Data-Driven Analytics for Automated Cell Outage Detection in Self-Organizing Networks

Ahmed Zoha*, Arsalan Saeed †, Ali Imran‡, Muhammad Ali Imran† and Adnan Abu-Dayya*
*QMIC, Qatar Science and Technology Park, Doha, Qatar 210531
†CCSR, University of Surrey, Guildford, UK, GU2 7XH
‡University of Oklahoma,Tulsa, USA 71435
ahmedz@qmic.com, arsalan.saeed@surrey.ac.uk, ali.imran@ou.edu

***Abstract*—In this paper, we address the challenge of autonomous cell outage detection (COD) in Self-Organizing Networks (SON). COD is a pre-requisite to trigger fully automated self-healing recovery actions following cell outages or network failures. A special case of cell outage, referred to as Sleeping Cell (SC) remains particularly challenging to detect in state-of-the-art SON, since it triggers no alarms for Operation and Maintenance (O&M) entity. Consequently, no SON compensation function can be launched unless site visits or drive tests are performed, or complaints are received by affected customers. To address this issue, we present and evaluates a COD framework, which is based on minimization of drive test (MDT) reports, a functionality recently specified in third generation partnership project (3GPP) Release 10, for LTE Networks. Our proposed framework aims to detect cell outages in an autonomous fashion by first pre-processing the MDT measurements using multidimensional scaling method and further employing it together with machine learning algorithms to detect and localize anomalous network behaviour. We validate and demonstrate the effectiveness of our proposed solution using the data obtained from simulating the network under various operational settings.**

***Keywords*—Anomaly Detection, Cell Outages, Low-Dimensional Embedding, LTE, Self-Organizing Networks, Sleeping Cell, MDT**

I. Introduction

The increased demands of high throughput, coverage and end user quality of service (QoS) requirements, driven by ever increasing mobile usage, incur additional challenges for the network operators. Fueled by the mounting pressure to reduce capital and operational expenditures (CAPEX & OPEX) and improve efficiency in legacy networks, the Self-Organizing Network (SON) paradigm aims to replace the classic manual configuration, post deployment optimization, and maintenance in cellular networks with self-configuration, self-optimization, and self-healing functionalities. A detailed review of the state-of-the-art SON functions for legacy cellular networks can be found in [1]. The main task within self-healing functional domain is autonomous cell outage detection and compensation. Current SON solutions generally assume that the spatio-temporal knowledge of a problem that requires SON-based compensation is fully or at least partially available; for example, location of coverage holes, handover ping-pong zones, or congestion spots are assumed to be known by the SON engine. Traditionally, to assess and monitor mobile network performance manual drive test have to be conducted. However, this approach cannot deliver the stringent resource efficiency and low latency, and cannot be used to construct dynamic models to predict system behavior in live-operation fashion.

978-1-4799-7795-6/15/$31.00 ©2015 IEEE

This is particularly true for a *Sleeping Cell* (SC) scenario, which is a special case of cell outage that can remain undetected and uncompensated for hours or even days, since no alarm is triggered for Operation and Maintenance (O&M) system [2]. A SC either cause deterioration of the service level or a total loss of radio service in its coverage area, due to a possible software (SW), firmware or hardware (HW) problem. SC can only be detected by means of manual drive tests or via subscriber complaints. These solutions are not only time and resource consuming but also require expert knowledge to troubleshoot the problem. As future cellular network have to rely more and more on higher cell densities, manual or semi-manual detection of SC can become a huge challenge. Therefore, automatic cell detection has become a necessity so that timely compensation actions can be triggered to resolve any issues.

The reported studies in literature that addressed the problem of cell outage detection are either based on quantitative models [3], which requires domain expert knowledge, or simply rely on performance deviation metrics for detection [4]. Until recently, researchers have applied methods from the machine learning domain such as clustering algorithms [5] as well as Bayesian Networks [6] to automate the detection of faulty cell behavior. Coluccia *et al.* [7] analyzed the variations in the traffic profiles for 3G cellular systems to detect real-world traffic anomalies. In particular, the problem of sleeping cell detection has been addressed by constructing and comparing a visibility graph of the network using *Neighbor Cell List (NCL)* reports [8].

Compared to aforementioned approaches, the solution proposed in this paper differs in various aspects. Our proposed Cell Outage Detection (COD) framework adopts a model-driven approach that makes use of mobile terminal assisted data gathering solution based on minimize drive testing (MDT) functionality [2] as specified by 3GPP. MDT functionality allows eNBs to request and configure UEs to report back the key performance indicators (KPIs) from the serving and neighboring cells along with their location information. To accurately capture the network dynamics, we first collect UE reported MDT measurements and further extract a minimalistic KPI representation by projecting them to a low-dimensional embedding space. We then employ these embedded measurements together with machine learning algorithm to autonomously learn the "normal" operational profile of the net-

work. The learned profile leverages the intrinsic characteristics of embedded network measurements to intelligently diagnose a sleeping/outage cell situation. This is in contrast to state-of-the-art techniques that analyze one or two KPIs to learn the decision threshold levels and subsequently apply them for detecting network anomalies. In addition, the COD framework further exploits the geo-location associated with each measurement to localize the position of the faulty cell, enabling the SON to autonomously trigger cell outage compensation actions.

Furthermore, in this study, we compare and evaluate the performance of density and domain based anomaly detection approaches: *Local Outlier Factor based Detector* (LOFD) and *One Class Support Vector Machine based Detector* (OCSVMD), respectively, while taking into consideration the acute dynamics of the wireless environment due to channel conditions as well as load fluctuations. To the best of our knowledge, no prior study examines the use of OCSVMD and LOFD in conjunction with embedded MDT measurements for autonomous cell outage detection. Secondly, the framework provided paves a way towards developing a fully automated cell outage management solution via integrating self-healing functionality in the proposed architecture for the emerging (LTE) as well as future (5G) self organizing networks. In view of this, the proposed solution is validated with simulations that are setup in accordance with 3GPP LTE standard. The remainder of this paper is structured as follows: Section II presents the system architecture for proposed COD framework, which also includes a brief description of LOFD and OCSVMD techniques that are used to profile, detect and localize anomalous network behaviour. In Section III, we provide details of our simulation setup and evaluation methodology, as well as present extensive simulation results to substantiate the performance of our proposed COD framework. Finally, Section IV concludes this paper.

