Temporal Prediction of RF-EMF Exposure Using Surrogate Modeling With Environment-Based Clustering

Paul Lagouanelle, Shanshan Wang, Yukun Liu, Yarui Zhang, and Joe Wiart

Abstract – This article addresses RF electromagnetic field (RF-EMF) exposure modeling using data-driven surrogate models as an alternative to costly real-world measurements. It relies on data from large-scale EMF sensor networks deployed in urban environments. Although neural network-based models showed promising results, they are insufficient to capture complex spatiotemporal exposure. Therefore, alternative approaches such as Kriging and polynomial chaos are explored but require prior data clustering. This clustering is based on environmental complexity extracted from OpenStreetMap, using variables such as building density and urban occupation rate. Although current models achieve acceptable safety-level accuracy, further improvements are needed for high-precision exposure mapping and complete time series prediction.

1. Introduction

The RF electromagnetic fields (RF-EMF) from wireless networks have become an integral part of everyday life, with mobile connections now far exceeding the global population and smartphone usage reaching very high levels. Although international exposure guidelines [1] and standardized compliance testing methods exist, public concern about potential health risks from RF-EMF remains strong, particularly with the recent deployment of 5G networks. This has fueled societal debate and motivated research efforts aimed at better understanding mobile network characteristics, including signal propagation, coverage, and performance across different environments. The assessment of human exposure to RF-EMF can be performed through direct measurements that require many temporal samples or by simulation that needs a strong understanding of the system and its environment to be accurate. Thus, surrogate modeling is perfectly suited to accurately predict human exposure to RF-EMF because it can be quickly computed without the need for deep theoretical analysis.

This article aims to exploit data from a network of EMF sensors deployed across French cities [2]. These sensors designed by EXEM measure the spatial components of the electric field in the range of 250 kHz to 6 GHz. Every 2 h, the sensors are activated to conduct one 6 min measurement. Sensor networks are particularly useful for the temporal forecasting as well as spatial mapping of human exposure due to the large available data set of measurements they provide. Over recent decades, significant work has been devoted to RF-EMF exposure assessment and monitoring through these fixed probes and spot measurements using convolutional neural networks (CNNs) models. [3, 4]. Although particularly suited to process large volumes of data, CNNs have been efficient at estimating the spatial distribution and magnitude variations of electric field levels across multiple frequency bands but are still unable to independently predict time series data. Indeed, the amount of available training data is insufficient for building a neural network (NN) model for predicting both spatial and temporal variations of RF-EMF exposure. Thus, the work presented here aims to expand these recent developments to another surrogate model for the prediction of time series data: the polynomial chaos-based Kriging (PCK) metamodel [5]. Unlike CNNs, the PCK surrogate model used here requires smaller training data sets obtained through prior data clustering for enhanced prediction performance. Another advantage of PCK over NNs is its ease of use in performing rapid sensitivity analysis using Sobol indices [6].

2. Environment-Based Clustering

After studying several pure statistical approaches using the signal distribution to cluster the time series from the 106 sensors, the resulting metamodels were unable to properly approximate the temporal variation of the electric field on most sensors. As the temporal variation of the EMF is highly dependent on the environment surrounding the sensor, the most conclusive approach aims to directly exploit the environmental complexity to group data into disjoint training sets. A sensor's environment can be easily extracted using the collaborative tool OpenStreetMap, which allows data to be collected within a 400 m^2 around each sensor, including the number of

Manuscript received 17 April 2026.

This work was supported, in part, by the Horizon Europe Research and Innovation Programme project GOLIAT (grant 101057262).

Paul Lagouanelle, Shanshan Wang, Yukun Liu, and Joe Wiart are with Chaire C2M, Information Processing and Communications Laboratory (LTCI), Télécom Paris, Institut Polytechnique de Paris, 19 place Marguerite Perey, 91120 Palaiseau, France; e-mail: paul.lagouanelle@telecom-paris.fr, shanshan.wang@telecom-paris.fr, yukun.liu@telecom-paris.fr, joe.wiart@telecom-paris.fr.

Yarui Zhang is with Laboratory of Systems and Applications of Information and Energy Technologies (SATIE), ENS Paris-Saclay, CNRS, Université Paris-Saclay, 4 avenue des Sciences, 91190 Gif-sur-Yvette, France; e-mail: yarui.zhang@ens-paris-saclay.fr.



(a) Bordeaux_01 sensor
n_b = 2186 *OCC* = 0.2163



(b) Saint-Louis-de-Montferrand_01
sensor n_b = 116 *OCC* = 0.0377

Figure 1. Example of two different sensor environments in the Bordeaux region (buildings are shown in red).

buildings and for each building: its average height, its surface area, and its function (e.g., residential, commercial, religious, agricultural). After an in-depth analysis of these environmental data, two environmental variables were chosen to partition the data: the number of buildings n_b and the occupancy rate of the urban space in volume:

$$\mathrm{OCC} = \frac{\sum \text{Building Volume}}{400\ \mathrm{m}^2 \times \text{Max Building Height}} \tag{1}$$

An example of two different environments within the Bordeaux region with the associated environmental variables is shown in Figure 1: the highly urbanized center of the city of Bordeaux and a rural village.

