Measurements clustering for robustness improvement of indoor WLAN propagation models

Rida Tahri, Valery Guillet, Patrice Pajusco

Measurements Clustering for Robustness Improvement of Indoor WLAN Propagation Models

R. Tahri, V. Guillet and P. Pajusco
France Telecom, 6 Avenue des usines, Technopôle, 90000 Belfort

Abstract - For indoor Wireless Local Area Networks (WLANs) planning, an accurate propagation modelling is required. Semi-empirical models represent an efficient approach to the indoor coverage prediction. Their parameters are estimated from measurement results. In the case named ill-conditioned, the Least Square (LS) regression leads to a bad estimation of the model parameters and thus to a numerical instability. In this paper, we present an approach based on the K-means clustering method which allows us to increase the estimation stability. Clustering is used to select the measurement points locations in a strategic way for having a robust estimation.

Index Terms – K-means, Radio Channel Modelling, Indoor Propagation, WLAN.

I. INTRODUCTION

Over the last few years, the market for wireless service has grown at an unprecedented rate. The commercial success achieved by the introduction of cellular mobile radio phones has generated strong interest in the development of other wireless communications systems.
This development is being driven primarily by the transformation of what has been largely a medium for supporting voice telephony into a medium for supporting other services, such as the transmission of video, images, text, and data. Thus, the demand for new wireless capacity is growing. WLANs based on IEEE 802.11 popularly known as WiFi provide mobility and flexibility to users and data transmission theoretically up to 54 Mbps.

WLANs operate mainly in indoor environments. One of the most important characteristics of the radio waves propagation is their path loss. An accurate estimation of the propagation losses provides a good basis for a proper selection of the access points (APs) locations and a proper assignment of the radio frequencies. By knowing propagation losses, one can efficiently determine the field signal strength, signal-to-noise ratio (SNR), carrier-to-interference (C/I) ratio, etc.
However it is very difficult to predict how a radio wave travels in indoor. Generally we do not have any accurate information about the building materials and the furniture, their electromagnic properties are not a priori known. So there is a need for developing a propagation model to predict indoor radio coverage more accurately.
An accurate prediction of the field strength level is a very complex and difficult task. Up to now, various field strength prediction methods have been proposed in the literature [5].
Due to their simplicity, their computational efficiency, and their reliability, semi-empirical models are applied for WLAN indoor coverage predictions. They proved good performances with optimised estimated parameters. Taking into account the indoor environments diversity, such estimation requires the undertaking of extensive measurements campains for each studied environment. However, this is not always possible because of the high cost of the required experiments..
The multiple regression method is usually used in the least square (LS) sense to estimate the model parameters. One of the sever limitations of this approach is that it can suffer from the problem of collinearity [1,2]. Collinear measurement points lead to numerical instability in the model parameters estimation.
It is therefore necessary to define a process to estimate the model parameters using a limited number of measurement points overcoming the LS regression instability.
We explain in this paper how the use of the K-means clustering method allows selecting only a few measurement points' locations for having a robust estimation.
The paper is organised as follows. Section II provides the context of the study; it provides the formulation and the definition of all the coefficients which will be used in our statistical study. Section III shows the principle of the K-means clustering method and its implementation for selecting the tuning points. In section IV we show the experimental setup, the software and hardware used for measurements, and the propagation environment in which measurements were done. In section V we describe the statistical analysis used to proof the advantages of the K-means approach.

II. CONTEXT OF THE STUDY

A. Motley-Keenan model

Thanks to its performances (short computation time and relatively good accuracy), the Multi-Wall model [3-6] is applied for WLAN indoor coverage predictions. It proved good performances with optimised parameters. This is a semi-empirical approach based on fixing empirical parameters for the attenuation produced by the obstacles

present in the building. This path loss model is given by the following analytical formula:

$$L_{MK} = \beta + \sum_i k_i Att_i + \alpha \log_{10}(d) \qquad (1)$$

Where:

- L_{MK} is the predicted average path loss [dB]
- d is the distance between the transmitter and the receiver [m].
- k_i is the number of crossed walls of type i
- Att_i is the loss of wall of type i [dB]
- β is a constant loss [dB]
- α is the coefficient associated to the distance effect

Usually, the multiple regression method [4,5] (based on LS estimation) is used to predict α, β and the different attenuation parameters, from measurement data collected on the site under study.

