Finding Forensic Artefacts in Long-Term Frequency Band Occupancy Measurements Using Statistics and Machine Learning

Bart Somers[1], Asanka Sayakkara[2], Darren R. Hayes[3], and Nhien-An Le-Khac[1(✉)]

[1] University College Dublin, Dublin, Ireland
bart.somers@ucdconnect.ie, an.lekhac@ucd.ie
[2] University of Colombo, Colombo, Sri Lanka
asa@ucsc.cmb.ac.lk
[3] Pace University, New York, USA
dhayes@pace.edu

Abstract. Wireless real-time communication between users is a key function in many types of businesses. With the emergence of digital systems to exchange data between users of the same spectrum, usage of the wireless spectrum is changing and increasing. Long-term frequency band occupancy measurements, carried out in accordance with the requirements of the International Telecommunication Union, can be used to measure and store informative values for further forensic investigation. In the existing literature, there is very limited research on using that information for a forensic investigation due to a lack of relevant datasets, examination methods and valuable artefacts. In this paper, we present a new approach to identify forensically sound deviations, often referred to as outliers, from using a monitored frequency band. We present the medcouple method for statistically detecting and classifying outliers. Furthermore, we created two datasets of long-term frequency band occupancy measurements that were used to evaluate our approach. We also evaluated our datasets with different machine learning techniques, which demonstrate that Random Forest has the highest classification accuracy and sensitivity to automatically detect outliers. These datasets will also be made publicly available for further research.

Keywords: Long-term Frequency Band forensics · statistical analysis · machine learning · outlier detection · digital forensics · signal intelligence

1 Introduction

For decades, real-time communication between users has been a key function in many types of businesses. For a long time, this was only possible by using

S. Goel and P. R. Nunes de Souza (Eds.): ICDF2C 2023, LNICST 570, pp. 227–248, 2024.
https://doi.org/10.1007/978-3-031-56580-9_14

radio-techniques in several forms. With the emergence of digital techniques, the landscape is shifting towards digital transmissions that deliver real-time communication between users. Within this landscape, one can find all kinds of communication devices, ranging from analogue 2-way radio systems, to advanced 4G and 5G networks. One thing all of these techniques have in common is the usage of the radio-spectrum to transfer information between users. With the current frequency allocations, a specific type of spectrum with clear borders in frequency is designated to a specific service. This is called an allocated frequency band. Usage of a specific frequency band can be monitored by interested parties or by the regulatory office responsible for allocating the spectrum. By measuring the occupancy of the specific frequency band, one can give estimations about the usage and the possible remaining capacity. This information can be used to give insight into usage or possible shortages of capacity in a frequency band.

All frequency bands have a designated allocation for usage. This is the designated task of a regulatory office in a country. An allocation of frequency band defines the legal type of usage of this band and the permitted users. Based on this allocation, some information is already known before monitoring: one can assume that wide-band broadcast FM signals are not to be expected in a frequency band, allocated for 4G data services for mobile communications [1]. When a specific frequency band is monitored using this knowledge, a certain "baseline" or normal type of usage will be gathered over time. Inspecting the usage in the band without decoding the content[1], the gathered data reveals parametric data over the usage. When the usage changes, variations will occur in these parametric values. If, for example, new active transmitters with the designated usage of the band are placed nearby, the amount of transmissions will increase. Or, if transmitters are being used for purposes, other than the designated usage, other parametric values can change, including duration of transmissions or maximum power of received signals. A sudden increase or decrease in one or more of the parametric values can indicate a change of usage. This can be caused by both legal and illegal usage of the monitored frequency band. In order to interpret these changes in usage, it is necessary to carry out a forensic investigation. i.e. by using a scientific method of solving crimes, involving examining the objects or substances that are involved in the crime. In this specific case, the change in usage of the radio spectrum is an indication for further investigations.

A recent example is the current war in the Ukraine: close to the border between Russia and Ukraine, radio communications could be detected from the other side of the border. When applying long-term frequency band occupancy measurements on bands of interest, a normal type of usage can be identified. This could be caused by normal training procedures of the different armies. With the upcoming invasion of the Ukraine, it is likely that the radio usage

[1] As channel usage, modulation types, coding schemes and possible encryption can change without notification, or collisions can occur, the only reliable method of detecting usage, is inspecting the RF power, to prevent false negatives due to the incapability of decoding the transmitted information. Compared to the OSI model, this analysis is executed on layer 1.

was changing, due to the build-up of the army. During and after the invasion, it is assumed that the usage changes again. The types of messages, which are exchanged via radio, will probably change as the information sent to and from different units in the army changes. This can influence the average transmission time. Furthermore, the amount of messages exchanged will probably increase. And as the invasion evolves, transmitters may come closer to the monitoring station, which will increase the received power values on the monitoring station. All of these parametric values can be used to detect changes in a very early point in time or over a longer period of time, depending on the values and method used to measure. With knowledge of the spectrum, these changes can be identified as outliers, and thus be labeled as forensic artefacts, as they identify differences in frequency band usage.

Another example would be smuggling drugs over sea: Due to the lack of mobile phone service, chances are that maritime radio systems are used to communicate between the smugglers. Long term monitoring could reveal outliers in usage, which can be an indicator of smugglers trying to import the drugs.

In terms of existing academic studies, while the literature is useful for frequency band allocations or primary and secondary user separation in cognitive radio systems [2,3], there are very few approaches to identifying outliers or artefacts used to investigate a frequency band. Detecting and classifying outliers or artefacts in the measured occupancy is left up to the investigator to inspect the data.

The goal of this paper is to identify outliers in each dataset of the spectrum, being investigated, using statistical methods and also uncover efficient machine learning (ML) algorithms by identifying these outliers in terms of the highest classification accuracy and sensitivity. The main contribution of this work can be summarized as follows:

- A statistic-based approach to forensically detect outliers and to analyse the skewed datasets from frequency band occupancy measurements. These confirmed outliers are labeled as forensically sound outliers in the dataset.
- An efficient, supervised ML method to detect outliers, i.e. relevant forensic artefacts for future examination.
- Two new validated datasets of long term frequency band occupancy measurements, publicly available for further research.

2 Long-Term Frequency Band Data

The reason to execute frequency band occupancy measurements, both in short-term or long-term, is to gain insight in usage of the frequency band under investigation. Hopefully, this measurement gives insight from different perspectives.

