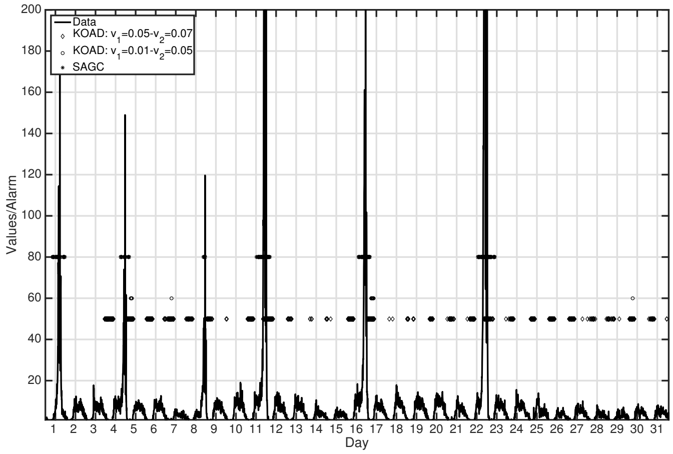

Fig. 8. Traffic data and alarms triggered by SAGC and KOAD with two different threshold settings – December 2013, BS 5738.

less meaningful matches. This phenomenon suggests that the presence of very high peaks over the time leads to an increase of the MD events. The introduction of the spatial filtering on the GT does not improve this aspect because we have developed the GTSF algorithm for reducing the FA.

On the contrary, the SA strategies are able to overcome this problem, as shown by the results displayed in the figure. Indeed, SA compares the traffic variation (by means of the DCs) among the different BSs of the network. This strategy triggers alarms when there are areas of the network (i.e. a subset of BSs) that vary their traffic more than other ones. The strategy overcomes the problem highlighted by TA. As shown in the figure, the alarms triggered by SAG and SAGC are observable in all considered days. The number of these alarms depends on the importance of the event. Indeed, the less important events, i.e. the matches of December 4 and 8, present a lower number of alarms. Furthermore, the results reported in the figure show that in most cases the SAGC alarms are in a medium–size time interval close to the start of the match. The length of the time interval is proportional to the number of spectators. A high number of spectators implies more street traffic in the area of the stadium. This scenario lasts for more time. Consequently, the cellular traffic shows an abnormal activity for a longer time interval. In the study with the KOAD, we set the low and high threshold to $\nu_1 = 0.05$ and $\nu_2 = 0.07$ respectively. The study reported in [22] shows that these threshold values provide the highest detection rate, with the lowest number of missed events. Furthermore, these values imply the highest number of FA (see Table I in [22]). In all cases of our scenario, the KOAD generates alarms only when the traffic is of few units. We have carried out further simulations changing the threshold settings. In particular, we have investigated the settings $\nu_1 = 0.01$ and $\nu_2 = 0.05$, which lead to having zero FA and one of the lowest values of the number of detected events. Figure 8 shows the results obtained during the whole observed period for the BS 5738. We selected this BS because it covers mainly the San Siro Stadium. “Abnormal” traffic peaks can be easily related to football matches.

The points at ordinate 60 are the alarms triggered by the KOAD with the thresholds settings $\nu_1 = 0.01$ and $\nu_2 = 0.05$, while those at 50 refer to the settings $\nu_1 = 0.05$ and $\nu_2 = 0.07$. The figure confirms the ability of the SAGC algorithm to detect the “abnormal” traffic peaks. As concerning the KOAD, we note that the first settings generate few alarms, which cannot be related to social events. On the contrary, the second settings trigger alarms every day during the early morning hours when the traffic is negligible. In summary, the KOAD shows poor performance in the particular scenario of our study. As reported in [22], the choice of traffic measurement and the feature space has a strong impact on the performance of KOAD algorithm and determines what type of anomalies can be detected.

B. *The strike of the Milan public transportation system*

One of the key features of the proposed solutions is the ability to detect the alarms in presence of trends and periodicity in the traffic, without a dedicated pre-processing of data. To highlight this property, we focus our attention on the December 16, the day of a meaningful strike of the Milan public transportation system. Actually, during our analysis, we have found a lot of alarms located in areas near to the most important access ways to the Milan Centre. Then, searching on the newspapers of December 16, we have found the news on the important strike. The strike constrained the workers to use their cars. Thus, all the main ingoing/outgoing roads of Milan Centre experimented an important traffic growth. In the BSs affected by this event, the traffic pattern is normally characterised by short peaks in the morning and in the evening, during the working days. During the no-working days, big changes with respect to the "normal" pattern can be observed during particular events in the Milan Centre. In the working days of December, we observed the significant changes during the day of the strike. To show this phenomenon, in the time period 19:30-21:00 of December 16, we analyse the alarms over the whole network. Figure 9 depicts the GTSF alarms, highlighting that groups of BSs have triggered alarms. The detected groups are the following.