B. Least Square method

The usual estimation procedure for the model parameters is the Least Square (LS) one. That is an unbiased estimation with minimum variance [2].
Indeed for n measurement points of path losses (L_i) and for p variables thought to affect all the measured values Li, the multiple regression model (1) that relates an individual value L_i of the path loss L to the p variables can be expressed by:

$$L_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i \qquad (2)$$

X_{ij} are the elements of the data matrix X (made up in our case by the logarithm of distances between the transmitter and the receiver and the number of crossed walls of type j). The parameters β_j, are the regression coefficients corresponding to the p variables; ε_i is the error or residual, and β_0 is a constant which can account for all other unconsidered variables.

The predicted value of the dependent variable L_i, can be expressed by:

$$\hat{L}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij} \qquad (3)$$

The LS procedure consists in minimising the model mean square error. This error is estimated as the difference between the measurement path loss values and those estimated by multiple regression using (3). An unbiased estimation of the root mean square error is given by [1,2]:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (L_i - \hat{L}_i)^2}{n - p} \qquad (4)$$

Let us write the equation (2) under a matrix form as a standard model for a multiple linear regression model.

$$Y = X\beta + \varepsilon \quad , \quad (X_{i,0} = 1 \ \forall \ i) \qquad (5)$$

Where Y is the measurement points vector.
We suppose that the noise vector ε is a multi-normal distributed random variable with standard deviation σ: ε~N(0, σ^2I). Under this hypothesis the estimator $\hat{\beta}$ is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (6)$$

And, the covariance matrix of $\hat{\beta}$ is given by:

$$\mathrm{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \qquad (7)$$

Where the elements on the main diagonal determine the variance of the estimator: $s^2(\hat{\beta})$.
The LS method provides the optimal estimates of the model parameters (unbiased one). Indeed we cannot find estimators with smaller variances (theorem of Gauss-Markov [1]).
The LS estimation requires the explicit calculation of the matrix $(X^T X)^{-1}$. This estimation procedure is a good one only if $X^T X$ is not nearly singular. When the matrix X used for tuning contains collinear data, the matrix $X^T X$ becomes ill-conditioned for inversion. The resulting estimator $\hat{\beta}$ has typically inflated variance and is unstable [1,2]. The model parameters estimates are thus highly sensitive to the tuning data, and may give poor predictions.
An important issue in regression analysis is assessing its overall quality. A statistical analysis of the root mean square errors (σ) is a useful way for evaluating the regression quality. The goodness of the regression analysis can also be assessed using the adjusted coefficient of determination R_a^2 [1,2] which expression is given by:

$$R_a^2 = 1 - \frac{\sum_{i=1}^{n} (L_i - \hat{L}_i)^2}{\sum_{i=1}^{n} (L_i - \overline{L})^2} \times \frac{n-1}{n-p} \qquad (8)$$

Where $\overline{L}$ is the mean value of L, and n the number of measurement points used for tuning $R_a^2 \in [0,1]$ measures the proportion of variation in L explained by a regression model. If R_a^2=1, all the variations have been explained by the model but if it is equal to zero, the regression model is useless.
Furthermore, for evaluating the robustness of the regression model, it is also necessary to check the significance of the partial regression coefficients $\hat{\beta}_j$. This is done by analysing

the variance $s^2(\hat{\beta})$, but also by means of the t-test [1,2]. The appropriate statistic of this test for the j^{th} variables, is:

$$t_j = \left| \frac{\hat{\beta}_j}{s(\hat{\beta}_j)} \right| \quad (9)$$

For having a good robustness, the computed t_j should be greater than the tabulated $t_{n-p,\alpha}$ for a α level of significance. More details about these statistical coefficients are given in [1,2,5]. In [6] we have made use of these statistical coefficients to study the influence of the tuning points' choice on the stability and accuracy of a regression model.

