Application of Gaussian Mixture Modeling Methods to Analysis and Prediction of Cellular Communications Pathloss Distributions

Walter Kuklinski
The MITRE Corporation
Bedford, MA
wskuklin@mitre.org

Evan R. Ding
The MITRE Corporation
McLean, VA, USA
eding@mitre.org

Jeffrey Correia
The MITRE Corporation
Bedford, MA
jcorreia@mitre.org

Jared R. Burdin
The MITRE Corporation
Bedford, MA, USA
jburdin@mitre.org

Michael Bowman
Defense Spectrum Organization
DISA
Annapolis, MD, USA
michael.w.bowman55.civ@mail.mil

***Abstract*— The increased demand for more commercial cellular spectrum has led to the repurposing of spectrum currently used by Federal Spectrum Dependent Systems (SDS) for commercial access, to both exclusive commercial use and shared commercial/Federal use. In spectrum sharing scenarios, spectrum coexistence analysis is performed to ensure Federal SDSs and commercial cellular networks can operate without receiving interference which would degrade system performance. The models used in this analysis rely on the accurate classification of proposed cellular sectors based on their expected emission profiles. This paper introduces a Gaussian Mixture Model (GMM) approach to the problem of classifying LTE sectors based on their uplink (UL) emission profile. A large set of Key Performance Indicator (KPI) data collected from diverse cellular markets within the contiguous United States is used to train a GMM which calculates an optimal set of nominal UL pathloss distributions by which LTE sectors can be characterized. We also introduce a regression-based algorithm that assigns each LTE sector in a network to one of the nominal UL pathloss distributions using only network configuration information (including radio location, antenna height, antenna elevation and azimuth beamwidth, and the census bureau land use morphology associated with each base station). The resulting UL pathloss predictions are more accurate than existing methods which classify sectors by their land use morphology.**

Keywords— Gaussian Mixture Models, Bayesian Information Criterion, regression analysis, cellular network interference prediction, RF, interference, spectrum sharing, Advanced Wireless Services 3 (AWS-3)

I. OVERVIEW

The increased demand for more commercial cellular spectrum has led to the repurposing of spectrum currently used by Federal Spectrum Dependent Systems (SDS) for commercial access, to both exclusive commercial use and shared commercial/Federal use. Repurposing spectrum for exclusive use has become increasingly difficult as Federal systems must operate daily, performing testing, training, and critical missions for our Nation. Spectrum sharing between commercial wireless networks (i.e., 4G/LTE) and Federal systems has emerged as a means to increase the utilization of the spectrum while meeting both commercial and Federal needs.

The Advanced Wireless Service (AWS) spectrum auction of 2014–2015 repurposed the 1755-1780 MHz frequency band from exclusive Federal use and allowed commercial cellular access to spectrum in the 1755–1780 MHz frequency band. Federal systems have incrementally transitioned out of the band and must share spectrum within the transition timelines. Portions of the spectrum will be shared indefinitely.

This auction introduced the possibility that Federal incumbent SDSs using those frequencies could experience interference from cellular User Equipment (UE) (Fig. 1). As a result, some restrictions must be placed on the deployment of cellular networks to ensure interference levels would not reach a point of impacting Federal SDS operation. To maximize spectrum usage, these restrictions are informed by a pre-deployment coexistence analysis which predicts the magnitude of the interference experienced by the Federal SDS using only network laydown and configuration information.

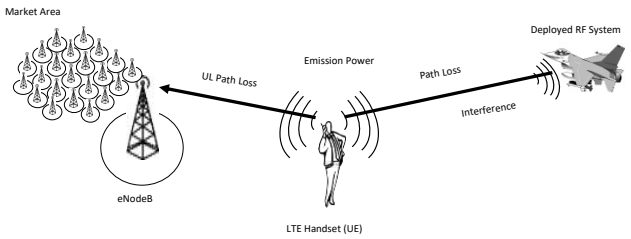

Fig. 1. Interference prediction situations of interest.

