Outlier Detection and Spectrum Feature Extraction Based on Nearest-Neighbors Correlation and Random Forest Algorithm

1st Rodney Martinez Alonso
ESAT - WaveCore - Networked Systems
KU LEUVEN
Leuven, Belgium
ORCID: 0000-0003-2529-5944

2nd David Plets
INTEC - WAVES
imec / Ghent University
Ghent, Belgium
ORCID: 0000-0002-8879-5076

3rd Sofie Pollin
ESAT - WaveCore - Networked Systems
KU LEUVEN
Leuven, Belgium
ORCID: 0000-0002-1470-2076

4th Luc Martens
INTEC - WAVES
imec / Ghent University
Ghent, Belgium
ORCID: 0000-0001-9948-9157

5th Wout Joseph
INTEC - WAVES
imec / Ghent University
Ghent, Belgium
ORCID: 0000-0002-8807-0673

***Abstract*—Most spectrum surveys conducted worldwide demonstrate that the radio-electric spectrum in use at any given location and instant of time is below 25%. Current spectrum management policies and spectrum utilization inefficiency is becoming unsustainable for future development of radio technologies and services. In this context, dynamic spectrum access is a promising technique for improving spectrum utilization efficiency. A key scientific gap is identifying inaccurate spectrum data from hidden nodes that is not homogeneously distributed in the spatial domain and dynamically vary in time and frequency. For bridging this gap, our paper presents the research results of a spectrum feature extraction algorithm based on multi-correlation and Random Forest. Our algorithm is capable of estimating the spectrum utilization pattern in the spatial and frequency domain with a minimum reliability of 92% for a real heterogeneous networking scenario.**

***Index Terms*—Spectrum Sensing, Spectrum Sharing, Spectrum White Spaces, Machine Learning, Random Forest**

I. Introduction

The radioelectric spectrum is a limited resource and the most important utility for wireless networks. The traditional assignment of spectrum has lead to bottlenecks and delays for allocating spectrum sub-bands to new services and operators. For instance in the case of Europe by early 2022 only 56% of the 5G harmonized spectrum was allocated [1]. This negative trend is having an impact also on the energy efficiency of wireless networks and the operators' cost-efficiency. Paradoxically, spectrum surveys performed worldwide demonstrate that the spectrum utilization efficiency, as the ratio between the allocated spectrum and the spectrum really in use at any given location and instant of time, is lower than 25% [2], [3] and below 11% in rural areas [4]. In this context next generation technologies (e.g., 6G) will require a more sustainable and efficient paradigm for spectrum management and allocation based on Dynamic Spectrum Access (DSA) [5].

For guaranteeing multi-band coexistence, cognitive radio technologies must evolve to an intelligent enabled network capable of determining the spectrum features in the spatial, temporal, and frequency domain [5]. Despite the advances on sensing and estimation, traditional cognitive radio relies on distributed decision mechanisms based on data that is not always reliable because of environmental conditions or across multiple use cases. For instance, hidden nodes (i.e., locations in the shadow from at least one transmitter) have a critical impact on interference [6]. Research on cooperative spectrum sensing have addressed this issue [7]–[9]. However, most algorithms presented in the literature allocate the spectrum based on a per-channel and per-location decision-making mechanism [7], [8], [10]. This kind of mechanism might not perform well in heterogeneous networking scenarios and requires some heuristics for multi-band assessment, increasing their computational complexity. In addition to the hidden node effect, when the signal level is near the decision threshold for allocating the spectrum, algorithms based on a per-channel estimation accounting only for the total channel power, lose accuracy as they lack a mechanism for correlation in the frequency domain. Generally, beside a signal level pattern between neighboring locations, within the channel bandwidth the power allocation follows a pattern in the frequency domain (e.g., OFDM signals).

