

A Reproducible Analysis of RSSI Fingerprinting for Outdoor Localization Using Sigfox: Preprocessing and Hyperparameter Tuning

Grigorios G. Anagnostopoulos, Alexandros Kalousis

Geneva School of Business Administration, HES-SO Geneva, Switzerland

Email: {grigorios.anagnostopoulos, alexandros.kalousis}@hesge.ch

Abstract—Fingerprinting techniques, which are a common method for indoor localization, have been recently applied with success into outdoor settings. Particularly, the communication signals of Low Power Wide Area Networks (LPWAN) such as Sigfox, have been used for localization. In this rather recent field of study, not many publicly available datasets, which would facilitate the consistent comparison of different positioning systems, exist so far. In the current study, a published dataset of RSSI measurements on a Sigfox network deployed in Antwerp, Belgium is used to analyse the appropriate selection of preprocessing steps and to tune the hyperparameters of a kNN fingerprinting method. Initially, the tuning of hyperparameter k for a variety of distance metrics, and the selection of efficient data transformation schemes, proposed by relevant works, is presented. In addition, accuracy improvements are achieved in this study, by a detailed examination of the appropriate adjustment of the parameters of the data transformation schemes tested, and of the handling of out of range values. With the appropriate tuning of these factors, the achieved mean localization error was 298 meters, and the median error was 109 meters. To facilitate the reproducibility of tests and comparability of results, the code and train/validation/test split used in this study are available.

Keywords—IoT, Fingerprinting, Sigfox, Localization, Positioning, Reproducibility, Preprocessing, Machine Learning, knn

I. INTRODUCTION

The recent emergence of Internet of Things (IoT) technologies has made so that a plethora of low power devices make their appearance worldwide, in people's everyday life. The concept of smart cities becomes familiar to the broad public, and numerous applications are being proposed, implemented and deployed in domains such as massive gathering of sensor measurements, automatic control, asset tracking, etc. One domain of great interest concerning the IoT technologies is the offering of Location Based Services.

Outdoor positioning is generally considered a solved problem, as various Global Navigation Satellite Systems (GNSS),

such as the commonly known Global Positioning System (GPS), Galileo, GLONASS and BeiDou, have made outdoors positioning an everyday reality for most users of smartphones and custom positioning devices. These systems achieve an impressive accuracy in their estimates. Nevertheless, the battery consumption of their chipsets is considerable, and when it comes to low power IoT mobile devices, their usage is problematic. Therefore, an alternative way of localizing such devices is needed.

The proliferation of IoT devices has been facilitated by the increasing marketization of different LPWAN technologies, such as Sigfox or LoRaWAN. The localization capabilities of these technologies have been tested in practice. The LoRa alliance has released a geolocalization white paper, presenting an overview of the methods tested and used by its members [1]. Similarly, the Sigfox company is advertising the capabilities of its localization service [2]. A detailed comparative analysis of the most prominent LPWAN technologies can be found in [3] and references therein.

The architecture of these networks is straightforward. Bases-tations, which are connected to a central server, are deployed statically in urban and rural areas. Depending on the use-case, the low power devices might be statically deployed as, for instance, sensors repeating measurements in fixed locations, or might be mobile. In the latter case, they could be mounted on vehicles so that they report sensor measurements in various locations, or they could be used for asset tracking. The devices transmit messages formatted according to the protocol of the technology used, which is received by the basestations in range. All messages are centrally gathered to a central server.

Apart from the content of the message which depends on the use case and the task assigned to the mobile devices, several other types of information concerning the transmission are being reported. These types may be: the Received Signal Strength Indicator (RSSI), the Time of Arrival (ToA), the Time Difference of Arrival (TDoA), the Logarithmic Signal over Noise Ratio (LSNR), etc. This information can be utilized by ranging techniques of localization, such a multilateration, to offer position estimates. Ranging techniques have the advantage that they do not require a surveying phase, and can be used directly, assuming knowledge of the basestations' locations. Nevertheless, since they do not inherently contain information related to the particularities of the environment

Code: <https://doi.org/10.5281/zenodo.3228752>

Data: <https://doi.org/10.5281/zenodo.3228744>

Preprint

©2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

over which they are used, they have a disadvantage, in terms of accuracy, when compared with fingerprinting methods.

Fingerprinting methods rely on datasets that are collected throughout the area of interest. These datasets contain measurements of signal reception values characterizing known locations. These datasets of fingerprints recorded in known locations are used in order to build models which predict the unknown location of new signal receptions. A disadvantage of fingerprinting techniques is that the creation and maintenance of an up-to-date fingerprint database requires considerable effort and cost. One practical way to record such a dataset outdoors is presented in the work of Aernouts et al. [3]. In that work, a dataset is made publicly available and its collection methodology is presented. A total of 20 cars of the Belgian postal services were equipped with low power devices communicating via Sigfox and LoRaWAN with a central server. At the same time, the location of the car, as estimated by a GPS device, was also reported. In these datasets, the GPS estimates are considered as the spatial ground truth of the location of the message transmission.