The experienced interference is an aggregate of all the interference contributions from individual sectors in the network. An LTE sector consists of a particular LTE base station (eNodeB) radio that is deploying a logical "cell" at a particular frequency as well as all the UEs that are being serviced by the

This technical data deliverable was developed using contract funds under Basic Contract No. W56KGU-18-D-0004.

deployed cell. To accurately predict the aggregate UL interference, it is necessary to develop accurate models for the UL emission powers in real-world LTE sectors. One way to do this is to create accurate distributions for the Effective Isotropic Radiated Power (EIRP) of a typical UE's emission from a sector. The UE EIRP captures the effective interference power at the source. The interference experienced at the SDS will depend on the pathloss between the UE and the Federal SDS as well as the hardware characteristics of the SDS, but those factors must be tailored toward specific operational environments.

In an operational cellular network, the spatial distribution and temporal behaviors of UEs influence the UE EIRP distribution of each sector. Detailed information about the behavior and distribution of UEs in individual sectors is difficult to obtain. Instead, aggregate statistics collected by eNodeBs are leveraged to infer trends across the network. Specifically, a small set of nominal UL pathloss distributions can be used to approximate the diverse UE distribution and behavior within the network. Each LTE sector is assigned one of the nominal UL pathloss distributions based on sector configuration characteristics.

In this work, we developed an unsupervised clustering process based on the Gaussian Mixture Model (GMM) method to determine an optimal set of nominal UL pathloss distributions. The clustering process leverages a dataset consisting of Key Performance Indicator (KPI) data collected from diverse AWS cellular markets within the contiguous United States (the multi-market dataset). We also developed a method to assign nominal UL pathloss distributions using only sector configuration information. The sector configuration information available in the multi-market dataset included eNodeB locations, antenna heights, antenna elevation and azimuth beamwidths, and effective sector radii. The census bureau land use morphology associated with each eNodeB was computed from the eNodeB location.

The Gaussian Mixture Model method is an unsupervised clustering method that assumes observed data has been generated by the weighted sum of two or more Gaussian processes. In this study the observed data consists of M-dimensional vector representations of UL pathloss occurrences within predefined ranges. Generating a GMM requires solving a nonlinear optimization problem where the likelihood function of the data given the model is maximized as a function of the $C(1 + M + M(M+1)/2)$ model parameters, consisting of C mixture coefficients, the C M-dimensional mean vectors, and the C $M \times M$ covariance matrices, each of which, due to symmetry, have $M(M+1)/2$ parameters. An iterative solution to this optimization problem was obtained using an Expectation Maximization (EM) algorithm. In formulating the GMM procedure, one additional parameter, the number of classes (nominal pathloss distributions) used to represent the data set, must also be determined. In this study, the optimal number of classes was determined using the Bayesian Information Criterion (BIC). The BIC quantifies each model's ability to fit the data (where a model in this case is a GMM solution for a specific number of classes) as a function of model complexity with the optimal model corresponding to the lowest BIC value. A five class GMM yielded the lowest BIC value for the multi-market UL pathloss data.

The performance of the resulting GMM model to predict specific sector pathloss distributions and to compare its performance relative to other nominal class determination processes was quantified by computing the mean square of the difference between each sector's measured UL pathloss distribution and the UL pathloss distribution of the GMM class to which it belongs (i.e., the mean pathloss distribution for all sectors in that class). Comparisons were made between the predictive performance of the GMM and other pathloss assignment methods in which the UL pathloss classes were based on either a rural or urban morphology designation of a sector (identified from US Census Bureau designations) and a method based on effective sector radius. In this study, effective sector radius was defined as the 90% threshold of sector UE to eNodeB distance histograms contained in the multi-market database. Averaged over the entire multi-market data set, the GMM class assignment method was found to be superior, that is, the mean squared differences between the actual sector UL pathloss distributions and the UL pathloss distributions predicted by the urban/rural morphology or the 90% radius threshold were larger than those predicted by the GMM nominal UL pathloss distributions.

The second issue addressed in the study was how to assign nominal UL pathloss distributions to sectors based only on sector configuration information. The sector configuration information available in this study included the locations of all eNodeBs and information regarding the eNodeB antennas. The sector assignment method developed in this study computed an optimal mapping between the sector feature space and the GMM UL pathloss feature space. This method mapped sectors to a point in the 3-dimensional GMM UL pathloss feature space based on their location in an N-dimensional sector feature space using a regression approach. The resulting mapping procedure was able to predict UL pathlosses more accurately than methods based on each sector's urban/rural morphology designation.