We consider that spectrum fingerprinting by machine learning based on the correlation of spectrum vectors that include the features of multiple frequencies (i.e., signal levels as a function of frequency), will jointly lead to a more efficient use of the spectrum and reducing the interference probability. Rather than a channel-by-channel estimation, we consider

VLIR-UOS Project: UOSTEA2022001201

the whole spectrum as a fingerprint for each location. In this paper we present a classifier learning algorithm based on Random Forest for identifying outliers. In the context of our research, an outlier corresponds to inaccuracies in the spectrum data. The outcome of this algorithm is used for extracting the spectrum utilization features considering data from a real heterogeneous network scenario. For evaluating the algorithm, we performed a large-scale spectrum survey in a real heterogeneous networking scenario collecting 69,182 signal level samples in the spectrum between 170 and 1000 MHz from a total of 71 locations, and considering two different experimental setups. The novelty of our method is that we simultaneously consider the spatial and frequency domain in a multi-channel vector for the extraction of the spectrum features in order to achieve a higher estimation accuracy.

II. Method

The dynamic spectrum access by cognitive radio devices is unreliable when the allocation decisions are fully distributed and independently taken by each BS based only on the reported information from their end-terminals. Our aim is validating the hypothesis that up to a certain spatial distance the spectrum occupancy is correlated enough for inferring a certain pattern based on the data from multiple neighbors. In this context hidden nodes play a critical role in false spectrum allocation when some end-devices report inaccurate data due to the effect of shadowing [11]. For improving the trade-off between spectrum occupancy and interference [3], we present a method based on the Random Forest algorithm in order to identify outliers from the spectrum dataset. Locations reporting inaccurate data (outliers) follow a different spectrum pattern compared to their neighbors. Their trend is to report lower signal levels on certain frequencies. If this data is not identified in the spectrum dataset a false detection of white spaces occurs (i.e., spectrum wrongly identified as not occupied), maximizing the interference probability.

First, we performed a spectrum survey in a rural scenario for collecting the signal levels as a function of the geo-locations and sensed frequencies (Subsection II-A). Second, for a certain given maximum distance between locations we find the spatial correlation between the spectrum vectors (signal levels as a function of the frequency) and the Cumulative Distributed Function (CDF) of the correlations (Subsection II-B). Finally, we tag the Spectrum Occupancy as a function of the measured signal levels for each collected sample in the survey. A Random Forest algorithm is applied for finding outliers between samples in the spatial and frequency domain in order to improve the estimation of spectrum utilization patters for DSA applications (Subsection II-C).

A. Scenario and Experimental Measurement Setup

For the spectrum measurements we consider a rural scenario in Nevele, Belgium. This is a mostly flat area with detached isolated houses and farms. We defined a grid of different locations for measuring the signal levels [dBm] across the spectrum between 170 MHz and 1000 MHz, during a period of 0.5 h [4]. The measurements across the different locations are not synchronized and therefore not correlated in the time domain. The maximum signal fading recorded during the time of the measurements was approximately 5 dB. Therefore, at each measurement location we account for the 99^{th}-percentile of the signal levels through time. This is done for accounting for the maximum levels and removing the temporal variations caused by fading.

TABLE I
Measurement Setups

Parameter	Setup 1	Setup 2	Unit
Number of Locations Nloc	50	21	
Sweep Band	170-1000	170-600	MHz
		600-1000	MHz
Resolution Bandwidth	100	3	kHz
Spectrum Vector Frequency Resolution	1.32	0.68	MHz
Distance Resolution	1	0.5	km
Signal Level Percentile	99	99	
Noise Floor	-112	-116	dBm

Some settings of the measurement setup have an important impact on the results. In particular those parameters related to the noise floor might impact the final outcome of the experiment. For instance, we have found that when the noise floor is too close to the decision threshold for the spectrum allocation or the recorded signal levels, most machine learning algorithms will not provide any accurate estimation as it will learn the signal noise rather than extracting any signal feature. For this reason we defined two setups for the measurements. Table I lists the settings for each measurement setup.

