WiMAX traffic analysis and base stations classification in terms of LRD

Cristina Stolojescu-Crisan,1 Alexandru Isar,1 Sorin Moga2 and Philippe Lenca2

1. Introduction

During the last years, the role of wireless technology has expanded significantly, becoming more affordable and easy to use. There has been a significant increase of the demands for delivering telecommunication services to home residences and business premises; therefore, the objectives of telecommunication equipments are to provide high quality services to customers and to remain economically competitive. Because of the Internet expansion, huge amounts of data are available, requiring adapted analysis methods as the data mining methods are.

Network traffic data are rich in information because these handle a large amount of data, follow a strict time ordering and enable specialists to study a variety of network characteristics. More, traffic data analysis helps network operators to manage their network monitoring, network traffic engineering, capacity planning and protocols evaluation (Andersen & Feamster, 2006).

Worldwide Interoperability for Microwave Access (WiMAX) is a quite new telecommunication technology, based on IEEE 802.16 standard, capable of delivering advanced IP applications, such as voice, video and data, over the microwave radio frequency spectrum, to stationary and moving users.

The performance of wireless communication networks depends on an efficient architecture [correct positioning of base stations (BS)]. The understanding of the network topology could be performed by using data mining methodologies (Calcada et al., 2012).

Our purpose was to analyse the traffic in a WiMAX network, to evaluate the correctness of the BS positioning in the network topology (to identify those BS with an incorrect localization in the topology of the network). The strategy chosen for this purpose is based on the long-range dependence (LRD) analysis of traffic, which became quite used in data analysis (Karagiannis et al., 2004). LRD, also known as long memory or strong dependence, is a characteristic of some random processes, which have a persistent correlation. The goal of this paper was to prove that LRD analysis of traffic time series, acquired in a WiMAX network, can be used to evaluate the network topology. More, we will classify the BS in terms of the number of days in which the uplink traffic is long-range dependent. A fusion of those results with the results already reported for downlink channel, obtained for the same network and in the same time interval, will reveal the final conclusions about BS localization in the network architecture (Stolojescu, 2010).

The LRD of a traffic time series expresses the heaviness of traffic. The time series associated with light traffic have short-range dependence. The database analysed in this paper contains a large amount of data. So, the LRD analysis proposed in this paper searches new knowledge form large volume of data, following the definition of data mining, given at the beginning of Section 4. Hence, our objective is specific for data mining.

The present paper represents one of the first attempts to identify the BS incorrectly positioned in a WiMAX network, based on a traffic analysis, implemented by using an original data mining methodology. More precisely, we analysed the LRD of traffic, for each BS composing the considered WiMAX network. There are other data mining methodologies which can be used for the estimation of the heaviness of traffic, as for example the traffic forecasting (Stolojescu et al., 2009). So, the identification of the BS with an incorrect position in the architecture of a WiMAX network could be carried out using other approaches as well. The statement of the problem proposed in this paper is the following: considering the traces obtained by monitoring
the traffic in a WiMAX network, composed by 67 BS, during 8 weeks, and with a sampling step of 15 min, we must identify the BS that have an incorrect position.

In Section 2, some related works are briefly presented. Our approach is described in Section 3. Section 4 describes the dataset used in this study. In Section 5, we present a short introduction to the topic of LRD in communication networks, the definition of the Hurst parameter, which measures the degree of LRD, and some methods used to estimate it (Clegg, 2005). In Section 6, the results of the LRD analysis are presented, and conclusions and perspectives on future work are enumerated in the last section of the paper.

2. Related work
Communication networks produce large amounts of data that can reveal a variety of network characteristics, useful for better understanding the network behaviour (Andersen & Feamster, 2006). Data mining is appropriate for the extraction of those network characteristics from this large amount of data.

The analysis of traffic measurements, which are time series, from various communication networks, has proved that the traffic is long-range dependent or fractal (self-similar) (Karagiannis et al., 2004; Clegg, 2005). An overview of what the community has learned during 10 years of research has been made by Karagiannis et al. (2004).

The LRD is often found in network traffic aggregated from multiple sources. One of these possible origins is that LRD arises from the network topology (Clegg, 2004). This means that, if a BS exhibits strong LRD, it should be repositioned when the next session of network maintenance will take place. We have not found in the literature many studies about the LRD of traffic in WiMAX networks. The WiMAX downlink traffic was analysed in terms of LRD by Stolojescu (2010). It has been shown that WiMAX traffic exhibit LRD. Other related work, concerning different sources of LRD, will be mentioned in Section 5.3.