II. Gaussian Mixture Model Clustering

Gaussian Mixture Modeling is an unsupervised clustering method that assumes the observed data, M-dimensional vectors of measured UL pathloss in this study, has been generated as the weighted sum of two or more Gaussian processes. The form of the assumed probability density function is given as:

$$\rho(\bar{x}) = \sum_{i=1}^{C} \alpha_i \mathcal{N}(\bar{x}\,;\, \bar{\mu}_i\, \Sigma_i)$$

where C is the number of components, $\bar{x}$ is an M-dimensional data vector, and α_i is the mixture coefficient/weight factor for component i.

The multivariate Gaussian probability density function for component i is given as:

$$\mathcal{N}(\bar{x};\, \bar{\mu}_i\, \Sigma_i) = \frac{1}{2\pi^{\frac{M}{2}}} |\Sigma_i|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\,(\bar{x} - \bar{\mu}_i)^T \Sigma_i{}^{-1} (\bar{x} - \bar{\mu}_i)\right\}$$

where $\bar{\mu}_i$ is the mean vector for component i, and Σ_i is the covariance matrix for component i.

This formulation allows the Gaussian Mixture Model process to be cast as a nonlinear optimization problem where the

likelihood function of the data given the model is maximized as a function of the $C(1 + M + M(M+1)/2)$ model parameters, comprised of C mixture coefficients, the C M-dimensional mean vectors, and the C $M \times M$ covariance matrices. An iterative solution to this optimization problem is obtained using an Expectation Maximization (EM) algorithm. In formulating the GMM procedure one additional parameter, the number of components that will be used to represent the data set of interest, must be determined. In this study the GMM optimization was performed for a range of components and the BIC was used to select the optimal number of components. The BIC allows selection among a finite set of data models, quantifying each model's ability to fit the data as a function of model complexity with the optimal/best model corresponding to the lowest BIC value

$$BIC = p\ln(N) - 2\ln(\hat{L})$$

where p is the number of model parameters.

Using a 3-dimensional representation of the pathloss distributions, described in more detail below, each Gaussian component required 9 parameters, 3 for the mean feature vector and 6 for the associated covariance matrix. In the BIC calculation, N is the number of data (pathloss distributions) samples and $\hat{L}$ is the likelihood function, $\hat{L} = \Pr\{x|\hat{\theta}, GMM\}$, where x are the observed data (pathloss distributions) and $\hat{\theta}$ are the model parameter values that maximize the likelihood function. The negative of the natural logarithm of the likelihood function ranges between:

$$0 \leq -\ln(\hat{L}) < \infty$$

with 0 representing absolute certainty that the model generated the observed data and ∞ representing absolute inability of the model to generate the observed data. Using the resulting optimized GMM the class membership of any data vector can be computed using a Bayesian estimation procedure.

For the problem of interest, rather than use the raw UL pathloss distributions that consisted of occurrences/fractions of UL pathlosses for 20 quantized dB bins, a dimensionality reduction process was used to improve the numerical stability of the GMM algorithm. Raw UL pathloss distributions were mapped to a lower (three dimensional) GMM feature space, consisting of the mean, variance, and skewness of the raw UL pathloss distributions. The resulting three-dimensional representation of the multi-market database pathloss data is seen in Fig. 2.

Using the BIC, five UL pathloss distribution components/classes was determined to be optimal for the full multi-market database, as indicated by the asymptotic behavior of the BIC versus number of proposed classes seen in Fig. 3. Using five classes the resulting representation of the multi-market UL pathloss distributions can be seen in Fig. 4, where each sector that belongs to a given class is rendered in the same color. Nominal UL pathloss distributions for each class were computed by averaging the UL pathloss distributions within a given class. These nominal UL pathloss distributions are shown in Fig. 5. The five nominal GMM pathloss distributions are shown along with the nominal pathloss distributions for the sectors labeled as urban (green curve) or rural (red curve) morphologies. Since existing methods of predicting SDS interference use a set of nominal EIRP distributions obtained from field measurements of sector emissions from locations with either rural or urban US Census Bureau designations, the nominal urban and rural UL pathloss distributions are included for comparison with the GMM derived nominal UL pathloss distributions.