In the first setup (*Setup 1*), each measurement location in the grid is located at a distance of approximately 1 km (spatial resolution), while for the *Setup 2* the spatial resolution is 0.5 km. For reducing the noise floor in the second setup we paralleled the measurements by splitting the total evaluated bandwidth in two segments in order to reduce the resolution bandwidth of the measurement device (*FSH8 R&S*). For this, we considered the characteristics of the services according to the regulator assignment. In the sub-600 MHz sub-bands there is a predominant High-Power High-Tower infrastructure of broadcasting services and above 600 MHz the predominant services are Low-Power Low-Tower from mobile infrastructure. The output from the measurement campaign is a matrix containing the geolocation of each measurement location in the grid (*GPSx [°]; GPSy [°]*), the sensed frequency (*Freq [MHz]*) and the signal level corresponding to the setup defined percentile across the time domain (*L [dBm]*).

B. Spectrum Vectors: Correlation and Fingerprinting

From the collected data at each location, we define a set of spectrum vectors S of signal levels as the function of the sampled frequency for any given location. Therefore the spectrum vector of a certain location i can be defined as:

$$S_i = \{Li_{f_1}; Li_{f_2}; ...; Li_{f_n}\} \quad (1)$$

where L_i is the 99^{th}-percentile of the signal level measured for the frequency f_n at the location i.

For the whole dataset we find the correlation between all the spectrum vectors that are within a given maximum distance d_{max} of 0.5 km or 1 km. The distance d_{ij} between any given pair of locations i and j for including them in the same cluster of radius d_{max} is calculated based on their geolocation and the haversine function. The correlation coefficient (ρ) between the spectrum vectors (S_i, S_j) of any pair of locations within d_{max} is defined by Equation 2:

$$\rho(S_i, S_j) = \frac{\mathbb{E}\left[(S_i - \mu_{Si}) \cdot (S_j - \mu_{Sj})\right]}{\sigma_{Si} \cdot \sigma_{Sj}} \tag{2}$$

where $\mathbb{E}$ is the expectancy (weighted mean), μ is the mean, and σ is the standard deviation. In this way the correlation is not referred to the deviation of the mean but a weighted mean where a group of values might be more representative (higher weight) than others in the vector.

Further a cumulative distribution is applied to the obtained correlations for all the possible combinations of locations. This allows estimating a spectrum fingerprinting if the spectrum correlation is strong enough for most of the sub-clusters (i.e., $CDF > 90\%$). Notice that if the data within any given sub-cluster is not enough correlated a Random Forest algorithm might retrieve random outcomes as it is not possible defining a spectrum utilization pattern within any given area. For a best case, if the spectrum vectors within a cluster are not strongly correlated the number of trees for finding a pattern will be too high, leading to an exponential decrease of the algorithm computational performance.

C. Outlier Detection and Estimation of Spectrum Occupancy Pattern

The end goal of our algorithm is finding the data from certain measurement locations, that based on the pattern identified by an absolute majority of their neighbors, seems to be inaccurate (i.e., outliers). For this we will use the classification features of the Random Forest algorithm for finding the outliers. This is performed by estimating and further quantifying the *distance* between any data sample on any given decision tree (proximity). For instance, if it seems that two samples belong to the same *leaf* on a certain percentage of the decision trees it is assumed that the two samples are identical or similar enough. Once we are able to find any outlier in both the spatial and frequency domain we can use the outliers as a mask for filtering the initially estimated Spectrum Occupancy from the recorded signal levels at each independent location. Algorithm 1 defines the proposed algorithm for finding the outliers and estimating the spectrum occupancy pattern.

The algorithm includes as inputs the geo-location of each measurement ($GPSx; GPSy$) and each sensed frequency (*Freq*). The output Y is a binary tag defining the spectrum occupancy (*SpectrumO*) based on the individual signal levels L for each given input X (combination of $GPSx; GPSy; Freq$). Equation 3 defines the tag assignment Y as a function of the signal level:

$$\begin{aligned} Y(X) &= 1 \quad \forall L(X) \geqslant -95dBm \\ Y(X) &= 0 \quad \forall L(X) < -95dBm \end{aligned} \tag{3}$$

We consider an occupancy threshold of $-95\ dBm$ based on traditional cognitive radio technologies and previous research findings in [2], [3], [12]. Nevertheless, any other threshold can be considered if there is enough margin between the detected signals and the noise floor for the chosen learning method.