- BS 1741. This area is near the "Bypass (Tangenziale) West of Milan", where we have the highway A50 and A7, with the A7 that arrives within Milan Centre.
- BS 3075. This area corresponds to the "Bypass (Tangenziale) East of Milan". In this case, we have the termination of the A1 highway that connects Milan with the South and South-East of the Milan neighbours.
- BS 3161, which covers the area of "Opera" and Via Virgilio Ferrari, another important road linking the "Bypass (Tangenziale) of Milan" and the city centre.
- BS 7625, which covers the route connecting city centre with the highway A4 and A8 (North-West area of Milan)
- BS 9485, which covers the A4 highway toll gates area, which is used by the workers moving from/to the West Milan neighbours.

The Milan GRID of the dataset divides the Milan Area in stripes of equal size of 100 BSs. Each BS ID reported in the above list represents the reference of a group. For each group, the set of involved BS IDs can be completed adding/subtracting multiple of 100 (movement towards North/South) and adding/subtracting few units (movement towards East/West) to the reference BS ID. As an example, we focus the attention on the set of BSs around 9485, reporting the alarms in figure 10. We can observe that the alarms are triggered in BS groups having IDs that differs of multiple of 100.

For a deep analysis, among these particular groups, we select the area of one of the A4 highway toll gates located at the North of Milan. In particular, we present the analysis of the BS with Id in the range $[9483, 9491]$. The considered BSs are located in parallel to the highway, as shown in figure 11, which reports the map and the GPS points of some considered BSs. The BS 9484 is located close to the toll gates area, while the furthest are 9483 (towards West) and 9491 (towards East).

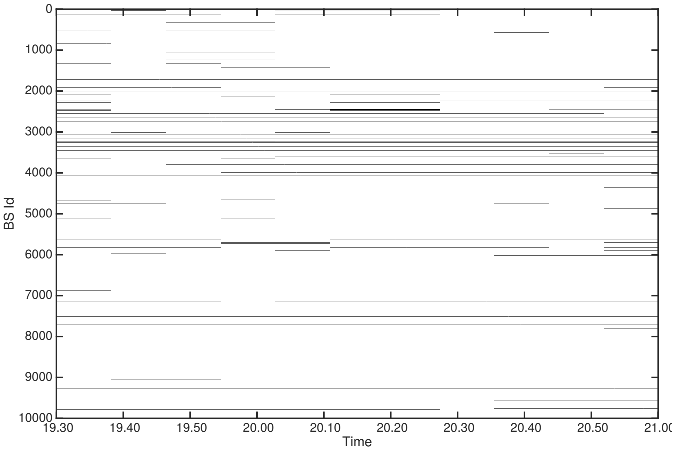

Fig. 9. Alarms vs. Time over the whole network. Observation period: December 16 hours 19.30-21.00 - SAG

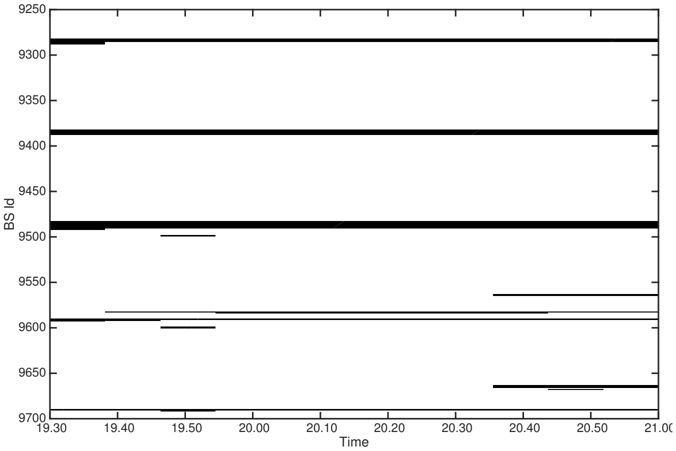

Fig. 10. Alarms vs. Time over the area with BS ID in the range (9250-9700). Observation period: December 16, hours 19.30-21.00. SAG

The distance between BS 9483 and 9491 is about 2 Km. About 7 km far from the considered A4 highway toll gates, towards East, we have the big industrial area of Agrate. The workers of this area usually employ the public transportation services, but the strike led to the utilisation of their cars. The considered toll gates area is about 12 Km far from the Milan city centre, towards North-East.

The figure 12 shows the alarms observed in the considered BSs at December 16, hours 05.00-23.00. Given that the BSs are parallel to the highway, the alarms provide interesting information for the rough estimation of the queuing length at the highway toll gates area. The figure highlights when at the toll gates the queue length grows in the direction of entering the Milan metropolitan area (i.e. the BSs with Id higher than 9485 have triggered alarms) or in the opposite direction (i.e. the BSs with alarms have Id lower than 9485).

The analysis of the figure shows the presence of alarms up to the BS 9493, which is roughly more than 1.8 Km away from the A4 toll gates area. These alarms are related to the queueing at the A4 gate ingoing to Milan and appear around 18:20.