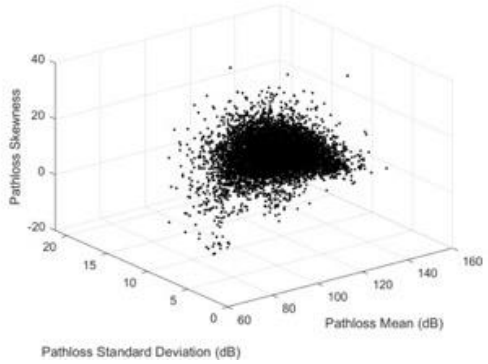

Fig. 2. Feature space representation of multi-market database UL pathloss distributions.

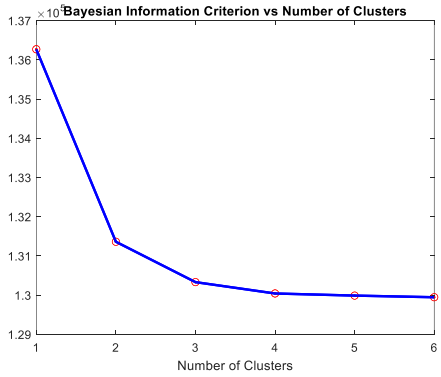

Fig. 3. BIC versus number of proposed clusters/clusters for the multi-market database UL pathloss distributions.

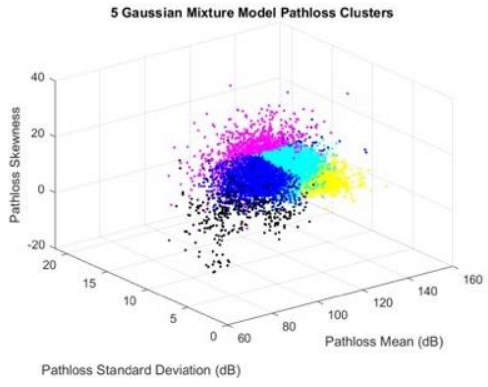

Fig. 4. Feature space representation of multi-market database UL pathloss distributions colored by GMM cluster class.

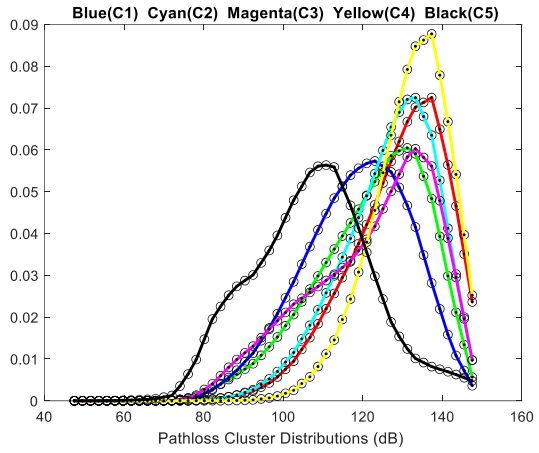

Fig. 5. Gaussian Mixture Model nominal UL pathloss distributions and nominal UL pathloss distributions for the sectors labeled as urban (green curve) or rural (red curve).

As expected, the nominal rural morphology sector UL pathloss distribution peaks at a higher value, approximately 138 dB, than the urban designated sectors, with a peak at approximately 130 dB. Larger pathloss values are more probable in rural sectors given the larger distances between base stations and UE in rural sectors as compared to urban sectors. The differences between the urban and rural nominal UL pathloss distributions are associated with corresponding differences in EIRP behavior.

The relative similarity of urban and rural nominal UL pathloss distributions is in stark contrast to the nominal UL pathloss distributions of the GMM classes, which exhibit significant variation. It is interesting to note that two of the GMM classes, the cyan (C2) and magenta (C3) curves in Fig. 5, produced nominal UL pathloss distributions very similar to the urban and rural nominal UL pathloss distributions. The remaining three nominal UL pathloss distributions correspond to sectors that have markedly different characteristics from the nominal urban and rural UL pathloss distributions. The GMM black (C5) and blue (C1) nominal UL pathloss distributions represent sectors that have significantly lower peak pathloss and exhibit "enhanced urban behavior" i.e., most UE being near the sector base station. The GMM yellow (C4) nominal UL pathloss distribution represent sectors that exhibit "enhanced rural behavior", with most UEs experiencing large pathlosses due to being far away from the base station.