First, the algorithm creates a matrix $[X; Y]$ from the inputs and tagged output (line 3 to 5 in Algorithm 1). The machine learning tree bagger function will create up to 100 random decision trees for classifying any given output as a function of the three-dimensional input (line 6 in Algorithm 1). The samples are randomly distributed across the trees for having a low correlation on the prediction of each tree, maximizing the reproducibility of the algorithm results. Otherwise, certain combinations of inputs will have a higher chance of being correctly classified than others. As any given tree might lead to a different outcome, for classification tasks a decision based on a majority vote (or a certain percentile) is typically implemented. Before assessing the proximity between samples in the trees the algorithm compacts the implemented model for achieving a higher computational efficiency.

The proximity or similarity between any given sample in any built decision tree is later quantified by the Outlier Measurement function (line 8 in Algorithm 1). The proximity quantification considers both the spatial and frequency inputs (three-dimensional distance). This means that not only the correlation between neighboring locations is considered but also the self-correlation within any portion of the spectrum vector. This is particularly important when the received signal significantly varies across the spectrum occupancy decision threshold leading to inconsistent occupancy tags. Based on the proximity assessment we define the outliers as the samples for which the proximity value is not within the 80^{th} percentile (line 9 in Algorithm 1). Here TF contains a binary array defining the outliers while U, L, C (Upper threshold, Lower threshold, and Center value) are scalars quantifying a certain

Algorithm 1 Machine Learning Algorithm

Inputs: *GPSx*, *GPSy*, *Freq*
Output: *SpectrumO*

```
1:  Nloc: Number of measured locations
2:  Nfreq: Number of frequency samples per location
3:  for i = 1:Nloc
4:      [X(i) , Y(i)] = [Inputs(i) , Output(i)]
5:  end for;
6:  ML_TB = TreeBagger (100, X, Y, classification, OOBpredictor);
7:  CML_TB = compact (ML_TB);
8:  Proximity = Outlier_Measure (CML_TB,X);
9:  [TF,U,L,C] = isoutlier (Proximity, Percentiles = [0 80]);
10: for i = 1:Nloc
11:     for j = 1:NFreq
12:         n = NFreq*(i-1)+j;
13:         if (Y(n) == 1) OR (TF(n) == 1);
14:             SpectrumO(i) = 1;
15:         else
16:             SpectrumO(k) = SpectrumO(k)
17:         end if;
18:     end for;
19: end for;
```

scale of the median average deviation. The U, L, C are only used for fitting and optimizing our model (e.g., fitting the number of trees).

Because of the relative small size of our dataset, we did not accounted for a higher percentile (e.g., $90^{th} - 95^{th}$) because that causes that a single pair of locations within a cluster might define the proximity function and therefore the outliers. For the 80^{th} percentile, the combination of more than 2 pairs of locations is needed for defining the proximity function used as reference for extracting the outliers. In addition a higher percentile might lead to over-fitting due to the limited total number of spatial samples and neighbors around each location. For larger datasets a higher percentile might be considered (i.e., lower proximity between samples are considered as outliers). Nevertheless, over-fitting still can occur. An over-fitting in the frequency domain will be equivalent to a very narrow-band allocations (as small as the frequency resolution), which could be impractical for most DSA applications. For instance, for our scenario a $> 95^{th}$ percentile in the outlier criteria leads to a ten times higher U, L, C, which is equivalent to include in the outliers, samples that are not correlated enough either in the spatial or frequency domain. Instead of defining the outliers based on a majority vote for a certain percentile of the outliers measured proximity, *Grubbs* method can also provide a fine-tuned result without over-fitting but at a cost of a lower computational performance as the outliers are extracted in a one-by-one basis.