The International Telecommunication Union (ITU) [4] is the United Nations specialized agency for information and communication technologies and sets, among other things, the standards and guidelines for spectrum monitoring. Thus, we require that the tool and method used to collect the data must implement a

frequency band occupancy measurement method, according to the guidelines [5] from the ITU. Using this method, the data collection is method scientific proven and can be used as a solid base for forensic investigations.

The long term frequency band data should be multiple dimensions. This includes, but is not limited to, energy detection, duration of transmissions, amount of observed transmissions and the total amount of time the frequency band being observed is in use. Thus, with single value measurements the gained insight in usage is very limited. If, for example, the usage of the spectrum under investigation is changing in duration for transmissions, and only energy detection is used for monitoring, this change might go unnoticed.

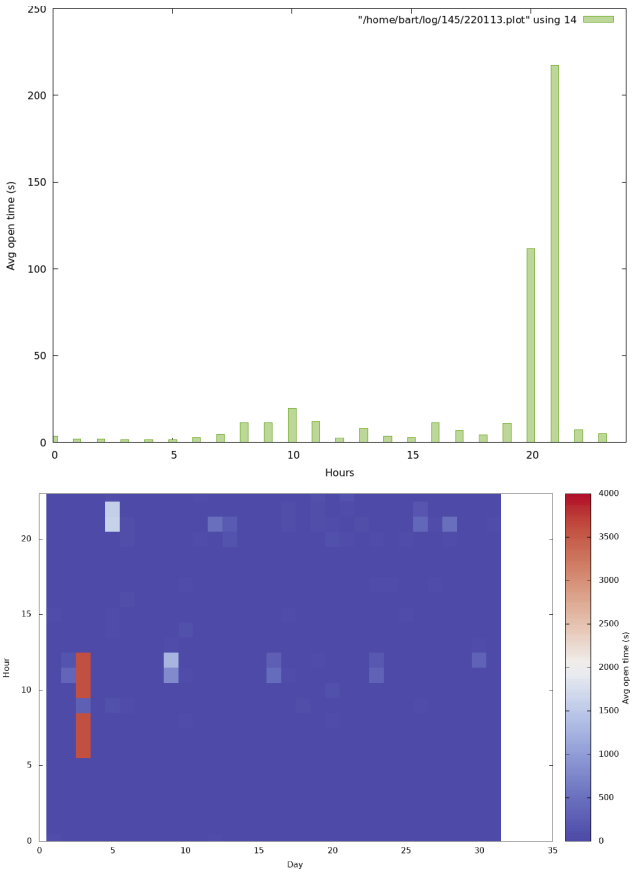

Fig. 1. Example of short and long term view of average duration of transmissions: Daily view (upper), Monthly view (lower)

There are multiple tools available for spectrum monitoring and frequency band occupancy measurements. Figure 1 shows an example from the Specmon tool [6] with the different views.

3 Related Work

In [7] the authors proof that normal statistical outliers detection, based on the Inter Quartile Range (IQR) values, is unusable for skewed data. They presented a method to detect outliers where the skewness of the data is used as a correction for the outlier limits.

Spectrum forensics is defined by the authors in [8]. While they have the focus on sensing the spectrum using electrical engineering methods, without analysis of the measurements, it sets the definition for scientific proof of the measurements and the upcoming analysis.

Authors in [9] proposed a platform for both long-term spectrum monitoring and a ML platform to analyse the data. This research primarily focused on cognitive radio and uses long term spectrum monitoring to build a database and predict unused spectrum resources, which can be used by secondary users. While the basis for their research is long-term monitoring, the proposed solution is not focused on single usage of a specific band and artefacts cannot be determined in a forensic way using this method as their proposed method does not monitor the band continuously. Furthermore, their proposed method only monitors the minimum, maximum and average signal power, which results in a single dimension for the ML method proposed. The goal of their research is to forecast the usage of the primary user for the monitored spectrum, in order to predict unused time-slots for a secondary user.

In [10] authors proposed multiple ML techniques, which could be used in a reliable way to sense available spectrum. Although the research was focused on cooperative radio systems and cognitive radio, the authors selected multiple ML techniques that are used to identify forensic artefacts. The data used for the statistical analysis and ML methods is based on classical energy detection. Because only energy detection is monitored, the proposed method can be identified as a single dimensional solution. The proposed ML methods are adopted in this research and verified for forensic usage.

A measured spectrum occupancy database (MSOD) is presented in [11]. This database holds the data for long-term frequency occupancy measurements and multiple display options of the usage are presented. The model is used by the National Telecommunications and Information Administration (NTIA) and suggestions for several frequency band occupancy measurements are provided. In this system, relatively low bandwidth amplitude-versus-frequency data are sent from sensors to MSOD for the purpose of computing occupancy statistics every 10 or 15 min for relatively long-term intervals. Although this is useful for spectrum usage data, it lacks some additional dimensions which are useful to detect artefacts in usage.

The authors of [12] used a one-class Support Vector Machine (SVM) for outlier detection and proposed the use of a linear kernel function in the SVM

model for ML. A comprehensive list of dependencies is presented in this research on the usage of several kernels in outlier detection. Although they focused on unsupervised learning from unlabeled data, the presented techniques are usable for selecting multiple kernels in the SVMs to detect outliers.

In literature, using ML techniques has been widely studied for the digital forensic domain with varied methods such as Random Forest, Multi-Layer Neural Network, Convolution Neural Network [13], Regression Decision Tree, SVM [14] or even transfer learning [15]. However, most of the approaches focused on the classification of artefacts rather than the detection of outliers, which are also considered as forensic artefacts in our research context. Moreover, very few of them focused on the forensic analysis of spectrum data.

In [2] authors investigated the usage of ML to sense TV spectrum and determine available free space for cognitive radio usage. In this research the authors focused on Energy Detection (ED) and the usage of a threshold on the sensed data to detect usage on the spectrum under investigation. These extracted values are the parameters or dimensions used in the ML algorithms. Multiple supervised ML algorithms are compared by the authors, based on the information extracted from the detection of a TV signal.