II. System Architecture

The COD framework aims to detect the network performance deterioration, whenever a problem occurs within a network. Firstly, to profile the normal operational behaviour of the network, our solution collects KPIs from the network leveraging MDT functionality. The goal is to use the learned profile to perform problem identification and localization autonomously, during the monitoring period.

The MDT reporting schemes have been defined in LTE Release 10 during 2011 [2]. The release proposes to construct a data base of MDT reports from the network using *Immediate* or *Logged* MDT reporting configuration. In this study, the UE's are configured to report the cell identification and radio-measurement data to the target eNB based on immediate MDT configuration procedure as shown in Figure 1. The signaling flow of MDT reporting procedure consists of configuration, measurement, reporting and storing phase. The UE is first configured to perform measurements periodically as well as whenever an A2 event (i.e., serving cell becomes worst than a *threshold*) occurs. Subsequently, it performs KPI measurements: serving and neighbors Reference Signal Received Power (RSRP), serving and neighbors Reference Signal Received Quality (RSRQ), as specified in Table I, and further reports it to the serving eNB. The eNB after retrieving these

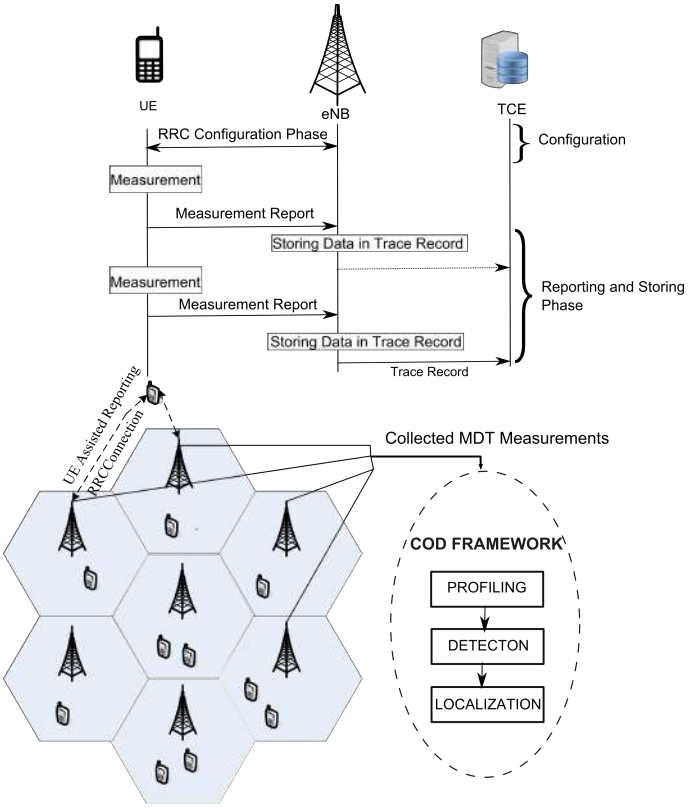


Fig. 1: System Model for Cell Outage Detection

Measurements	Description
Location	longitude and latitude information
Serving Cell info	Cell Global Identity (CGI)
RSRP	Reference Signal Received Power in dBm
RSRQ	Reference Signal Received Quality in dB
Neighboring Cell Information	Three Strongest intra-LTE RSRP, RSRQ information

TABLE I: MDT Reported Measurements

measurements further appends time and wide-band channel quality information (CQI) and forwards it to Trace Collection Entity (TCE). TCE collects and stores the trace reports which are subsequently processed to construct a MDT database. In this study, the trace records obtained from the reference scenario (i.e., fault-free) act as a benchmark data and is used by the anomaly detection models to learn the network profile. These models are then employed to autonomously detect and localize outage situations. The proposed framework as shown in Figure 1 consists of profiling, detection and localization phases, as detailed in the following subsections.

A. Profiling Phase

The next step after collecting measurements from the network is to perform data transformation. Each trace record is processed to extract a KPI vector V that contains the RSRP and RSRQ KPIs of the serving as well as of the three strongest neighbouring cells along with the CQI augmented to form a measurement vector as shown in Equation 1

$$V = \{RSRP_S, RSRP_{n_1}, RSRP_{n_2}, RSRP_{n_3}, RSRQ_S, RSRQ_{n_1}, RSRQ_{n_2}, RSRQ_{n_3}, CQI\} \quad (1)$$

where *S* and *n* stands for serving and neighboring cells, respectively. The 9-dimensional feature vector V corresponds to one measurement sample which is further embedded to only three dimensions in the Euclidean space using Multi-Dimensional Scaling (MDS) method [9]. MDS provides a low-dimensional embedding of the target KPI vector V while preserving the pairwise distances amongst them. Given, a $t \times t$ dissimilarity matrix Δ^X of the MDT dataset, MDS attempts to find t data points $\psi_1 ... \psi_t$ in m dimensions, such that Δ^Ψ is similar to Δ^X. Classical MDS (CMDS) operates in Euclidean space and minimizes the following objective function

$$\min_{\psi} \sum_{i=1}^{t} \sum_{j=1}^{t} ({\delta_{ij}}^{(X)} - {\delta_{ij}}^{(\Psi)})^2 \tag{2}$$

where ${\delta_{ij}}^{(X)} = \parallel x_i - x_j \parallel^2$ and ${\delta_{ij}}^{(\Psi)} = \parallel \psi_i - \psi_j \parallel^2$. Equation 2 can be reduced to a simplified form by representing Δ^X in terms of a kernel matrix using Equation 3

$$X^T X = -\frac{1}{2} H \Delta^X H \tag{3}$$

where $H = I - \frac{1}{t} e e^T$ and e is a column vector of all 1's. This allows us to rewrite Equation 2 as

$$\min_{\Psi} \sum_{i=1}^{t} \sum_{j=1}^{t} ({x_i}^T x_j - {\psi_i}^T \psi_j)^2 \tag{4}$$