3. Proposed method
We aim to analyse the localization of BS in the topology of the network. First, we analyse the uplink traffic using the methodology proposed by Stolojescu et al. (2011), and next, we combine the results obtained with the results corresponding to downlink channel, to identify the BS with an incorrect positioning.

Our approach supposes several steps. We deal with a database obtained by monitoring the traffic in a WiMAX network, composed by traces corresponding to each BS and having a duration of 8 weeks. The first step of our approach was to understand the initial data collection, by learning the structure of the database. The second step of our approach was the data cleaning and preprocessing, to reconstruct missing data. The third step was to find useful information in order to represent data. First, we have estimated the Hurst parameter for all the time series from the database. Next, we have segmented those time series in sequences with the duration of 1 week and have estimated again their Hurst parameters. Finally, we have segmented all the time series from the database in sequences with the duration of 1 day and have estimated the Hurst parameters. The fourth step of our approach was to create the final dataset, which is composed by binary values corresponding to daily values of the Hurst parameter, for each BS. The final step of our approach was to search for patterns of interest and to interpret the discovered patterns. We considered the frequency of apparition of daily LRD in uplink as decision criterion about the correct positioning of the corresponding BS. The steps already mentioned are the following: understanding the initial data collection, data cleaning and preprocessing, finding features to represent data, creating the final data set, and searching for patterns of interest and interpreting the discovered patterns, which composes the structure of any data mining project (Chapman, 2006), as it is mentioned in Section 4. So, our approach belongs to a large research domain, data mining.

4. Mining data
Knowledge Discovery is a domain that searches new knowledge about an application domain. One of its branches is data mining, which is an analytic process designed to explore and to extract useful information from large volume of data. Mining time series data is one of the most challenging problems in data mining research (Yang & Wu, 2006).

The database used for this work was obtained by monitoring the traffic in a WiMAX network developed by Alcatel Lucent Timisoara, Romania. We collected values for two particular Management Information Base objects (virtual database used for managing the entities in a communications network). These values represent the amount of traffic measured in bytes per second, or in packets per second, for all the links of a BS, for all the BS in the WiMAX network, during a period of 8 weeks. The data were organized in .xls files, corresponding to a particular week. Each of these files contains information about 67 BS. So, we deal with 67 traces, 8 weeks long. The values were recorded every 15 min, so it can be easily deduced that for each BS, we have 96 samples/day, 672 samples/week and a total number of 5376 samples.

The first step of the proposed approach is to understand the structure of the WiMAX traffic database. The second step supposes data cleaning and preprocessing. According to CRISP-DM methodology, data preparation includes selecting data to be used for analysis, data clearing, such as identification of the potential errors in data sets, handling missing values, and removal of noises or other unexpected results that could appear during the acquisition process (Chapman, 2006). The database was affected by 10% of missing values. We have reconstructed our data by cubic interpolation. From this point, we have used the time series obtained after interpolation. We have started with a preliminary analysis of these time series. At this stage, the input data were analysed to find if it contains large spikes and valleys, indicating periodicities. The simple plot of the traffic curves proved the existence of periodicities in the traffic. To verify the presence of these periodicities, we computed the Fourier Transform of the signal and analysed the power spectral density. The analysis indicates that the
most dominant period across all traces is the 24 h one. We have also detected a secondary period of 7 days. Another result of the preliminary analysis is that the traffic for each BS is non-stationary. The tendency of traffic was estimated for each BS by a forecasting methodology by Stolojescu et al. (2009). It was shown that the tendency of traffic, which can be interpreted as a mean value of a random variable, varies in time. So, the traffic is non-stationary.

The third step of our approach is to find a useful feature for representing the degree of heaviness of WiMAX traffic. We propose the use of the daily uplink traffic Hurst parameter as feature for the representation of the degree of heaviness of WiMAX traffic. There are multiple sources for the LRD of network traffic, as it will be shown in Section 5.3. The effects of those sources are different in uplink and downlink. An LRD source, common for uplink and downlink, is the incorrect positioning of BS. The strategy chosen in our approach is to separate and to analyse the effects of this LRD source.

The fourth step of our approach is to create the final dataset. It is composed by binary values, which correspond to the presence or absence of LRD, produced by the incorrect positioning of a BS in a given day.

The last step of the proposed approach is to search for patterns of interest and to interpret them. We consider as pattern of interest the frequency of apparition of LRD, caused by the incorrect positioning of BS. We will classify the BS, following this criterion, as incorrectly positioned, if the frequency of apparition of daily LRD is high. We will give supplementary details about this approach in Section 6.