Recent works in the field of indoor and outdoor positioning [3], [4], [5] have underlined the fact that it is indispensable for the publications of the field to favour the reproducibility of the experiments and the comparability of the results. In this spirit, we utilize a publicly available Sigfox dataset [3], and we share with the community the code of the experiments of the current study (DOI:10.5281/zenodo.3228752), and the train/validation/test set split used in said dataset (DOI: 10.5281/zenodo.3228744), to facilitate a consistent comparison of results in future works.

In the current study, we present a detailed examination of the hyperparameter tuning and of the process of selecting the most appropriate preprocessing methodologies for RSSI fingerprinting on an urban Sigfox setting. A systematic examination of the data preprocessing and the hyperparameter tuning steps can optimize the performance of a localization system. Consequently, it may offer an evaluation of the capabilities of the technology used. This work aims to exemplify such a process, and to characterize and report the capabilities of a Sigfox-based localization system on a well defined urban setting.

The rest of this paper is organized as follows. In Section II, the work relevant to this subject is discussed. Section III presents in detail the dataset used in this work. In Section IV the preprocessing steps analysed in this work are presented. After a concise presentation of the experimental setup in Section V, an extensive presentation and discussion of the results is developed in Section VI. Finally, conclusions drawn and ideas for future work are presented in Section VII.

II. RELATED WORK

Fingerprinting techniques have been common ground for the indoor positioning community for the last two decades [6]. Particularly, RSSI has been the main measurement type that is used [6]. Wi-Fi and Bluetooth have been some of the most commonly used technologies for indoor fingerprinting [7].

Indoor localization using smartphones has been a main field of application of fingerprinting techniques, in which context fingerprinting is either used as a standalone method or combined with other methodologies in hybrid systems [7].

The proliferation of Low Power Wide Area Networks (LPWAN), such as Sigfox and LoRaWAN, has brought a new domain of application of the fingerprinting methods. A recent study [8] has experimentally verified the intuitive assumption that fingerprinting methods outperform, in terms of accuracy, proximity or ranging positioning methods, in a Sigfox setting.

Several works have studied localization methods using LPWAN technologies. Plets et al. [9], have evaluated experimentally RRS and TDoA ranging positioning methods using a LoRaWAN network, reporting median errors of 1250 and 200 meters for RRS and TDoA respectively. Similarly, Podevinj et al. [10] present a tracking algorithm which uses a TDoA ranging method for its initial raw estimates that feed the tracking module, reporting a similar median performance (200m) on the raw localization estimates. The two aforementioned works however do not report the mean error of their systems. Other works [11],[12], have focused on rather specific settings over which they evaluate positioning methods. The first work [11] evaluates an RSSI fingerprinting algorithm in the confined area of a parking lot, with four basestations installed in this area, reporting a mean error of 24 meters. It is noteworthy that since knn fingerprinting methods cannot predict a location outside the area of the fingerprints of the training set, a very limited area of fingerprint collection will drastically affect the error statistics. The later work [12], also deals with the subject of car parking, but with a different approach, as in this case the cars were equipped with tags that exchanged messages among them, only later reporting the RSS to a server. In a small-scale test presented, a maximal error of 8 meters is claimed.

In a series of studies [3], [8], [13], the authors of the IDLab of the University of Antwerp have presented a publicly available Sigfox dataset and have elaborated on the proper configuration of a kNN fingerprinting algorithm. In the publication in which the dataset used in this study was presented to the research community [3], the authors exemplified the usage of the dataset, reporting a 689 meter mean error in this first approach of utilizing the dataset. In a following publication of the IDLab by Janssen et al. [13], the authors present a detailed study of numerous distance metrics and data preprocessing (or data representation) methods, using the same dataset and achieving a mean error of 340 meters. The current study builds on top of the ground these works have set, further examining certain parameters of the preprocessing steps and presenting the impact that an appropriate tuning of these parameters can have in the reduction of the positioning error.

III. THE DATASET USED

Aernouts et al. [3] have made publicly available 3 fingerprinting datasets of Low Power Wide Area Networks. Two of these datasets were collected in the urban area of Antwerp, one using Sigfox and another using LoRaWAN. The third dataset was collected in the rural area between the towns

of Antwerp and Ghent, using Sigfox. The authors underline their motivation by mentioning in their work that: ‘*With these datasets, we intend to provide the global research community with a benchmark tool to evaluate fingerprinting algorithms for LPWAN standards.*’