As shown in [9], that the Ψ can obtained by solving $\Psi = \sqrt{\Lambda} V^T$, where V and Λ are the matrices of top m eigenvectors and their corresponding eigenvalues of $X^T X$ respectively. The m dimensional embedding of the data points are the rows of $\sqrt{\Lambda} V^T$, whereas the value of m is chosen to be 3 in our case. The embedded KPI representation V^e has several advantages. First, it makes the framework generic allowing it to incorporate new KPI's and network-centric features such as call drop ratios, data traffic etc without imposing higher computational requirements. Moreover, the interrelationships of high-dimensional databases can be explored in a lower-dimension space. Secondly, given the growing complexity of the networks, particularly in case of SON, it is challenging to identify few KPIs that accurately capture the behavior of the system. The network-level intelligence can be inferred through low-dimensional representation of large volume of network measurements. The embedded space reveals a hidden structure by mapping similar measurements close to each other and vice versa, that naturally isolates high and low data density regions. This makes it easier to detect the underlying patterns that are representative of network dynamics. The learning algorithm leverages embedded network measurements to learn a optimal decision rule and subsequently during the monitoring phase apply it to classify observed network measurements as anomalous or normal, as discussed in detail below.

To construct a reference database D_M, we apply an MDS based data transformation as explained, on the network measurements collected from a fault-free operating scenario. The D_M also includes samples of Radio Link Failure (RLF) events, in addition to periodical MDT measurements, as expected in a realistic environment. As shown in Figure 2(a), the D_M acts as a training database for the anomaly detection algorithm, enabling it to learn the "normal" network behaviour. This involves learning a decision function 'f' and a corresponding threshold 'θ', which is used to differentiate between normal and abnormal network measurements. Thus, it can be treated as a binary classification problem which can formally be expressed as follows:

$$f(x_i) = \begin{cases} Normal, & \text{if } f(x_i, D_M) \leq \theta \\ Anomalous, & \text{if } f(x_i, D_M) > \theta \end{cases} \tag{5}$$

where x_i is the test measurement. Two state-of-the-art algorithms from the machine learning domain: OCSVMD and LOFD are examined for modeling the dynamics of network operational behaviour. The brief working description of the two detection algorithms are summarized as follows:

*1) Local Outlier Factor based Detector (***LOFD***):* The LOFD method [10] adopts a density based approach to measure the degree of outlyingness of each instance. In comparison to nearest neighbor based approaches, it works by considering the difference in the local density ρ of the sample to that of its *k* neighbors; instead of relying on distance estimation alone. A higher score will be assigned to the sample, if ρ is highly different from the local densities of its neighbor. The algorithm starts by first computing the distance of the measurement x to its k^{th} nearest neighbor denoted by d_k, such that

$$\begin{aligned} d(x, x_j) \leq d(x, x_i) \quad & \text{for at least } k \text{ samples} \\ d(x, x_j) < d(x, x_i) \quad & \text{for at most } k-1 \text{ samples} \end{aligned} \tag{6}$$

The subsequent step is to construct a neighborhood $\mathcal{N}_k(x)$ by including all those points that fall within the d_k value. The following step is to calculate the reachability distance of sample x with respect to rest of the samples

$$d_r(x, x_i) = \max\{d_k(x_i), d(x, x_i)\} \tag{7}$$

The local reachability density ρ is the inverse of average d_r and can be defined as

$$\rho(x) = \frac{\mid \mathcal{N}_k(x) \mid}{\sum_{x_i \in \mathcal{N}_k(x)} d_r(x, x_i)} \tag{8}$$

Finally, the $\mathcal{S}^{(LOFD)}$ represents a local density-estimation score whereas value close to 1 mean x_i has same density relative to its neighbours. On the other hand, a significantly high $\mathcal{S}^{(LOFD)}$ score is an indication of anomaly. It can be computed as follows:

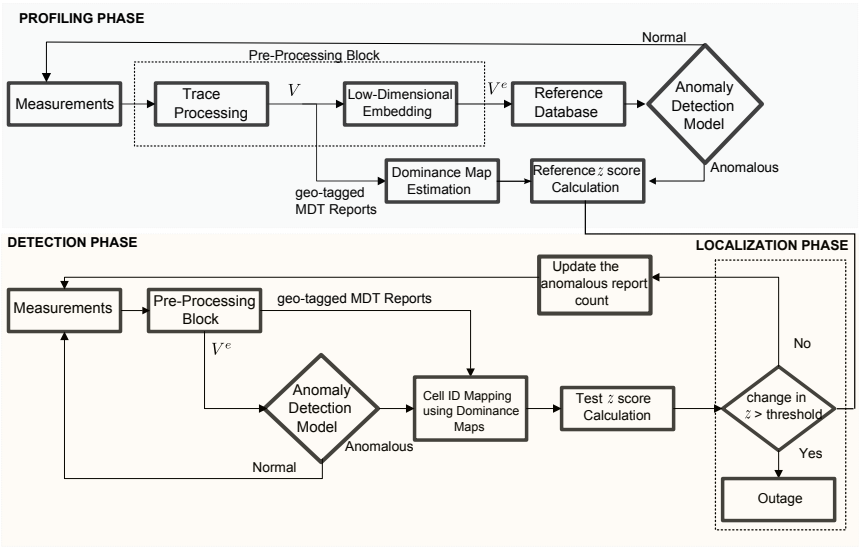
$$\mathcal{S}^{(LOFD)}(x) = \frac{\sum_{x_i \in \mathcal{N}_k(x)} \frac{\rho(x_i)}{\rho(x)}}{\mid \mathcal{N}_k(x) \mid} \tag{9}$$

Since, $\mathcal{S}^{(LOFD)}$ is sensitive to the choice of k, we iterate between k_{min} and k_{max} value for each sample, and take the maximum $\mathcal{S}^{(LOFD)}$ as described in Algorithm 1.

