Automatic Calibration of Propagation Models on Railway Communications

João Pedro Rebelo Martinho

Abstract — This work shows that by associating a Personalized K-Means clustering algorithm to Genetic Algorithms, the calibration of propagation models in railways for different kinds of environments and characteristics, produces parameters that minimize the existing error solutions in adjusting the curves and hence obtaining a reduction of the error in estimating the attenuation, compared to the values obtained through a global optimization [3]. The information on which the proposed algorithm was based, consists of actual measurements made along the railways, including the information of the clutter.

Keywords — Clustering; Genetic Algorithms; Clutter; Automatic Calibration of Propagation Models; Railway Communications.

I. INTRODUCTION

GSM-R technology (Global System for Mobile Communications - Railway) arose from the need of creating a wireless digital communications system to fulfill the objective of technological standardization across the rail network in Europe, which led to the design of a specific mobile communications system for the railway network [1].
The set of specifications developed for the implementation of GSM-R technology increased the requirements in terms of quality of service in radio networks. This technology operates at frequencies 876-880 MHz (uplink) and 921-925 MHz (downlink) and it's based on a robust technology, secure and with fast access, satisfying the special needs of the rail infrastructure operators, in terms of professional communications of voice and data [2].

The radio signal coverage prediction is one of the key steps in planning a radio mobile communication network. In rail environment, this estimate requires precision and higher accuracy compared to public networks, given the constraints arising from safety requirements. It is therefore essential to calibrate the propagation models used for different kinds of environments and characteristics of the railroad. The process of adjusting the parameters of a given model, requires the use of automatic optimization techniques, which from test samples, produce parameters solutions that minimize the error in the setting of the curves.

This paper was submitted for review on October 15th, 2016.
João Martinho
Instituto Superior Técnico, Lisboa, Portugal.
joaormartinho@tecnico.ulisboa.pt

The use of genetic algorithms has shown to be valid on the optimization of calibration parameters in propagation models, when applied to radio coverage prediction in railways. However, it highlighted the difficulty in obtaining an overall optimization in terms of signal modulation behavior for different types of environments as well as the non-utilization of clutter information, in a general way [3].

It is proposed to associate the advantages of the use of propagation models based on radio coverage prediction with a Personalized K-Means clustering algorithm (PKM) allowing to obtain in advance the classification of the types of environments, in order to reduce the overall error in the prediction. Thus, it is necessary to study and test various clustering schemes, analyze the parameters characterizing the geographical location as well as the clutter information, to obtain a more efficient classification and determine the number and the final characteristics of the types of environments.

For the different types of environments / classes, we use the most suitable estimation models, including the Okumura-Hata [4] and [5], which showed good results in radio coverage prediction in railways [6]. Thus, making it necessary to perform an analysis to different propagation models applied to the respective embodiments, determining the best designs based on different classes of identified environments and use the clutter information to improve the models accuracy.

This paper is structured as follows: section II describes the propagation models which were used for the prediction of radio signal, the clutter information, the operating principle of Genetic Algorithms and the phases of a clustering process; section III describes the methodology used for the completion of the Proposed Algorithm; section IV provides the final configuration of the developed algorithm, as well as the cluster analysis. Section V presents the conclusions and future work drawn from this project.

II. THEORETICAL FOUNDATIONS

On the following section the propagation models, used for the prediction of radio signal, are presented as well as the clutter information, the operating principle of Genetic Algorithms and the phases of a clustering process. The section is finished addressing the problems in performing clustering in high-dimensional scenarios.

A. Propagation Models

The Okumura-Hata model provides the median value of attenuation, which is influenced by parameters such as frequency, f, the distance from the receiver to the base transceiver station, d, and the height of the receiver antenna, hm (considering the cabin radio scenario, with an isotropic antenna, installed at a height of $4m$ [7]). The attenuation median value is given by the following equation [8]:

$$L_{p[dB]} = 69.55 + 26.16\ log\left(f_{[MHz]}\right) - 13.82\ log\left(h_{be[m]}\right) + \left[44.90 - 6.55\ log\left(h_{be[m]}\right)\right] log\left(d_{[km]}\right) - H_{mu[dB]}(h_m, f) - \sum correction\ factors \quad (1)$$

Where for a basic suburban environment is presented as follows:

$$H_{mu[dB]} = \left[1.10\ log\left(f_{[MHz]}\right) - 0.70\right]h_{m[m]} - \left[1.56\ log\left(f_{[MHz]}\right) - 0.80\right]. \quad (2)$$

The across path correction is considered, concerning the orientation between the antenna and the railroad. When both elements have the same orientation, the amount of their attenuation is given by:

$$K_{ac}(\theta)_{[dB]} = 2.1\ log\left(d_{[km]}\right) - 6.3. \quad (3)$$

The height of ground undulation, Δh_b, is obtained from the difference between the 10 percentile and 90 percentile of the respective terrain height. The attenuation of this undulation is given by:

$$K_{th}(\Delta h_b)_{[dB]} = -3\ log^2\left(\Delta h_{b[m]}\right) - 0.5\ log\left(\Delta h_{b[m]}\right) + 4.5 \quad (4)$$

When the mobile terminal location, on the ground undulation, is known, this attenuation is obtained by the following equation:

$$K_{hp}(\Delta h_{b[m]})_{[dB]} = -2\ log^2\left(\Delta h_{b[m]}\right) + 16\ log\left(\Delta h_{b[m]}\right) - 12 \quad (5)$$

Where Δh_b is, in this case, the average height of the land undulation, which is obtained by averaging the difference between the 10 percentile and 90 percentile of the terrain height.