Authors in [16] described a method for spectrum sensing based on the autocorrelation of received samples. Their goal is to determine free space for secondary users in cognitive radio systems, where the monitored spectrum is considered as unoccupied when the Signal to Noise Ratio (SNR) drops below a certain value. This method is compared with classical energy detection and it proved to have a high autocorrelation in usage of the sensed spectrum. Although their research focuses on spectrum sensing and the detection of usage of the spectrum, there are no descriptions of long term observations of a spectrum under investigation. As correlations of sensed data can reveal usage of the spectrum, it does not reveal if this is normal usage or a possible forensic artefact. In their research, detecting this type of artefact is irrelevant, they focus on informing the secondary user to not use the spectrum when the primary user starts transmitting.

3.1 Summary

Based on an overview of related work in Table 1 we can conclude that there is no research based on the guidelines from the International Telecommunication Union, using an open data model as well as a forensic basis for the research. The presented method in [12] is usable for forensic anomaly detection, although it has no focus on spectrum monitoring in any way. Despite the fact, it is the only research that uses multiple features from a dataset to detect anomalies.

Therefore, the research gap is a multi-dimensional, ITU based measurement method, with an open data, able to determine forensic outliers, both using statistics and ML.

Table 1. Comparison of related work

Methods	[2]	[7]	[8]	[9]	[10]	[11]	[12]	[13–15]	[16]
Outlier detection	N	Y	N	N	N	N	Y	N	N
ITU method for measurement	N	N	N	N	N	N	N	N	N
Forensic Methods	N	N	Y	N	N	N	N	Y	N
Spectrum Monitoring	Y	N	Y	Y	Y	Y	N	N	Y
Long Term Monitoring	Y	N	N	Y	Y	Y	N	N	N
Machine Learning	Y	N	N	Y	Y	N	Y	Y	Y
Multiple Features	N	N	N	N	N	N	Y	Y	N
Open data	N	N	N	N	N	N	N	Y	N

4 Methodology

In this section, we present a methodology to address the research gap mention in Sect. 3.1. Our approach includes the data pre-processing, statistical and supervised ML methods for outlier detection. The goal of this approach is to determine the parameters, usable for the outlier detection and consequently find the forensic artefacts in the usage of the frequency band. Then, using the ML methods, we investigate automated methods for future forensic investigations on this type of data.

4.1 Feature Selection

The features, in which the forensic artefacts are to be identified, must have a direct relation with the observed use of the spectrum under investigation. The relevance of the features are individually evaluated, as we approach the features in this research as univariate. The observations in this type of measurement consist only on a single characteristic of the spectrum usage, per measured value, or feature, in the dataset.

We are using the data created by Specmon. This dataset is in plain text and the values are space separated. An example of the data format is displayed in Table 2.

When inspecting Table 2, we can conclude that not all the features in the original dataset are useful. For example, one of the values is the threshold used by the software and method, and thus describing the measurement setup. This is also a fixed value which does not change in the entire dataset. Hence, this value does not add knowledge to the data and is not selected.

4.2 Data Classification

After the feature selection, we need to classify the outliers. Outliers, or forensic artefacts when they are identified and labeled as such, are data points in the

Table 2. Fields in the Specmon data

U	Hour of the measurement, ranging 0–23
V	Number of seconds transmissions are observed
P	Percentage of usage, based on seconds in one hour
T	Threshold value used during the measurement
O	Number of openings, or transmissions observed
D	Day of month
A	Average duration of openings
M	Maximum value received
W	Day of the week, ranging 0–6
L	Average low value (this is a highly experimental value)

dataset that attract attention and can be a starting point for further investigation. The artefacts should be identified in a reproducible method, so that equal events in the future are identified in the same manner. This prevents arbitrary choices for artefacts and thus random investigations on events, i.e. false alarms. The first step taken is to recognize outliers, as such, by manually inspecting the data and graphs plotted from the data. During this phase, a normal view of the spectrum under investigation can be recognized by the investigator and measurements deviating from this identified normal view of the spectrum, are considered as outliers. In the second phase, the identified data points are investigated as outliers or artefacts and the mathematical methodologies are tested to prove these data points as outliers or forensic artefacts. Both methods are compared to build a solid base in order to prove the outliers in a forensic way. Outlier detection is investigated in the following order: (i) Manual method; (ii) IQR method and (iii) Medcouple method.

After successful detection of outliers and the normal data, the data can be labeled with the one class value, indicating the outliers and normal data. With this additional information on the outliers, the data is classified.

Manual Method. As outliers are deviations from the normal usage, one can only determine the outliers after inspecting the normal usage of, in this case, the spectrum under investigation. Long-term spectrum monitoring is required to measure for a prolonged period of time and collect parametric data over the usage of the spectrum. Based on these datasets, an overview of the usage can be created. This overview is gained by inspecting the results from the same tool that is used for long-term spectrum monitoring. The graphical results of the datasets can be manually inspected. This step is required for the research to determine the amount of outliers, or artefacts in the datasets. No quantitative values can be extracted from the data using this manual method as this process relies solely on the experience of the investigator.

IQR Method. The next step after the manual detection of outliers is to inspect the distribution of the selected features using violin plots. Violin plots are an extension to the box and whisker diagrams, which visualize univariate extreme values. Only if the data distribution satisfies a normal distribution, the IQR method can be applied and descriptive statistics [17] can be calculated for selected features in the datasets. Outliers are investigated using the Inter Quartile Range outlier detection [18], based on 1.5IQR method. These calculated outliers are then checked against the dataset and graphs from the datasets.

Medcouple Method. For non-Gaussian distributions, outlier detection based on 1.5IQR method is not a reliable method, as the outlier factor must be corrected according to the skewness of the data distribution. The medcouple method provides this correction [7]. This method determines outliers based on the skewness of the data and is considered to be a robust method to calculate possible outliers. We chose to treat the different variables as independent variables. Next, the calculated values are inspected and confirmed against the graphs from the used tool. Accordingly, the outlier limits are calculated:

if $MC > 0$ **then**
$[Q1 - 1.5e^{-4MC}IQR, Q3 + 1.5e^{3MC}IQR]$
else
if $MC \leq 0$ **then**
$[Q1 - 1.5e^{-3MC}IQR, Q3 + 1.5e^{4MC}IQR]$
endif
endif

The above functions clearly shows that the outliers are corrected with an outlier factor: the "Normal" IQR value is increased or decreased according to the medcouple value. If the medcouple value is positive, the data is right-skewed. If the medcouple value is negative, the data is left-skewed and in both cases the outlier limit is corrected accordingly.