On the basis of these two environmental variables, each sensor can be associated with one of the five clusters defined by a classical k-means method that has been proven quite efficient for partitioning data from various urban sources [7]. As observed in Figure 2, the two previous environments are clearly reflected in two

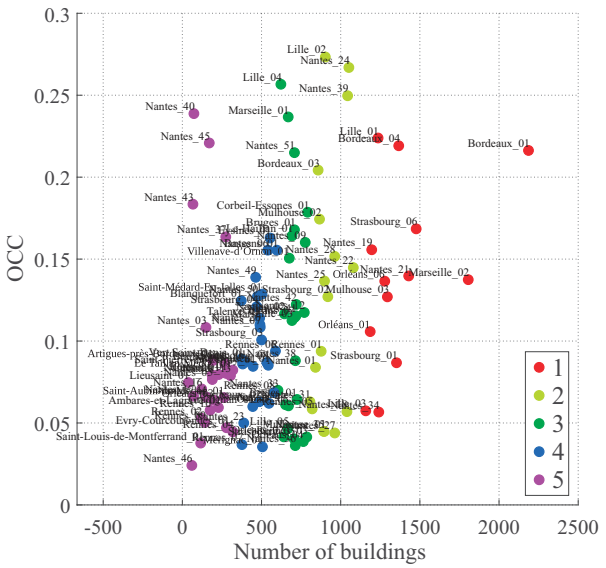


Figure 2. Partitioning of the 106 sensors into five different clusters according to the occupancy rate OCC and the number of buildings around each sensor.

very different clusters (1 and 4). However, this graph clearly shows that the clusters seem complex to define because many data pairs are close to each other. One could imagine mixing distribution-based and environmental-based clustering by adding extra layers for computing clusters on the basis of the mean or the variance. However, due to the high influence of the environment on the moments of the distribution, this does not provide any improvement to the existing clusters of Figure 2.

3. Surrogate Modeling

On the basis of the clusters defined in Figure 2, sensors from each cluster are separately used to train five different surrogate models over a short time period: December 9, 2022, to January 9, 2023. For each sensor, the training data set consists of

$$\text{Input: } \left[n_b, \mathrm{OCC}, t_{\text{date}} = \begin{pmatrix} \text{09 Dec 2022 00h01} \\ \cdots \\ \text{09 Jan 3 23h59} \end{pmatrix} \right] \tag{2}$$

$$\text{Output: } E_{Vm} = \begin{pmatrix} E_m(\text{09 Dec 2022 00h01}) \\ \cdots \\ E_m(\text{09 Jan 2023 23h59}) \end{pmatrix} \tag{3}$$

Although each sensor provides a time series made of 384 samples for the training, the amount of data points is insufficient for the use of classical NN surrogate models. Hence, to build an exact interpolator, the chosen surrogate model combines the advantages of the well-known Kriging and chaos polynomial surrogate models: the PCK metamodel. The Kriging part is used to interpolate local variations in the model output, whereas the

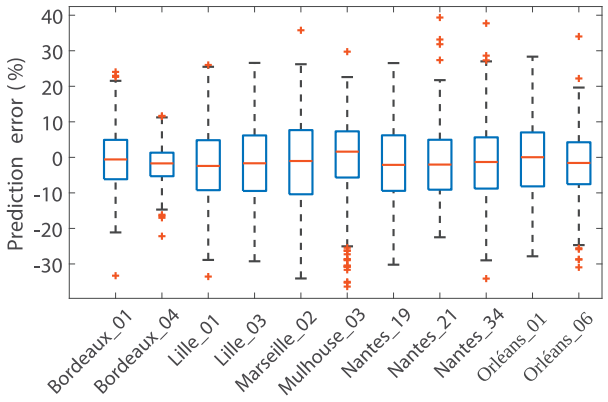


Figure 3. Box plots of prediction rates of the CV surrogate models in the first cluster. This plot shows the median (black line), the IQR (box: 25th to 75th percentiles), and whiskers extending to 1.5 × IQR, with outliers beyond this range shown as red dots.

polynomial chaos expansion approximates the overall behavior of the system. A PCK metamodel is defined by [5]

$$\mathcal{M}(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} y_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) + \sigma^2 Z(\boldsymbol{x},\omega) \tag{4}$$

where $\sum_{\boldsymbol{\alpha}\in\mathcal{A}} y_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x})$ is a weighted sum of orthonormal polynomials (the trend of the PCK) and σ^2 is the variance of the stationary Gaussian process $Z(\boldsymbol{x},\omega)$.