Our methodology is sufficiently general to be independent of other parameters of the considered WiMAX network, as for example the strategies of traffic control, or the number of users who exploit a given BS, because our database was constructed by collecting data from a functional network with strategies of traffic control already established and users already deserved.

5. Long-range dependence in communications networks

In this section, we will present the definition of LRD, some considerations about its detection in time series and some connections of this concept with the traffic behaviour in communication networks.

5.1. Definitions of LRD

A stationary random process, \( X_t \), has LRD if its autocorrelation function (ACF), \( \rho(k) \), has a slow decay (Samorodnitsky, 2007). The speed of decaying of an ACF can be evaluated by using the sum of the series with the general term given by the considered correlation. A random process has LRD if the corresponding series is divergent:

\[
\sum_{k=-\infty}^{\infty} \rho(k) = \infty
\]  

(1)

The White Gaussian Noise process has the fastest decaying of the ACF. Indeed, its autocorrelation is proportional with the unit pulse \( \rho(k) = \delta(k) \), and the sum of the series having this ACF as general term, is equal to 1. If the asymptotic behaviour of the ACF is

\[
\rho(k) \sim c_r k^{-z}
\]  

(2)

where \( c_r \) is a constant and \( z \in (0,1) \), the divergence in equation (1) is guaranteed (Grossglauser & Bolot, 1998). The random processes with the ACF satisfying equation (2) are long-range dependent. Hence, the LRD can be detected estimating the value of \( z \) and verifying if it belongs to the interval \((0,1)\).

On the basis of the previous observation, the LRD of a stationary process, \( X_t \), can be detected in the frequency domain as well (Clegg, 2004). By computing the Fourier Transform in discrete time of the ACF in equation (2), the asymptotic behaviour of the power spectral density of an LRD random process, \( X_t \), can be obtained:

\[
f(\lambda) \sim c_f |\lambda|^{-\beta}
\]  

(3)

with \( \lambda \to 0 \), \( \beta \in (0,1) \) and \( c_f \) is a constant, in conformity with Wiener–Hinckin theorem:

\[
f(\lambda) = \sum_{k=-\infty}^{\infty} \rho(k)e^{-ik\lambda}
\]  

(4)

So, the LRD can be detected in the frequency domain by estimating the value of \( \beta \) and by verifying if it is included in the interval \((0,1)\). A different asymptotic behaviour in the frequency domain, which permits the detection of LRD, is presented in equation (5) and includes a pole for the power spectral density:

\[
f(\lambda) \sim c_f (\cos \lambda - \cos \lambda_0)^{-\beta}
\]  

(5)

where \( \lambda \to \lambda_0 \), \( \lambda_0 \in [0,\pi] \), \( \beta \in (0,1) \) and \( c_f \) is a positive constant. The degree of LRD of a time series can be estimated with the aid of its Hurst parameter, which will be defined in Section 5.2.

5.2. Estimation of Hurst parameter

The predominant way to quantify LRD is the estimation of the Hurst parameter \( H \). The Hurst parameter of a process, \( X_t \), can be defined by using the parameter \( z \) from equation (2):

\[
H = 1 - z/2
\]  

(6)

or the parameter \( \beta \) from equation (3):

\[
H = (1 + \beta)/2
\]  

(7)

In conformity with equation (2), the considered process has LRD if \( z \in (0,1) \). The corresponding interval for \( H \) is \([0.5,1]\). The degree of LRD increases as \( H \to 1 \). A value of \( H \) smaller than 0.5 indicates that there is a short-term dependence between the samples (the ACF is absolutely summable).

There are several methods used to estimate \( H \) (Taqu & Teverovsky, 1998; Clegg, 2005). On the basis of equations (2) and (3), the Hurst parameter estimators can be classified...
into two categories: estimators operating in the time domain and estimators operating in the frequency domain. The estimators operating in the time domain are the following: rescaled adjusted range (R/S) method, aggregated variance method and absolute value method. The estimators operating in the frequency domain are the following: periodogram, Whittle estimator and wavelet-based estimators.