The correction factor that considers the average slope of the land is given by:

$$K_{sp}(\theta)_{[dB]} = \begin{cases} -0.0025\ \theta^2_{[mrad]} + 0.204\ \theta_{[mrad]}\ ,\ (d < 10km) \\ -0.648\ \theta^{1.09}_{[mrad]}\ ,\ (d < 30km) \\ -0.0012\ \theta^2_{[mrad]} + 0.840\ \theta_{[mrad]}\ ,\ (d < 60km) \end{cases} \quad (6)$$

The parameter $\beta = \frac{d_s}{d}$ describes the relationship between the distance of the path where there is water, d_s, and total distance of the path between the base station and the mobile terminal, d. The corrective factor that supports this type of route is given by:

$$K_{mp}(\beta)_{[dB]} = \begin{cases} \begin{cases} -11.9\beta^2 + 4.7\beta\ ,\ d > 60km \\ -7.8\beta^2 + 5.6\beta\ ,\ d < 30km \end{cases} A \\ \begin{cases} -12.4\beta^2 + 27.2\beta\ ,\ d > 60km \\ -8.0\beta^2 + 19.0\beta\ ,\ d < 30km \end{cases} B \end{cases} \quad (7)$$

Where A considers the situation in which the water is located far from the base station relative to the mobile terminal, being B the inverse situation.

Diffraction losses are determined by using a model consisting of an approximation assuming that the obstacles have a blade geometry, as considered in the P.526 model recommendation [9]. The attenuation is given by:

$$L_{ke[dB]} = 6.4 + 20\log\left(v + \sqrt{v^2 + 1}\right), v > -0.7 \quad (8)$$

where $v = h\sqrt{\frac{2d}{\lambda d_t d_r}}$ (9),

v is the parameter defined by Fresnel-Kirchhoff, where h is the obstacle height, whether it is above or below the direct radius between the transmission and reception antennas, d is the total link distance, d_t is the distance between the base station and the obstacle, d_r is the distance between the obstacle and the mobile terminal, and λ is the wavelength.

B. Clutter Data

The weather and geoclimatic data, along with the morphological characteristics of the ground surface and clutter database, are resources that can be used by propagation models to improve the effectiveness of prediction loss between transmission and reception antennas [10].

Clutter refers to a classification of surface characteristics that influence the propagation of radio waves. Clutter is usually produced from multispectral satellite images where different classes of surface features may be designed through spectral homogeneity, among other characteristics.

In order to account for the characteristics of the studied scenarios, we use clutter information, in which each pixel is associated with a code that defines the characteristics of that pixel [11].

C. Genetic Algorithms

Figure 1 presents the pseudo-code of a GA.

```
start
    t ← 0
    Randomly initiate P(t)
    Evaluate P(t)
    while (not optimization criteria)
    do
        t ← t + 1
        P(t) ← Seleção P(t − 1)
        Modify P(t)
        Evaluate P(t)
    do
end
```

Fig. 1, GA pseudo-code.

A GA it's a probabilistic algorithm, which maintains a population $P(t) = \{x_1^t, \dots, x_n^t\}$ for iteration t.

Each P element represents a possible solution of the problem, being each individual measured according to a given criterion. Then, it generates a new population from this one, being replaced by their descendants, a subset of selected individuals, possessing the most gifted, a higher probability of being included in this selection. These descendants are obtained by applying genetic operators, such as crossover and mutation. After several iterations, the algorithm converges to an optimal point.

The implementation of a GA is characterized by a genetic representation of the solutions to a given problem; by the initial population creation process; by a given classification function simulating an environment to assess the individual's fitness; by genetic operators, to change the composition of the descendants of the population; and values for the various parameters that a GA uses (population size, probability of the use of genetic operators, etc.) [12].

D. Steps of a Clustering Process

The process of grouping a set of objects into groups (clusters) of similar objects is called Clustering. A cluster is a collection of data elements that show similarities between elements of the same group and differences between elements of other clusters. The intrinsic basic steps to a clustering process are illustrated in Figure 2 [13].

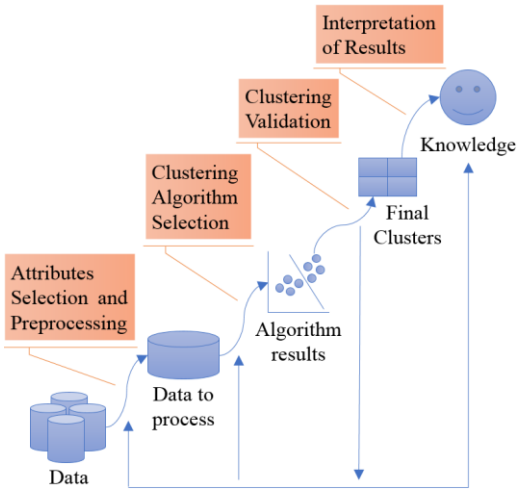

Fig. 2 Steps of a clustering process.

The phase of selection and preprocessing of data elements attributes, has the goal of correctly select the characteristics on which the clustering must be performed to encode as much information as possible, for the ultimate goal. The techniques of data preprocessing are applied to raw data, making them consistent and free from noise, improving the accuracy of the clustering algorithm.

The choice of a clustering algorithm, focuses on the definition of a proximity measure and a grouping criteria. This definition characterizes a clustering algorithm. The proximity measure quantifies the similarity between data elements.

The accuracy of the results of a clustering algorithm is verified using validation indices. Since clustering algorithms define assemblies that are not known in advance, regardless of the clustering methods, it is imposed an assessment of the final data partition.

The interpretation of results, typically incorporates the clustering results with other experimental evidence, with the objective of analyze and extract useful information [14].

E. Clustering in High Dimensional Scenario

With the increase of data size, the distances lose their effectiveness, as well as their statistical significance, due to irrelevant attributes.

The idea focuses on the fact that the attributes characterized by small fractions, will remain relevant with the increase of data size, providing the loss of distance definition, as well as the increase of concentration effect, due to the behavior of irrelevant attributes.

The effect of concentration refers to the situation in which a high number of noisy attributes or not correlated, causes a scenario in which all distances between points, become similar [30].

