A Signal Strength Fluctuation Prediction Model Based on the Random Forest Algorithm

P. Dohnalek[2], M. Dvorsky[1], P. Gajdos[2], L. Michalek[1], R. Sebesta[1], M. Voznak[1]
[1]*Department of Telecommunications, FEECS, VSB-Technical University of Ostrava, 17. listopadu 15, Ostrava, Czech Republic*
[2]*Department of Computer Science, FEECS and IT4Innovations, VSB-Technical University of Ostrava, 17. listopadu 15, Ostrava, Czech Republic*
pavel.dohnalek@vsb.cz

***Abstract*—This article describes the effect of the weather on radio wave propagation in a mobile telecommunication network. The research is focused on urban and countryside environments where a correlation between the received signal power level and weather conditions is found using the Random Forest algorithm as a signal level approximator. The results achieved in this paper clearly indicate that it is possible to predict the behaviour of the received power level in relationship to atmospheric phenomena.**

***Index Terms*–GSM, radioclimathology, prediction, adaptive network-based fuzzy inference system, random forest.**

I. INTRODUCTION

The Global System for Mobile Communication (GSM) network is part of the outer environment that people use every day via mobile phones, when browsing the internet or receiving data from distant devices such as flood control sensors. The network uses the air as a transfer medium. To be able to describe the reliability of a GSM service, one needs to distinguish which weather attributes affect the propagation of a GSM signal most. Then it is possible to predict significant changes in the propagation of the signal such as the Receive Level on the side of a mobile phone. On the GSM side of our system, the power level of the received signal in idle mode has been detected. The crucial part of our work is to detect dependencies between the received level and meteorological conditions.

Mobile network providers need to plan and optimize their networks to find a compromise between the radio coverage in different types of environment, the quality of services and the costs associated with construction and operation of the network. Received level deterioration in the ultrahigh frequency (UHF) band affects multipath propagation, reflection and path loss. In digital radio communications, these deteriorations cause errors and affect the quality of communication. Moreover, the magnitude of these degradations varies with time. Therefore, it could be significantly interesting and scientifically useful to predict path loss and use the output of the prediction as the input for coverage planning, mobile radio network optimization and radio link adaptation [1].

Cell sizes depend on many factors, such as terrain profile, type of environment (urban, suburban and countryside) and transmission parameters (such as transmitted power) of the base stations. One of the factors which also affect the cell size (the radio coverage) is weather conditions such as rain [2], [3] or sand and dust storms [4]. Measurements taken during the rainy season led to the modification of the Okumura Hata wave propagation model [5]. The effect of wet ground on radio propagation has been studied in [6]. The factors affecting path loss due to rain and snow have been implemented into the propagation model [7]. Some practical improvements in the existing models for macrocells, microcells and signal prediction in the indoor environment, as well as some new models, have been studied in [8]. The behaviour of electromagnetic waves propagating through densely arboreous environments [9] and through sparsely arboreous areas has been studied in [10]. The effect of the environment and altitude on the UHF band has been studied in [11]. In [12] a radio link was augmented by a radio channel state prediction method. In [13] we presented a K–means method that was used to decide which parameter related to weather affects the propagation of radio waves in a mobile telecommunication network.

This research is focused on proposing a way to approximate the signal strength based on its analysis with regards to possible effects by atmospheric phenomena such as humidity and temperature. The secondary contribution would be to create such an approximation for urban and countryside environments, including possible comparison.

II. DATA ACQUISITION

The data which are analysed consist of two subsets. The first subset is related to the identifiers associated with GSM technology while the second subset is related to the identifiers associated with weather. The data related to GSM are parameter values that are transferred between the mobile station (MS) and the mobile network through a signalling

Manuscript received October 21, 2013; accepted December 17, 2013.

The research leading to these results received funding from the grant of SGS No. SP2014/72 VSB-Technical University of Ostrava, Czech Republic, and was partially supported by the European Regional Development Fund in the IT4 Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and by the Development of human resources in research and development of the latest soft computing methods and their application in practice project (CZ.1.07/2.3.00/20.0072) funded by the Operational Programme Education for Competitiveness, co-financed by ESF and the state budget of the Czech Republic.