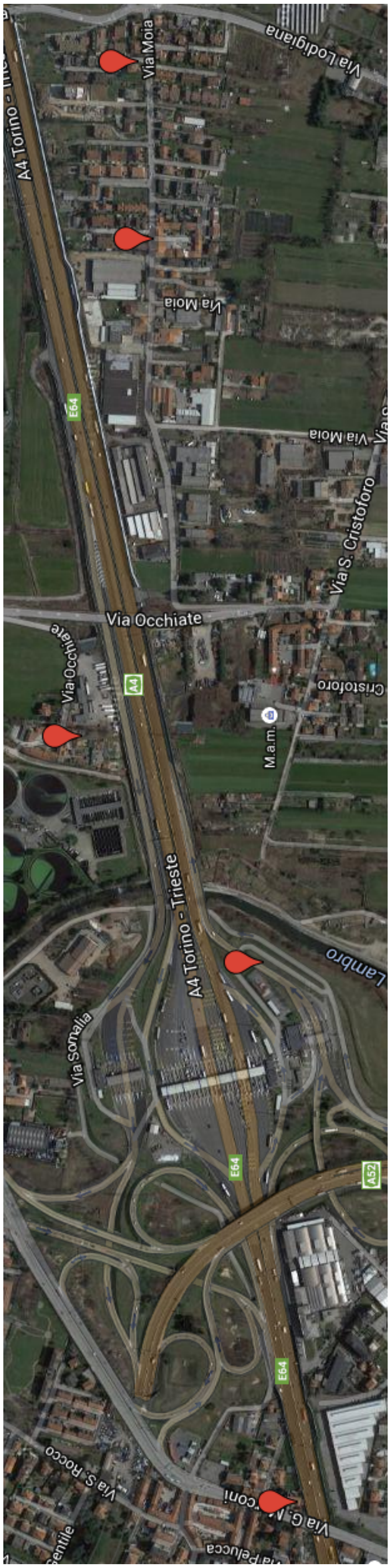

Fig. 11. Localization of the BS with Ids in the range 9483-9491. The gps points refers to BS 9483, 9484, 9487, 9490, and 9491, from the bottom to the top.

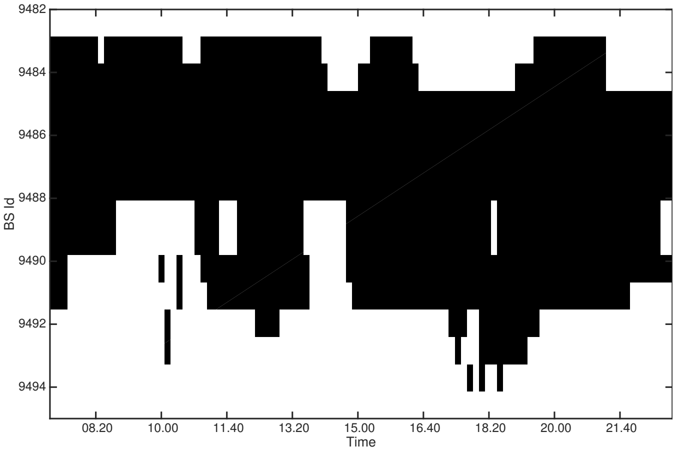

Fig. 12. Alarms vs. Time over the area with BS ID in the range (9482-9495). Observation period: December 16 hours 5.00-23.00. SAG

Other alarms for BS with Id higher than 9491 are observed at 7:10, 12.30, and 17.30. In the opposite direction, we observe alarms at BS 9483 (i.e. roughly 500 m from the gate) from the 7.10 to the 14.10. Other traffic peaks are at 15.00 and in the early evening at 20.00.

Figure 13 depicts the traffic patterns and the triggered alarms of the subset of these BSs. We can observe that all algorithms trigger alarms around the traffic peaks, except the GTSF and the KOAD. The KOAD confirms the problems pointed out in the previous results. The GTSF strategy generates only a low number of alarms concentrated during the peaks of the late afternoon. Furthermore, in the case of BS 9483, GTSF generates no alarm. The spatial filtering considers the number of BSs that simultaneously detect the alarm in a square area of side 2750 m. This strategy does not permit to confirm all the alarms detected by the GT in each BS of the area. We can observe this condition in the early morning of December 16 (and for the whole day in the BS 9483). On the contrary, the other mechanisms trigger alarms in these periods. This observation confirms that the spatial filtering applied to the GT leads to MD.

The SA algorithms detect a traffic variation also during the night hours. These hours are characterised by an average traffic activity negligible with respect to that of the working hours. In this condition, only the SAG and SAGC algorithms trigger alarms. These algorithms have the peculiarity of setting the thresholds for each TS. In this manner, they are able to dynamically follow the traffic variation over the whole network independently of the absolute values. This strategy allows to detect anomalies also when the traffic values are negligible w.r.t. the peak or the average.