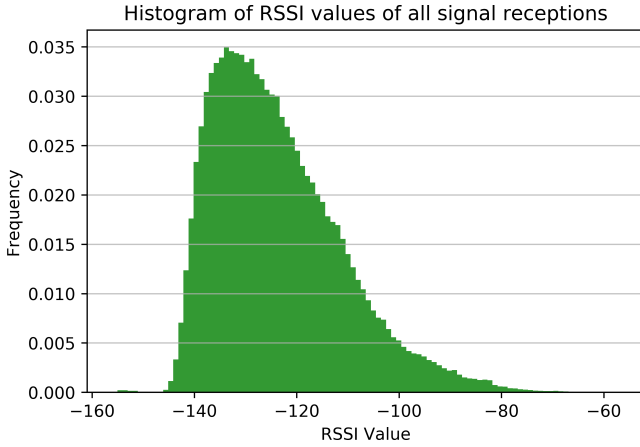


Fig. 1. The histogram of RSSI values of all signal receptions.

In this work, we have used the first dataset, containing Sigfox messages in the urban area of Antwerp. The fingerprints are collected in an area of approximately 53 square kilometers, though the majority of them lay in the central area of Antwerp which is approximately half the size of the full area. A total number of 14378 messages are reported in the dataset. Each message contains the following information: the RSSI value of the transmitted signal by each of the 84 base stations, and the spatial ground truth of the signal’s transmission location, as estimated by a GPS device, alongside the Sigfox devices. Undeniably, the fact that an estimate which is subject to error is used as ground truth introduces bias. Nevertheless, the error of GPS is at the order of a few tens of meters while the localization accuracy ranges at the order of several hundreds of meters.

In order to obtain a feeling of the distribution of the RSSI values of the dataset, the histogram of all 317126 RSSI values that are present in the dataset is plotted in Figure 1. A big part of the distribution (more than 60% of the data) is concentrated in the $[-140, -120]$ range, having almost 10% of the data in the value range above -100 . In cases where a basestation did not receive a message, an out of range RSSI value of -200 was inserted, in order not to leave an empty value and also for differentiating this entry from the minimum RSSI values actually received (-156). The out of range values of -200 were not included in the histogram’s creation.

For the purposes of our study, we have split the dataset into a training, a validation and a test set containing 70%, 15%, and 15% of the sample respectively (10063, 2157 and 2157 entries, in absolute numbers). The spatial distribution of the ground truth locations of the three aforementioned sets can be seen in Figure 2.

In the spirit of verifiability, reproducibility and comparability of results, it is important not only to report performance metrics over a publicly available dataset, but to also report the specific way the dataset is split into a training, a validation and a test set. In this way, researchers will be able to train models on the same training set, make decisions over the optimal tuning at the same validation set, and most importantly report unbiased performance results on the same test set. The three subsets that are used in the current study are made available to the research community for future reference. For the cases that a different training and validation strategy is desired, or an entirely different dataset is to be used, the full code implementation is available so that the same tests can be reproduced in these different settings.

IV. DATA PREPROCESSING

In a highly cited work, Torres-Sospedra et al. [14] have presented a systematic study of data preprocessing methods, and of distance and similarity metrics for Wi-Fi fingerprinting in indoor positioning systems. In that work, the authors have shown the great significance of an appropriate data preprocessing step, and have presented four alternative methods of preprocessing the fingerprint data. These four data representation methods are the following:

As defined in [14], *positive* values data representation subtracts the minimum RSSI value from all the entries.

$$Positive_i(x) = \begin{cases} (RSS_i - min) & \text{if basestation } i \text{ received the message and } RSS_i \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where RSS_i is the RSS from the i_{th} basestation, min is the lowest RSS value minus 1 among all RSS values of the database [14]. Also, τ is a threshold value, used so that basestations with intensities lower than τ are considered as not-detected, and the lowest possible value is assigned to them.

It is worth emphasizing here that, since utilizing in the training set’s preprocessing any kind of information coming from the validation or the test set would introduce *information leakage*, the min should only be calculated by the training set data. Calculating the min , or normalizing according to the data of the full dataset, would imply a breach of protocol of the train/validation/testing scheme, where the validation and testing sets are supposed to be completely unknown during the preprocessing phase.

- Replace all values $RSS_I < \tau$ with τ , where $\tau \geq min$
- Define $Positive_i(x)$ as:

$$Positive_i(x) = RSS_I - \tau \quad (2)$$

The *normalized* data representation, normalizes the *positive* values data representation into the $[0,1]$ range. (To be noted that the RSSI values are negative)