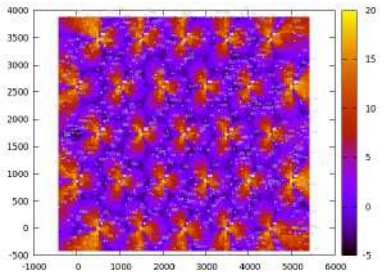
*2) One-Class Support Vector Machine based Detector (***OCSVMD***):* One-Class Support Vector Machine by Schölkopf et al. [11] maps the input data/feature vectors into a higher dimensional space in order to find a maximum margin hyperplane that best separates the vectors from the origin. The idea is to find a binary function or a decision boundary that corresponds to a classification rule

$$f(x) = < \mathbf{w}, \mathbf{x} > + b \tag{10}$$

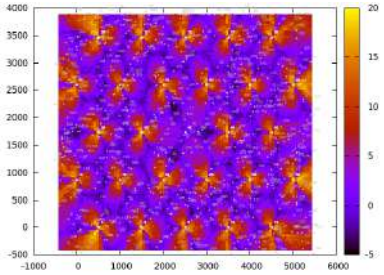
The $\mathbf{w}$ is a normal vector perpendicular to the hyperplane and $\frac{b}{\|\mathbf{w}\|}$ is an offset from the origin. For linearly separable cases,



(a) Profiling, Detection and Localization Phases of COD Framework



(b) Normal/Reference Scenario



(c) Outage Scenario

Fig. 2: An overview of COD Framework

Algorithm 1 Local Outlier Factor Based Detection Model

Input Data $\mathcal{X} = \{x_j\}_{j=1}^{N}$, k_{min}, k_{max}
for $j = 1, 2, \ldots, N$: **do**
  for $k =k_{min}$ to k_{max}: **do**
    Find $d_k(x_j)$ from Equation 6
    Find the neighborhood $\mathcal{N}_k$ of x_j
    Calculate $d_r(x_j, x_i)$ from Equation 7
    Calculate $\rho(x_i)$ from Equation 8
    Calculate $\mathcal{S}^{(LOFD)}$ from Equation 9
  end for
  $\mathcal{S}^{(LOFD)} = \max(\mathcal{S}^{(LOFD)}{}_{k_{min}}, \ldots, \mathcal{S}^{(LOFD)}_{k_{max}})$
end for

the maximization of margin between two parallel hyperplanes can be achieved by optimally selecting the values of w and b. This margin, according to the definition is $\frac{2}{\|\mathbf{w}\|}$. Hence, the optimal hyperplane should satisfy the following conditions

$$\begin{aligned} &minimize\ \frac{1}{2}\|\mathbf{w}\|^2 \\ &subject\ to: y_i(\langle \mathbf{w}, x_i \rangle + b) \geq 1 \\ &for\ i = 1, \ldots, N \end{aligned} \tag{11}$$

The solution of the optimization problem can be written in an unconstrained dual form which reveals that the final solution can be obtained in terms of training vectors that lie close to the hyperplanes, also referred to as support vectors. To avoid overfitting on the training data, the concept of *soft decision* boundaries was proposed, and slack variable ξ_i and regularization constant ν is introduced in the objective function. The slack variable is used to soften the decision boundaries, while ν controls the degree of penalization of ξ_i. Few training errors are permitted if ν is increased while degrading the generalization capability of the classifier. A *hard margin* SVM classifier is obtained by setting the value of $\nu = \infty$ and $\xi = 0$. The detail mathematical formulation for SVM models can be found in [11]. The original formulation of SVM is for linear classification problems; however non-linear cases can be solved by applying a kernel trick. This involves replacing every inner product of $x.y$ by a non-linear kernel function, allowing the formation of non-linear decision boundaries. The possible choices of kernel functions includes polynomial, Gaussian radial basis function (RBF), and sigmoid. In this study, we have used the RBF kernel: $\kappa(x, y) = exp(-\|x - y\|^2/2\sigma^2)$, and the corresponding parameter values of the model are selected using cross validation method, as described in Algorithm 2.

As shown in Figure 2(a), using the benchmark data, we compute a reference z-score for each target eNB in the network. The z-score is calculated as follows:$z_b = \frac{|n_b - \mu_n|}{\sigma_n}$, where n_b is the number of MDT reports labeled as anomalies for the eNB b, and variables μ_n and σ_n are the mean and standard deviation anomaly scores of the neighbouring cells. In the profiling phase, we also estimate the so called dominance area, i.e., for each cell, we define the area where its signal is the strongest. This is to establish the coverage range for each cell by exploiting the location information tagged with each UE measurement. The dominance estimation is required to determine a correct cell and MDT measurement association during an outage situation. This is because as soon as the SC situation triggers in the network, the malfunctioning eNB either becomes completely unavailable or experience severe performance issues. This triggers frequent UE handovers to the neighboring cells, and as a result the reported measurements from the affected area contains the neighbor cell E-UTRAN Cell Global Identity (CGI), instead of the target cell. Hence, CGI alone cannot be used to localize the correct position of faulty cell during an outage situation. The detection and localization phase of our COD framework make use of estimated dominance map and reference z-score information established in the profiling phase to detect and localize faulty cell as

discussed in the following subsection.