4.3 Machine Learning Method

In the previous section, we identified outliers in the dataset using several statistical methods. Both the outliers and the normal data are labeled, and the results in datasets with added knowledge on the normal usage of the spectrum under investigation, and outliers of usage of the same spectrum.

We also used multiple supervised ML methodologies to train the data with classified outliers. The purpose of this is two-fold: (i) validating the usability of created datasets with ML approaches to detect the outliers; (ii) finding the efficient ML techniques in terms of the high classification accuracy and sensitivity that can be used for the outlier detection from similar datasets.

The following algorithms were compared in this research: (i) k-Nearest Neighbors (kNN); (ii) Logistic Regression (LR); (iii) Naive Bayes (NB); (iv) Neural Network (NN); (v) Random Forest (RF); (vi) SVM-RBF kernel (SVM-R); (vii) SVM-Polynomial kernel (SVM-P) and (viii) SVM-Linear kernel (SVM-L).

We used two performance metrics. The classification accuracy, which is calculated with the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is the True Positive, TN is the True Negative, FP is the False Positive and FN is the False Negative. The second metric is the sensitivity, or recall of the model which is calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

As this research focuses on the detection of outliers, and the outliers identified with the medcouple method above are labeled with class 1, the sensitivity is calculated for this class only. Both the classification accuracy and the sensitivity are calculated using a 10 fold stratified cross-validation method.

Validate Classification Task. In the previous sections we described the multiple steps taken to label the data and used the labeled data to train ML methodologies. Depending on the classification accuracy and sensitivity, it is possible to select the best performing ML methodology. When both the classification accuracy and sensitivity are lower than 1, that is not 100% accuracy, the ML methodology generated some mis-classified outliers. When selecting these mis-classified outliers, we can validate them to the original data. This might suggest reasons for mis-classification. We followed the following steps to map mis-classified outliers back to the original data: (i) Select the mis-classified outliers; (ii) Locate the feature values of the mis-classified outlier in the normalized data; (iii) Record the position of the outlier in the normalized data and (iv) Find the same position in the original data.

The values from the original data could now be validated with the result of the manual outlier selection and the results of the selected statistical method.

5 Experiments

5.1 Experimental Environments

Tool and Method. Specmon is usable for Frequency Band Occupancy measurements using software defined radio. This tool is capable of measuring multiple features of the frequency band being monitored. The measured data is visualized in multiple graphs, both in 2D for short-term observations and in 3D for long-term observations. Using these visualizations, we can manually identify outliers and cross-reference these outliers with outliers determined with the other methods presented.

Frequency Bands Investigated. We have chosen to create two separate datasets using Specmon, so that the method can be tested against different datasets, to make sure that the method is not a "single targeted solution" but a broad usable method.

868 MHz EU ISM Band, Main Part for the Usage of LoRa Network
A Raspberry Pi model 3 with an Airspy R2 SDR was used. The SDR was tuned to 868.35 MHz with a sample rate of 2.5 MSps. This results in 2.5 MHz of bandwidth centered around 868.35 MHz. The Airspy is connected to an FM-Broadcast band-stop filter to prevent front-end overloading from nearby transmitters and then the filter is connected to a Diamond D-130 discone antenna on top of the building. With this low-gain omni-directional antenna, nearby transmissions can be received. The software used is Specmon with FFT-size 2048, resulting in a frequency-resolution of 1220 Hz per FFT-bin. As an average transmission using the LoRa standard is 125 KHz wide [19], depending on the spreading factor used, this monitoring resolution matches the ITU requirements [5,20].

145 MHz Amateur Radio Band
A Raspberry Pi model 3, with a RTL-SDR V3 SDR, was used. The SDR was tuned to 145 MHz with a sample rate of 2 MSps. This resulted in 2 MHz of bandwidth which was the entire amateur radio band, ranging from 144 to 146 MHz. The RTL-SDR was connected to an FM-Broadcast band-stop filter to prevent front-end overloading from nearby transmitters and the filter was connected via the VHF-port from a Diamond MX-3000 triplexer to a Diamond V-2000 vertical antenna on top of the building. Due to the gain, designed frequency, and antenna-height, transmissions from a wide area could be received. The software used was Specmon with FFT-size 2048, resulting in a frequency-resolution of 977 Hz per FFT-bin. As the amateur radio band was dedicated for experimental usage, there was no single type of usage in this spectrum. The usage ranged from ultra narrow band modes to broadband data transmissions. A frequency resolution of 977 Hz was assumed to cover most of the transmissions, according to the ITU guidelines [5,20].

The platform for the data analysis and ML experiments was Debian 11 with Orange 3.31.1 [21], installed via miniconda.

5.2 Datasets

With the aforementioned setups, two different long-term frequency band occupancy measurements were recorded. Using setup 1, six months of data was collected from the usage of the 868 MHz LoRa frequencies. The measurement period was from October 2021 up to and including March 2022. Using setup 2, five months of data was collected from the usage of the 145 MHz amateur radio band. The measurement period for this second setup was from November 2021 up to and including March 2022.

Every measurement was executed for one hour and then restarted. The results per hour were aggregated by the Specmon software and 24 aggregated lines

Table 3. Example data from Specmon

U	00	V	1449.56	P	40.27	T	10.00	O	1030	D	01	A	1.41	M	46.24	W	6	L	1.59
U	01	V	1418.47	P	39.40	T	10.00	O	1034	D	01	A	1.37	M	45.88	W	6	L	1.62
U	02	V	1419.11	P	39.42	T	10.00	O	1026	D	01	A	1.38	M	45.61	W	6	L	1.66
U	03	V	1481.41	P	41.15	T	10.00	O	1021	D	01	A	1.45	M	45.98	W	6	L	1.61
U	04	V	1431.47	P	39.76	T	10.00	O	1002	D	01	A	1.43	M	45.93	W	6	L	1.69
U	05	V	1366.83	P	37.97	T	10.00	O	1001	D	01	A	1.37	M	46.46	W	6	L	1.73
U	06	V	1427.98	P	39.67	T	10.00	O	1009	D	01	A	1.42	M	45.80	W	6	L	1.64
U	07	V	1412.10	P	39.22	T	10.00	O	972	D	01	A	1.45	M	45.60	W	6	L	1.80
U	08	V	1373.73	P	38.16	T	10.00	O	992	D	01	A	1.38	M	45.68	W	6	L	1.66
U	09	V	1450.59	P	40.29	T	10.00	O	1039	D	01	A	1.40	M	45.89	W	6	L	1.63

were combined per day. The aggregated lines consisted of 10 values from which an example is shown in Table 3 and the description of the field in the data is displayed in Table 2. Finally the measurement results per day in a single month were combined to a single file per month. This resulted in 672 measurements in February, with 28 d and 744 measurements in months with 31 days. An example of the data is displayed in Table 3. The dataset for the 868 MHz consisted of 4800 data points and the 145 MHz dataset consisted of 3216 data points.