The ability to predict a specific sector pathloss distribution using the nominal UL pathloss distributions produced by the GMM process was quantified by computing the mean squared difference between the nominal UL pathloss distribution for a given class, i.e., the mean pathloss distribution for all sectors in that class, and the actual pathloss for a given sector. The resulting UL pathloss error as a function of pathloss is the green curve in Fig. 6. To evaluate the GMM sector class assignment method relative to a method that assigns a nominal UL pathloss to each sector based on either a rural or urban morphology designation of the sector, the error function for the urban/rural sector assignment method was calculated and plotted as the red curve in Fig. 6. In addition, the performance of a sector assignment method based on an effective sector radius is seen in the black curve in Fig. 6. In this specific example the effective sector radius was defined as the 90% threshold for the sector UE to base station distance histograms that were included in the multi-market database.

As previously mentioned in our approach, predicting the interference produced by a proposed cellar network is a two-step process. The first step is class assignment where each sector is assigned one of the nominal UL pathloss distributions based on available characteristics of the sector. The second step uses the UL pathloss distributions to predict the EIRP and subsequently determine the interference level at the SDS. In this two-stage process two figures of merit are of interest. The first, presented as our Fig. 6. error curves, quantifies how well the sector assignment methods predict individual sector UL pathloss distributions given detailed information about network operation such as sector UE to base station distance histograms and UL pathloss distributions. The distance histogram and pathloss distribution data was available in the multi-market database. While this data can only be known after a cellular network has been deployed and operated for a considerable period of time, comparing the predictive performance of various sector assignment methods can provide insight into which sector assignment process has the potential to be most effective in operational settings where less information is available to perform the sector assignment process.

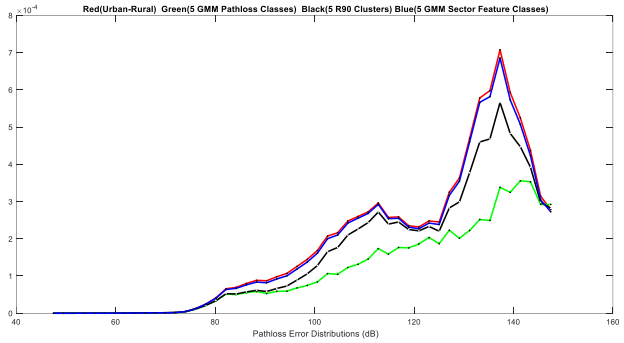

Fig. 6. Pathloss distribution prediction errors versus pathloss, urban/rural assignment process (red curve), effective sector radius assignment process (black curve), GMM class assignment process (green curve), assignment using sector feature space to GMM UL pathloss feature space regression mapping method (blue curve).

For the entire multi-market data set, the GMM class assignment (green curve in Fig. 6) was superior. The mean squared differences between the actual sector pathloss distributions and the pathloss distributions predicted by the urban/rural morphology class assignment method (red curve in Fig. 6) and the 90% radius threshold assignment method (black curve in Fig. 6) were both larger for all pathloss values than those of the nominal UL pathloss distributions predicted by GMM.

III. Sector GMM Class Membership Prediction from Sector Feature Information

For a given set of nominal UL pathloss distributions the outstanding issue is how to optimally assign one of the nominal