$$Normalized_i(x) = Positive_i(x)/(-min) \quad (3)$$

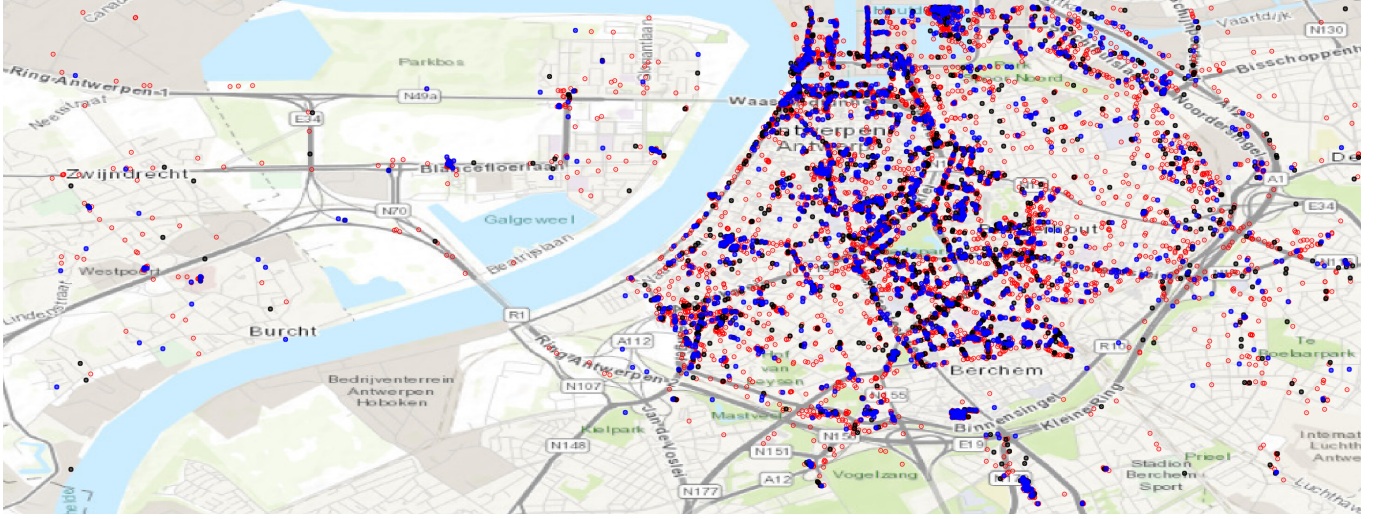


Fig. 2. The spatial distribution of the data points of the dataset, as reported by the GPS devices used to establish the ground truth. The train, validation and tests sets that were created in this work are depicted in red, black and blue respectively.

The *exponential* and *powered* representations are the result of the intention to go beyond the linear handling of the RSSI values, since the RSSI values correspond to a logarithmic scale. More particularly, the *exponential* and *powered* representations are defined as follows:

$$Exponential_i(x) = \frac{\exp(\frac{Positive_i(x)}{\alpha})}{\exp(\frac{-min}{\alpha})} \quad (4)$$

$$Powered_i(x) = \frac{(Positive_i(x))^\beta}{(-min)^\beta} \quad (5)$$

The proposed default values for the parameters α and β are $\alpha = 24$ and $\beta = e$, where e is the mathematical constant. Nevertheless, since these parameters were selected upon testing with Wi-Fi signals in an indoor setting, which has a different range and distribution of RSSI values, it is a good advice to adjust them in accordance with a new setting. This ground setting analysis of data processioning presented in [14], has been used as a reference in the work by Janssen et al. [13], and is studied in further detail in the current work. The impact of these preprocessing steps and of the appropriate tuning of their parameters on the achieved positioning accuracy is extensively discussed in Section VI, along with all results of this study.

V. EXPERIMENTAL SETUP

In this study, a detailed examination of various preprocessing methods is presented, along with a hyperparameter tuning of the k-nearest neighbours (kNN) method. For the experiments presented in this study the free machine learning library for the Python language, scikit-learn, was used. Particularly, scikit-learn version 0.19.1 and Python 3.5.5 were used. The Haversine formula has been used for measuring distances on the reported experiments. Upon completion of the experiments, the Vincenty formula, which gives the shortest

geodesic distance on an ellipsoid modeled earth, was also tested. It is straightforward that the comparisons among the performance of models remain consistent if either of the two formulas is used. The absolute distances measured with the two formulas differ by less than 0.5%.

The common train/validation/test set division methodology is used on the analyses presented in this work. The train set is used to train the model. In the case of knn, no training process needs to take place per se, thus the training set practically creates the space of neighbors among which each new fingerprint is compared against. The validation set is used for evaluating different candidate models and checking the appropriateness of the selected hyperparameter values and preprocessing steps. Based on the performance on the validation set, the optimal model configuration is chosen. As both train and validation sets have participated in the configuration of the final model, an unbiased final evaluation of the model's performance needs to take place using a third set, containing previously unseen data: the test set. Therefore, during all preprocessing, training and tuning steps, the test set is inaccessible, and no information stemming from it is to be used.

Initially, similarly to the work by Janssen et al. [13], we examine various distance metrics (the full list of the Distance Metrics class of scikit-learn) and tune the hyperparameter k for each distance metric. We perform this analysis twice: once using the dataset as it is with the out-of-range -200 value unchanged, and then again, replacing it with the experimentally found minimum of actually received RSSI values in the training set minus one (-157). Moreover, we examine the performance of a wide range of candidate values of threshold value τ of Equation 2. Lastly, concerning the parametric preprocessing representations, *exponential* and *powered*, a tuning of their respective parameters (α and β) is performed.