Finally, the spectrum occupancy pattern is generated by using the TF vector of outliers as a mask of the initially considered occupancy from each independent spectrum vector. In this way if either a certain location within a cluster has a frequency identified as occupied or the frequency is marked as outlier, the specific frequency is considered occupied, otherwise as white space (line 10 to 19 in Algorithm 1).

D. Simulations for Algorithm Validation

For validation purposes, we also performed a total of 20 simulations with a pseudo-random generated dataset and a known outcome (ground truth). The signal levels for the dataset are pseudo-randomly generated with a maximum deviation from the noise floor equivalent to the maximum fading from our measurements. The spectrum occupancy is then tagged considering Equation 3. Here we apply *Grubbs* method as it is not possible to define for this dataset a fair percentile for the voting decision. For each simulation we recursively remove the identified outliers samples from the original dataset and the new dataset if used for the next iteration of the algorithm.

Simulation ground truth: As the signal levels are pseudo-randomly generated the correlation should be near 0. Also the tags will be pseudo-randomly distributed with the same probability. The algorithm will find some random outliers but after each new simulation the gradient of the number of outliers will decrease and the number of identified outliers will trend to a constant value. For the dataset of our measurements, if the correlation is strong enough, in a few simulations after extracting the outliers, the number of new outliers will trend to 0. As a consequence, the gradient descent of the number of outliers identified after each simulation will be significantly higher for our experimental dataset than for the pseudo-random generated dataset. Therefore, the ground truth for the validation can be defined as:

$$\lim_{n \to n_m} N_{TF}(n) = 0 \quad \forall \rho(Y_i; Y_j) \to 1 \tag{4}$$

and:

$$\lim_{n \to n_m} N_{TF}(n) = k \quad \forall \rho(Y_i; Y_j) \to 0 \tag{5}$$

where N_{TF} is the percent of identified outliers as the function of the simulation index n, $k > 0$ is a constant, and n_m is the maximum number of simulations and satisfies that the minimum length of the vector $TF(n)$ is at least 50% of the original dataset size (length of $TF(0)$). The ground truth condition also fulfills that:

$$\frac{\partial N_{TF}(n)}{\partial n} > \frac{\partial N'_{TF}(n)}{\partial n} \tag{6}$$

where $N'_{TF}(n)$ is the percent of outliers corresponding to the artificially generated dataset.

III. Results

A. Nearest Neighbors Correlation

Fig. 1 shows the cumulative distribution function of the spectrum correlation between all the neighboring locations up to a distance *d*, for the analyzed measurement setups and frequency sub-bands. Results findings in Fig. 1 show a strong correlation between the spectrum vectors from each neighboring location for both distances *d=0.5 km* and *d=1 km*, except for the spectrum sub-band between 600 MHz and 1000 MHz for the measurement *Setup 1* at a distance between neighbors up to 1 km. This is caused by the combination of two factors. First, for the measurement *Setup 1* there is a lower spatial and frequency resolution, and a higher noise floor due to the filter resolution bandwidth. With the *Setup 1* also for the frequency range from 170 MHz to 1000 MHz we did not obtain the best correlation. Second, in the upper bands analyzed there is a predominant Low-Tower Low-Power infrastructure with smaller cells compared to the predominant High-Tower High-Power infrastructure in the lower sub-bands. As the minimum distance between measured locations is equivalent to the cell size there is a higher variability between spectrum vectors at different neighboring locations. Therefore, the effect of the limitations of *Setup 1* has a higher impact on the spectrum correlation between neighbors for the sub-band between 600 MHz and 1000 MHz.

An important result is that in the surveyed area for both segments of spectrum we obtained a correlation higher than 0.85 between any neighboring location up to a distance of 1 km with a 95% probability (considering the measurement *Setup 2*). Based on these results a relatively small and computationally efficient machine learning might filter the data for detecting the uncorrelated data between the spectrum vectors at any given location compared to their neighbors (i.e., finding outliers).