To assess the PCK metamodel performance, leave-one-out cross-validation (CV) surrogate models have been built among the different clusters: for each sensor, a surrogate model is trained using the other sensors belonging to its cluster and validated using the sensor time series. The evaluation metrics used to assess model performance are the root mean square error (RMSE) and mean absolute scaled error (MASE), which are defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(E_{\mathrm{pred}}[i] - E_{\mathrm{true}}[i]\right)^2} \tag{5}$$

and

Table 1. RMSE and MASE for the CV surrogate models in the first cluster

Sensor	RMSE	MASE (%)
Bordeaux_01	0.184	58.1
Bordeaux_04	0.0385	52.7
Lille_01	0.0716	68.6
Lille_03	0.0912	70.3
Marseille_02	0.0943	72.9
Mulhouse_03	0.229	66.6
Nantes_19	0.0952	69.5
Nantes_21	0.0935	69.7
Nantes_34	0.100	66.3
Orléans_01	0.226	65.7
Orléans_06	0.0693	64.2

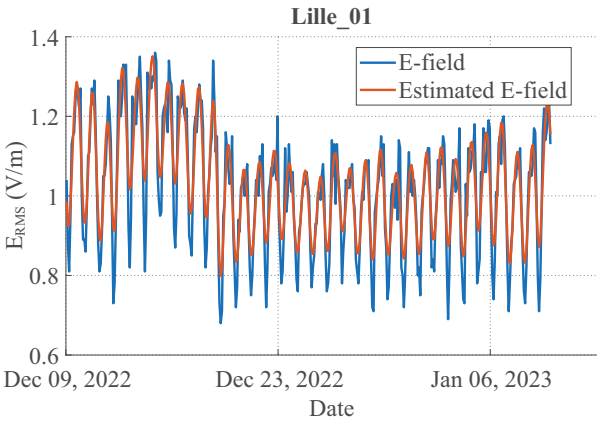


Figure 4. Estimated (blue) and true values (red) of the EMF variation for the Lille_01 sensor (RMSE = 0.0716 and MASE = 68.6%).

$$\mathrm{MASE} = \frac{\frac{1}{N}\sum_{i=1}^{N} |\, E_{\mathrm{pred}}[i] - E_{\mathrm{true}}[i] \,|}{\frac{1}{N-1}\sum_{i=2}^{N} |\, E_{\mathrm{true}}[i] - E_{\mathrm{true}}[i-1] \,|} \times 100\% \tag{6}$$

where $i \in [1, N]$ is the sample index of E_{pred} (the predicted total electric field) and E_{true} (the true total electric field).

4. Results

The CV results for the first cluster, consisting of 11 sensors, are presented in Figure 3, which shows the median (black line), the interquartile range (IQR; box: 25th to 75th percentiles), and whiskers extending to 1.5 × IQR, with outliers beyond this range shown as red dots. Table 1 displays the MASE and RMSE for each CV surrogate model.

When looking at the validation metrics, we find that their values are sufficient to define the safety of the environment for human exposure for each sensor (an example is displayed in Figure 4 for estimating electric field values for the Lille_01 sensor). Thus, the global metamodel built on the full cluster with 11 sensors is able to accurately predict the temporal variation of any environment within the cluster defined by n_b and the OCC.

The same work can be performed on the four other clusters defined in Figure 2. Figure 5 summarizes the results for the five surrogate models with MASE CV errors on each sensor. The color scale displays the MASE errors in the plane (n_b, OCC), ranging from green values indicating excellent prediction values (as low as 18% and up to 50%) to sufficient values in black (between 50% and 100%), with poor predictions being marked as red and limited to 200% on the scale.

The promising results from the first cluster in predicting the time series appear with even better results in most of Cluster 5. Although the prediction accuracy

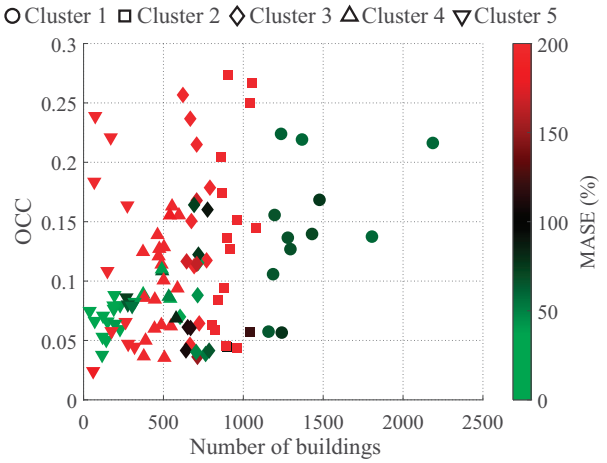


Figure 5. MASE CV errors among the different sensors. This plot shows sensors with good predictions (green), mediocre predictions (in black), and poor predictions (in red) according to the MASE values.

is quite poor for some sensors, similar environments are well predicted, with some MASE errors ranging around 20%, for example, in the centered area of Cluster 5, which consists of sensors placed in rural areas with a few small buildings in the countryside or on the outskirts of the cities.