channel. A set of these parameters is called a Measurement Report (MR). An MR includes, among others, parameters such as *"Receive Level, Absolute Radio Frequency Channel Number"* (ARFCN) and *"Base Station Identity Code"* (BSIC) [14]. The parameters are acquired by a GSM modem and stored on a computer at periodic intervals. An application that synchronously stores the current values of GSM parameters and current values related to weather has been developed. The data related to weather were acquired by two professional meteorological stations located in both urban (Poruba district of the city of Ostrava) and countryside (the village of Bukovec) environments located in The Czech Republic. Given the slightly different conditions of the stations in each location, the parameters measured vary slightly. Parameters common for both locations are *"date, time, temperature, humidity, dew point, wind direction, wind speed, wind chill, barometric pressure, heat index"* and *"precipitation per hour"*. Parameters specific for the Poruba station are *"THW index, falling"* and the time of *"sunrise"* and *"sunset"* while the sole Bukovec-specific parameter is the *"pressure development"* over 24 hours. The *"THW index"* uses humidity, temperature and wind to calculate a seeming temperature that incorporates the cooling effects of the wind into our perception of temperature. *"Falling"* represents the trend of barometric pressure.

III. PREDICTION MODEL

In order to approximate the power level of the BTS, a prediction model needs to be constructed. In this paper, the Random Forest algorithm that groups Classification and Regression Trees (CARTs) into a majority voting classifier was used.

A. Classification and Regression Tree

Originally proposed by Breiman in [15], CART is a tree structure capable of solving both classification and regression problems. Basically, a CART is a binary tree that uses a set of yes/no questions to construct its nodes by splitting an observation into two parts that are as homogenous as possible and then repeating the process for each resulting part until complete decomposition of the observation is achieved. The classification and regression modes of the CART use different algorithms that govern the splitting.

Common for both nodes is the use of an *impurity function* to evaluate the homogeneity of the split [16]. Given that the impurity of a parent node is always constant when computing its child nodes, the impurity of the child nodes is computed as a *change* in impurity of the computed node, $\Delta i(t)$. Let t_p be a node that is the parent to its left node t_l and right node t_r and P_l and P_r be the probabilities of the respective nodes. The change in impurity can then be computed as

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r). \quad (1)$$

Annotating an observation variable x_j and the best splitting value of the variable x_j^R, where $j = 1, \ldots, M$ and M is the number of variables, CART can be observed as solving the following maximization problem in each of its nodes

$$\arg\max_{x_j \le x_j^R} \left[i(t_p) - P_l i(t_l) - P_r i(t_r) \right]. \quad (2)$$

The difference between the classification and regression mode of the CART is in the way the impurity function is calculated. In the classification mode, the most commonly used impurity function follows this formula, known as the Gini splitting rule or Gini index

$$i(t) = \sum_{k \ne l} p(k|t)\, p(l|t), \quad (3)$$

where $k, l = 1, \ldots K$, K is the number of classes and $p(k|t)$ is the conditional probability of class k given that the current node is node t. Incorporating the Gini rule, CART solves the following maximization problem

$$\arg\max_{x_j \le x_j^R} [-\sum_{k=1}^{K} p^2(k|t_p) + P_l \sum_{k=1}^{K} p^2(k|t_l) + P_r \sum_{k=1}^{K} p^2(k|t_r)]. \quad (4)$$

As regression trees have no classes, their output is a response vector Y containing response values for each observation. The Gini rule cannot be applied due to the absence of classes; instead a *squared minimization residual* problem is solved. Defined as

$$\arg\min_{x_j \le x_j^R} \left[P_l Var(Y_l) + P_r Var(Y_r) \right], \quad (5)$$

where $Var(Y)$ is the response vector of the corresponding child node, the problem attempts to minimize the expected sum variance for the two resulting nodes. Here, $x_j \le x_j^R, j = 1, \ldots, M$ is the optimal splitting question capable of satisfying the above formula.

While the Gini rule cannot be applied directly, it can quite easily be adapted for the regression mode. Let objects of class k be assigned the value of 1 and objects of other classes the value of 0. The variance of these values would be

$$p(k|t)\left[1 - p(k|t)\right]. \quad (6)$$

Summation over the number of classes then yields the impurity function commonly used in regression trees

$$i(t) = 1 - \sum_{k=1}^{K} p^2(k|t). \quad (7)$$

Once a tree is constructed using these rules appropriately, it can be very large, especially in regression mode. Therefore, pruning is employed to reduce the tree size to the desired value. Once pruning is complete, the tree is

ready for use in classification and regression problems.