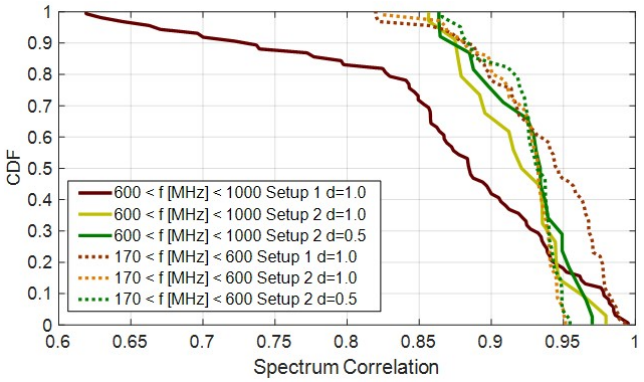

Fig. 1. Cumulative distribution of the spectrum correlation between neighboring locations. f: Frequency range, d: Radius of the cluster.

B. Outliers Detection

Fig. 2 shows the outliers identified by the Random Forest algorithm across the spectrum vectors at each measured location for the *Setup 2*. For a more clear representation of the results the figure presents a near 2D-projection over the frequency axis. In the spectrum mainly assigned to mobile services there is a higher variation in the identified outliers for different neighboring locations (in Fig. 2b). There are two factors influencing the variance in the received levels across the neighbors. First, poor coverage of mobile services in rural areas particularly of the 4G services (e.g., down-link allocated to band B20 from 791 MHz to 821 MHz). Second, sporadic access to the services by the end-terminals (e.g., up-link assigned to B20 from 832 MHz to 862 MHz). This is the contrary to what happens in locations with higher population density where the access probability (emission probability) of the end-devices is higher and more uniformly distributed in both the temporal and spatial domains.

These results reveal that the strict masking method proposed in Section II should be applied for DSA if there is uncertainty or inconsistency in the sensed data between neighboring locations. In general, for scenarios with higher heterogeneity of infrastructure sharing the same spectrum or when the coverage is poor, the probability of false detection of white spaces, and therefore interference, is higher.

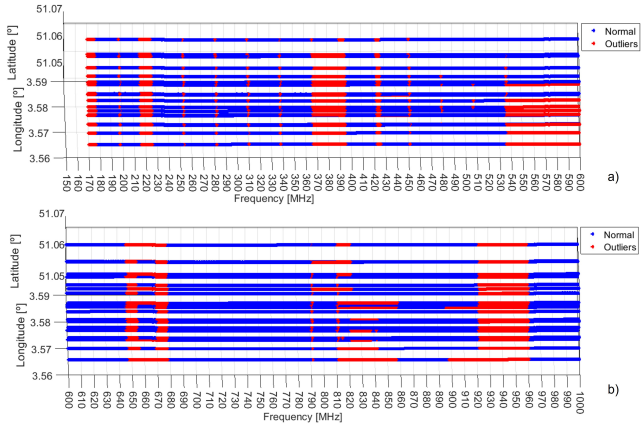

Fig. 2. Findings of outliers within the Spectrum Vectors at each geolocation. a) Spectrum sub-band $170 < f[MHz] < 600$, b) Spectrum sub-band $600 < f[MHz] < 1000$

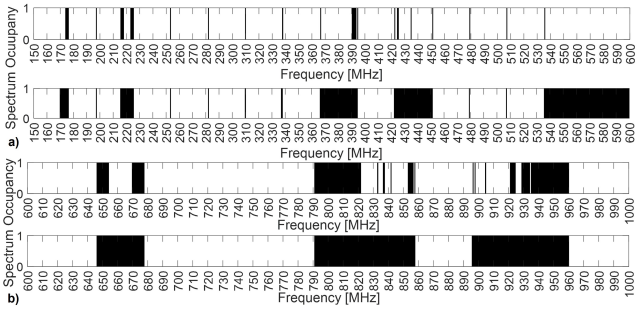

Fig. 3. Spectrum occupancy comparison between traditional cognitive radio allocation algorithm (upper tiles) and our algorithm (lower tiles).