The model robustness highly depends on the measurement points chosen for tuning especially when we have the constraint to work with a small number of points (typically less than 50 points). As it is not obvious to define on a building map the measurement points' locations leading to an accurate modelling, we have searched to do that in a strategic way rather than using randomly chosen locations. In order to make a good tuning points' choice and to deal with the problem of collinearity, we present in the following the K-means clustering method principle. We explain how this method will be used in our context to choose non-collinear tuning measurements points allowing a robust estimate.

III CLUSTERING APPROACH

A. *Principle of the K-means algorithm*

The K-means clustering method [7-9] divides a set of points into groups, or clusters, such that points within each cluster are similar to one another (for a metric, usually the Euclidean distance), and points in different clusters are far from one another.

The iterative K-means algorithm begins with a set of k reference points whose initial values can be chosen by the user. The better choice is to place them as much as possible far away from each other.

All the data points are partitioned into k clusters by assigning each point to the cluster with the nearest reference point. Adjustments are made by calculating the centroid for each cluster and then using these new centroids as reference points for the next partitioning of all the data points.

For data point x in cluster i, if the centroid z_i is the nearest reference point, no adjustments are made and the algorithm proceeds to the next data point. However, if the centroid z_j of the cluster j is the reference point closest to data point x, then x is reassigned to cluster j, the centroids of the "losing" cluster i (minus point x) and the "gaining" cluster j (plus point x) are recomputed, and the reference points z_i and z_j are moved to their new centroids. After each step, every one of the k reference points is a centroid, or mean, hence the name "K-means." The k reference points change thus their location step by step until no more changes are done. In the next section we will see how the K-means approach will be used to select the tuning points for which the model parameters estimation is accurate and robust.

B. *K-means implementation for measurement points choice*

In our context, the selection of the measurement points' locations is made from simulation data. In fact the number of crossed walls and the logarithmic of distance between the transmitter and the potential receivers are deduced by tracing rays between transmitters and receivers points. The idea is to build a matrix X for a very large number of points distributed regularly in the simulated environment. This is done by using a ray tracing model [10]. Then the data matrix is standardized (in order to mitigate the problem of differences between the values of different data matrix elements). The next step is to cluster the X matrix points into a limited number k of clusters. k is defined by the user to take into account the tuning points' number sufficient for a good estimation.

Finally we select from each cluster the nearest point from its centroid. Indeed the K-means algorithm builds the clusters, such that centroids of each cluster are as far as possible from one another. We build thus a sample of points which have a maximum distance (and thus a minimum collinearity) between them. This solves the problem of the ill conditioning of the data matrix X.

The resulting points' locations given by the clustering approach are then used as the locations where measurements will be carried out, in order to tune the model in a given environment and to have an estimation of the model parameters with a good robustness. The advantage of this approach will be presented in the section (V).

IV MEASUREMENT SCENARIO DESCRIPTION

Field strength measurements were carried out in an office building in Belfort city. Measurements have been realized in the 2.4GHz frequency band using 802-11b/g equipments and a software tool developed by France Telecom for data acquisition.

The transmitters were placed at four locations within the building. In Fig 1 the four green points show the locations of the transmitters while the crosses represent the locations of some measurement points.

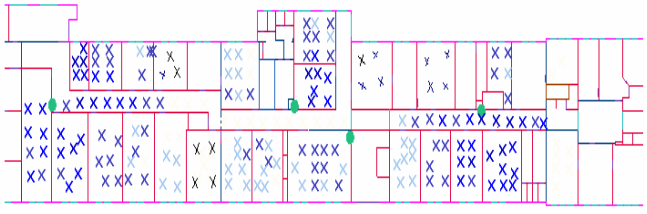

Figure 1: Location of the transmitters' antennas in the measurement environment ~ (100m×18.5m).

Both transmitters and receivers use an omnidirectionnal antenna. The transmitters are placed at a height of 2.5m near the ceiling (the ceiling height was about of 2.88 m). The locations of the reception points were at a height of 1.2m above the floor level. The measurements were processed in line of sight (LOS) and non line of sight (NLOS).
Internal walls were made of 7 cm thick plasterboard with 3 cm thick wooden doors. The outer walls are made of 44 cm thick concrete with double-glazing. Floor is made of concrete. Most of the furniture is made of agglomerated wood and metallic cupboards. The building has a rectangular shape, with a S-shaped corridor separating the different rooms.
In order to remove the fast fading effects, 35 field strength measurements, around each considered reception point, were performed and averaged. The whole measured points represent a set of 350 values.