B. Random Forest

Random Forest (RF) could be considered a majority voting classification and regression method as it combines a number of CARTs into a larger structure. For each tree in the forest, a random combination of a predefined number of input parameters is chosen and used to construct it. Testing samples are then evaluated against conditions in each node and propagated throughout the tree. When the sample reaches a leaf node, it is then assigned the class or value to which the samples in that node belong. In the forest, this is performed by all trees, providing a response from each of them. The testing sample is then assigned the class that was suggested by the majority of trees. Commonly, a binary tree with logical conditions is used, as was the case in this paper. Given that an RF consists of CARTs, it is capable of working in two modes, *classification* and *regression*, depending on the task it is expected to solve.

IV. EXPERIMENTS

The following section provides a description of the data used to train and evaluate the approximator, of the approximator itself as well as the experiments performed and results measured.

A. Data Preparation

Before the data could be used as an approximator's input, optimization procedures had to be performed. The parameter set contains the Receive Level values from one service station and up to 6 neighbouring Base Transceiver Stations (BTS). These values are continually sorted from the maximum to minimum value. Unfortunately, since the value of the Receive Level is affected by fast Rayleigh's fading, the position of the cell related to the first one or one of the neighbours varied. Therefore, it was necessary to select the cell which is identified by ARFCN.

Once selected, a data matrix was assembled from the collected data for each of the selected BTSs in the common row-sample, column-value format. To construct the matrix, dates and times were parsed into 3 numeric values each, and text parameters (like *wind direction*) were converted into integer numbers. Overall, the Poruba meteorological station provided 21 separate parameters against 16 input values per sample from Bukovec. For each BTS, a different number of samples were measured, however, every BTS had an abundance of data to perform the experiments (millions of samples).

As mention in the previous subsection, the RF algorithm requires training and therefore a training set. For this purpose, a small portion of data, *60000* samples, was separated. This was the lowest number of training samples that provided the maximum accuracy. Increasing the number of training samples had no effect on the accuracy while after decreasing it the effect was detrimental. The training samples were chosen randomly with normal distribution. The rest were then used to evaluate the Random Forest method.

B. Experimental Settings

Overall, 6 different cells were used for the evaluation, two for the Poruba location with CellID *75F4* and *76B3* and four for Bukovec with CellID *A863*, *A864*, *A865* and *AEDD*. To measure the RF performance, the Root Mean Squared Error (RMSE) between the expected and resulting output was used. The approximation accuracy was measured for two different approximation scenarios. The first scenario, *scenario1*, allows RF trees to choose from all possible input parameters. The second scenario, *scenario2*, limits the number of parameters to a mere three, each representing a water-related weather attribute – temperature, humidity and dew point. This scenario is a natural reaction to the empirical observation suggesting the parameters that most closely correlated to the BTS power levels are related to water.

The settings used for the experiments were as follows: *6* randomly selected parameters per tree and *1000* trees in the forest for *scenario1*, *2* parameters and 3 trees for *scenario2*, *regression* mode. While the number of trees for *scenario2* seems low, 2 parameters out of 3 can only be combined 3 different ways. Thus, more trees would result in redundant identical trees in the forest which could, according to the majority voting principle of the algorithm, influence the performance in favour of one of the combinations. In *scenario1*, the parameters can be combined in thousands of different ways.

C. Results

Given the random nature of choosing parameters during RF tree constructions, the experiments for each BTS were performed in 10 trials and the results were averaged. Each trial was made with a different, randomly chosen training set. Table I shows the resulting RMSEs for *scenario1*. It can be seen that while RF is a random algorithm that, given the number of possible input parameters in *scenario1*, cannot cover all possible combinations, it is quite robust in its performance. The particular results in individual trials vary only insignificantly.

TABLE I. RESULTING RMSES FOR THE FIRST SCENARIO.

	Cells for Bukovec				**Cells for Poruba**	
Trial	**A863**	**A864**	**A865**	**AEDD**	**75F4**	**76B3**
1	1.2415	1.2485	0.8525	1.3392	1.8638	1.9980
2	1.2415	1.2485	0.8559	1.3453	1.8668	1.9985
3	1.2431	1.2512	0.8536	1.3402	1.8576	1.9965
4	1.2426	1.2489	0.8545	1.3443	1.8639	1.9947
5	1.2429	1.2485	0.8541	1.3457	1.8601	1.9978
6	1.2432	1.2487	0.8540	1.3441	1.8555	1.9970
7	1.2413	1.2479	0.8535	1.3411	1.8647	1.9983
8	1.2422	1.2489	0.8535	1.3436	1.8613	1.9983
9	1.2415	1.2485	0.8533	1.3376	1.8639	2.0010
10	1.2426	1.2487	0.8534	1.3439	1.8625	1.9961
Average	**1.2423**	**1.2488**	**0.8538**	**1.3425**	**1.8620**	**1.9976**
Deviation	7.21e-4	8.94e-4	9.18e-4	2.77e-3	3.45e-3	1.69e-3

Table II presents the results for *scenario2*, where the average approximation RMSE was, as expected given the much lower number of input parameters, higher.