UL pathloss distributions to a sector using only sector configuration information. The sector configuration information considered in this study included the locations of all eNodeBs and information regarding the eNodeB antennas. To accomplish this assignment, a method that directly computed an optimal mapping between a given sector feature space and the GMM pathloss feature space is investigated. This method directly maps sectors to a point in the 3-dimensional (mean, variance and skewness) GMM pathloss feature space based on an N-dimensional sector feature space vector. In our studies a regression approach is utilized. For a given set of functional relationships between the sector feature space of interest and the 3-dimensional GMM feature space, the problem of determining the optimal set of weights that map the sector feature space to the corresponding point in the 3-dimensional GMM feature space is formulated as an overdetermined linear regression problem. For example, if 4 functions of the sector feature vectors are used in a case with k sectors, the mapping between the sector feature space and the GMM feature space would be:

$$
\begin{aligned}
G_1^1 &= C_1 f_1^1 + C_2 f_1^2 + C_3 f_1^3 + C_4 f_1^4 \\
G_1^2 &= C_5 f_1^1 + C_6 f_1^2 + C_7 f_1^3 + C_8 f_1^4 \\
G_1^3 &= C_9 f_1^1 + C_{10} f_1^2 + C_{11} f_1^3 + C_{12} f_1^4 \\
&\cdots \\
&\cdots \\
G_k^1 &= C_1 f_k^1 + C_2 f_k^2 + C_3 f_k^3 + C_4 f_k^4 \\
G_k^2 &= C_5 f_k^1 + C_6 f_k^2 + C_7 f_k^3 + C_8 f_k^4 \\
G_k^3 &= C_9 f_k^1 + C_{10} f_k^2 + C_{11} f_k^3 + C_{12} f_k^4
\end{aligned}
$$

where: G_i^j is the j^{th} GMM feature for sector i, and f_i^l is the l^{th} function of the sector features for sector i. The functions of the sector features could be just the "raw" sector features themselves, products of any number of "raw" sector features, or in fact any non-linear function of any of the "raw" sector features. In the matrix form of this mapping, the resulting least squares estimate of the regression coefficients and the corresponding estimate of the sector features in GMM feature space are given as follows:

$$
\begin{aligned}
F\underline{C} &= \underline{G} \\
\underline{\hat{C}} &= (F'F)^{-1}F'\underline{G} \\
\underline{\hat{G}} &= F\underline{\hat{C}}
\end{aligned}
$$

For the results in this paper, three sector features were considered: effective sector radius calculated using the physical antenna geometry, antenna azimuth beam width, and antenna height above mean sea level. The mapping is implemented as a 9-coefficient linear transformation between these three features and the mean, variance, and skewness of the UL pathloss distribution. The resulting mapping procedure is able to predict sector pathlosses (blue curve, Fig. 6) more accurately than the urban/rural morphology assignment method (red curve, Fig. 6).

IV. Market-by-Market Studies

To further quantify the performance of the GMM class membership sector assignment process, a set of market-by-market studies were conducted. The GMM feature space representation of the pathloss data on a market-by-market basis are shown in Fig. 7.

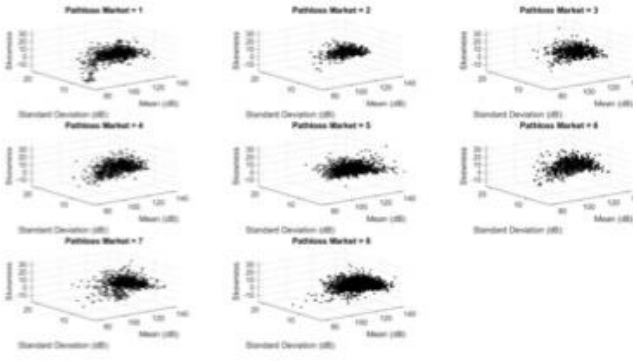

Fig. 7. Market-by-market feature space representation.

The corresponding nominal GMM class UL pathloss distributions, the sector feature class pathloss distributions and the resulting pathloss error prediction distributions are presented in Fig. 8, Fig. 9, and Fig. 10, respectively.

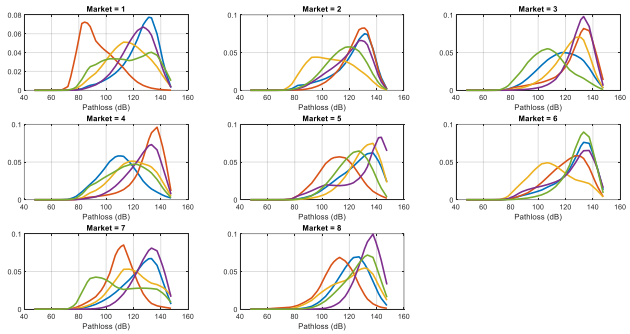

Fig. 8. Market-by-market nominal GMM class assignment UL pathloss distributions.