In addition to the different datasets per month, a full dataset was created with all the measurement data in a single datafile, per setup. These datasets were named "full" and contain five or six months of continuous data.

5.3 Data Preprocessing

Feature Selection. The features selection method from the original data is explained in Sect. 4.1. The final data consists of five features:

openings: This value is the number of times the threshold, during the measurement period, was triggered. This shows the number of transmissions observed. (**O**)

maximum open value: This value is the maximum value measured during one transmission and shows the maximum relative power level of the transmission. (**M**)

average open time: This value is the average duration of observed transmissions during the measurement period. (**A**)

open seconds: This value counts the total amount of time in seconds that the measurement values were above the threshold, during the measurement period. (**V**)

class: This value is default set to zero for later classification. (**C**)

5.4 Measurement Results Outliers

In this section we present the results from both the manual method and the medcouple method and compare the results of these two methods. The measurement data from the month January 2022 is selected from the 868 dataset for

manual inspection and classification. As a first step, the output from Specmon is visually inspected and manually searched for outliers.

The 3D plots from the 868 MHz measurements are shown below in Fig. 2 where in Fig. 2 (upper) the percentage of usage is displayed. Using the same technique, other parameters are shown, such as in Fig. 2 (lower) where the maximum measured signal strength is displayed.

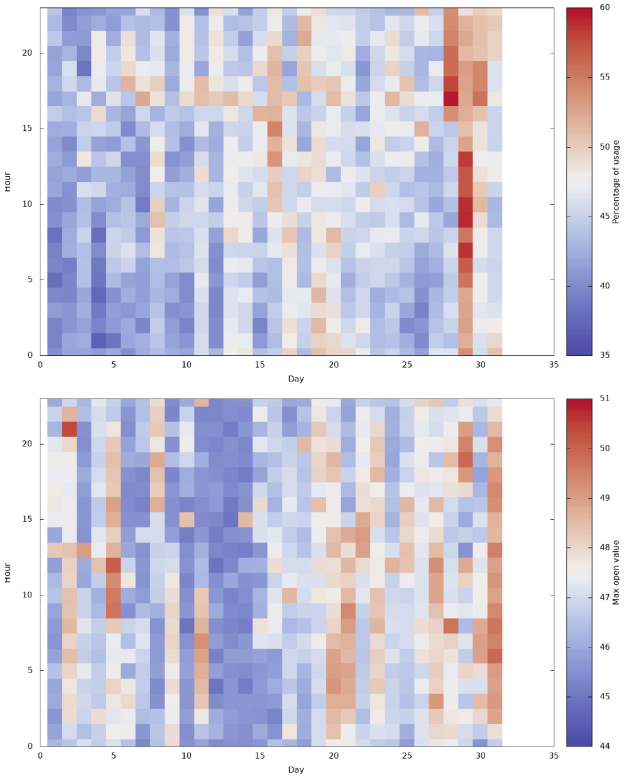

Fig. 2. Percentage (upper) and max open value (lower) plots from Specmon

Subsequently two other 3D plots are generated from the same dataset, as displayed in Fig. 3 where Fig. 3 (upper) shows the average duration in seconds of an observed transmission and Fig. 3 (lower) displayed the 3D graph with the number of observed transmissions, both per hour.

In order to identify outliers, all the graphs must be inspected and possible outliers must be recognized in combination with the other graphs. As outliers on the high side have a more red color, they can be identified with relative ease. On the other hand, outliers on the lower side have a color difference in blue scale and this is more difficult to identify. Some examples are visible in Fig. 3 (upper): on January between 01:00 and 15:00 h, the entire scale of blue colors is visible but it is not clear when the values are low enough to identify as outlier. Inspecting the darker blue color on January 14th, 21:00 in the same figure, might reveal

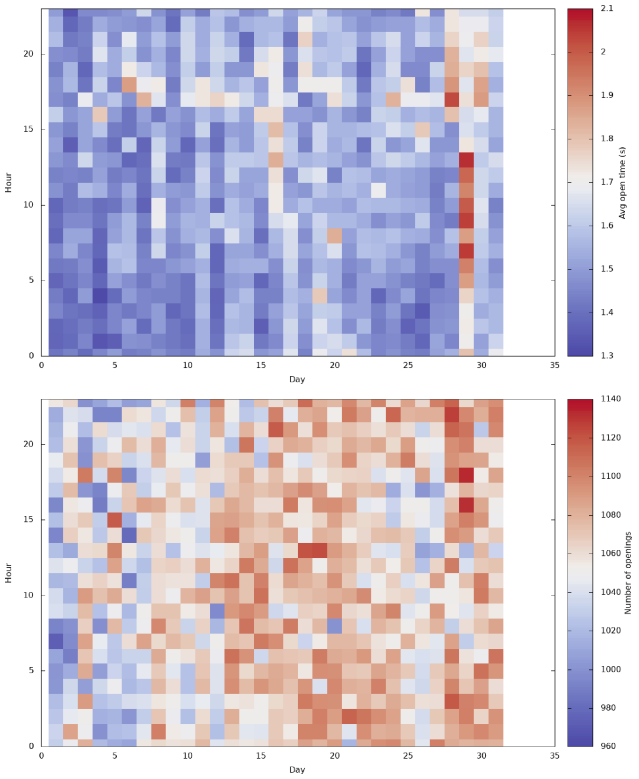

Fig. 3. Average opening values (upper) & number of openings (lower) plots from Specmon (Color figure online)

an outlier on the lower side. The same problem appears when inspecting Fig. 3 (lower): The bright red colors from January 28th, 16:00 onwards for 24 h show clearly that there was a change in usage. But the light red colors on 29th from 01:00 to 04:00 including, are probably within the limits of normal usage. By only visually inspecting the graphs, one is really dependent on the experience of the investigator to identify the subtle differences in colors in the graphs, to identify a measurement as an outlier. Furthermore, four different graphs with different color scales from the same measurement need to be inspected and validated. A python script was written to calculate the medcouple values from the different features. The medcouple value is slightly negative which means that the selected data in this example is skewed and the result of this correction in the medcouple method is visible, as the values for the IQR method and the values for "mean ± (3 * standard deviation)" and the medcouple method differ. The method "mean ± (3 * standard deviation)" is not used further in this research but only shown here to visualize the difference with a default method for outlier detection.