In clustering algorithms based on distance, the noise and the effect of concentration are problematic in two ways:

- An increase in noise caused by irrelevant attributes, can cause errors in distance of representation and, consequently, promote a wrong representation of the distances between objects.
- The concentration effect, encouraged by irrelevant dimensions, leads to a reduction in the statistical significance of results from distance based clustering algorithms.

One of the premises to tackle these problems, focuses on controlling the data size, selecting the attributes, regarded as the most influential, as well as on applying a proximity function, which offers better data contrast, in the calculation of distances between points.

III. Combining Clustering with Optimization

This section describes the procedure of each step involved in the implementation of the PA (Proposed Algorithm), including the clustering process, as well as the combination of GA with the developed algorithm. The section is finished with an explanation of the method for interpretation of clustering results, this being based on clutter information from each cluster and on the comparison between the final statistics from each of the paths shown in Figure 3.

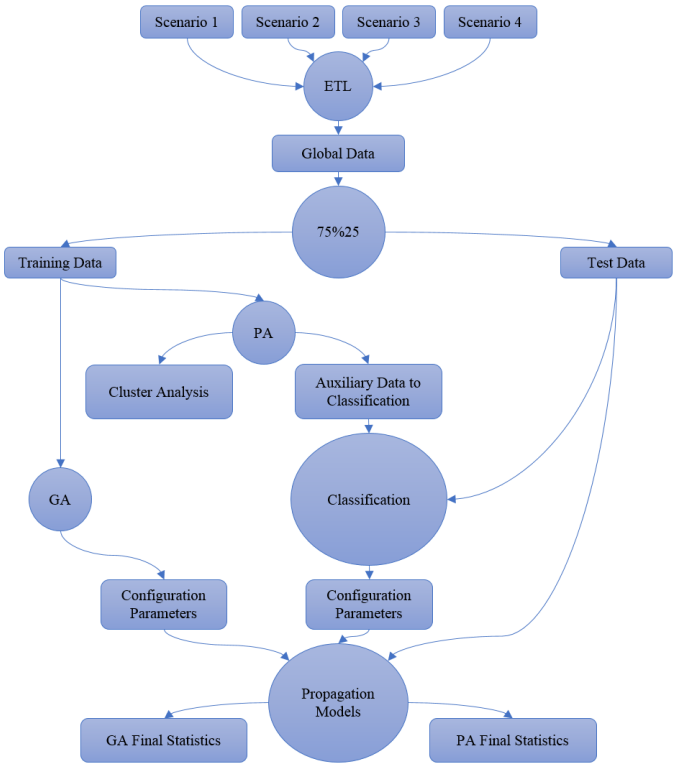

Fig. 3 Implementation's block diagram.

The collection of measures in railway environment, enabled the creation of the above represented scenarios. This geographical information, which includes the clutter information, is subjected to an ETL process (Export, Transform and Load) [16], with data being subsequently stored in a file container.

After ETL process, it is performed a random division for the phases of training (75%) and test (25%), whose implementation has the objective of validate the developed algorithm.

In training phase, is selected the information to be consumed by GA, being the resulting information, a set of configuration parameters, optimized for the global case, to be used by propagation models, in test phase.

On the other hand, is also selected the information to be consumed by PA, being the resulting information, the characterization of clusters, which were obtained using the Personalized K-Means clustering algorithm (PKM).

Moreover, from PA results a classification instrument data set. After the classification of test elements, a set of configuration parameters is returned, previously optimized for a given cluster, to be used by the propagation models, in test phase.

Then, the geographical information from test data, along with the configuration parameters of both algorithms (GA and PA), are introduced into the propagation models and a final prediction is obtained.

In order to compare the prediction, coming from both algorithms, with the previously collected measures, we use first order statistics (the absolute mean error, ME, the root of the mean square error, RMSE and the error standard deviation, ESD), as well as the correlation coefficient (RE). Lastly, the final statistics of GA are compared with the final statistics of PA.

A. Geographical Information

The geographical characteristics/morphological characteristics of the studied scenarios (Algarve, Cascais, Sintra and Vendas Novas), collected by BTS, intrinsic of each railway point, are the following:

- Distance between a BTS and a railway location point;
- Effective height of the BTS antenna;
- Parameters relating to the 3 main obstacles;
- Distance travelled on vegetation;
- Distance travelled on water;
- Terrain undulation height;
- Average height of terrain undulation.

In railway environment, the used metric, either for distance or for referencing a given occurrence or installation, is referenced as PK. There is no numerical method to associate a PK with a geographic point, thus it's necessary to use a file with this information.

Along with this information was added the height of the mobile antenna, the frequency and the clutter information, consisting of 19 classes.

The collected information described so far, it's subject to an ETL process. As the name suggests, this process has the objective of extract, transform, and store data from an external source, to a certain file container.

The information is extracted from the respective containers, being converted to a matrix format, and subsequently transformed. This transformation is performed in such a way that each data element is represented by a line and by a number of columns, equivalent to the number of attributes.

The resulting structure of the data set, is represented by an array of $n - by - p$ (n elements by p attributes). These attributes correspond to the above described information characteristics (distance, height of the antennas, classes of clutter, etc).

Then, the data are stored in Global Data file, to facilitate the selection of data elements, either for GA, either for PA.

B. Training

During the training process, is selected the information to be consumed by PA, being the resulting information, the data to be classified, which are constituted by K sets of configuration parameters and by K centroids locations. This process also

leads to the characterization of each cluster, which were obtained using the developed clustering algorithm (PKM).

On the other hand, is selected the information to be consumed by GA, which corresponds to the training data elements, characterized by the attributes presented in *A - Geographical Information*, excluding the information of clutter. The resulting information from this process is a set of configuration parameters, globally optimized.

In this subsection are reported the applied strategies, which I considered to be the most advantageous for the ultimate goal. Figure 4 illustrates a zoom in of PA.