B. Detection and Localization Phase

In the detection phase, the trained detection model is employed to classify network measurements as normal or anomalous. The output of the detection models allow us to compute a test z-score for each eNB. To establish a correct cell measurement association, the geo-location of each report is correlated with the estimated dominance maps. In this way, we can achieve detection and localization by comparing the deviation of test z-score of each cell with that of reference z-score, as illustrated in Fig 2(a).

III. Simulation Results

In this section, we demonstrate the performance of our COD framework by presenting the simulation results obtained under different network operating conditions.

A. Simulation Setup

To simulate the LTE network based on 3GPP specifications, we employ a full dynamic system tool. We set up a baseline reference scenario that consists of 27 eNBs having an inter-site distance of 1000m, with a cell load of 10 users. To model the variations in signal strength due to topographic features in an urban environment, the shadowing is configured to be 8 dB. Normal periodical MDT measurements as well as RLF-triggered data due to intra-network mobility, reported by UE's to eNBs, is used to construct a reference database for training outage detection models. To simulate a hardware failure in the network, at some point in the simulation the antenna gain of cell 11 is attenuated to -50 dBi that leads to a cell outage in a network. The measurements collected from the outage scenario are then used to evaluate the detection and localization performance of the proposed COD framework. The SINR plots of the reference and SC scenario has been already shown in Figure 2(b) and 2(c), respectively. The detailed simulation parameters are listed in Table II. The detection performance of the outage detection models is also examined for different network configurations, obtained by varying the simulation parameter settings for ISD, load and shadowing.

Parameter	Values
Cellular Layout	27 Macrocell sites
Inter-site Distance (ISD)	1000m
Sectors	3 Sectors per cell
User Distribution	Uniform Random Distribution
Path Loss	$L[dB] = 128.1 + 37.6log_{10}(R)$
Antenna Gain (Normal Scenario)	15 dBi
Antenna Gain (SC Scenario)	-50 dBi
Slow Fading Std	8 dB
Simulation Length	420s (1 time step = 1ms/1TTI)
BS Tx Power	46 dBm
Network Synchronization	Asynchronous
HARQ	Asynchronous, 8 SAW channels, Maximum Retransmission = 3
Cell Selection Criteria	Strongest RSRP defines the target cell
Load	10 users/cell
MDT Reporting Interval	240 ms
Traffic Model	Infinite Buffer
HO Margin	3dB

TABLE II: Simulation Parameters

Algorithm 2 Parameter Estimation using CV Method

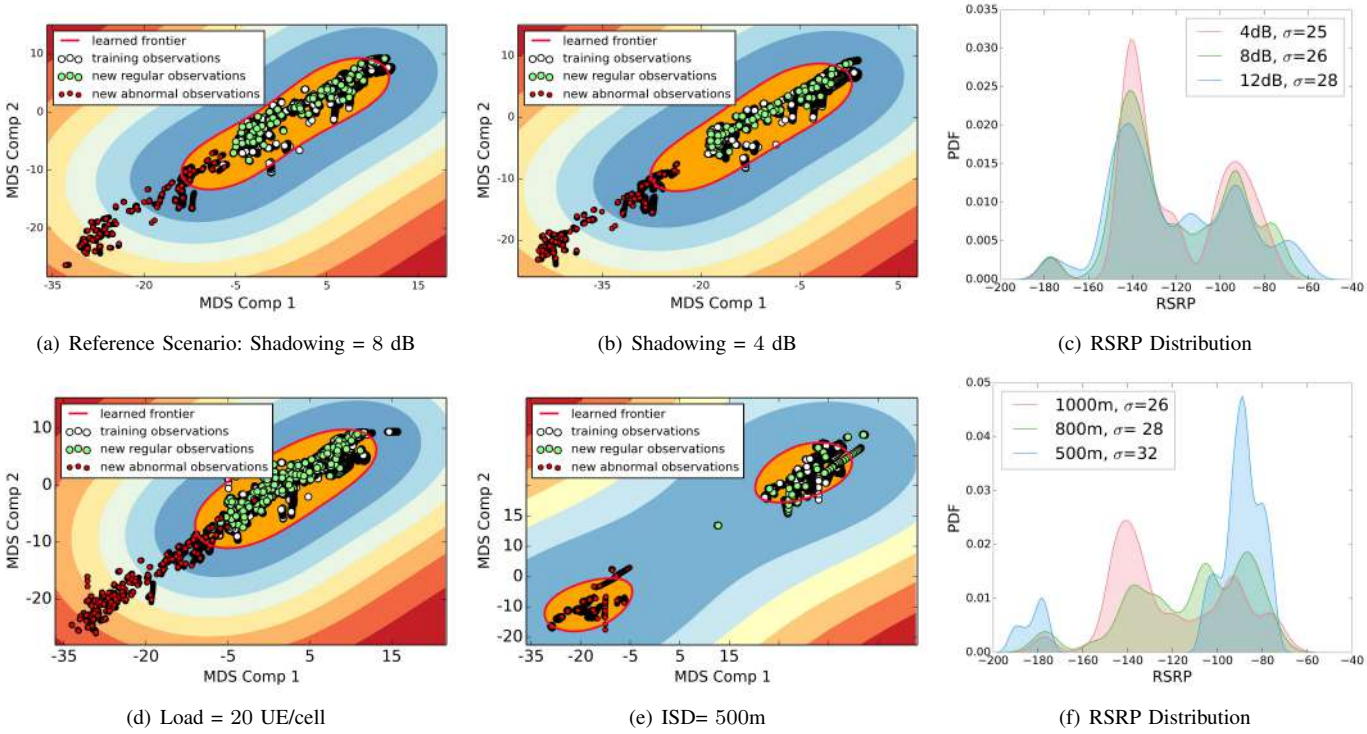
1: Define parameter combination C_i, $i = 1, \ldots, M$
2: **for** $i = 1, 2, \ldots, M$: **do**
3: Split the target dataset D_M into K chunks.
4: **for** $l = 1, 2, \ldots, K$: **do**
5: Set D_{val} to be the l^{th} chunk of data
6: Set D_{train} to be the other $K-1$ chunks.
7: Train model using C_i, D_{train} and evaluate its performance P_l on D_{val}.
8: **end for**
9: Compute average performance P_i over K chunks
10: **end for**
11: **Parameter Selection**: Select C_i corresponding to highest P_i
12: **Performance Estimation**: Evaluate the performance of the model $M(C_i, D_M)$ on D_{test}