5.2.1. Time domain estimators One of the oldest and better known methods for estimating the parameter $H$ is the Rescaled Adjusted Range Method (R/S), proposed by Hurst (Hurst, 1951). This method uses the rescaled range statistic (R/S statistic), which is the range of partial sums of deviations of a time series from its mean, rescaled by its standard deviation. The R/S plot is obtained by representing the logarithm of the R/S statistic, versus the logarithm of the number of points. Finally, linear regression is used to fit a straight line to the R/S plot, the slope of this line being an estimate of $H$ (Willinger et al., 1998). The aggregated variance method is one of the easiest methods to estimate Hurst parameter. It consists in the segmentation of the time series analysed in the aggregation of those segments and in the estimation of the variance of the aggregates. The absolute value method exploits the same idea of segmentation, followed by aggregation, as the aggregated variance method, but it computes the absolute moment of the original time series, which depends on the aggregated series.

The disadvantage of the time domain Hurst parameter estimators is given by its asymptotic nature [equation (2)], which implies a great volume of data. This disadvantage can be counteracted by the use of frequency domain Hurst parameter estimators.

5.2.2. Frequency domain estimators In conformity with equation (3), LRD determines the spectrum of a process to behave as a power law for frequencies close to zero. One of the simplest estimators for the power spectral density of a process, based on the Fourier Transform, is the periodogram. It is a good estimator for stationary processes. In the case of non-stationary processes, the quality of the estimation performed by the periodogram decreases, as a consequence of the bad time localization of the Fourier transform. So, the periodogram can be improved by substituting in its definition the Fourier transform with the short-time Fourier transform (STFT). A new spectral estimator is obtained by averaging smoothed periodograms computed with STFT, on different segments of the time series, which must be analysed. To diminish the drawbacks of the Hurst parameter estimator based on periodograms, given by the difficulties of making a precise spectrum analysis at very low frequencies, Whittle proposed in 1953 the minimization of a likelihood function, which is applied to the periodogram of the time series.

The discrete Whittle (D-Whittle) estimator of the Hurst parameter consists of two analytic approximations to the exact Gaussian maximum likelihood estimator, to avoid the huge computational complexity of the exact algorithm. The first approximation basically replaces the covariance matrix by an integral of a function of the spectrum. Because the computational difficulties remain after this approximation, a second approximation is performed. It consists in the discretization of the frequency domain integration rewritten in terms of the periodogram. The constant-bandwidth spectral analysis performed by the STFT used for the computation of the periodogram does not match with the structure of the power spectral density of a LRD process, which requires higher resolution at very low frequencies. The wavelet-based spectral estimators perform a constant relative bandwidth spectral analysis that matches with the structure of the power spectral density of an LRD process, enabling the increase of resolution at very low frequencies.

Abry and Veitch proposed a Hurst parameter estimator, at each scale of the wavelet decomposition of the random process that must be analysed (Abry et al., 2003). Basically, they estimate the variance of the wavelet coefficients from any decomposition level. In the design of the Abry–Veitch estimator is assumed that a continuous-time input process is available. There are numerous cases, as is our case, where only discrete time observations of the input process are available. To extend the Abry–Veitch estimator to these cases, a discretization procedure is required. A discretization method based on generalized quadrature variations has been proposed by Istaš and Land (1987). By using this method, the discretized Abry–Veitch estimator (DAV) was obtained (Abry et al., 2003). This Hurst parameter estimator is implemented in Matlab by the function HEST. It represents the best solution for the evaluation of the LRD for non-stationary discrete time series (which constitutes our problem), as we will show in a following section.

5.2.3. A Comparison of some Hurst parameter estimators based on simulations For implementation reasons, in the following, we will compare two software products, which perform Hurst parameter estimation: SELFIS and Matlab. This comparison is necessary because we have not found in the literature specifications about the precision of these implementations of the estimators of Hurst parameter in the case of non-stationary discrete time series.

The Hurst parameter estimators used in the next simulations are the following: aggregate variance, R/S, periodogram, and absolute value (implemented in SELFIS) and generalized quadrature variations estimator (also called DAV) (implemented in Matlab$^{8}$ – function HEST). As already mentioned, these estimators of the Hurst parameter could have different performance for the time series that compose our database. We deal with non-stationary discrete time series, having different lengths. In the following simulations, we will use two types of random processes, with known values of $H$, and we will check the estimated values given by different estimation methods. For the first simulations, we considered a White Gaussian Noise as input process, for the simple verification of the implementations of Hurst parameter estimators considered. The corresponding value of $H$ must be 0. The simulation results are presented in Table 1.

By analysing the results in Table 1, it can be observed the increasing of the bias of R/S estimations with the decreasing of the length of the input sequences. The superiority of the wavelet-based estimator (HEST) is obvious.