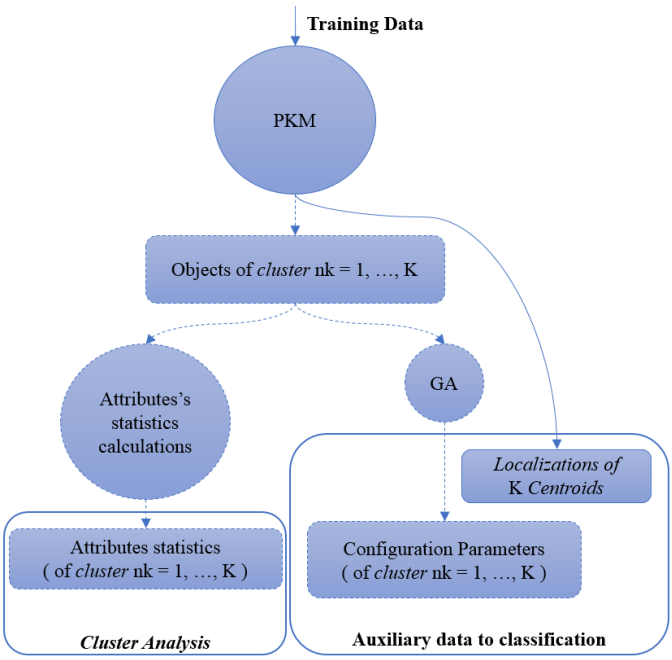

Fig. 4 Zoom in of PA process.

The input arguments of PKM are a matrix X and a positive integer value K. The lines of X correspond to observations / points and the columns correspond to variables / attributes; while K, corresponds to the number of desired clusters.

PKM returns a vector composed by the cluster indices of each observation.

The algorithm's name is inspired by the fact that the strategies of data elements selection and preprocessing, as well as the clustering validation technique that was used to estimate K's value [17], were added to the original K-Means algorithm.

Thus, becoming a Personalized K-Means clustering algorithm, adjusted to the characteristics of the used data set, with the goal of obtaining the best possible grouping.

From Global Data is performed the data elements selection and preprocessing, for the PKM clustering algorithm.

The application of a preprocessing technique on the attributes, on which the clustering algorithm is executed, has the objective of improving the quality of grouping. As such, the standardization of data elements attributes is performed.

Min-Max is the method that presents better results for the used data type, and, as such, is chosen for the data preprocessing. The Min-Max standardization is given by the following equation [15]:

$$MinMax(X_{ij}) = \frac{X_{ij} - X_{min}}{X_{max} - X_{min}} \quad (10)$$

Several grouping strategies have been implemented with the goal of achieving an optimal selection of attributes, which promotes the best statistical result, that is, the greater reduction of attenuation estimation error, compared to values obtained in [3].

The selection of data elements, characterized by a number of attributes, aims to collect the more influential particularities of each observation, for clustering.

In order to avoid possible concentration effects, listed in *E - Clustering in High-Dimensional Scenario*, influences below a certain threshold were eliminated, in the containing images (pixels) of each railway point. This cleansing, along with a reduction of clutter classes, aims to improve clustering accuracy.

The strategies for reducing the data size are based on the definitions of the 19 classes of clutter. This reduction process consists in the creation of new attributes, built through the junction of clutter classes that share similarities, i.e, that have similar characteristics in radio signal spread context (Table 1).

1	Water
2	Vegetation
3	Urban
4	Open

Table 1 Final clutter classes.

The final classes of clutter along with the parameter v1, obtained using Deygout model, proved to be the best selection of attributes, for the implementation of PKM.

Concerning clustering validation, the chosen index for validation is the percentage of variance explained, which is the ratio of between-group variance with total variance. The smaller the value of this index, the greater the dispersion within a cluster. The higher the value of the variance

explained, the lower the dispersion within a cluster (higher compactness).

The premise of the implemented approach consists in choosing the best clustering scheme, based on a set of predefined schemes, characterized by different values of K.

The goal is to find the value of K that best fits the data set.

Running PKM multiple times, for a range of K values, and keeping the best variance explained values corresponding to each value of K (nk), the graph of variance explained, in function of K, was drawn (Figure 5).

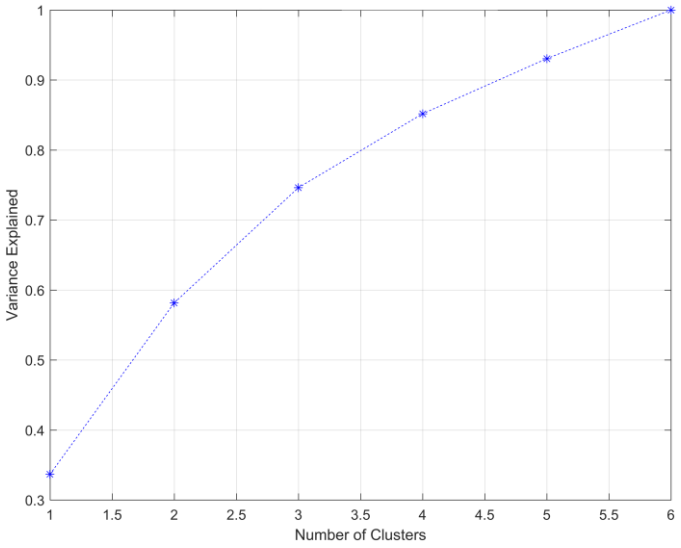

Fig. 5 Clustering validation to estimate K.

Analyzing the above chart, it appears that from K=5, although visually, is not easily perceived, the variance explained value undergoes a significant change, and the fact that correspond to a result higher than 90%, it is concluded to be the location corresponding to the number of clusters, underlying the data set.

For the data elements to group, characterized by the 4 clutter classes and by the parameter corresponding to the main obstacle, $v1$, the K's value which resulted from the validation process, is equal to 5.