Parameter Estimation and Evaluation: The parameter selection for LOFD and OCSVMD is performed using a combination of grid search and cross-validation (CV) method as listed in Algorithm 2. Initially, a grid of parameter values are specified that defines the parameter search space. For example, the hyperparameters of OCSVMD ν and kernel parameter γ is varied from 0 to 1 with 0.05 interval to determine different combinations. Subsequently, for every unique parameter combination C_i, CV is performed as follows: The D_M is divided into training D_{train} and validation dataset D_{val}, and subsequently performance of the model is evaluated using K-folds approach as shown in Algorithm 2. The value of K is chosen to be 10 in our framework. The performance estimate of the model over K folds is averaged and iteratively this process is repeated until all the parameter combinations are exhausted. The C_i yielding the highest performance estimate is selected as an optimal parameter combination for the target model. The value of k_{min} and k_{max} for LOFD is found out to be 5 and 14. In case of OCSVMD, RBF kernel is employed and the values of the hyperparamters ν and γ is found out to be 0.3 and 0.25, respectively. Finally, the test data D_{test} from the outage scenario, has been used to estimate the performance of the trained models.

In our study, the quality of the target models is evaluated using Receiver Operating Characteristic (ROC) curve analysis. The ROC curve plots the true positive rate or detection rate (DR) (i.e., a percentage of anomalous measurements correctly classified as anomalies) against the false positive rates (FPR) (i.e., a percentage of normal cell measurements classified as anomalies) at various threshold settings. An Area under ROC curve (AUC) metric is used for model comparison, whereas a AUC value of 1 or close to it, is an indicator of higher discriminatory power of the target algorithm.

B. Cell Outage Detection Results

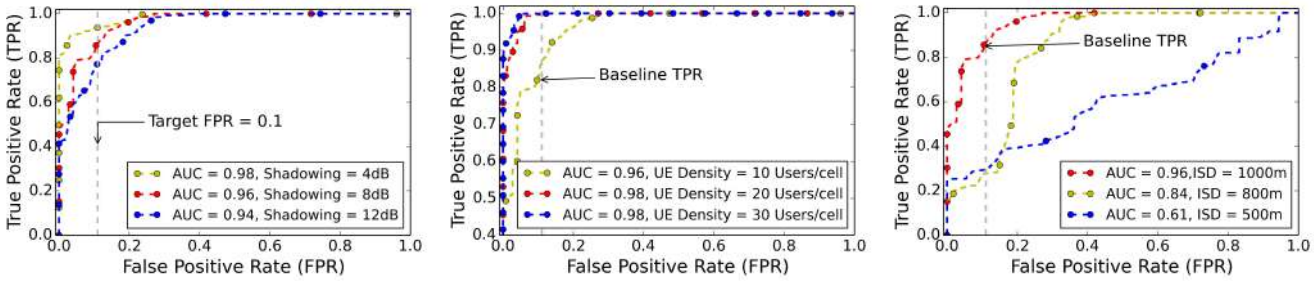
The training database D_M contains pre-processed embedded measurements from the reference scenario as discussed in Section II-A. The database is subsequently used to model the normal operational behaviour of the network. The database measurement also includes RLF-triggered samples, since even in the reference scenario UE's experience connection failures due to intra-LTE mobility or shadowing. The test data collected



(a) Reference Scenario: Shadowing = 8 dB (b) Shadowing = 4 dB (c) RSRP Distribution

(d) Load = 20 UE/cell (e) ISD= 500m (f) RSRP Distribution

Fig. 3: (**a**) OCSVMD learned network profile for Reference Scenario (**b**) Low-shadowing case (**c**) Distribution of RSRP values for all shadowing cases (**d**) Medium Traffic case (**e**) smaller ISD case (**d**) Distribution of RSRP values for all ISD cases



(a) Shadowing (b) Cell Load (c) ISD

Fig. 4: OCSVMD ROC Curves for shadowing, traffic and ISD cases

from the outage scenario is used to evaluate the performance outage detection models.

The diagnosis process has been tested in twelve scenarios by changing the shadowing, user-density and inter-site distance (ISD) parameters of the baseline simulation setup as listed in Table II. We have evaluated the detection performance of the OCSVMD and LOFD against every target network configuration. Figure 3(a), illustrates the MDS projection of MDT measurements from the normal and the outage scenario using the baseline network operational settings. It can be observed that the abnormal measurements belonging to SC scenario lie far from the regular training observations. As discussed earlier in Section II-A, MDS tries to maximize the variance between the data points and consequently dissimilar points are projected far from each other, allowing the models to compute a robust dissimilarity measure for outage detection. The goal of OCSVMD is to learn a close frontier delimiting the contour of training observations obtained from the non-outage scenario. In this way, any observation that lie outside of this frontier-delimited subspace (i.e. representative of the normal state of the network) is classified as an anomaly or an abnormal measurement. However, the inlier population (i.e. measurements that lie inside the OCSVMD frontier) is contaminated with RLF events, which ultimately elongates the shape of the learned frontier. As a result, during the detection phase, the observations from the outage scenario exhibiting similarity to RLF-like observations are positioned within the frontier-delimited space as shown in Figure 3(a), and hence wrongly classified as normal. The shape of the learned frontier determines the precision of the model for detecting anomalous network measurements.