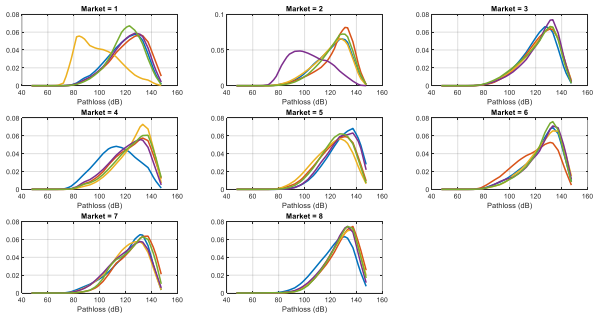

Fig. 9. Market-by-market enhanced regression mapping nominal UL pathloss distributions.

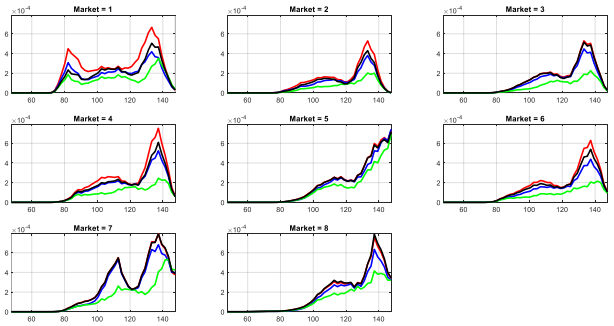

Fig. 10. Market-by-market pathloss prediction error distributions urban/rural(red), sector features(black), 90% radius threshold(blue), GMM(green).

In Fig. 8 the individual curves are the nominal GMM UL pathloss distributions for each market. As was the case for the entire multi-market analysis, GMM sectors with markedly different UL pathloss distributions exist both within a given market and in comparison between different markets. The sector feature UL pathloss distributions seen in Fig. 9 were obtained using an enhanced regression method that scaled the set of mapped sector feature GMM vectors such that their second order statistics equaled the second order statistics of the corresponding target GMM vectors. In this method the mean vectors and covariance matrices for the original GMM feature space data and the sector feature space mapped GMM data were computed as: $\{\mu_{\underline{G}}\ R_{\underline{G}}\ \mu_{\underline{\hat{G}}}\ R_{\underline{\hat{G}}}\}$ .The corresponding eigenvectors and diagonal eigenvalue matrices for the covariance matrices were calculated: $\{\varphi_{\underline{G}}\ \Lambda_{\underline{G}}\ \varphi_{\underline{\hat{G}}}\ \Lambda_{\underline{\hat{G}}}\}$. Each mapped sector feature GMM vector, $\underline{\hat{G}_k}$, was normalized via the following procedure:

$$\underline{\tilde{G}_i} = \Lambda_{\underline{G}_i}^{\frac{1}{2}} \varphi_{\underline{G}_i}^T \left[\Lambda_{\underline{\hat{G}}_i}^{-\frac{1}{2}} \varphi_{\underline{\hat{G}}_i}^T \left(\underline{\hat{G}_i} - \mu_{\underline{\hat{G}}_i}\right) + \mu_{\underline{G}_i}\right.$$

In computing the market-specific scaled GMM UL pathloss feature vectors using the enhanced regression method, additional sector features that where not used when analyzing the multi-market dataset as a single entity where included. These market-specific sector features were derived from the spatial distribution of base station locations within each market.

One set of sector features were derived from the Voronoi Diagrams associated with base station locations for each of the markets. In two dimensions, given a set of n seed points (base station locations in this study) Voronoi Diagrams represent the partition of the plane into n regions such that each point lies in only one region. As well, every point in each region will be closer, i.e., have a smaller Euclidian distance, to the seed point for that region than any of the other n-1 seed points. A representative Voronoi Diagram for one market from multi-market data base is seen in Fig. 11. The areas of the Voronoi cell for each base station were used as sector features in the market-specific analysis.