After determining K, the involved steps in the implementation of PKM clustering algorithm are initiated, starting by initializing K centroids.

The choice of centroids initialization process is a fundamental step of the K-Means based algorithm.

When centroids are randomly chosen, different executions of the clustering algorithm, produce different results compared to the sum of the quadratic error. Adding the fact that the resulting clusters are typically poor in terms of cohesion, and in terms of useful information extraction [33].

According to Arthur and Vassilvitskii [18], k-Means++ improves the execution time of Lloyd's algorithm (K-Means), as well as the quality of the final solution.

Arthur and Vassilvitskii demonstrated, by using a simulation study of various cluster guidelines, that K-Means++ achieves a faster convergence, getting clusters more compact, compared to Lloyd's algorithm. As such, K-Means++ is used as the initialization method, in the implementation of PKM.

The assignment of components to the closest centroids is performed using a proximity measure function, in order to quantify the concept of "closest" in relation to the elements of the used data set.

In high dimensional scenarios, the ratio between the nearest and the more distant point, approaches 1, i.e., the points become evenly apart from each other.

In [19] is provided theoretical and experimental demonstration, on the analysis of the dependence of L_m standard, regarding the value of m.

It is shown that the relative contrasts, of the distances to a consultation point, are heavily dependent on the L_m metrics choice.

Thus, for a high dimension, d, data set, ($d \geq 3$), it is advantageous to use low values of m. Which means that L_1 metric (distance of Manhattan), offers higher contrast, compared to L_2 (euclidean distance), and therefore the it's the chosen one.

The cluster indices corresponding to each data element, as well as the location of the K centroids, are the resulting information from PKM clustering algorithm.

The collected geographical information, along with the model parameters, enables the development of a prediction made by propagation models.

Based on the error between the prediction and the measures, GA provides new model parameters. The new parameters generate a new prediction, which is again evaluated by the algorithm. This process repeats itself until it hits a stop condition, either for having been reached a certain amount of error, either for having been reached the maximum number of iterations.

GA optimization process developed in [3], is applied globally to all training data elements, as well as, partially, by cluster.

The information resulting from the global optimization is a set of model configuration parameters, optimized for the totality of training data elements.

The optimization per cluster is obtained, taking advantage of the cluster indices corresponding to each element, returned by PKM, allowing the selection of the required attributes,

referring to information contained in each cluster, for the achievement of the optimization process.

The information resulting from the optimization per cluster, is a set of model configuration parameters, optimized for the elements present in each cluster.

C. Test

The remaining 25% of the randomly sampled information is forwarded for test phase (Figure 6).

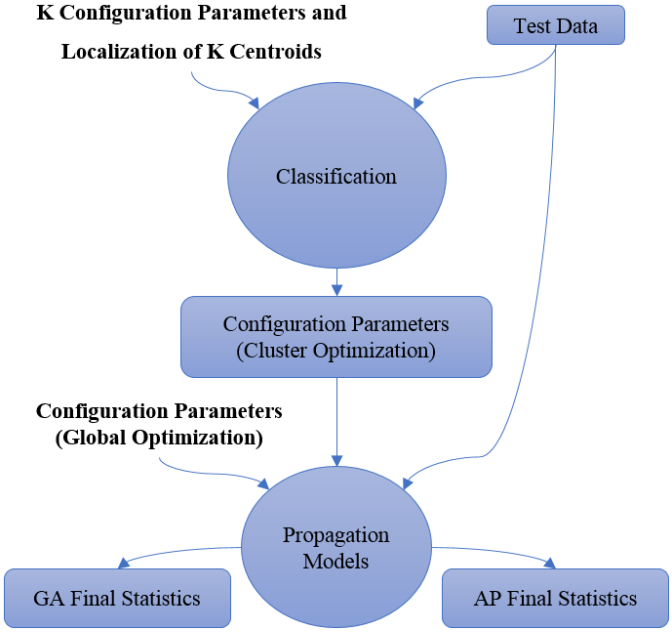

Fig. 6 Parcial block diagram

After the classification of the test elements, in their respective clusters, the previously collected geographical information, along with the configuration parameters, from both algorithms, are introduced into the propagation model and a final prediction is calculated.

It is then carried out a comparison of predictions, coming from both algorithms, with the previously performed measures, using first order statistics and the correlation coefficient.

The classification is the process of finding a model that describes and distinguishes a data element, with the aim of using that model to predict the category of elements, whose description / label is unknown. The derived model is based on the analysis of a training data set [20].

The calculation of the distance between a test element and the K locations of centroids is performed using the same metric that was used during clustering process (Manhattan). The smallest of the K distances resulting, reveals the group to which the element of test is belonging. Being the set of configuration parameters, corresponding to the resulting cluster from classification, applied to that test element.

The model of the Okumura-Hata does not account for the losses due to the resulting diffraction of obstacles, therefore, for such purposes, it is considered a model that does.

The model used for the calculation of radio coverage prediction in GSM-R, is composed by the Okumura-Hata with all the corrective factors, by the Deygout method, in order to account for the additional losses due to diffraction grating, allowing for greater accuracy in the calculation of the total losses, and also by the method based on the ITU-R P.1546 recommendation, which proved to be the most beneficial for the determination of the base station antenna height [3].

With respect to the statistics, a set of model configuration parameters is considered more optimized, the smaller the deviation resulting from the prediction, calculated using these parameters, for real measurements.

For a better comparison between the prediction and the measures, first order statistics along with the correlation coefficient are calculated.