However, for the other three clusters, the time series is only poorly predicted. Indeed, when looking at the global distribution of the MASE errors displayed in Table 2, we find that the only cluster with excellent prediction results is Cluster 5, which is trained on data from extremely similar environments.

In other clusters, due to the variety of environments considered, the same conclusion from the validation metrics on the first cluster cannot be directly drawn. Indeed, the occupancy rate and the number of surrounding buildings are not sufficient to grasp the whole environmental complexity. Some similarities have been observed between certain locations where the model did not accurately perform. Notably, these include small and narrow streets in city centers, open areas inside city centers, and the presence of one or several buildings (such as train stations, schools, universities, or administrative buildings) with many people passing by during the day.

The general exposure limits defined by the International Commission on Non-Ionizing Radiation Protection (at 400 MHz $\approx$ 27.5 Vm^{-1} at the lowest) are extremely high compared with the worst prediction of the surrogate models, even with a poor model, the predicted field falls under the limit. Thus, the surrogate models are acceptable enough to define the safety of human exposure to RF-EMF on the time period considered on only half of the sensors (45 of 106 sensors).

5. Conclusion

This article shows the great potential of combining environment-based data partitioning with PCK surrogate modeling to predict the time series of RF-EMF exposure measured by sensors in various urban environments. Environment-based clusters have been defined on the basis of the number of buildings and the occupancy rate in volume around each sensor to split data into five groups used to build five different metamodels. Although some sensors are completely out of the predictive range, almost half of the sensors have been well predicted for exposure studies with a naive k-means clustering on a simple environmental analysis.

Table 2. MASE performance distribution per cluster

Cluster	Good (<50%)	Decent (50% to 100%)	Poor (>100%)
1	0	12	0
2	0	1	16
3	4	5	16
4	4	2	19
5	14	3	10
Total	22	23	61

Many improvements will be explored in the future on both the surrogate modeling and the clustering strategy. In particular, additional environmental variables are needed to better distinguish environments of the poorly predicted clusters. When observing some inaccurate sensors, especially in city centers, we find that some city structures are very similar from one location to another. Notably, the street network has a direct influence on the exposure levels, as do vegetation and open areas such as parks. Those structural organizations could be analyzed to create one or more additional variables that would help to define sparser clusters, thus improving both the computation time of the surrogate models and their prediction performance.

Finally, because most of the computation time for the prediction is based on the surrogate model computation (from a few minutes to several hours, depending on the cluster size), an alternative, instead of defining clusters, is to estimate new areas with surrogate modeling built only from their nearest neighbors in the environmental analysis parameter space, which could greatly decrease the construction time of these metamodels to minutes in most cases. This could enable the use of such predictive models on a day-to-day basis for RF-EMF exposure.

6. References

1. International Commission on Non-Ionizing Radiation Protection, "Guidelines for Limiting Exposure to Electromagnetic Fields (100 kHz to 300 GHz)," *Health Physics*, **118**, 5, May 2020, pp. 483-524.
2. O. Jawad, E. Conil, J.-B. Agnani, S. Wang, and J. Wiart, "Monitoring of the Exposure to Electromagnetic Fields With Autonomous Probes Installed Outdoors in France," *Comptes Rendus Physique*, **25**, S1, May 2024, pp. 41-61.
3. S. Wang, T. Mazloum, and J. Wiart, "Prediction of RF-EMF Exposure by Outdoor Drive Test Measurements," *Telecom*, **3**, 3, June 2022, pp. 396-406.

4. Y. Zhang, S. Wang, and J. Wiart, "ExposNet: A Deep Learning Framework for EMF Exposure Prediction in Complex Urban Environments," *IEEE Transactions on Machine Learning in Communications and Networking*, **4**, July 2026, pp. 1229-1243.
5. R. Schöbi, B. Sudret, and J. Wiart, "Polynomial-Chaos-Based Kriging," *International Journal for Uncertainty Quantification*, **5**, 2, May 2015, pp. 171-193.
6. T. Crestaux, O. Le Maître, and J.-M. Martinez, "Polynomial Chaos Expansion for Sensitivity Analysis," *Reliability Engineering & System Safety*, **94**, 7, July 2009, pp. 1161-1172.
7. L. T. T. Nguyen, T. T. D. Nguyen, Q.-T. Bui, and B. Vo, "Geospatial Data Clustering in Network Space: A Survey," *WIREs Data Mining and Knowledge Discovery*, **15**, 2, May 2025, p. e70023.