VI. RESULTS

A. Distance Metrics, Data Preprocessing and k

In this test, we examine various distance metrics and tune the hyperparameter k for each distance metric. In the experiments of the current study, all four data representations defined in Equations 2-5 were tested as a preprocessing step. The results of these tests, presented in Tables I and II, have verified a fact that was observed in previous works as well ([14], [13]): the *exponential* and *powered* representations clearly outperform the *positive* and *normalized* data representations, for all distance metrics used. For simplicity, in Tables I and II we only report the results of the two most performant methods, *exponential* and *powered*. It should be underlined that selection of the optimal hyperparameter value k and the localization error statistics reported in Tables I and II are calculated on the validation set. Moreover, in the context of this test, the default values of the parameters ($\alpha = 24$ and $\beta = e$) of the *exponential* and *powered* representations are used.

Concerning the distance metrics, apart from those reported in Tables I and II, another family of distance metrics available in scikit-learn was evaluated, but proven to be entirely unsuitable. That family of metrics such as the Jaccard, Matching, Dice or Kulsinski distance, are intended for boolean-valued vector spaces, setting as True any non-zero entry. Consequently, those representations utilize a binary type of information stating if each basestation has received the message or not. Overall, the results show that, in accordance with the previous works, the Bray-Curtis metric (equivalent to the Sørensen metric mentioned in previous works [14], [13] offers the best results in terms of accuracy.

Comparing the results of Tables I and II, the following conclusions can be drawn. In Table I, *exponential* representation is more performant than *powered*, for all distance metrics except for the Canberra distance. On the contrary, in Table II it is *powered* representation that offers better results, for all but one distance metrics. In Table I the best performance is achieved by the Bray-Curtis metric and the *exponential* data representation, with a mean and median error of 344 and 148 meters respectively. For the results of Table II, the best performance is achieved by the Bray-Curtis metric and the *powered* data representation, with 319 and 123 meters of mean and median error respectively. The corresponding mean and median values of error on the test set are, 301 and 109 respectively.

While the results obtained by the *positive* and *normalized* representation were not reported in Tables I and II, it is worth to briefly discuss the performance they achieve. The Bray-Curtis distance metric remains the best performing one for both data representations. Since the *normalized* representation is just a rescaled version of the *positive* data representation, both methods achieve identical results in this setting. When the dataset is used as is, meaning with $\tau = -200$, the mean error on the validation set is 552 meters, while for $\tau = -157$, the error is 400 meters. A first observation is that an appropriate transformation of the RSS values may change entirely the level

TABLE I
LOCALIZATION ERROR ANALYSIS ON THE VALIDATION SET, REPORTING THE OPTIMAL k VALUE OF ALL DISTANCE METRICS AND TWO PREPROCESSING STRATEGIES, WITH THE DEFAULT REPLACEMENT OF THE MISSING VALUES WITH THE $\tau = -200$ VALUE

DistanceMetric	Exponential RSS			Powered RSS		
	k	mean	median	k	mean	median
euclidean	9	355	155	8	387	182
manhattan	5	360	145	6	389	183
chebyshev	7	428	222	3	506	285
hamming	10	1100	950	7	1058	906
canberra	6	516	316	11	572	368
braycurtis	8	344	148	6	364	168

TABLE II
LOCALIZATION ERROR ANALYSIS ON THE VALIDATION SET, REPORTING THE OPTIMAL k VALUE OF ALL DISTANCE METRICS AND TWO PREPROCESSING STRATEGIES, HAVING ADJUSTED THE MISSING VALUES TO THE EXPERIMENTAL MINIMUM RSSI VALUE, MINUS 1 ($\tau = -157$)

DistanceMetric	Exponential RSS			Powered RSS		
	k	mean	median	k	mean	median
euclidean	8	345	148	8	343	141
manhattan	7	348	138	6	343	136
chebyshev	4	409	191	4	399	175
hamming	10	1100	950	7	1065	912
canberra	6	420	205	8	470	275
braycurtis	5	341	126	6	319	123

of accuracy, leading from an initial error of 552 meters in the linear handling of the RSS values, to a 344 meters error. Furthermore, adjusting the τ value, and therefore, the way out-of-range values are set, may further improve the performance, thus the 319 meters mean error observed in Table II.

In an attempt of an interpretation of this difference in performance among the two different threshold values used, the following arguments are presented. It appears that the selection of an out-of-range value may drastically affect the performance of the positioning method. In the studied dataset, the -200 out-of-range value is quite distinct to the experimental minimum RSSI found in the dataset (-156). This artificial gap between these two values may assign a significant importance to the distinction between a very distant gateway receiving the signal and a non-receiving one. On the other hand, bridging this gap may treat these two cases as more similar, and favor the distinction of RSSI values among closer detected basestations, improving the distance measurement between fingerprints, and consequently the efficient selection of the closest neighbours.