Aside from the absolute and average RMSE of each BTS, both tables also show the standard deviation of the results. The deviation across trials is miniscule and therefore insignificant, showing the robustness of Random Forest when trained using different samples.

Figure 1 illustrates Random Forest's capability of fitting its output to the expected values. The horizontal axis is not important as it only shows the index of a sample; the vertical axis, however, expresses the output parameter, the received power level. The light-coloured polyline illustrates the

expected value while the darker polyline is the approximation of 200 randomly chosen test samples.

TABLE II. RESULTING RMSES FOR THE SECOND SCENARIO.

	Cells for Bukovec				Cells for Poruba	
Trial	**A863**	**A864**	**A865**	**AEDD**	**75F4**	**76B3**
1	2.7929	2.7522	2.1431	2.6781	3.5231	3.6464
2	2.7943	2.7527	2.1432	2.6867	3.5393	3.6368
3	2.7878	2.7541	2.1455	2.6825	3.5262	3.6340
4	2.7919	2.7543	2.1357	2.6817	3.5218	3.6293
5	2.7984	2.7568	2.1428	2.6842	3.5207	3.6359
6	2.7889	2.7496	2.1445	2.6869	3.5097	3.6356
7	2.7935	2.7551	2.1399	2.6846	3.5204	3.6399
8	2.8020	2.7461	2.1401	2.6834	3.5210	3.6473
9	2.7930	2.7568	2.1344	2.6816	3.5206	3.6436
10	2.7898	2.7609	2.1397	2.6834	3.5250	3.6380
Average	**2.7932**	**2.7539**	**2.1409**	**2.6833**	**3.5228**	**3.6387**
Deviation	4.31e-3	4.11e-3	3.66e-3	2.59e-3	7.31e-3	5.69e-3

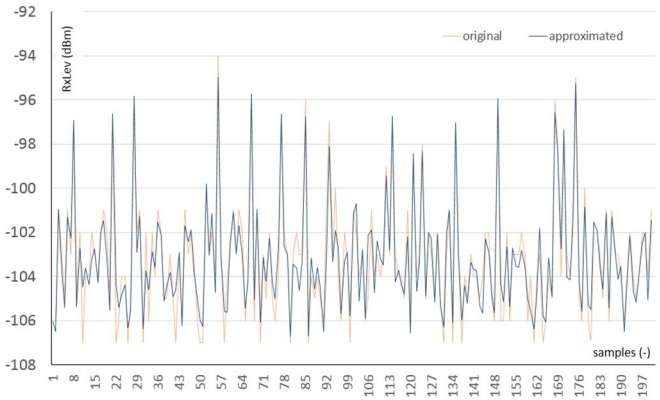

Fig. 1. An illustration of the Random Forest results fitting the expected values.

The results proved that it was more difficult to approximate the power level in the urban rather than in the countryside environment. A sparsely populated countryside environment indicates different propagation characteristics than a densely populated urban environment. Countryside path loss values are lower than urban path loss values because the countryside areas are composed of open land with small buildings and plain areas. Moreover, in countryside and open areas, the range of slow fading is lower than that in suburban and urban areas. These facts contribute to a more complicated and difficult approximation, as shown above.

V. CONCLUSIONS

Based on the results in this paper, we are able to predict the behaviour of the received power level affected by atmospheric phenomena. The average RMSE proved that the proposed model can be applicable in a method for increasing the efficiency of power consumption for a base station. It is well known that the main source of energy consumption in cellular mobile networks is the base station. Therefore, reducing the energy consumption of BTSs as the main energy consumers is extremely important. The research in this paper suggests that the radiated power level of BTS can be adjusted while considering the atmospheric conditions, which leads to an improvement in the energy efficiency of the mobile radio network.

In their further work, the research team are to focus on the same issues in different frequency bands related, for example, to Digital Audio/Video Broadcasting technology or the new generation of mobile network such as Long Term Evolution (LTE). Formalizing and mathematically describing the relationship between the received power level and mentioned weather parameters, implementing the researched technique into a hardware solution, combining different kinds of input data or developing whole new approaches to solve the problem, possibly based on hybrid paradigms, are other topics of both future and currently ongoing research.