The statistics are intended to assess the overall error of radio signal prediction and are given by the following equations:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left|P_{measi} - P_{predi}\right| \quad (11)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|P_{measi} - P_{predi}\right|^2} \quad (12)$$

$$ESD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|P_{measi} - P_{predi}\right| - ME\right)^2} \quad (13)$$

Where P_{measi} is the signal level (dBm) of the measured signal at point i, being n, the total number of points and P_{predi}, the corresponding prediction value. The calculation of the correlation coefficient is given by:

$$RE = \frac{\sum_{i=1}^{n}(P_{measi}-\bar{P}_{meas})(P_{predi}-\bar{P}_{pred})}{\sqrt{\sum_{i=1}^{n}(P_{measi}-\bar{P}_{meas})^2}\sqrt{\sum_{i=1}^{n}(P_{predi}-\bar{P}_{pred})^2}} \quad (14)$$

After applying clustering to the data set, the resulting clusters are characterized by the attributes of the elements belonging to these clusters. Allowing the classification of an unknown element in a specific cluster, based on the similarity between their attributes and those of, already defined, clusters.

It thus becomes possible to extract useful knowledge of the initial data. In order to evaluate the presence of each attribute, in terms of data variation, statistical calculations are applied, in particular the mean and standard deviation, to the initial values of the attributes, present in each cluster.

The mean provides a central location of a data set. The standard deviation describes the dispersion of data, as well as its distribution around the mean [21].

IV. Results

After being discovered the ideal arguments, representative of the final clustering configuration of PKM algorithm, was performed the selection and preprocessing of the data elements to be grouped in 5 clusters. The 5 centroids were initialized using the method of Arthur and Vassilvitskii and the assignment of elements to the closest centroids, was performed based on the distance of Manhattan.

A. *Final Clustering Scheme*

The best result was obtained through the selection of 4 clutter classes, along with the parameter v1. Figure 7 shows the comparison between the statistics ($ME, ESD, RMSE\ and\ RE$), corresponding to this selection, from PA, and the final statistics, from GA.

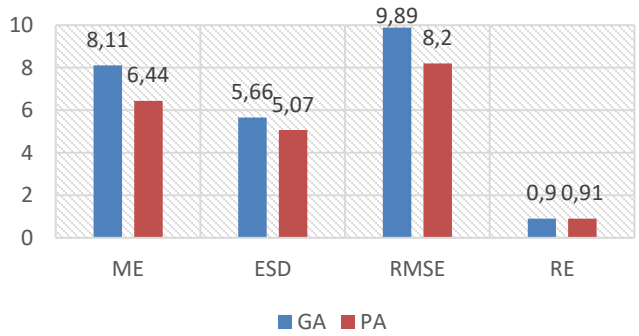

Fig. 7 Comparison between the statistics of GA and the best statistics of PA.

The combination of 4 clutter classes, with the parameter v1, promotes the best statistical result, that is, the greater reduction of error in estimating the value of the attenuation, compared to the values obtained in [3].

Being the clustering configuration that best fits the used data set, defined by the following points:

- Mapping of 19 clutter classes, to a minimum number of classes (4), filtering influences below 15%;
- Standardization of attributes, for parameter $v1$, using the Min-Max method;
- Attributes selection from the data set:
  - 4 clutter classes (Water, Vegetation, Urban and Open) and parameter $v1$.
- Estimating the value of K, using variance explained as a method for clustering validation;
- Initialization of K centroids using the k-Means++ algorithm;
- Clustering using the metric of Manhattan, as a function of proximity,
- Multiple executions of the PKM algorithm;

B. *Cluster Analysis*

The resulting clusters, as has already been mentioned above, are characterized by the attributes of the elements belonging to these clusters.

Bellow are presented the figures that illustrate the characterization of the constructed clusters.

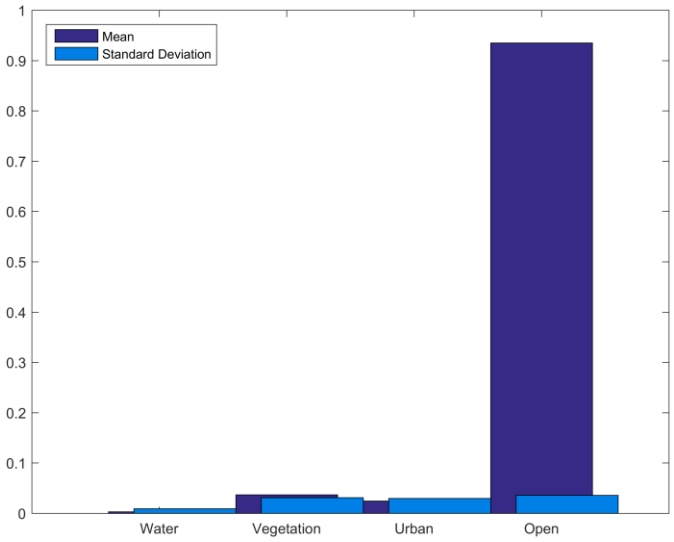

Fig. 8 Cluster 1.

Evaluating the resulting statistics values of each attribute, as well as the relationship between both indicators, it is concluded that the first cluster is characterized mainly by the presence of open terrain areas ("Open"), resulting in a uniform presence of data, considering the low value of the respective standard deviation.

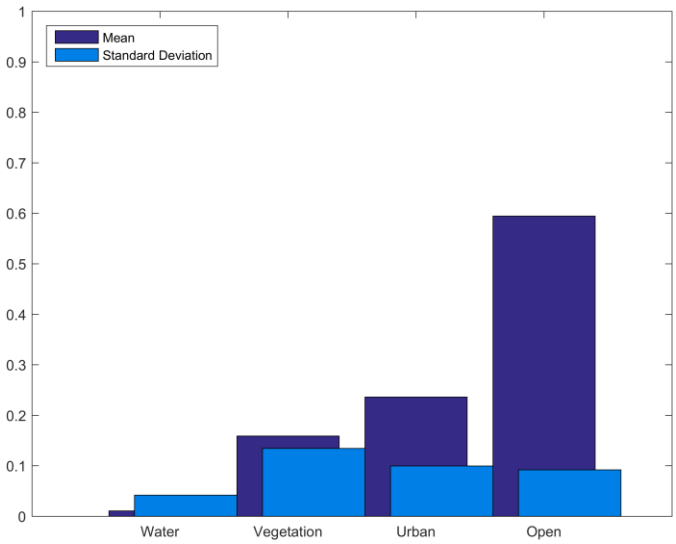

Fig. 9 Cluster 2.