B. Threshold Value τ

In this test, the impact of the value of the τ threshold on the localization performance is examined. The best configuration found so far is used ($k = 6$, with a *powered* data representation), for examining all τ values in the range $[-200, -130]$. Setting τ with a value higher than the experimental minimum of -156 , will replace all values in the range $[-156, \tau]$ with τ . Out-of-range values that are lower than the experimental minimum, are simply set to τ .

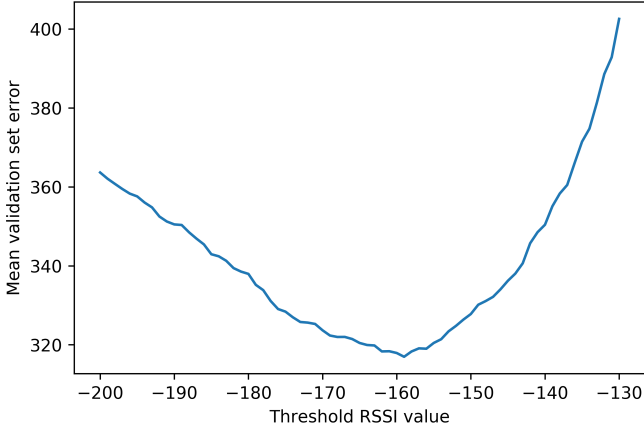


Fig. 3. The mean error on the validation set for different values of the threshold τ of Equation 2.

The results of this analysis are reported in Figure 3. It is evident that the best performance comes from values around the experimental minimum. In particular, the lowest mean error in the validation set was 317 meters, and it was given by $\tau = -159$. The corresponding performance in the test set was 298 meters mean error and 109 meters median error. The optimal value $\tau = -159$ is just below the experimental minimum of received RSSI values, which suggests that putting the out-of-range value just below the experimental minimum is the best option, according to this test.

C. Parameters a and b of the Exponential and Eowed Data Representations

As presented in Section IV, the *exponential* and *powered* data representations rely on a parameter (α and β respectively). The study that introduced these two representations [14] recommends default values for these parameters, upon experimentation with data coming from indoor WiFi measurements. The range and distribution of those values are different than those of the dataset used in this work. Therefore, an appropriate adjustment of the values of parameters α and β would adapt these data transformations to optimally fit this different setting.

1) *Parameter α of the Exponential Data Representation:* Regarding the parameter α of the *exponential* data representation, a range of candidate values has been tested. It is reminded that the default value of α is 24. Integer values in the range [10,40] have been evaluated in the best configuration found so far, concerning the *exponential* data representation. Thus, the Bray-Curtis distance has been used with $k = 5$ and $\tau = -157$. The results are presented in the plot of Figure 4.

The value $\alpha = 19$ provides the lowest mean validation error of 339 meters. The corresponding test set performance is characterized by a mean error of 318 meters and a median of 117 meters. In the plot of Figure 4 it can be observed that there is a significant difference in performance comparing to the default value ($\alpha = 24$).

In Figure 5, the results of a similar examination are depicted. This time, the test spans the space of candidate values of both

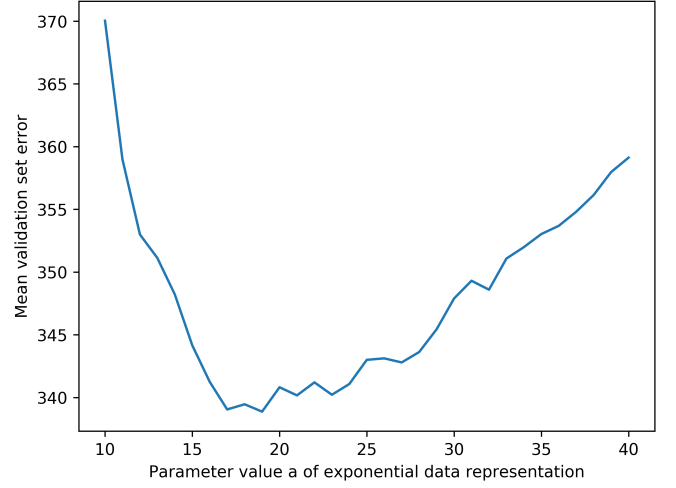


Fig. 4. The mean error on the validation set for different values of the parameter α of the *exponential* data representation.