ACKNOWLEDGMENT

The research results were also achieved thanks to the *Institute of Geoinformatics, Faculty of Mining and Geology, VSB-Technical University of Ostrava,* which provided the meteorological data.

REFERENCES

[1] P. Ziacik, V. Wieser, “Mobile radio link adaptation by radio channel state prediction”, *Elektronika ir Elektrotechnika*, no. 8, pp. 27–30, 2011. [Online]. Available: http://dx.doi.org/10.5755/j01.eee.114.8.689

[2] D. A. Shalangwa, A. Abdulrazak, G. Jerome, “Influence of atmospheric parameters on global system mobile communication (GSM) outgoing calls quality in Mubi Adamawa State Nigeria”, *Journal of Mobile Communications*, vol. 3, pp. 56–58, 2009.

[3] A. Sharma, P. Jain, “Effects of rain on radio propagation in GSM”, *Int. Journal of Advanced Engineering & Applications*, 2010.

[4] E. M. Abuhdima, I. M. Saleh, “Effect of sand and dust storms on GSM coverage signal in southern Libya”, in *Proc. Electronic Devices, Systems and Applications (ICEDSA)*, pp. 264–268, 2010.

[5] J. C. Ogbulezie, M. U. Onuu, J. O. Ushie, B. E. Usibe, “Propagation models for GSM 900 and 1800 MHz for Port Harcourt and Enugu, Nigeria”, *Network and Communication Technologies*, vol. 2, no. 2, 2013.

[6] S. Helhel, S. Ozen, H. Goksu, “Investigation of GSM signal variation depending weather conditions”, *Progress in Electromagnetics Research B*, vol. 1, pp. 147–157, 2008. [Online]. Available: http://dx.doi.org/10.2528/PIERB07101503

[7] N. Sah, T. Thakur, “Impact of clutters on quality of service in mobile communication using Walfisch-lkegami propagation model”, in *Proc. IEEE Personal Wireless Communications (ICPWC)*, pp. 290–294, 2005.

[8] A. Neskovic, N. Neskovic, G. Paunovic, “Modern approaches in modeling of mobile radio systems propagation environment”, *IEEE Communications Surveys & Tutorials*, vol. 3, pp. 2–12, 2000. [Online]. Available: http://dx.doi.org/10.1109/COMST.2000.5340727

[9] J. C. Rodrigues, S. G. C. Fraiha, H. S. Gomes, G. P. S. Cavalcante, A. R. O. de Freitas, G. H. S. de Carvalho, “Channel propagation model for mobile network project in densely arboreous environments”, *Journal of Microwaves and Optoelectronics*, vol. 6, no. 1, pp. 236–248, 2007.

[10] M. Bitirgan, Z. E. Yoruk, S. Celik, O. Kurnaz, S. Helhel, S. Ozen, “Generation of an empiric propagation model for forest environment at GSM900/GSM1800/CDMA2100”, *General Assembly and Scientific Symposium XXXth URSI*, 2011, pp. 1–4.

[11] H. Elshafie, N. Fisal, Y. Baguda, H. Sayuti, Y. Abdulrahman, M. Hafizal, N. Ramli, M. Abbas, “Measurement of UHF signal propagation loss under different altitude in hilly environment”, *Applied Mechanics and Materials*, vol. 311, pp. 37–42, 2013. [Online]. Available: http://dx.doi.org/10.4028/www.scientific.net/AMM.311.37

[12] *Digital cellular telecommunications system (Phase 2+) - Mobile radio interface layer 3 specification*, ETSI Standard TS 100 940. 2003.

[13] J. Skapa, M. Dvorsky, L. Michalek, R. Sebesta, P. Blaha, “K-mean clustering and correlation analysis in recognition of weather impact on radio signal”, in *Proc. 35th Int. Conf. on Telecommunications and Signal Processing*, Prague, 2012, pp. 316–319.

[14] J.-S. Roger Jang, “ANFIS: adaptive-network-based fuzzy inference systems”, *IEEE Trans. Systems, Man, and Cybernetics*, vol. 23, pp. 665–685, 1993. [Online]. Available: http://dx.doi.org/10.1109/21.256541

[15] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and Regression Trees.* Chapman and Hall/CRC, 1984.

[16] R. Timofeev, “Classification and regression trees (CART) theory and applications”, M.S. thesis, Humboldt University, Berlin, 2004.