C. Spectrum Feature Extraction: White Space Availability

Based on the results findings from the previous section we applied the outliers mask as a boolean AND function to the spectrum allocation individually performed from the sensed data at each location. If a sensor has detected a certain frequency as white space but the mask determine it is an outlier, then it is considered as occupied spectrum. Fig. 3 shows the spectrum occupancy individually determined based on traditional cognitive radio decision (upper tile) compared to our proposed feature extraction based on the identification of outliers by machine learning and Random Forest algorithm, for a) the frequency range from 170 MHz to 600 MHz and b) from 600 MHz to 1000 MHz.

For the sub-bands in Fig. 3a allocated mainly to broadcasting services (e.g., T-DAB and DVB-T/T2), a traditional estimation of the spectrum occupancy by cognitive radio devices (upper tile) might classify nearly 95% of the spectrum as white spaces. However, because of potential interference, the coexistence is not guaranteed, particularly for the sub-bands with higher heterogeneity of infrastructure characteristics and variability of coverage. Our algorithm allocated approximately 29% less white spaces (lower tile) but shielded the access to bands with lower data reliability. In the sub-bands assigned mainly to mobile services (Fig. 3b) the feature extraction by our algorithm was capable to identify as occupied the spectrum in use by the mobile operators (791-862 MHz and 890-960 MHz including up-link, down-link and band-guards) with a reliability of 92%. The segment from 880-890 MHz was allocated by the regulator for uplink, but not detected in use by any device at the time of the measurement campaign. We did not consider that segment in the reliability estimation as it is not related to the performance of our algorithm.

D. Simulations and Validation

Fig. 4 shows the simulations performed for validating the designed algorithm (Algorithm 1) considering the method and ground truth presented in Subsection II-D. After just 2

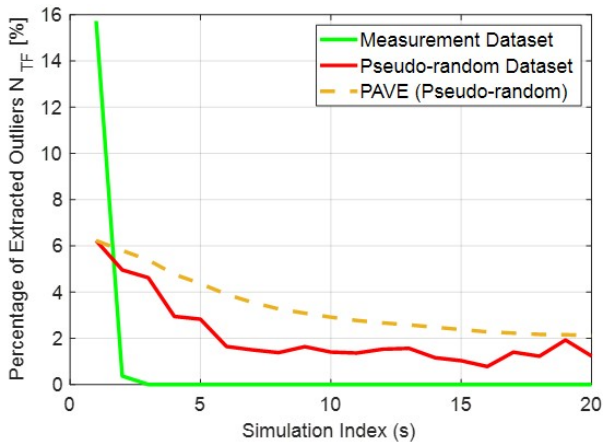

Fig. 4. Algorithm Validation

iterations there is a clear trend towards 0 of the number of extracted outliers from the highly correlated dataset of our measurements (criteria from Equation 4 is satisfied). At the same time, the moving average of the extracted outliers from the pseudo-random generated dataset shows a tendency to a constant value (approximately towards 2% extracted outliers per simulation). As the maximum correlation for this dataset is also near 0 ($|\rho| < 0.1$) the criteria from Equation 5 is also satisfied. Finally, the gradient for the measurement dataset is almost 10 times higher than for the pseudo-random dataset after just 3 simulations (validation criteria from Equation 6 is also satisfied).

E. Model Error, Inputs Relevance and Computational Performance

The mean error of the machine learning model was 1.8% for 100 trees. The permuted delta error of the machine learning inputs shows that there is a high spatial correlation (> 0.85), as expected from the results presented in subsection III-A. The frequency input has a higher relevance for the model prediction in the order of 12:1 compared to the spatial inputs (geo-location). This is likely because the self-correlation in the frequency domain is more limited, particularly for sub-segments with a larger signal variation. This is caused either by the randomness and sporadic use of some services (uplink) or variations across the signal level threshold defining the spectrum occupancy. The average computational time of the Algorithm 1 was approximately 62 seconds considering a computational capacity of only 16 GFLOPS. For 10 trees the computational time decreases to 14 seconds but at a cost of a higher mean error by approximately 2%.