For the second set of simulations, we considered a fractional Brownian motion (fBm) input process, containing 10000 samples with the following values of $H$: 0.5, 0.6, 0.7, 0.8 and 0.9. The goal of the second set of simulations is to
evaluate the performance of the Hurst parameter estimators implemented in SELFIS and Matlab, respectively, for time series with known value of $H$, as similar as possible with the time series composing our data base. In the second set of simulations, we have not analysed the influence of the length of the input signal on the bias of the Hurst parameter estimators as in the first set of simulations, we have used this time a fixed length of 10000 samples, appropriate for the accurate generation of the fBm input signal. The results are shown in Table 2.

The results presented in Table 2 prove that the precision of Hurst parameter estimators implemented in SELFIS depends on $H$ values. For small values of $H$, the R/S estimator is the best. Indeed, for $H=0.5$ (the signal fBm05), the value estimated by using the R/S estimator is 0.463, closer to 0.5 than the values obtained by using the other estimators from SELFIS: aggregate variance, periodogram and absolute value. For intermediate values of $H$ (fBm07), the absolute value estimator is the best one. For high values of $H$, the best estimator used by SELFIS is the aggregate variance. Once again, the results obtained by applying the DAV estimator are better than the results obtained by using estimators implemented in SELFIS.

The two types of simulations already analysed recommend the use of the DAV estimator for $H$ parameter estimation. These results show that the DAV estimator is practically not polarized, robust and efficient and recommend the use of this estimator for the estimation of daily WiMAX traffic $H$ parameter. The appropriateness of this estimator for the LRD analysis of traces from our database can be explained adding two supplementary reasons. The first reason is the discrete nature of the traces from our database (the number of packets of data is acquired every 15 min), because the DAV estimator was conceived for discrete events. The second reason is the non-stationary nature of the traces from our data base. These traces are discrete, non-stationary, random processes. The source of non-stationarity is the traffic overall tendency (Stolojescu et al., 2009). This non-stationarity does not affect the precision of the estimation made by the DAV estimator.

<table>
<thead>
<tr>
<th>Table 1: WGN input process</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of samples</td>
</tr>
<tr>
<td>100 000</td>
</tr>
<tr>
<td>10 000</td>
</tr>
<tr>
<td>4096</td>
</tr>
<tr>
<td>1024</td>
</tr>
</tbody>
</table>

5.3. Sources of LRD in communication networks

In the literature, several possible origins for LRD in communication networks are commonly cited (Veres & Boda, 2000; Clegg, 2004; Mistra, 2004; Gong et al., 2005). One of them is given by the hidden periodicities in the time series (Stolojescu et al., 2011).

The existence of video traffic coded with variable-bit rate induces LRD (Clegg, 2004). In this case, the LRD arises from the encoding mechanism, whereby video is encoded as a series of differences between frames with occasional full updates.

 Aggregate traffic is made up of many connections, which arrive randomly. Each connection is characterized by its ‘size’, representing the number of packets, and by the ‘rate’ of transmitted packets. The distributions of connections have very long tails. The LRD results by the aggregation of heavy-tailed data streams (Mistra, 2004).

Another potential cause of LRD is the feedback mechanisms in the Transmission Control Protocol (TCP). Let us consider the transmission of a packet between a sender–receiver pair on a network. The data are sent usually according to a reliable transport protocol, like TCP. The release of packets to the network is decided by the flow and congestion control mechanism. The TCP traffic was simulated by using a Markov model in (Gong et al., 2005). The conclusion of this simulation was that the multiple time-scale nature of traffic generation, coupled with transport protocol issues, makes the appearance of LRD-like behaviour inevitable. ‘TCP congestion control creates self similar traffic (…) showing both short-range and long-range dependence depending on system parameters’ (Veres & Boda, 2000).

The LRD also arises from network topology or routing algorithms (Clegg, 2004). The traffic in a network with incorrect topology is heavier than the traffic in a network with correct topology because the quality of the communication channels established in the network with incorrect topology is inferior to the quality of communication channels established in the network with correct topology. Even the efficiency of the routing algorithms decreases with the decreasing of the quality of communication channels. The references already cited concern wireline communication networks. The principal difference between wireline and wireless technologies is given by the ad hoc nature of wireless networks. The number of users in a cell of a wireless network, administrated by a BS, varies in time, increasing the dynamic of its traffic, as compared with the dynamic of the traffic in a server, in a wireline network. Another difference between wireline and wireless technologies is given by the architecture of the first layer of the stacks of protocols used. The TCP/IP protocols stack is used in the case of wireline networks. The architecture of the first layer (the link layer) is different in the case of wireless networks. The protocols belonging to the link layer of a wireless network are designed to enhance the communication adaptability to the transmission environment parameters. The dynamic of the traffic in a wireless network is higher than in a wireline network. The WiMAX technology increases the dynamic (variance) of the traffic, as compared with wireline technology. For these reasons, we suppose that the LRD is present in wireless (WiMAX) traffic as well. The increased dynamic of the WiMAX traffic makes possible the alternation of traffic.
sequences with different degrees of LRD. A sequence of WiMAX traffic with a very low value of $H$ can be followed by another sequence of WiMAX traffic with a very high value of $H$. So, the results of the LRD analysis of traffic depend on the adopted transmission technology.