To study the impact of different radio propagation environment on the detection performance, we varied the shadowing parameter from 8 dB to 4 dB and 12 dB cases. Under low-

shadowing conditions (i.e., 4dB), it can be observed from Figure 3(b) that inlier population exhibits wider separation from anomalous observation in comparison to reference scenario. This is because higher shadowing conditions affects the spread of the KPI measurements, as indicated in Figure 3(c). It can be inferred from the ROC analysis of OCSVMD, that detection performance deteriorates as the shadowing effect is varied from low to high. As shown in Figure 4(a), at target false positive rate of 10%, the model reports the highest detection rate (i.e.TPR) of 93% under low-shadowing conditions. Likewise, the AUC score has also decreased from 0.98 to 0.94 for high-shadowing scenario (i.e., 12 dB). Moreover, we also analyzed the OCSVMD detection performance under varying traffic conditions. Figure 3(d) depicts the distribution of measurements in the MDS space for a user density of 20 per cell. The higher user density implies an increase in the number of training observations that leads to a more accurate estimate of the frontier shape. This explains the slight improvement in the AUC score for OCSVMD with the increase in the cell load as shown in Figure 4(b). A notable detection rate improvement of 10% is observed for high traffic scenario (i.e., 30 users per cell) in comparison to the baseline OCSVMD.

As for different ISD configurations of a network, we see a significant change in the values of KPI measurements. This is expected since there is a strong correlation between UE reported KPI's and their distance from the eNB. Figure 3(f) shows the distribution of UE reported RSRP values for three different ISD cases. In case of ISD=500m, we see a distinct peak of RSRP values around -90dBm. Likewise, at the farther left end we see a small peak around -180dBm that is mainly due to RLF-like observations. In contrast, when ISD=1000m, the highest peak value is observed at around -140 dBm, and the observed measurements have lower data spread as indicated in Figure 3(f). As already highlighted, the shape of the learned frontier by OCSVMD is directly affected by the distribution of observations in the embedded space. This becomes evident in Figure 3(e) which shows that the OCSVMD learns two decision frontiers instead of one, since there exists two distinct modes in the data distribution, for the case of ISD=500m. As a result, OCSVMD interprets a region where RLF-like event are clustered, as inliers, which leads to an inaccurate network profile. The ROC analysis shown in Figure 4(c), clearly indicate the degradation of OCSVMD performance for lower ISD values.

Similar to OCSVMD, the performance of LOFD is also evaluated for all target network configurations. As explained in Section II-A, LOFD derives a measure of outlyingness of an observation (i.e., $\mathcal{S}^{LOFD}$), based on the relative data density of its neighborhood. Figure 5(a) illustrates the labels assigned by LOFD to the observations obtained from the baseline scenario. It can be observed that LOFD classifies some of the test instances that even lie close to the vicinity of training observations as anomalous. Due to such instances LOFD receives a high outlying scores $\mathcal{S}^{LOFD}$, since the local density around them is highly different from the density of its neighborhood. To further illustrate the impact of the variation and spread of the data on the values of $\mathcal{S}^{LOFD}$, we plot a cumulative distribution function (CDF) for different shadowing scenarios, as shown in Figure 5(b). It can be seen that for low-shadowing scenario, almost 80% of the observations obtain $\mathcal{S}^{LOFD}$ value less than 50. However, as the shadowing increases we see a gradual increase in the value of $\mathcal{S}^{LOFD}$. Likewise, a similar behaviour is observed with the increase of ISD, as shown in Figure 5(c). The shadowing and ISD parameters influence the distribution and spread of the data as explained earlier, and consequently the value of $\mathcal{S}^{LOFD}$. This leads to a low detection performance of LOFD, since it generates an increased number of false alarms.

As shown in Figure 6(a), the AUC score for LOFD decreases for high-shadowing scenario. On the other hand, the increase in the cell load also increase the spread of the data, which consequently affect the detection performance of LOFD. As shown in Figure 6(b), at false alarm rate of 10%, the highest detection rate of 81% is achieved for a network scenario in which load configuration is set to be 10 users per cell. Similarly, the change in the ISD has a severe effect on the model performance and low detection performance of 60% and 30% is achieved for 800m and 500m ISD configurations, as shown in Figure 6(c).

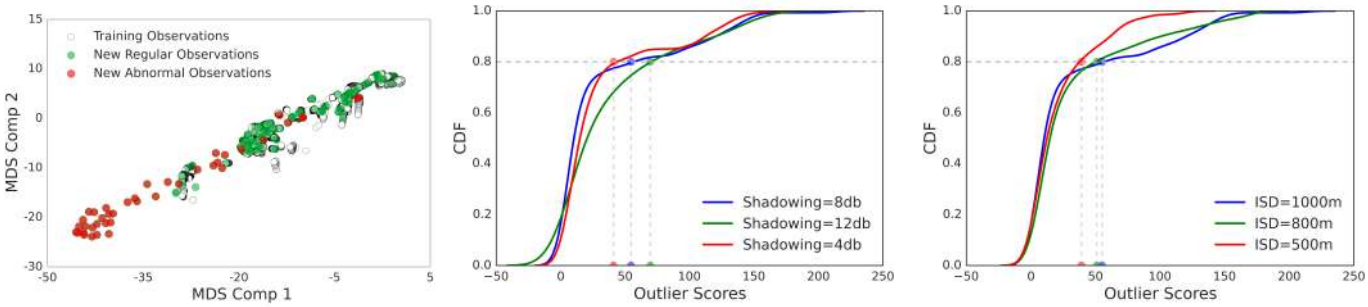
In summary, we can conclude from the reported results that OCSVMD under most cases achieves a better detection performance in comparison to LOFD. The outage detection models yields worst performance scores particularly for low ISD network configuration. The performance issue of the target outage detection models can be addressed as follows: For OCSVMD, in the pre-processing step the RLF-like events must be filtered before constructing a training database. This would help decrease the spread of the data and the model would only learn frontier that corresponds to normal operational network behaviour. In case of LOFD, incremental drift detection schemes can be incorporated to re-tune the model parameters in order to minimize the false alarm rate.