Figure 14 shows the results related to all Mondays of December for the BS 9491 (the furthest from the toll gates towards East). This figure grows our confidence in the correlation between the triggered alarms and the "Special" event of the strike. Given its distance from the A4 toll gates, we can observe the cellular traffic growth only when we have a

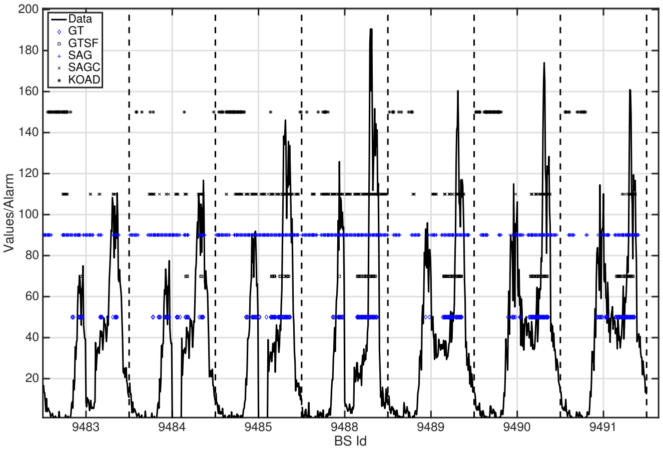

Fig. 13. Set of BSs around the highway A4 toll gates area at North of Milan. December 16, the day of the strike of the public transportation

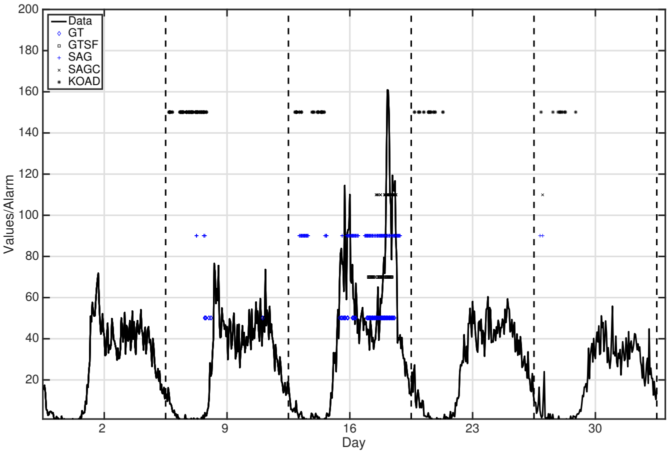

Fig. 14. Observation of all Mondays of December, BS 9491

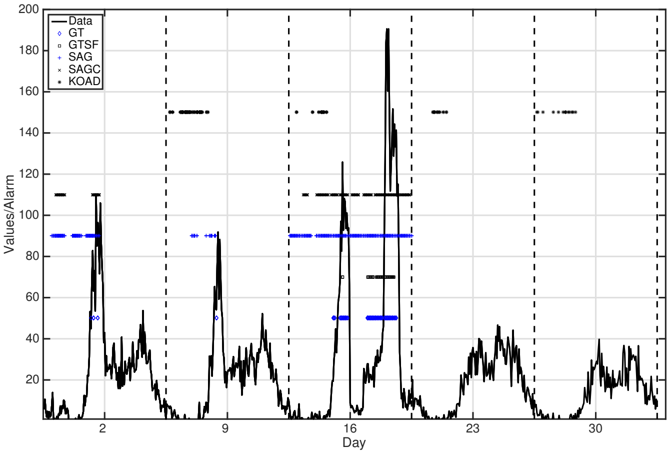

Fig. 15. Observation of all Mondays of December, BS 9488

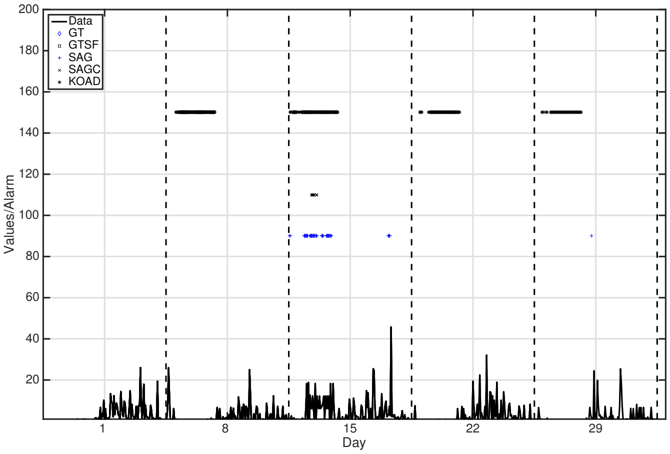

Fig. 16. Observation of all Sundays of December, BS 7151

long queue in the highway. The distance between the A4 toll gates and the BS 9491 is about 1.5 Km. The Figure 15 shows the results for the BS 9488, which is very close to the A4 toll gates. In this case, we notice a lot of alarms because a short queue length is more often observed. In this scenario, we have more often the traffic growth and consequently a higher number of alarms.