The second cluster has a disperse presence of areas covered by planting and/or trees canopies and a slightly uneven presence of areas with urban characteristics. This cluster is characterized mainly by the presence of open terrain areas and by the absence of areas covered with water.

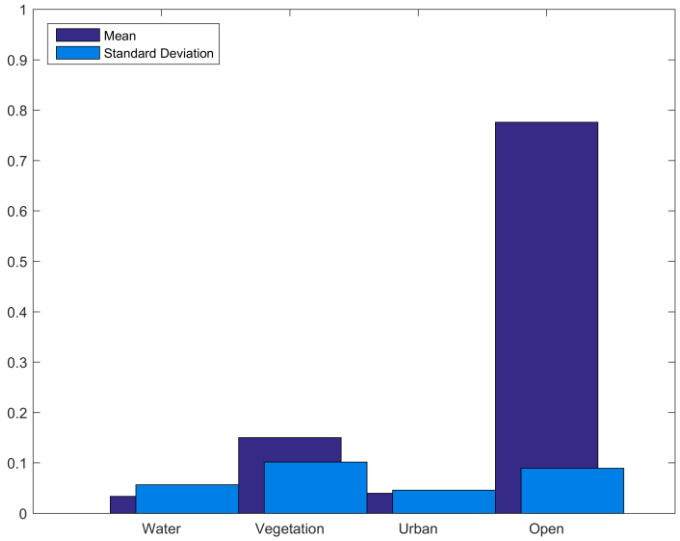

Fig. 10 Cluster 3.

The third cluster is characterized mainly by a uniform presence of open terrain areas and the shortage of areas with urban characteristics and covered by water. The attribute "Vegetation" has a small standard deviation, compared with its mean value, reflecting an influence of approximately 15% of areas covered by plantations and/or trees.

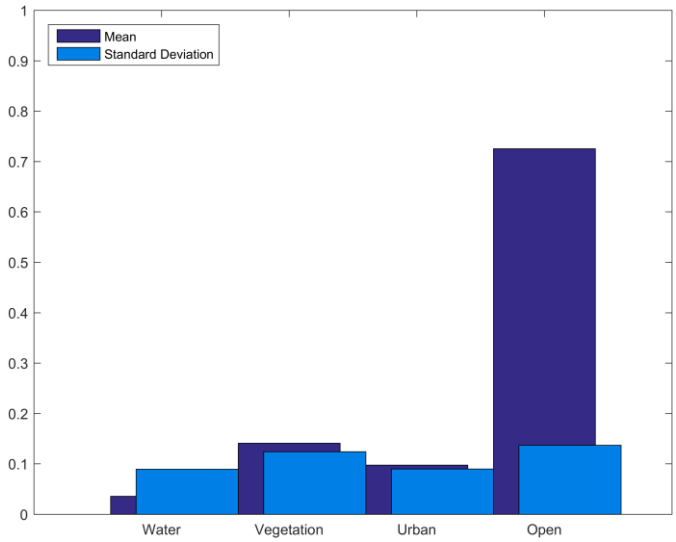

Fig. 11 Cluster 4.

The fourth cluster, presents a disperse presence of areas covered with vegetation. The standard deviation of the "Urban" influence is close to its mean value, indicating a high dispersion of data, revealing a high dispersed presence of areas covered by buildings or urban characteristics. This cluster is characterized mainly by areas of open terrain.

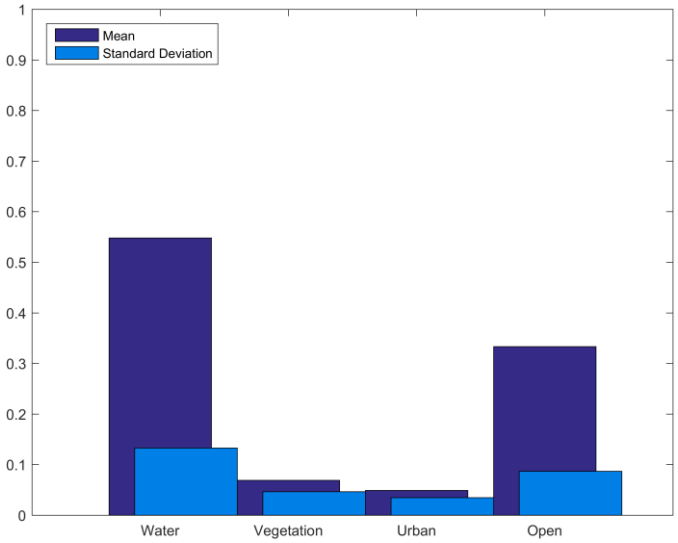

Fig. 12 Cluster 5.

The fifth cluster is characterized mainly by a uniform presence of flooded areas or covered with water. The presence of areas characterized by open land has, on average, an influence greater than 30%. This, represented by the attribute "Open", is translated into a uniform presence of data, considering the low value of the respective standard deviation. The presence of planting areas and/or trees, as well as areas within the urban perimeter, it is fairly scattered.

The statistical calculations relating to the characterization of the presence of obstacles using parameter $v1$, have not proved to be discriminating in the extraction of useful information, in terms of analysis of actual values used for the process of clustering.

However, the inclusion of $v1$ in the selection of attributes from the set of data elements, present in the final configuration of the clustering algorithm, proved to be advantageous, in terms of improvement of the results through the association of PKM with GA.