C. Localization

Since, OCSVMD model has outperformed LOFD for most test cases, it has been selected as a final model to compute per cell z-scores for the normal and SC scenario, as shown in Figure 7. It can be observed from Figure 7 that measurements are classified as anomalous even in the normal operational phase of the network due to occurrence of RLF events. This is particularly true for cell ID 1,5,11,16, and 19 whose n_b values are found to be 700,2000, 3000, 1500, respectively in the reference scenario. However, during an outage scenario, since cell 11 is configured as a faulty cell, the corresponding z-scores are significantly higher than the rest of the network. A simple decision threshold can be applied on the computed z-scores to autonomously localize faulty cells, and consequently an alarm can be triggered. In addition to cell outage localization, the change in the z-score can be used to identify performance degradation issues or a weaker coverage problems. This information can act as an input to self-healing block of SON engine, which can then trigger automated recovery process.

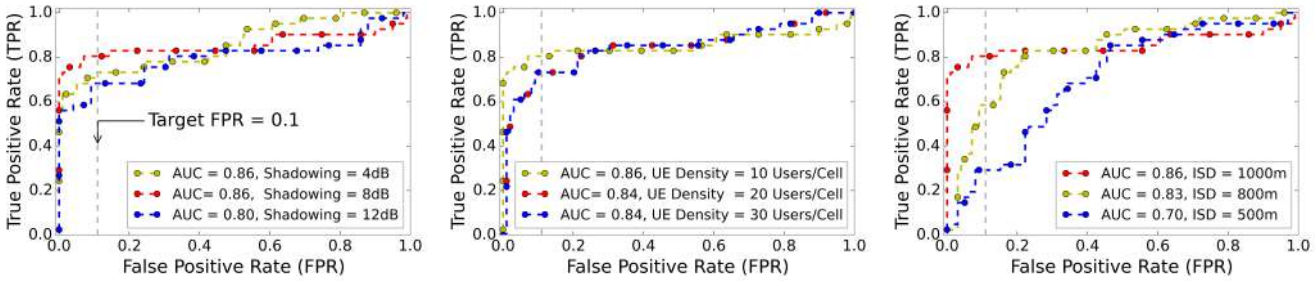
IV. Conclusion

This paper has presented a data-driven analytics framework for automating the sleeping cell detection process in an LTE network using minimization of drive testing functionality as specified by 3GPP in Release 10. The proposed approach builds a normal profile of the network behaviour in a low-dimensional embedding space. The measurements are labeled as anomalous if they deviate from the learned profile. For



(a) Reference Scenario: Shadowing = 8 dB (b) CDF of $\mathcal{S}^{LOFD}$ for Shadowing cases (c) CDF of $\mathcal{S}^{LOFD}$ for ISD cases

Fig. 5: Network profiling using LOFD



(a) Shadowing (b) Cell Load (c) ISD

Fig. 6: LOFD ROC Curves for shadowing, traffic and ISD cases

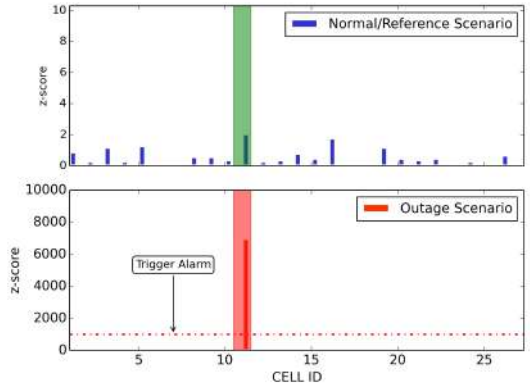


Fig. 7: Localization of SC based on per cell z-scores

this purpose, multi-dimensional scaling method in conjunction with domain and density based detection models: OCSVMD and LOFD, respectively, were examined for different network conditions. It was established that OCSVMD, a domain based detection model attained a higher detection accuracy compared to LOFD which adopts a density based approach to identify abnormal measurements. Finally the UE reported coordinate information is employed to establish the dominance areas of target cells which are subsequently used to localize the position of sleeping cell. The proposed cell outage detection framework can act as a foundation for next generation network monitoring tools, since it allows easy inclusion of other key performance indicators from the network and can be extended to detect other issues such as coverage holes, weak coverage as well as performance degradation problems.

ACKNOWLEDGMENT

This work was made possible by NPRP grant No. 5-1047-2437 from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 336–361, 2013.

[2] S. Hämäläinen, H. Sanneck, C. Sartori *et al.*, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*. John Wiley & Sons, 2012.

[3] R. Barco, V. Wille, and L. Díez, "System for automated diagnosis in cellular networks based on performance indicators," *European Transactions on Telecommunications*, vol. 16, no. 5, pp. 399–409, 2005.

[4] B. Cheung, S. G. Fishkin, G. N. Kumar, and S. A. Rao, "Method of monitoring wireless network performance," Sep. 21 2004, uS Patent App. 10/946,255.

[5] Y. Ma, M. Peng, W. Xue, and X. Ji, "A dynamic affinity propagation clustering algorithm for cell outage detection in self-healing networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2013, pp. 2266–2270.

[6] R. M. Khanafer, B. Solana, J. Triola, R. Barco, L. Moltsen, Z. Altman, and P. Lazaro, "Automated diagnosis for umts networks using bayesian network approach," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2451–2461, 2008.

[7] A. Coluccia, A. DAlconzo, 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.

[8] C. M. Mueller, M. Kaschub, C. Blankenhorn, and S. Wanke, "A cell outage detection algorithm using neighbor cell list reports," in *Self-Organizing Systems*. Springer, 2008, pp. 218–229.

[9] T. F. Cox and M. A. Cox, *Multidimensional scaling*. CRC Press, 2010.

[10] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[11] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.