C. *Politecnico of Milan area, during the Sunday*

To further demonstrate the ability of the SA approach to detect traffic anomalies also in time intervals where the traffic intensity is negligible with respect to that observed in other time periods, we report Figure 16. This figure shows the observed traffic and the associated alarms during all Sundays of December in the BS 7151. This BS mainly covers the Politecnico di Milan area, where we have only activities related to the university life. In the considered days, the traffic values are in general lower than 20. These values are negligible with respect to those measured during the same days in other city areas. However, the SA approach is able to detect an anomalous traffic growth at the early morning of December 15. On the contrary, TA generates no alarm because the traffic changes shown during all Sundays are negligible with respect to those observed during the whole monitoring period. As already emphasised, for each layer this situation does not allow the DCs to exceed the calculated thresholds. In the figure, we observe that KOAD generates alarms when very low traffic values are observed or during the "abnormal" period detected by the SA methods.

D. *Selected BSs during the Christmas Day*

We select the Christmas Day to analyse the compared algorithms when we have a "normal" traffic growth over the whole network. This scenario can hide "abnormal" variations related to events different to those associated with the Christmas Day. We first study the number of the alarms triggered by the compared algorithms over the whole network in the different hours of Christmas Day. Figure 17 depicts the results, which highlight the very high number of alarms of the TA approach. Even GTSF introduced to reduce the FA generates over 7000 alarms (i.e. more than the 70% of network nodes are in alarm) in the morning and around the lunch time. In these two time intervals of Christmas Day, the persons usually do a large

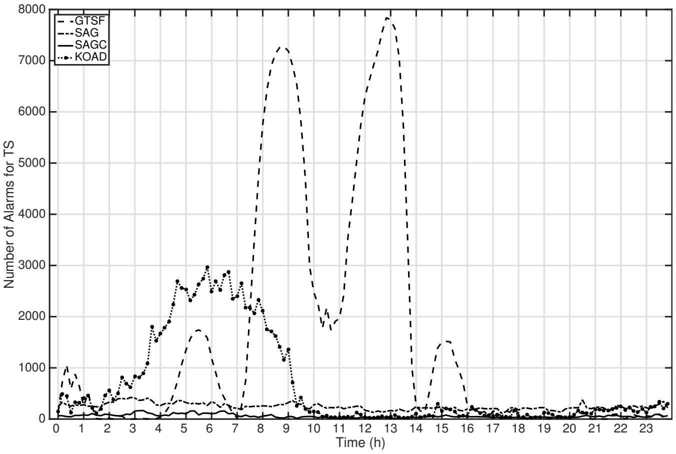

Fig. 17. The number of alarms triggered by the different approaches. Christmas Day

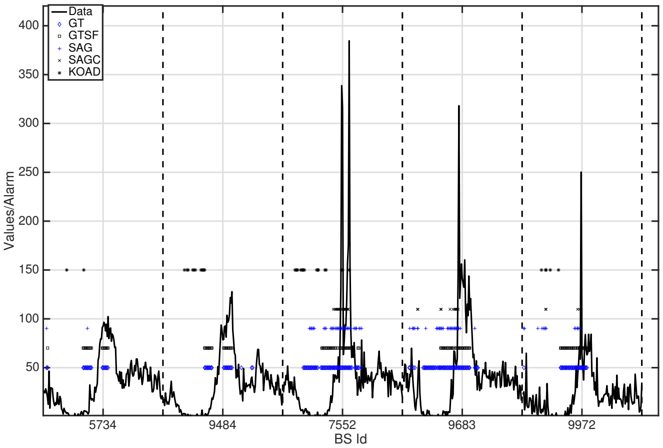

Fig. 18. Behaviour of a set of BSs through Christmas day

number of phone calls for exchanging the Christmas wishes with friends and relatives. Referring to the TA approach, this particular scenario implies the generation of a high number of alarms related to the "Christmas Day" event. These alarms prevent the detection of other events, which can happen in the city. On the contrary, the figure shows a low number of alarms detected by SAG and SAGC algorithms. This result points out that SA is able to account for traffic increase localised in a particular area also in the considered critical scenario. In particular, the algorithms detect the "abnormality" of traffic increase in some areas with respect to the average growth observed over the whole network. To quantify the differences, at noon the GTSF triggers 5750 alarms, while the SAG and SAGC detect 246 and 33 alarms respectively. Considering the SAGC, we observe the minimum number of BSs (i.e. 9) having triggered alarms at 11:10. These BSs are localised in three areas of Milan: the Parco Villa Litti at Affori (a public park), and two areas at North of Milan that cover the highway "Tangenziale Nord of Milan". The first one can be associated with an entertainment organised within the park, usually used for concert and social events. In the other two cases, considering the characteristic of the areas, we can assume that the alarms can be associated with events on the highway. Furthermore, the figure confirms that KOAD triggers alarms when we have a low traffic. Indeed, the number of alarms is higher in the period 5:00-7:00.