Table 4 is created using the medcouple method (MC) and by manual (Man) inspection of the graphs. In this table, the day and hour are shown and "1" in the table represents an outlier in that specific hour of the dataset. Furthermore, an extra column is added with a label added to the manual classification: False

Positive, False Negative and True Positive. The True Negatives are not listed in this table as they are the remaining measurements. January 2022 has 31 days with 24 h, so this results in 744 measurements. As there are 34 items identified in this month, using both the manual and medcouple method, the remaining 710 values are considered as True Negatives.

Table 4. Manual outliers and calculated outliers with medcouple method from January 2022 on 868 dataset.

day	hour	Man	MC	label
1	7	0	1	FN
2	14	0	1	FN
2	23	0	1	FN
4	1	0	1	FN
4	4	0	1	FN
4	6	0	1	FN
5	1	0	1	FN
5	16	1	0	FP
6	14	0	1	FN
7	18	1	0	FP
8	17	1	0	FP
12	17	1	0	FP
15	2	0	1	FN
16	13	1	0	FP
16	15	1	1	TP
20	8	1	0	FP
26	3	0	1	FN
28	17	1	1	TP
28	18	1	1	TP
28	19	0	1	FN
28	20	0	1	FN
28	21	0	1	FN
29	5	0	1	FN
29	6	1	1	TP
29	7	1	1	TP
29	8	1	1	TP
29	9	1	1	TP
29	10	1	1	TP
29	11	1	1	TP
29	12	1	1	TP
29	13	1	1	TP
30	17	1	1	TP
30	18	0	1	FN
30	19	0	1	FN

There was a noticeable difference in the manual and automated detection of outliers. At first hand, this seems to be a very big difference but when manually verifying the outliers, calculated using the medcouple method, they can all be confirmed using the graphs. When verifying the false positive manual outliers, that is the outliers manually identified but not confirmed using the medcouple method, these values are confirmed true false positives as the values are within the borders of the medcouple method. A plausible explanation can be that the color scale is not optimized for this type of manual inspection. This confirms that the difficult part here is to identify distinct differences in colors, which classifies an outlier, or not.

Using the functions explained, the classification accuracy and sensitivity of this manual method are calculated: (i) FN: 16; (ii) FP: 6; (iii) TP: 12; and (iv) TN: 710.

Table 4 holds the TP values, calculated with the medcouple method for this month of data. When we cross-reference these calculated outliers with the four different graphs made with Specmon, they can all be identified as outliers, in one or more of the graphs.

5.5 Measurement Results ML Methods

The labeled data can be used to train the different ML methods and calculate the performance metrics. 868 MHz dataset holds six months of continuous data and the dataset on the 145 MHz one holds five months of continuous data. In addition to these months, two larger datasets were compared: one for the 868 MHz dataset with the data from all six months and one for the 145 MHz dataset with all five months. This results in datasets for each month to train and calculate performance metrics. The performance metrics used to compare the ML algorithms were the classification accuracy and the sensitivity, or recall.

The sensitivity of the different ML models were calculated for the class 1 classification, as the class 1 value are the outliers or forensic artefacts. The class 0 value are TN, which are the values corresponding to the normal usage. Based on this, only the sensitivity of the forensic artefacts were calculated.

The results are displayed in the following tables for both the classification accuracy and the sensitivity, where abbreviations are used for the supervised ML algorithms.

868 MHz Dataset. This resulted in 672 measured values for the month of February to 744 measured values for the months with 31 days. The total datasets, with the data of all six months, contained 4,800 measurements, of which 220 were identified as outlier with the medcouple method (Table 5, Table 6).

145 MHz Dataset. Using the same settings and algorithms for the 145 MHz dataset, this resulted in 672 measured values for the month of February to 744 measured values for the months with 31 days. The total datasets, with the data of all five months, contained 3,216 measurements, of which 823 were identified as outlier with the medcouple method (Table 7, Table 8).

Table 5. Classification accuracy on 868 dataset

868	2110	2111	2112	2201	2202	2203	full
kNN	0.950	0.965	0.973	0.983	0.970	0.960	0.981
LR	0.925	0.942	0.949	0.968	0.951	0.950	0.961
NB	0.911	0.938	0.942	0.961	0.920	0.935	0.946
NN	0.972	0.974	0.972	0.984	0.967	0.964	0.991
RF	**0.997**	**0.994**	**0.993**	**0.999**	**0.993**	**0.991**	**0.999**
SVM-P	0.978	0.967	0.978	0.991	0.990	0.972	0.830
SVM-R	0.977	0.969	0.969	0.981	0.974	0.961	0.986
SVM-L	0.867	0.918	0.958	0.961	0.908	0.942	0.611

Table 6. Sensitivity values on 868 dataset

868	2110	2111	2112	2201	2202	2203	full
kNN	0.379	0.432	0.487	0.536	0.424	0.372	0.609
LR	0.034	0.045	0.026	0.143	0.000	0.140	0.264
NB	0.155	0.182	0.487	0.000	0.152	0.070	0.427
NN	0.672	0.568	0.487	0.571	0.333	0.372	0.841
RF	**0.966**	**0.909**	**0.897**	**0.893**	**0.879**	**0.884**	**0.977**
SVM-L	0.345	0.227	0.564	0.214	0.091	0.209	0.464
SVM-P	0.845	0.636	0.846	0.857	0.879	0.628	0.518
SVM-R	0.741	0.500	0.410	0.571	0.576	0.326	0.727

Table 7. Classification accuracy on 145 dataset

145	2111	2112	2201	2202	2203	full
kNN	0.843	0.831	0.888	0.917	0.806	0.927
LR	0.785	0.722	0.839	0.896	0.711	0.834
NB	0.782	0.720	0.860	0.853	0.758	0.822
NN	0.865	0.828	0.862	0.918	0.829	0.922
RF	**0.990**	**0.996**	**0.991**	**0.993**	**0.996**	**0.999**
SVM-P	0.875	0.390	0.465	0.917	0.351	0.275
SVM-R	0.849	0.477	0.849	0.920	0.576	0.421
SVM-L	0.808	0.367	0.386	0.917	0.336	0.317