V NUMERICAL RESULTS: ADVANTAGES OF THE K-MEANS METHOD

To investigate the K-means clustering method advantages we use the two following strategies. For the first one, we use K-means to cluster the 350 measurement points locations using the previous process (chap IV) into 50 clusters. From each cluster we select the nearest measurement point to the cluster centroid. We built thus a set of 50 points (designated by "K-fit sample") which have a minimum similarity between them. After withdrawing the "K-fit sample" measurement points from the "whole sample" we select a set of 50 measurement points randomly with uniform distribution among the 300 points remaining of the "whole sample". This set is designated by "R-fit sample". The model parameters and their variances estimates are then computed and compared using the "K-fit sample" and "R-fit samples". For validation, the RMSE and the adjusted coefficient of determination (R_a) (cf. Table 1) associated to these estimates are calculated for three "test samples" (approximately 100 points selected randomly with a uniform distribution among the 250 remaining points of the "whole sample").

Kind of fit sample	estimators variances sum [dB]	Test sample1		Test sample 2		Test sample 3	
		σ [dB]	Ra [%]	σ [dB]	Ra [%]	σ [dB]	Ra [%]
K-means	18.6	5.1	70	5.4	67	5.5	61
Random	28.6	6	60	6.3	55	6.4	50

Table 1: Overall and partial quality: Advantage of the K-means clustering method

By comparing the variances of the different model parameters, we can note a clear model stability improvement by using the "K-fit sample" tuning.

Although the number of points is the same, the choice of the points has a large influence on the accuracy and on the robustness of the model. Indeed, the model parameters' variances sum passes from 18.6 dB when we use the "K-fit sample" tuning to 28.6 dB when we use the "R-fit sample" one. On the other hand, the RMSE decreases (about 1 dB) and the R_a increases (about 10%) when we use the "K-fit sample" instead of the "R-fit sample" tuning.

A second strategy for the K-means method validation is described in the following. The model parameters are computed using the "K-fit sample". The RMSE (σ_i^K) between predicted path loss and measured path loss is now calculated for nine different "test samples". The index "K" corresponds to the "K-fit sample" when "i" corresponds to the i^{th} "test sample", $1 \le i \le 9$). Each one of the nine "test samples" contains approximately 50 points selected randomly from the "whole sample" after withdrawing the "K-fit sample" measurement points.
Using each “test sample”, we have tuned the model parameters (thus we have 9 corresponding models) and calculated the models RMSE (σ_i^R; $1 \le i \le 9$). Note that σ_i^R is by construction the smallest (theorem of Gauss Markov [1]) RMSE of the model we can get for the i^{th} "test-sample" with the LS approach. Figure 2 shows the comparison between σ_i^K and σ_i^R. We can note that the estimation carried out by tuning the model with a sample selected with K-means allows a great robustness of the model. Indeed, for each one of the nine "test samples" the path loss model tuned with the "K-fit sample" allows performances (σ_i^K) as good as those (σ_i^R) obtained when the model is tuned with "the sample test". The standard deviation of the difference $\sigma_i^K - \sigma_i^R$ is equal to 0.1 dB when its mean is equal to 0.14dB.

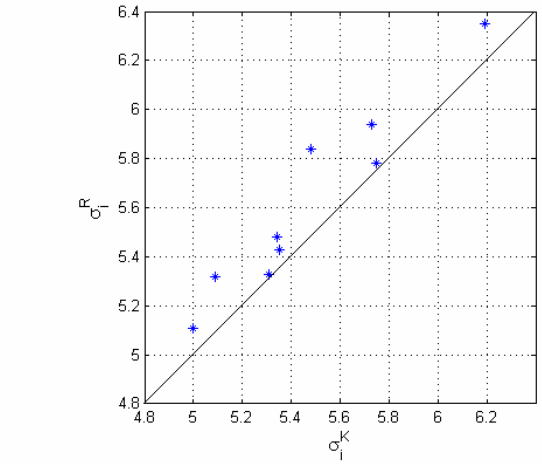

Figure 2: Comparison of the RMSE of the model fitted with the "K-sample" with the optimal model RMSE.