In the following sections, we will highlight the existence of the same LRD sources in the case of wireless networks, as in the case of wireline networks.

6. Results and discussions

In the following, we will present the implementation of the data mining methodology proposed for the identification of the BS incorrectly positioned.

After the understanding of the initial data collection, and after data cleaning and preprocessing, in the third step, we have decided that the Hurst parameter of the time series could be a useful feature for the estimation of the heaviness of traffic for each BS. As it was mentioned in Section 5.3, there are multiple sources of LRD in communication traffic: hidden periodicities, video traffic, feedback mechanisms of TCP and the network topology (the incorrect positioning of BS).

One strategy is to separate the last LRD source from the others. To do this, first, we separated the LRD caused by the hidden periodicities by segmenting each time series in sequences with the duration of 1 week and 1 day, respectively. We have observed that the values of the Hurst parameter for time series with a shorter length are smaller than the values of the Hurst parameter for time series with greater length. Hence, the effect of segmentation is to eliminate the source of LRD caused by hidden periodicities, with periods higher than the length of the segments. So, the daily time series does not contain LRD produced by hidden periodicities, and the first source of LRD was eliminated. For the reduction of the following two sources of LRD, the video traffic and the feedback mechanisms in the TCP protocol, we have compared the Hurst parameter of daily series in uplink and downlink. These two sources of LRD can be neglected in uplink, because video traffic is produced mainly by down-streaming applications and the volume of traffic is smaller in uplink than in downlink. Also, the feedback mechanisms in the TCP protocol manifest less frequently in uplink, because they depend on the volume of traffic, which is smaller in uplink. Our goal is to identify the BS which are incorrectly positioned. The incorrect positions of the BS represent a source of LRD common for both downlink and uplink transmissions. So, we have decided to use the values of the Hurst parameter obtained for daily uplink traffic time series for the classification of BS positioning, as input of the fourth step of the proposed data mining methodology. The goal of this step is to create the final dataset. To simplify the structure of the final dataset, this step consists of the binarization of the estimated values of the Hurst parameter. The value 0 indicates the absence of LRD in the traffic, and the value 1 indicates the presence of LRD. The last step of the proposed data mining methodology consists in the classification of the BS into two classes: correctly positioned and incorrectly positioned. We have used, as classification criterion, the frequency of apparition of daily LRD caused by the incorrect positioning of the considered BS. If a BS has a low frequency of apparition of LRD (inferior to 0.5), it was considered as correctly positioned. We have eliminated BS34 from the following analysis of the experimental results, obtained by applying the proposed method, because its trace contains many missing values.

6.1. Daily downlink traffic

This section presents the evaluation of $H$ in the case of WiMAX daily downlink traffic. Our goal is to identify the normal behaviour of the downlink traffic. The idea is to separate the BS with normal behaviour from the LRD perspective. We will consider that a day has LRD if the corresponding value of $H$ is bigger than 0.5. Indeed, downlink daily traffic exhibits LRD. On the basis of the obtained results, we realized the BS classification in terms of the number of days for which the downlink traffic exhibits LRD (number of values $H$ greater than 0.5). The results are shown in Table 3.

We can observe the extreme cases: BS63 with only 14 days with LRD traffic (the best case) and BS32 with 42 days (from a total of 56 days) with LRD traffic (the worst case).

6.2. Daily uplink traffic

We have analysed all the 66 traces of uplink traffic, using the DAV estimator, for each of the days of the 8 weeks. The uplink daily traffic exhibits LRD as well. In Table 4, we realized the classification of the 66 BS in terms of the number of days for which the uplink traffic exhibits LRD (number of values $H$ greater than 0.5).