Combined Results. Both datasets were processed with equal settings and using this method, the results of the calculations for the performance metrics can be compared. The results of both the classification accuracy and the sensitivity, or recall, were evaluated and the results with the highest values are printed in bold. When reviewing the results for both datasets, it becomes apparent

Table 8. Sensitivity values on 145 dataset

145	2111	2112	2201	2202	2203	full
kNN	0.548	0.694	0.408	0.352	0.583	0.746
LR	0.110	0.507	0.069	0.014	0.096	0.378
NB	0.329	0.730	0.523	0.141	0.512	0.452
NN	0.452	0.622	0.208	0.254	0.558	0.730
RF	**0.959**	**0.993**	**0.969**	**0.958**	**0.996**	**0.999**
SVM-L	0.219	0.789	0.777	0.282	0.817	0.950
SVM-P	0.562	0.776	0.669	0.268	0.812	0.978
SVM-R	0.384	0.862	0.192	0.254	0.929	0.987

that the Random Forest has the highest overall values for both classification accuracy and the sensitivity. Although both datasets have the same method for measurement, the usage of the different frequency bands differs. Even with this different type of usage, in both situations Random Forest remains the ML methodology with the highest results for classification accuracy and sensitivity. Furthermore the accuracy and sensitivity increases when the combined data for the different measurements are tested. Although this is more computationally expensive than calculating the accuracy and sensitivity for individual months of data, an increase in both the classification accuracy and the sensitivity, is visible.

If we compare the measurement results of both datasets, it is noticeable that SVM with the RBF kernel also provides high values for both the classification accuracy and the sensitivity. However, the values for the classification accuracy are slightly higher with Random Forest. The sensitivity with SVM with the RBF kernel also regularly reaches high values but has a large spreading in the results. This spreading is not observed when using Random Forest.

5.6 Validate Mis-Classifications

Despite the high values for classification accuracy and sensitivity, these values are not equal to 1, i.e. there is no 100% accuracy. Even with a classification accuracy of 0.999, a mis-classified outlier can occur. To gain insight into mis-classified outliers, some of these values are inspected and compared with the outlier values of the medcouple method with the validation method from Sect. 4.3.

Table 9. Medcouple values, 145 MHz dataset, March 2022.

Limits	lower value	upper value
max_open_value	34.1	62.6
avg_open_time	2.04	121.9
open_seconds	18.9	7230
openings	8.25	271.4

For this purpose, from the 145 MHz dataset the measurement results of March 2022 were assessed[2] and the mis-classified outliers are examined. The upper and lower limit values are displayed in Table 9.

In the next step, we selected the mis-classified values from the Random Forest algorithm and compared them with the classified outliers using the medcouple method. For this, the mis-classified outliers were selected in the flow chart used in Orange. However, this resulted in normalized values. In order to determine where the classification was different from the medcouple method, the original values have to be found because the medcouple method works on data that has not been normalized. Using Orange, four mis-classified outliers were identified in the selected data (Table 10).

Table 10. Misclassified values in 145 MHz dataset, March 2022 with original and normalized values

line	max_open_value	avg_open_time	open_seconds	openings	class (MC)	class (RF)
1A	−0.0526	−0.936	0.881	−0.882	0	1
1B	54.61	14.53	2557.58	176.0	0	1
2A	−0.087	−0.998	−0.922	−0.849	0	1
2B	34.11	4.6	170.13	37.0	0	1
3A	0.393	−0.997	0.338	0.382	1	0
3B	52.01	7.32	2416.05	330.0	1	0
4A	−0.083	−0.971	−0.897	−0.987	1	0
4B	34.23	53.72	214.86	4.0	1	0

In this table, the mis-classified outliers are displayed in lines 1A, 2A, 3A and 4A and the original values are in lines 1B, 2B, 3B and 4B. The first two mis-classified outliers, displayed in lines 1 and 2, are false positives, where the last two mis-classified outliers are false negatives. When comparing the original values with the medcouple upper and lower limits, the first and third mis-classified outlier are well within the limits for the medcouple method. There is no clear explanation why these values are mis-classified. However when inspecting the second and fourth mis-classified outlier and comparing the values to the medcouple upper and lower limits, it becomes clear that the lower limit for the max_open_value according to the medcouple method is 34.1 and the original values measured are 34.11 and 34.23. The difference between these two values and the medcouple method are both less than 0.1% and could be considered as outliers as the rounding of the values can possibly explain these differences. Line four however have four openings where the lower limit according to the medcouple value is 8.25 and this should be classified as an outlier. The explanation on the false negative mis-classification is for the future work.

[2] The upper value for open_seconds is outside the normal range for 1 h, or 3600 s. This means that no upper limit outlier on open_seconds can occur in this month.

6 Evaluation and Discussion

The focus of this research was to find forensic artefacts in long-term frequency band occupancy measurements. The base of this investigation are datasets from these types of measurements. They are not widely available so we created our own datasets for this research. Due to this limitation, it is difficult to compare to other datasets.

In both datasets, the time-frame is one hour. As proven by the result of this research a time-frame from one hour is usable to detect outliers. There can be other situations where one hour is too short or maybe too long. If the frequency band under investigation is designated for only very short transmissions, shorter time-frames can be investigated.

When we inspected the threshold values, calculated using the medcouple method over the several months, it shows that these values disperse. Based on this, one cannot use a fixed value for outlier detection over time. In this research we use one month of data as a base for calculating the medcouple value. This has a drawback that, for example, in the first week of the month outliers are not detected reliable due to the lack of data.

When we inspected the percentage of outliers in both the datasets, it appears that there was a significant difference in percentage of outliers between them. Where the 868 MHz dataset have an outlier ratio of approximately 4.5%, the 145 MHz dataset have an outlier ratio of approximately 25%. One can discuss if 25% of outliers still can be validated as outliers, although statistically seen, these values are proven outliers. When inspecting the dataset more closely, there are some time-based series of activities in the usage of this frequency band. It might, although this is not investigated fully, be that quiet (nightly?) periods and periods with peak activity are identified as outliers. This specific behavior does not occur in the 868 MHz dataset, which leads to the idea of time based series in usage.