Figure 18 shows the traffic and alarms observed in 5 BSs through Christmas Day. We have selected the two BSs with IDs 5734 and 9484 because the SA approaches trigger few alarms (no one in the case of BS 9484). The other BSs have been chosen from the three areas of Milan, where the SAGC generates alarms at 11.10. The part of the figure reporting the data of BS 5734 (the BS that covers mainly the San Siro Stadium) should be compared with that of the figure 6, which reports the traffic of the same BS during December 22, i.e. the day of the football derby. In figure 6 traffic values near to 120 lead to the alarms triggered by the four compared algorithms. On the contrary, in figure 18 similar values, i.e. about 100, do not present SAG and SAGC alarms. In both days, the BS exhibits traffic variation, but while on December 22 the growth is limited to some areas of Milan, during Christmas Day this is within the average increment exhibited by the whole network. We can derive a similar conclusion comparing the part of figure 18 associated with the BS 9484 with that related to the same BS in the figure 13.

The traffic data of the other BSs in figure 18 (i.e. the BSs with SA alarms at 11.10) are characterised by a fast growth near the noon. We observe that in the BS 7552 (covering the area of Parco Villa Litti at Affori) we have two different peaks, one before the noon and another one in the first hours of the afternoon. This behaviour suggests that the event of the park attracted people before and after the lunch. On the contrary, in the other two areas (i.e. BS 9683 and 9972), the data show only a peak near noon and then a traffic decrease towards the "normal" values. This observation suggests that we had a problem near noon that produced a traffic growth of short duration. While the problem was solving, the traffic intensity was decreasing to the "normal" values.

VII. Performance evaluation using artificial anomalies

The availability of ground truth is essential for the performance evaluation of anomaly detection mechanisms. When the ground truth lacks, as in our case, the most common solution is to label the data by means of the manual detection of anomalies. However, we have 10000 timeseries (one for each BS) each one of length higher than 4000 samples. Thus, the manual detection of anomalies is a time expensive operation. To deal with this problem, we have focused the first part of our analysis selecting some particular areas of Milan or some days of the December 2013 where we have information on social events likely leading to an "abnormal" cellular traffic growth. This first analysis has pointed out the problems of the KOAD and GT algorithms in the detection function. Hence, in this analysis we consider only the other considered methods.

To evaluate the performance of the compared strategies in terms of MD, we add artificial anomalies to the original dataset. In particular, we carry out 100 different simulation runs. In each one, we add a traffic anomaly event randomly in the space and in the time. In details, in each run we select randomly the reference BS and the reference TS, using a uniform distribution in both cases. We generate the anomaly in the area around the reference BS, in a time interval centred at the reference TS. The added anomaly is obtained multiplying for a selected constant the original traffic value.

The study considers different settings of the following parameters.

- P, which is related to the size of the area; the area is a square of side equal to $(2P+1)*250$ m having the reference BS in the centre.
- T, which defines the duration of the anomaly; the anomaly lasts for a time interval of size $(2T+1)*10$ min centred at the reference TS.
- C, which is the multiplier constant.

Table III summarises the results. In the "Configuration" column, the value assigned to each parameter is represented by the number reported after the related letter, i.e. P5 means $P=5$. The values reported in the other columns represent the results in terms of MD for each method.

The first two lines show the effects of the area size of the anomaly on the performance. Indeed, we maintain the same T and C, decreasing P from 5 to 2. The first value of P corresponds to the setting used for the spatial filtering of the GTSF method. The table shows that SAG and SAGC do not vary their performance, whereas we can observe a sharp increase of the MD given by the GTSF. This result points out the key role of the setting of the spatial filtering area on the performance of the GTSF. On the contrary, both SAG and SAGC do not have this problem.

The comparison of the lines 1 and 3 highlights that the reduction of the time interval interested by the traffic anomaly leads to the worsening of the performance. In this scenario, the performance differences among GTSF and SAGC are less evident. Furthermore, the SAG maintains a low number of MD. The worsening of the SAGC performance is because, halving the anomaly period, there is a lower probability to overcome the threshold in more than one wavelet layer. The doubling of the amplitude of the anomaly (i.e. the value of C) compensates the negative phenomena related to a shorter anomaly period. This conclusion is supported by the comparison of the results shown in lines 3 and 4. Reducing the amplitude of the anomaly (i.e. C sets to 2) the performance of all algorithms get worse, as shown by the results in the line 5. In this case, the lengthening of the anomaly period improves the performance, which however remains bad. The best result is an MD of 26%, achieved by the SAG. Increasing the number of C to 10, in all the 100 independent runs the SAG detects the added anomaly. On the contrary, the SAGC misses the anomaly in one run and the GTSF in 4.