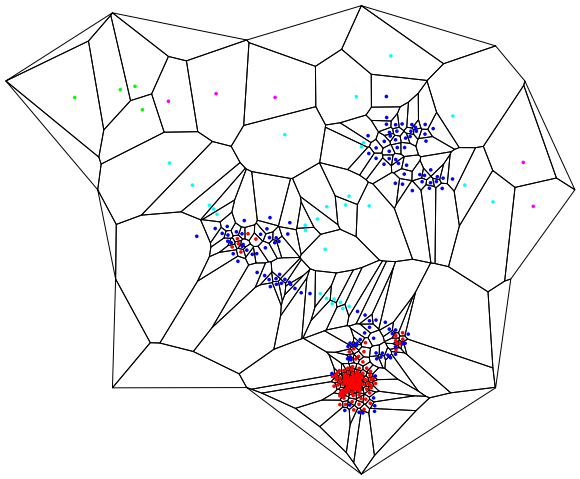

Fig. 11. Multi-market Voronoi partition example. Dots are base station cell tower locations color coded by sector feature determined GMM class. Black polygons define corresponding Voronoi cells.

An additional set of market-specific sector features were derived from the base station nearest neighbor distance distribution. Plots of the nearest neighbor distance distribution for the 10 nearest neighbors of each base station for the 8 markets are seen in Fig. 12. A linear approximation of the K nearest neighbor distance distribution (K = 10) was computed and both the intercept and slope of those approximations were used as sector features.

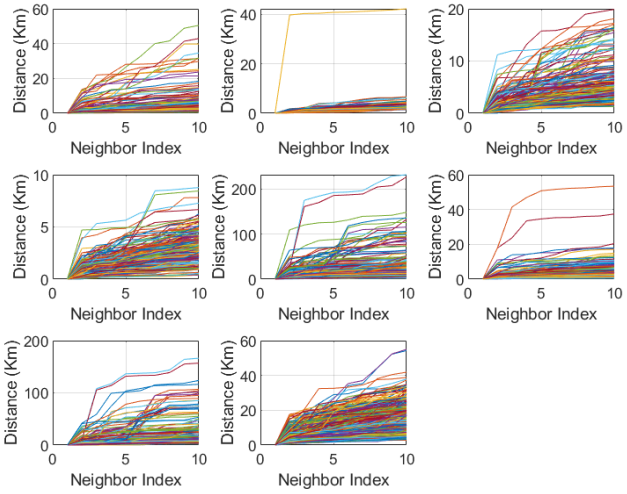

Fig. 12. K-Nearest neighbor distance distributions for K = 10 nearest neighbors of each base station on a market-by-market basis.

Although a complete analysis that would determine the best functional relationships between sector features and GMM feature space was not conducted, sector classifications were obtained using the following sector features: geometrical sector radius, sector antenna azimuthal beamwidth, Voronoi cell area, and K nearest neighbor distribution linear approximation slopes. These mappings produced the market-by-market nominal sector feature class pathloss distributions seen in Fig. 9. These estimated UL pathloss distributions are good approximations of

the exact/true GMM pathloss distributions seen in Fig. 8. In terms of the error plots seen in Fig. 10. these UL pathloss distributions (black curves in Fig. 10) were superior to sector UL pathloss distributions predicted by sector urban/rural morphology (red curves, Fig. 10).

V. Conclusion

In this work, the GMM method was applied to an extensive multi-market KPI dataset collected from deployed LTE networks throughout the contiguous United States to determine a set of five nominal UL pathloss distributions which capture the diversity of UE pathloss distributions in different LTE sectors. Averaged over the entire multi-market data, the GMM UL pathloss class assignments produced results superior to land use morphology based assignments. These UL pathloss distributions could be used to derive better LTE UE emission EIRP distributions for LTE UL interference modeling. Furthermore, this work also explored methods for assigning one of the GMM nominal UL pathloss distributions to a given sector using sector configurations. This is critical, as it enables the use of the GMM sector classes for making interference predictions regarding LTE networks that have not yet been deployed.

The mathematical techniques leveraged in this work can also be generalized to enable better data mining to support cellular emissions modeling for future spectrum repurposing studies. This will support higher fidelity predictive modeling of cellular interference, leading to more efficient spectrum sharing between commercial cellular and SDS.