7 Conclusion and Future Work

7.1 Conclusions

In this research, we explored datasets from long-term frequency band occupancy measurements to identify the usable and interpretable features in such datasets. Using these identified features, we investigated several methods to identify outliers in usage in this type of multi-dimensional data.

In this research we have proven that the *medcouple* method is a reliable statistical method to identify and classify outliers as artefacts in the data from long term frequency band occupancy measurements. The *medcouple* method has proven to be robust, even if the dataset does not have a normal, or Gaussian, distribution. We prove that this method is usable in a forensic sound way to identify and classify artefacts in the dataset and therefor add knowledge to the data. The used datasets are generated using long term frequency band occupancy measurements using the ITU guidelines.

We also showed that, when using both classified datasets to train supervised ML methodologies, forensic artefacts can be reliably identified. We demonstrated that the automated classification process can be executed with both a very high classification accuracy as well as a high sensitivity. We also compared multiple supervised ML algorithms, and Random Forest yields the highest accuracy and sensitivity with a sustained minimum accuracy level of 0.99 and a minimum sensitivity of 0.88 across both datasets.

Finally, we conclude that finding forensic artefacts in long-term frequency band occupancy measurements, using the proposed method, delivers forensic investigators a new method to identify outliers of usage.

7.2 Future Work

No real-time detection of forensic artefacts was investigated or implemented. We assume that real-time detection and forensic artefact alerts, after thorough training of the models can add value to this method. The usage of a moving period in stead of a fixed period of one month as base for outlier detection could also give the opportunity to detect outliers in near real-time.

This research was based on data from a single measurement setup. When using cooperative measurement setups in a given area, data from multiple sensors can be combined to either locate the transmitter or increase the reliability of the outlier detection.

References

1. Redmond, N., Tran, L.N., Choo, K.K.R., Le-Khac, N.A.: Long term evolution network security and real-time data extraction. In: Le-Khac, N.A., Choo, K.K. (eds.) Cyber and Digital Forensic Investigations. SBD, vol. 74, pp. 201–220. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-47131-6_9
2. Mohammad, A., Awin, F., Abdel-Raheem, E.: Case study of TV spectrum sensing model based on ML techniques. Ain Shams Eng. J. **13**(2), 101540 (2022). https://doi.org/10.1016/j.asej.2021.06.026
3. Molina-Tenorio, Y., Prieto-Guerrero, A., Aguilar-Gonzalez, R.: Real-time implementation of multiband spectrum sensing using SDR technology. Sensors **21**(10), 3506 (2021). https://doi.org/10.3390/s21103506
4. International Telecommunication Union, ITU. https://www.itu.int
5. Spectrum occupancy measurements and evaluation. International Telecommunication Union, R-SM.2256-1. https://www.itu.int/pub/R-REP-SM.2256-1-2016
6. Somers, B., Long Term Frequency Band Occupancy Measurements with Increased Bandwidth and Sensitivity using Specmon version 2 (2022). https://doi.org/10.13140/RG.2.2.17393.76640
7. Hubert, M., Van der Veeken, S.: Outlier detection for skewed data. J. Chemom. **22**(3–4), 235–246 (2008). https://doi.org/10.1002/cem.1123
8. Anderson, A., Wang, X., Baker, K.R., Grunwald, D.: Systems for spectrum forensics. In: Proceedings of the 2nd International Workshop on Hot Topics in Wireless, pp. 26–30 (2015)

9. Baltiiski, P., Iliev, I., Kehaiov, B., Poulkov, V., Cooklev, T.: Long-term spectrum monitoring with big data analysis and ML for cloud-based radio access networks. Wireless Pers. Commun. **87**(3), 815–835 (2016)
10. Tavares, C.H.A., Marinello, J.C., Proenca, M.L., Jr., Abrao, T.: ML-based models for spectrum sensing in cooperative radio networks. IET Commun. **14**(18), 3102–3109 (2020)
11. Cotton, M., et al.: An overview of the NTIA/NIST spectrum monitoring pilot program. In: 2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), pp. 217–222 (2015). https://doi.org/10.1109/WCNCW.2015.7122557
12. Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recogn. **58**, 121–134 (2016). https://doi.org/10.1016/j.patcog.2016.03.028
13. Sayakkara, A.P., Le-Khac, N.-A.: Electromagnetic side-channel analysis for IoT forensics: challenges, framework, and datasets. IEEE Access **9**, 113585–113598 (2021). https://doi.org/10.1109/ACCESS.2021.3104525
14. Serhal, C., Le-Khac, N.-A.: Machine learning based approach to analyze file meta data for smart phone file triage. Forensic Sci. Int. Digit. Invest. **37**, 301194 (2021). https://doi.org/10.1016/j.fsidi.2021.301194. ISSN: 2666-2817
15. Yasarathna, T.L., et al.: Crossed-IoT device portability of electromagnetic side channel analysis: challenges and dataset. arXiv:2310.03119. https://arxiv.org/pdf/2310.03119.pdf
16. Reyes, H., Subramaniam, S., Kaabouch, N., Hu, W.C.: A spectrum sensing technique based on autocorrelation and Euclidean distance and its comparison with energy detection for cognitive radio networks. Comput. Electr. Eng. **52**, 319–327 (2016)
17. Fisher, M.J., Marshall, A.P.: Understanding descriptive statistics. Aust. Crit. Care **22**(2), 93–97 (2009). https://doi.org/10.1016/j.aucc.2008.11.003
18. Vinutha, H.P., Poornima, B., Sagar, B.M.: Detection of outliers using interquartile range technique from intrusion dataset. In: Satapathy, S., Tavares, J., Bhateja, V., Mohanty, J. (eds.) Information and Decision Sciences. AISC, vol. 701, pp. 511–518. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-7563-6_53
19. Gao, W., Du, W., Zhao, Z., Min, G., Singhal, M.: Towards energy-fairness in LoRa networks. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), pp. 788–798 (2019). https://doi.org/10.1109/ICDCS.2019.00083
20. Handbook on spectrum monitoring. International Telecommunication Union, R-HDB-23-2011. http://handle.itu.int/11.1002/pub/80399e8b-en
21. Orange data mining, open source ML and data visualisation. https://orangedatamining.com/