Through the comparison of performed measurements, with the prediction of radio signal, obtained through the GA and with the obtained through the PA, it becomes possible to visualize an improvement in adjusting the curves. Figure 13, referring to a Cascais railway, it is an illustrative example of the comparison between the points concerning the measures (in red) and the corresponding curves of radio signal prediction, either by using GA (in blue), either by using the proposed association of PKM with GA (magenta).

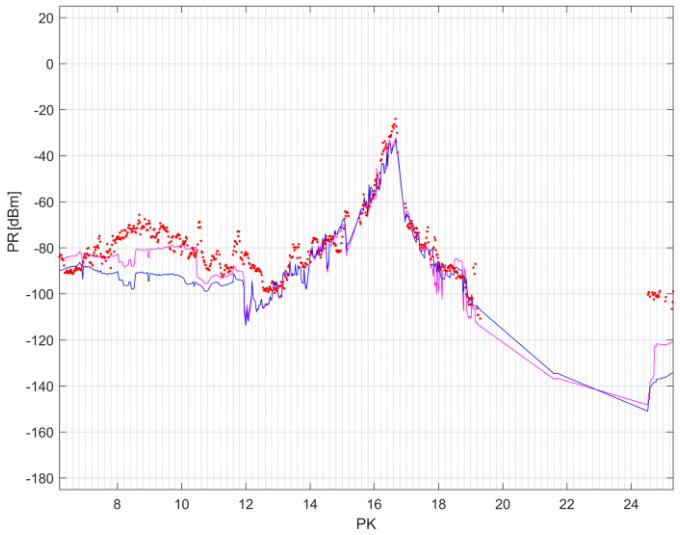

Fig. 13 Comparison between the points concerning the measures and the curves of prediction.

Watching the figure above, there is an improvement (around PK 10), in terms of matching the measures, in the adjustment of curves resulting from the developed algorithm, in relation to the adjustment resulting, using the GA.

V. Conclusions and Future Work

The application of the final clustering configuration, with the aim of collating all the data elements, into subsets that share geographic similarities / morphological characteristics, associated with the application of GA, to optimize the set of configuration parameters of the model, for the elements present in each of the groups obtained, produce parameters that minimize the error in estimating the attenuation, compared to the values obtained using the algorithm developed in [3].

Adding the fact that, through this association have been achieved a standard deviation of the error of radio signal prediction of approximately 5.1 dB. To reduce this statistic reveals the possibility of a reduction in the number of base transceiver stations, in the planning of the network and, consequently, a reduction in implementation costs.

As future work is proposed the use of SOM (Self Organizing Maps), a not supervised technique for data visualization, which can be used to view high dimensions sets of data elements in representations of lower dimensions (typically two). One of the main advantages of SOM, in terms of visualization, is based on the fact that this mapping preserves the topological relations, intrinsic to the original data [22].

References

[1] GSMR - Info. [online]. http://www.gsmr-info.com/, accessed on: September 2015.

[2] UIC Project EIRENE, System Requirements Specification. 2006 [Online]. http://www.uic.asso.f

[3] Beire, Ana; "Otimização de modelos de propagação utilizando Algoritmos Genéticos: Caso das Comunicações Móveis em Ferrovia", ISEL, December, 2013.

[4] Okumura, Y.; Ohmori, E.; Kawano, T.; Fukuda, K. "Field Strength and its Variability in VHF and UHF Land-Mobile Radio Service". Review of the Electrical Communication Laboratory, Vol. 16, Nº 9-10, October 1968, 16, pp. 825-73.

[5] Hata, Masaharu. "Empirical Formula for Propagation Loss in Land Mobile Radio Services". IEEE Transactions on Vehicular Technology, Vol. VT-29, Nº 3, August 1980, 29, pp. 317-25.

[6] Cota, Nuno; Serrador, António; Vieira, Pedro; Beire, Ana; Rodrigues, António; "On the Use of Okumura-Hata Propagation Model on Railway Communications," in Wireless Personal Multimedia Communications Symposium (WPMC2013), Atlantic City, New Jersey, USA, 2013.

[7] Cota, Nuno; Serrador, António; Franco, Nuno e Neves, José, "Planeamento Rádio em GSM-R: Metodologia e Caracterização do Sinal", URSI, Lisboa, 2009.

[8] Correia, Luís; "Sistemas de Comunicações Móveis – Modelos de Propagação". Lisboa, Portugal: IST, 2007.

[9] Recommendation ITU-R P.526-12, "Propagation by diffraction," January 2012.

[10] http://www.teleres.com.au/Terrain, accessed on: August 2016.

[11] Pahl, John; "Interference Analysis: Modelling Radio Systems for Spectrum Management"; pp. 100-156, April 2016.

[12] Holland, J. H. "Adaptation in Natural and Artificial Systems", Ann Arbor, MI: University of Michigan Press, 1975.

[13] http://www.cse.msu.edu/~jain/Clustering_Jain_Dubes.pdf vol. 3 e 4, accessed on: June 2016.

[14] http://web.itu.edu.tr/sgunduz/courses/verimaden/paper/validity_survey.pdf, accessed on: June 2016.

[15] http://maxwellsci.com/print/rjaset/v6-3299-3303.pdf, accessed on: July 2016.

[16] http://datawarehouse4u.info/ETL-process.html, accessed on: August 2016.

[17] http://www.mathworks.com/help/stats/kmeans.html, ccessed on: March 2016.

[18] http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf, accessed on: May 2016.

[19] https://bib.dbvis.de/uploadedFiles/155.pdf, accessed on: August 2016.

[20] http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf vol. 6 e 7, , accessed on: August 2016.

[21] https://statistics.laerd.com/statistical-guides/measures-of-spread-standard-deviation.php, acedido em: Setembro de 2016.

[22] pdfs.semanticscholar.org/3ffe/8f8a7b0d00297e0cd74d20b5d936349d6cbc.pdf, accessed on: September 2016.