The worst case is given by BS32, for which the number of days with LRD traffic is greater than the number of days

<table>
<thead>
<tr>
<th>Table 3: Base stations (BS) classification in downlink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of values $H &gt; 0.5$</td>
</tr>
<tr>
<td>------------------------------</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>36</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>34</td>
</tr>
<tr>
<td>33</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>31</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>27</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>14</td>
</tr>
</tbody>
</table>
without LRD traffic (44 days with LRD traffic). The BS with the smallest number of days for which the traffic manifests LRD are BS49 and BS63 (12 days with LRD traffic).

6.3. BS localization analysis in uplink and downlink

The goal of this section is the comparison between the daily uplink and downlink traffic for each BS.

This comparison could be made with the aid of Tables 3 and 4. Considering for example the BS12, it can be observed that it has more days with LRD in downlink (36 in Table 3) than in uplink (28 in Table 4). The same observation is valuable for the majority of the BS, which constitutes the considered network. So, the hypothesis that it is more LRD in downlink than in uplink is confirmed experimentally. We preferred to make the comparison with the aid of a ‘LRD map’ of the considered network, represented in Figure 1. The days without LRD are represented in white, the days with LRD in downlink are marked with grey, the days with LRD both in uplink and downlink are marked with black. The BS are arranged on the lines of Figure 1 and the days on the columns. For example, counting the grey squares and the black squares in Figure 1, we can build Table 3. The advantage of the representation in Figure 1 on the representation in Tables 3 or 4 is that we can compare the LRD component of different BS in the same day.

The number of days without LRD is greater than the number of days with LRD, both in uplink and in downlink, for a number of 22 BS. These BS are the following: BS7, BS49, BS61, BS3, BS4, BS8, BS9, BS16, BS17, BS46, BS48, BS51, BS53, BS58, BS59, BS60, BS62, BS63, BS64, BS65 and BS67. The normal behaviour of those BS is without LRD. Hence, they are correctly positioned.

The downlink traffic contains more days with LRD than the uplink traffic for 53 BS. So, the normal behaviour of one BS supposes more LRD in downlink than in uplink. As already mentioned, the downlink traffic contains more days with LRD than the uplink traffic for 53, from a total of 66 BS. This confirms the hypothesis that the best separation of LRD caused by the incorrect positioning of a BS is obtained in uplink. So, we can use the classification in Table 4. After the elimination of the BS for which the number of days without LRD is greater than the number of days with LRD both in uplink and in downlink, already mentioned, we find in Table 4 more 23 BS having a number of days with LRD smaller or equal than the total number of days (which is equal with 28). These BS are the following: BS19, BS52, BS2, BS5, BS10, BS11, BS20, BS22, BS27, BS59, BS60, BS61, BS62, BS63, BS64, BS65, BS66, BS67 and BS68. We can consider that they are correctly positioned as well.

For 15 BS, there are more days with LRD in uplink than days without LRD, in Table 4. These BS are the following: BS6, BS13, BS15, BS24, BS25, BS31, BS30, BS33, BS36, BS38, BS44 and BS50. We will consider these BS as incorrectly positioned.

The traffic of other six BS has an atypical behaviour, the number of days with LRD in uplink being greater than the number of days with LRD in downlink. These BS are following: BS23, BS26, BS32, BS35, BS41 and especially BS1. We consider that these BS could be repositioned at the next network release, which will take place at the next maintenance session. The BS are initially positioned by the manufacturer of the WiMAX network, when the network installation takes place. They use a procedure that combines measurements with simulations, performed by using special simulators. The performance of this positioning procedure can be improved when the next maintenance session will take place, by reprogramming the simulators, taking into account the results of the proposed LRD analysis.
The goal of this paper was to analyse the positioning of the BS in the architecture of a WiMAX network using the LRD analysis. In the case of uplink traffic, the LRD is influenced by the presence of some hidden periodicities in the time series. That is, the entire series of 8 weeks exhibits stronger LRD than each weekly series (there is a hidden periodicity with the period of a week and another with the period of a day).

We have inferred in the presentation of the proposed approach, in the introduction of Section 6, that the two sources of LRD given by the video traffic and the feedback mechanisms of TCP can be neglected in the uplink traffic. Then, we have estimated the Hurst parameter of the daily time series, for each BS, using the DAV estimator. Using these values, we have observed that the downlink traffic contains more days with LRD than the uplink traffic for 53 of 66 BS. Hence, the normal behaviour of BS supposes a stronger LRD in downlink than in uplink.