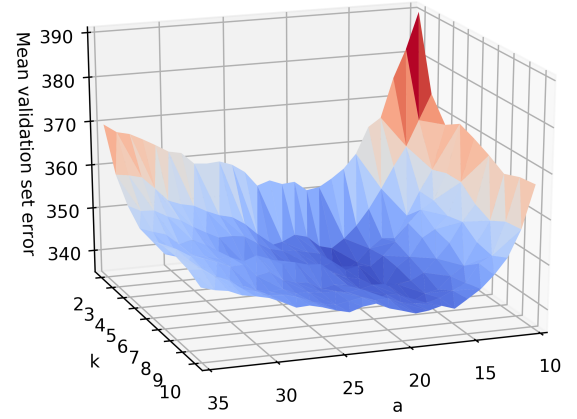


Fig. 5. The mean error on the validation set for different values of the parameter α of the *exponential* data representation and of the k hyperparameter of kNN.

the parameter α of the *exponential* data representation and of the k hyperparameter of kNN. The minimum mean validation error is provided by $\alpha = 18$ and $k = 4$, and is equal to 336 meters. The corresponding performance in the test set shows a mean error of 322 meters and a median error of 110 meters. There are several tuples of values of α and k in the proximity of the ones reported above that provide very similar results.

2) *Parameter β of the Exponential Data Representation:* An analysis similar to the previous one is performed for the parameter β of the *powered* data representation. The range [2,3] has been spanned with a granularity of 0.02. The default value of β is the constant e , which is equal to 2.718. The best configuration found concerning the *powered* data representation has been used. Thus, the Bray-Curtis distance has been used with $k = 6$. The results are presented in the plot of Figure 6.

The value $\beta = 2.6$ provides the lowest mean validation

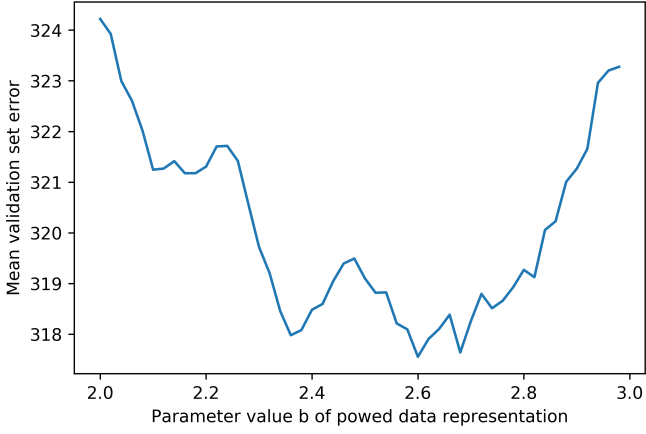


Fig. 6. The mean error on the validation set for different values of the parameter β of the *powered* data representation.

error of 318 meters. The mean error on the test set is 298 meters and the median 108 meters. In this case that parameter β is studied, the difference of performance of the best β value compared to the one of the default value, is less significant than in the relevant analysis of α that preceded. Indicatively, it is noted that the default value of β gave a mean error of 319 meters.

In Figure 7, the results for combinations of candidates values of both the parameter β of the *powered* data representation and of the k hyperparameter of kNN are reported. The minimum mean validation error is provided by $\beta = 2.6$ and $k = 6$, same as in the previous analysis of Figure 6. The convex form surface of Figure 7 reveals that the lowest values of the mean error appear in a close proximity, since there are several tuples of values α and k in the proximity of the optimal ones reported above that provide very similar results.

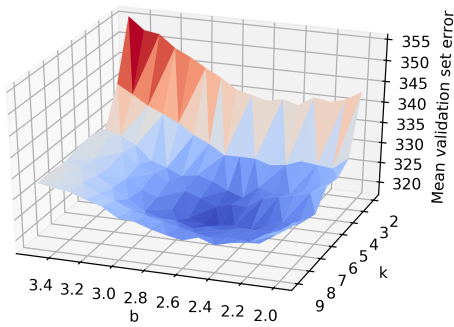


Fig. 7. The mean error on the validation set for different values of the parameter β of the *powered* data representation and of the k hyperparameter of kNN.

VII. CONCLUSIONS AND FUTURE WORK

In this work, we presented a detailed study of the process of selecting the most appropriate preprocessing methodologies and performing hyperparameter tuning for RSSI fingerprinting in an urban Sigfox setting. We have discussed the ways of appropriately preprocessing of the RSSI data, so as to improve the accuracy of the studied fingerprinting localization method. Moreover, identifying the limits of the achievable accuracy of a Sigfox-based positioning system deployed in an urban area has been a main motivation of this work.

The examination of the results of this study may offer several take-away messages. A linear handling of the RSSI values, with the *positive* and *normalized* data representations, achieved an above 500-meter mean error on the validation set. Introducing non-linear transformations, proposed by the relevant literature as more appropriate for the handling of RSSI values which correspond to a logarithmic scale, reduced the error to the level of 344 meters. To the best of our knowledge, this study is the first to go beyond these steps and further optimize the handling of out-of-range values and the tuning of the parameters α and β of the preprocessing data transformation, so as to match the particularities of this outdoors setting, which does not utilize a Wi-Fi system as the original paper [14], but another technology, namely Sigfox. The best performing setting proposed in this work achieves a mean error of 317 meters on the validation set. This setting achieves a 298-meter error on the test set, with the corresponding median error being 109 meters.