In summary, the SAG provides MD values lower than the 10% in all considered scenarios except the case with $C=2$. Indeed, when the anomaly manifests with the doubling of the "normal" traffic, the MD of the SAG is higher than 25%. In all scenarios, the MD of SAGC is about the double of the one given by the SAG. The GTSF shows the worst performance in all scenarios. Furthermore, the comparison of the results shown in line 1 and 2 of the Table indicates a critical point in the choice of the observation area size. Indeed, if this area is smaller than the one where we have the anomaly, the performance is poor.

Configuration	GTSF	SAG	SAGC
P5T6C5	17	3	9
P2T6C5	65	3	9
P5T3C5	21	10	18
P5T3C10	11	6	10
P5T3C2	42	41	75
P5T6C2	38	26	57
P5T6C10	4	0	1

TABLE III
NUMBER OF OBSERVED MD IN THE SIMULATION OF 100 RANDOM ARTIFICIAL TRAFFIC ANOMALIES

VIII. CONCLUSION

The study carried out with actual traffic shows the good performance achieved by the proposed approaches in detecting traffic anomalies generated by social events. The simulation results confirm that the SWT permits to highlight the traffic variation by easily filtering the periodic trends present in the cellular traffic. Starting from these results, we have presented two alternative strategies having the peculiar property of exploiting the spatial information available in the cellular traffic data. The alternative solutions differ on how they take into account this spatial information. The results show that significant traffic changes during the monitoring period in the considered BS considerably degrade the detection performance of the TA approach. The key weak point of TA is the comparison of the traffic variations over the time before using the spatial information.

To overcome this weak point, we have analysed the alternative strategy based on the comparison of the traffic variation observed in all network BSs at a particular TS. Its performance points out two key advantages. First, this strategy allows filtering the average traffic growth over the whole network in a particular TS. This characteristic reduces FA and MD when we have a special event implying a traffic growth over the whole network. The simulation results of the "Christmas Day" show this peculiarity. Second, it is able to detect the anomaly also in time intervals with traffic values negligible w.r.t. the average, as highlighted for example by the results related to the Politecnico of Milan area. Taking into account the simulation results, we can conclude that the strategy used by SA is the most suitable for exploiting the traffic spatial correlation when we aim at locating social events through the detection of cellular traffic anomalies.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

[1] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, “Data for development: the d4d challenge on mobile phone data,” *arXiv preprint arXiv:1210.0137*, 2012.

[2] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, “D4d-senegal: the second mobile phone data for development challenge,” *arXiv preprint arXiv:1407.4885*, 2014.

[3] Telecom Italia Mobile, “Open big data,” 2015. Available at https://dandelion.eu/datamine/open-big-data/.

[4] K. Wakita and R. Kawasaki, “Estimating human dynamics in cote d’ivoire through d4d call detail records,” *NetMob D4D Challenge*, pp. 1–3, 2013.

[5] C. Smith-Clarke and L. Capra, “Beyond the baseline: Establishing the value in mobile phone based poverty estimates,” in *Proceedings of the 25th International Conference on World Wide Web*, WWW ’16, pp. 425–434, 2016.

[6] V. D. Blondel, A. Decuyper, and G. Krings, “A survey of results on mobile phone datasets analysis,” *EPJ Data Science*, vol. 4, no. 1, p. 1, 2015.

[7] C. Chaparro and W. Eberle, “Detecting anomalies in mobile telecommunication networks using a graph based approach.,” in *FLAIRS Conference*, pp. 410–415, 2015.

[8] P. Casas, P. Fiadino, and A. D’Alconzo, “Machine-learning based approaches for anomaly detection and classification in cellular networks,” in *Traffic Monitoring and Analysis workshop (TMA)*, 2016.

[9] A. D’Alconzo, A. Coluccia, and P. Romirer-Maierhofer, “Distribution-based anomaly detection in 3g mobile networks: from theory to practice,” *International Journal of Network Management*, vol. 20, no. 5, pp. 245–269, 2010.

[10] P. Fiadino, A. D’Alconzo, M. Schiavone, and P. Casas, “Rcatool-a framework for detecting and diagnosing anomalies in cellular networks,” in *Teletraffic Congress (ITC 27), 2015 27th International*, pp. 194–202, IEEE, 2015.

[11] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[12] M. Ahmed, A. N. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016.

[13] C. Callegari, A. Coluccia, A. D’Alconzo, W. Ellens, S. Giordano, M. Mandjes, M. Pagano, T. Pepe, F. Ricciato, and P. Zuraniewski, “A methodological overview on anomaly detection,” in *Data traffic monitoring and analysis*, pp. 148–183, Springer, 2013.

[14] T. Ahmed, B. Oreshkin, and M. Coates, “Machine learning approaches to network anomaly detection,” in *Proceedings of the 2nd USENIX workshop on Tackling computer systems problems with machine learning techniques*, pp. 1–6, USENIX Association, 2007.