For further validation, figure 3 shows the comparison between the measured and the predicted (using the "K-fit sample") path losses for all the 300 remaining points of the "whole sample".

The result to retain is that we obtain with a "K-fit sample" of only 50 points an as good estimate as the one obtained by using all the points (350 points). For the first case the RMSE is equal to 5.39 dB while it is equal to 5.3dB for the second case.

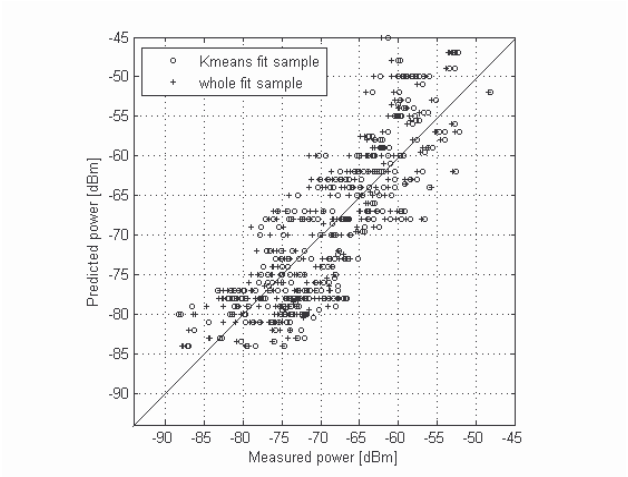

Figure 3: Measured versus predicted power values (Advantage of the K-means selection of tuning points).

VI CONCLUSION

The work presented in this paper allows optimizing an indoor semi-empirical propagation model from a reduced number of measurement points.

A clustering approach was studied to improve the model accuracy and robustness. The proposed process consists in using the K-means clustering method, starting from a large number of potential measurement points' locations, We have proved that it is then possible to select the measurements points' locations (by using a ray tracing in) in order to avoid a ill conditioned data matrix.

The established results showed that the tuning points' choice according to this method gives good path loss predictions with only a small number of points.

Further works would be useful to define more precisely the minimum number of clusters needed to reach a given significance level of the model parameters, rather than doing it empirically.

References

[1] Alan Stuart and J. Keith Ord, "Kendall's Advanced theory of statistics", Edward Arnold Fifth edition, 1991.

[2] Y. Dodge and V. Rousson, "Analyse de régression appliquée", DUNOD, 2ème édition, 2004.

[3] J. M. Keenan and A. J. Motley, "Radio coverage in buildings" Br. Telecom. Tech. J. vol. 8 no. 1, Jan. 1990, pp. 19-24.

[4] N. Papadakis, A. Economou, J. Fotinopoulou and P. Constantinou, "Radio Propagation Measurements and Modelling of indoor Channels at 1800 MHz" Wireless Personal Comm**.,** vol. 9, Kluwer Academic Publishers, 1999, pp. 95-111.

[5] A. Neskovic, N. Neskovic and G. Paunovic, "Modern approaches in modelling of mobile radio systems propagation", IEEE communication surveys, vol. 3rd quarter, Oct. 2000, pp. 2-12.

[6] R. Tahri, V. Guillet, J. Y. Thiriet and P. Pajusco, " Measurements and Calibration Method for Wlan Indoor Path Loss Modelling", 6th IEE International Conference on 3G & Beyond, London, UK. November 2005.

[7] R. Xu and D. Wunsch, "Survey of Clustering Algorithms ", IEEE Transactions on Neural Networks, vol. 16, no. 3, 2005, pp. 645-678.

[8] D. Zakrzewska and J. Murlewski, "Clustering Algorithms for Bank Customer Segmentation ", Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05).

[9] A. K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: a Review", ACM Computing Surveys, vol. 31, no. 3, Sept 1999, pp. 264-323.

[10] L. Chaigneaud, V. Guillet and R. Vauzelle, "A 3D ray-tracing tool for broadband wireless systems", IEEE Vehic. Tech. Conf., Atlantic city, Oct. 2001.