IV. Conclusions

In this paper we presented the machine learning spectrum feature extraction from a spectrum survey dataset. We presented an algorithm based on multi-correlation and Random Forest capable of finding outliers from spectrum data in both the spatial and frequency domain.

We found that in the surveyed rural area there is a 95% probability of having a strong correlation (higher than 0.85) between spectrum vectors up to a distance of 1 km between sensing locations. This shows that spectrum fingerprinting is reliable for extracting spectrum features like the spectrum occupancy or white space availability. Our proposed algorithm is capable of finding the outliers across the multi-domain inputs even when the signal levels are near the decision threshold defining the spectrum occupancy or in sub-bands with a higher heterogeneity of technologies' architecture and radiation footprints. Despite the randomness and sporadic access to the uplink sub-bands of the mobile services the proposed algorithm was capable to identify and fully shield the access to the mobile bands in use by the local operators, not being the case of a distributed allocation by traditional cognitive radio algorithms.

Acknowledgment

This research has been supported by VLIR-UOS funding UOSTEA2022001201. We would also like to acknowledge CESAM - Global Minds, Ghent University and *Rohde & Schwarz* for the support to the measurement campaign.

References

[1] European Commission, "Digital Economy and Society Index ," *Brussels, Belgium, 2022*.

[2] M. Höyhtyä, A. Mämmelä, M. Eskola, M. Matinmikko, J. Kalliovaara, J. Ojaniemi, J. Suutala, R. Ekman, R. Bacchus, and D. Roberson, "Spectrum occupancy measurements: A survey and use of interference maps," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2386–2414, 2016.

[3] R. Martinez Alonso, D. Plets, M. Deruyck, L. Martens, G. Guillen Nieto, and W. Joseph, "Multi-objective optimization of cognitive radio networks," *Computer Networks*, vol. 184, 2021.

[4] R. M. Alonso, A. C. Guerra, E. F. Pupo, D. Plets, G. G. Nieto, L. Martens, and W. Joseph, "Assessment of white spaces quality in rural areas: a large-scale spectrum survey," in *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2020, pp. 1–5.

[5] A. Ivanov, K. Tonchev, V. Poulkov, and A. Manolova, "Probabilistic spectrum sensing based on feature detection for 6g cognitive radio: A survey," *IEEE Access*, vol. 9, pp. 116 994–117 026, 2021.

[6] A. Aragón-Zavala, T. W. C. Brown, and G. Castañón, "Polarization and effects on hidden node/shadowing margin for tvws," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 46–54, 2016.

[7] J. Tong, M. Jin, Q. Guo, and Y. Li, "Cooperative spectrum sensing: A blind and soft fusion detector," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2726–2737, 2018.

[8] M. Golvaei and M. Fakharzadeh, "A fast soft decision algorithm for cooperative spectrum sensing," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 1, pp. 241–245, 2021.

[9] R. Sarikhani and F. Keynia, "Cooperative spectrum sensing meets machine learning: Deep reinforcement learning approach," *IEEE Communications Letters*, vol. 24, no. 7, pp. 1459–1462, 2020.

[10] L. Khalid and A. Anpalagan, "Adaptive assignment of heterogeneous users for group-based cooperative spectrum sensing," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 232–246, 2016.

[11] M. Hossain and J. Xie, "Third eye: Context-aware detection for hidden terminal emulation attacks in cognitive radio-enabled iot networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 214–228, 2020.

[12] R. Martinez Alonso, D. Plets, M. Deruyck, L. Martens, G. Guillen Nieto, and W. Joseph, "Dynamic interference optimization in cognitive radio networks for rural and suburban areas," *Wireless Communications and Mobile Computing*, 2020.