[15] P. Casas, J. Mazel, and P. Owezarski, “Unada: Unsupervised network anomaly detection using sub-space outliers ranking,” *NETWORKING 2011*, pp. 40–51, 2011.

[16] A. Soule, K. Salamatian, and N. Taft, “Combining filtering and statistical methods for anomaly detection,” in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, pp. 31–31, 2005.

[17] P. H. dos Santos Teixeira and R. L. Milidiú, “Data stream anomaly detection through principal subspace tracking,” in *Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 1609–1616, ACM, 2010.

[18] A. Lakhina, M. Crovella, and C. Diot, “Mining anomalies using traffic feature distributions,” in *ACM SIGCOMM Computer Communication Review*, vol. 35, pp. 217–228, ACM, 2005.

[19] A. Coluccia, A. D’Alconzo, and F. Ricciato, “Distribution-based anomaly detection via generalized likelihood ratio test: A general maximum entropy approach,” *Computer Networks*, vol. 57, no. 17, pp. 3446–3462, 2013.

[20] I. A. Karatepe and E. Zeydan, “Anomaly detection in cellular network data using big data analytics,” in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, pp. 1–5, VDE, 2014.

[21] C. Callegari, S. Giordano, and M. Pagano, “An information-theoretic method for the detection of anomalies in network traffic,” *Computers and Security*, vol. 70, pp. 351–365, 2017.

[22] T. Ahmed, M. Coates, and A. Lakhina, “Multivariate online anomaly detection using kernel recursive least squares,” in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pp. 625–633, IEEE, 2007.

[23] R. H. Riedi, V. J. Ribeiro, M. S. Crouse, and R. G. Baraniuk, “Network traffic modeling using a multifractal wavelet model,” in *European Congress of Mathematics*, pp. 609–618, Springer, 2001.

[24] D. Kwon, K. Ko, M. Vannucci, A. N. Reddy, and S. Kim, “Wavelet methods for the detection of anomalies and their application to network traffic analysis,” *Quality and Reliability Engineering International*, vol. 22, no. 8, pp. 953–969, 2006.

[25] S. Kim and A. Reddy, “Statistical techniques for detecting traffic anomalies through packet header data,” *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 562–575, 2008.

[26] P. Barford, J. Kline, D. Plonka, and A. Ron, “A signal analysis of network traffic anomalies,” in *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pp. 71–82, ACM, 2002.

[27] C. Callegari, S. Giordano, M. Pagano, and T. Pepe, “Wave-cusum: Improving cusum performance in network anomaly detection by means of wavelet analysis,” *Computers & Security*, vol. 31, no. 5, pp. 727–735, 2012.

[28] S. Novakov, C.-H. Lung, I. Lambadaris, and N. Seddigh, “Studies in applying pca and wavelet algorithms for network traffic anomaly detection,” in *High Performance Switching and Routing (HPSR), 2013 IEEE 14th International Conference on*, pp. 185–190, IEEE, 2013.

[29] X. Song, Y. Ouyang, B. Du, J. Wang, and Z. Xiong, “Recovering individual’s commute routes based on mobile phone data,” *Mobile Information Systems*, vol. 2017, 2017.

[30] G. Re Calegari and I. Celino, “Smart urban planning support through web data science on open and enterprise data,” in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1407–1412, ACM, 2015.

[31] C. Callegari, R. Garroppo, and S. Giordano, “Inferring social information on foreign people from mobile traffic data,” in *IEEE International Conference on Communications*, 2017.

[32] A. Janecek, D. Valerio, S. Ruehrup, K. Hummel, H. Hlavacs, F. Ricciato, B. Rainer, and W. Mullner, “Incident detection from cellular network signalling,” in *19th ITS World Congress*, 2012.

[33] D. Valerio, T. Witek, F. Ricciato, R. Pilz, and W. Wiedermann, “Road traffic estimation from cellular network monitoring: a hands-on investigation,” in *Personal, Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on*, pp. 3035–3039, IEEE, 2009.

[34] D. Valerio, A. D’Alconzo, F. Ricciato, and W. Wiedermann, “Exploiting cellular networks for road traffic estimation: a survey and a research roadmap,” in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, pp. 1–5, IEEE, 2009.

[35] I. Daubechies *et al.*, *Ten lectures on wavelets*, vol. 61. SIAM, 1992.

[36] J. Morlet, P. Tchamitchian, and J. Holschneider, “A real-time algorithm for signal analysis with help of wavelet transform,” *Wavelets, Time-Frequency Methods and Phase Space, Springer, Berlin*, 1989.

[37] J. E. Fowler, “The redundant discrete wavelet transform and additive noise,” *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 629–632, 2005.

[38] I. M. Johnstone, “Gaussian estimation: Sequence and wavelet models,” *Book Draft version, September*, 2015.

[39] T. Amhed and M. Coates, “Koad code,” 2006. Available at http://www.tsp.ece.mcgill.ca/Networks/projects/projdesc-anom-tarem.html.