The classification of BS for the daily uplink channel was made in terms of the number of days for which the traffic is LRD, using the DAV estimator. On the basis of this classification, the conclusion is that the normal behaviour of a BS corresponds to a number of days with LRD in uplink smaller than the number of days with LRD in downlink.

The results obtained for the daily traffic in uplink and downlink were compared, and the following conclusions were obtained. A number of 22 BS have more days without LRD than with LRD, both in downlink and in uplink. These BS can be considered as correctly positioned. For them, our LRD analysis presents a high level of confidence. For other 25 BS, the number of days for which the traffic presents LRD in uplink is smaller than the number of days without LRD. These BS are considered as correctly positioned as well. For these BS, the LRD analysis presents lower level of confidence than the LRD analysis of the first 22 BS already mentioned. The traffic of six BS has an atypically behaviour.

The limitations of the approach proposed in this paper are the following: the imperfect separation of the sources of LRD of traffic, the atypical comportment of some BS, the incertitude of decision in the case of some other BS and the variability of the decisions degree of confidence.

The approach based on LRD traffic analysis, proposed in this paper, could be useful for the enhancement of other functions of WiMAX network as well. For example, the value of $H$ parameter could be used for traffic anomalies detection, increasing the network security. The proposed approach, based on data mining methodology, requires a reduced volume of computations and is fast. The system that implements this approach can be installed ‘on place’, requiring only minor modifications of the network structure.

The identification of BS incorrectly positioned in a WiMAX network has an important ‘business’ impact on the functioning of the corresponding network. After the correction of the identified BS positions, the overall traffic of the network will be more uniform, all the BS being exploited similarly. As a consequence, the network will be able to deserve an increased number of users. The results presented in this paper can be integrated in the BS positioning procedure established by the installer of the WiMAX network, increasing the precision and decreasing the costs of this procedure.

As future work, we can mention the following directions: the application of the LRD analysis on data of different nature and the combination of the LRD analysis with other kinds of statistical analysis procedures.

Acknowledgements

The research reported in this paper was developed in the framework of a grant of the Romanian Research Council (CNCSIS) with the title ‘Using Wavelets Theory for Decision Making’ no. 349/13.01.09. The authors thank Alcater Lucent, Timisoara, for providing the WiMAX traffic database and for helpful discussions around WiMAX technology. We also acknowledge the quality of the work done by the anonymous reviewers.

References


References


The authors

Cristina Stolojescu-Crisan

Cristina Stolojescu-Crisan was born in Timisoara, Romania in 1983. She received the Engineering degree in electronics and telecommunications from the University ‘Politehnica’ of Timisoara, Romania, in 2007 and the PhD degree from the University ‘Politehnica’ of Timisoara, Romania and Telecom Bretagne, Brest, France in January 2012. She is currently an Assistant Professor at the Electronics and Telecommunications Faculty, ‘Politehnica’ University of Timisoara. She is the author of 20 articles in conference proceedings and journals. Her current research interests include signal processing, biomedical imaging, data mining and wavelets.

Alexandru Isar

Alexandru Isar was born in Lugoj, Romania, in 1957. He received the Engineering degree in electronics and communications and the PhD degree from the University ‘Politehnica’ of Timisoara, Romania, in 1979 and 1993, respectively. He is currently a Professor with the Communications Department, Electronics and Telecommunications Faculty, ‘Politehnica’ University of Timisoara. He is the author of about 70 articles in international journals and conference proceedings. He is a Reviewer for the Journal of Electronic Imaging, IEEE Transactions on Instrumentation and Measurements and Signal Processing. His current research interests include wavelets and signal and image processing.

Sorin Moga

Sorin Moga was born in Targu-Mures, Romania, Romania, in 1971. He received the BS degree in electronics and telecommunications from the University ‘Politehnica’ of Timisoara, in 1995 and the MSc degree in image and signal processing, and the PhD degree both from the University of Cergy-Pontoise, Cergy, France, in 1996 and 2000, respectively. He is currently an Associate Professor at Institut Telecom/Telecom Bretagne, France, and member of the laboratory CNRS LabSTICC UMR 6285. His current research interests focus on artificial intelligence and data mining fields.

Philippe Lenca

Philippe Lenca received the MS degrees in artificial intelligence, pattern recognition and applied informatics from the Universities of Paris V and VI, and the PhD degree in 1997 from University of Rennes I. He is currently Professor at Institut Telecom/Telecom Bretagne, France, and member of the laboratory CNRS LabSTICC UMR 6285, where he co-leads the DECision aid and knowleDge discovEry team. His current research interests focus on artificial intelligence, decision aiding and data mining.