A comparison with the performance of previous works would not be strictly consistent, even if the same dataset has been used, since the validation and test sets are not the same. Nevertheless, we could discuss the order of magnitude of the error achieved. In the initial work where the used dataset became available [3], the linear handling of the RSSI value resulted in a high mean error of 689 meters. A more recent work [13], has utilized the data preprocessing methods proposed by Torres-Sospedra et al. [14], reporting a mean validation error of 340. The results of Table I of subsection VI-A, report a similar mean error of 344 meters, for the same best data representation found in [13], the *exponential*.

The additional analysis of the threshold value τ , and the appropriate adjustment of the preprocessing parameters α and β proposed by this work, further improves the localization accuracy. Moreover, it is noteworthy that the appropriate adjustment of τ , sets a different preprocessing method as the best performing one. While in Table I of subsection VI-A, as well as in the work of Janssen et al. [13], the *exponential* data representation is suggested by the obtained results, the appropriate adjustment of τ sets the *powered* data representation as preferable. It is these preprocessing adjustments that give a mean error of 317 meters on the validation set and 298 meters in the test set, with 121 and 109 meters for the corresponding median errors.

The improvements in accuracy obtained by firstly using and secondly tuning these preprocessing steps indicate the

significance of these steps. An appropriate preprocessing of the data may have huge impact on the performance of the positioning system using them, and thus it is recommended as an indispensable step of the model selection process

A driving force of motivation of the authors has been the intention to encourage the sharing of material among the positioning community, which can facilitate consistent comparisons, and accelerate the advancement of the field. Being thankful for the public offering of the dataset to the community by Aernouts et al. [3], we proceed in sharing the dataset's split into train/test/validation sets, used in this work, as well as the code used for the tests.

As localization with LPWANs is a rather recent field of study, there aren't many publicly available datasets so far. Moreover, the size of the dataset plays a crucial role, when machine learning approaches are used for localization. A dense spatial sampling of the area of interest can positively affect the localization performance. Additionally, other types of measurements, apart from the RSSI value, such as ToA, TDoA or LSNR values, would be of great interest to be studied. The intention of the authors is to work in the direction of creating and sharing such datasets. We also invite the community to embrace this effort that can accelerate the improvement of the field.

Lastly, an interesting future direction for studies such as the current one may be the evaluation of the computational complexity of the machine learning methods and the distance metrics that are tested. In the current work, the focus has been on the performance of the positioning system in terms of accuracy. Nevertheless, in practical settings factors such as the computational complexity may play a crucial role in the selection of the model used.

ACKNOWLEDGEMENTS

This work was funded by the Commission for Technology and Innovation CTI, of the Swiss federal government, in the frame of the OrbiLoc project (CTI 27908.1 PFES-ES).

REFERENCES

- [1] LoRa Alliance Strategy Committee. Geolocalization whitepaper. Accessed on 24.05.2019. [Online]. Available: <https://lora-alliance.org/resource-hub/lora-alliance-geolocation-whitepaper>

- [2] Sigfox. Sigfox geolocalization. Accessed on 24.05.2019. [Online]. Available: <https://www.sigfox.com/sites/default/files/Sigfox%20Geolocation%20Flyer%20Digital%20A4.pdf>
- [3] M. Aernouts, R. Berkvens, K. Van Vlaenderen, and M. Weyn, "Sigfox and lorawan datasets for fingerprint localization in large urban and rural areas," *Data*, vol. 3, no. 2, 2018. [Online]. Available: <http://www.mdpi.com/2306-5729/3/2/13>
- [4] R. Montoliu, E. Sansano, J. Torres-Sospedra, and O. Belmonte, "Indoorloc platform: A public repository for comparing and evaluating indoor positioning systems," in *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Sep. 2017, pp. 1–8.
- [5] S. Adler, S. Schmitt, K. Wolter, and M. Kyas, "A survey of experimental evaluation in indoor localization research," in *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Oct 2015, pp. 1–10.
- [6] S. Xia, Y. Liu, G. Yuan, M. Zhu, and Z. Wang, "Indoor fingerprint positioning based on wi-fi: An overview," *ISPRS International Journal of Geo-Information*, vol. 6, no. 5, 2017. [Online]. Available: <http://www.mdpi.com/2220-9964/6/5/135>
- [7] P. Davidson and R. Piché, "A survey of selected indoor positioning methods for smartphones," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 1347–1370, Secondquarter 2017.
- [8] M. Aernouts, B. Bellekens, R. Berkvens, and M. Weyn, "A comparison of signal strength localization methods with sigfox," in *2018 15th Workshop on Positioning, Navigation and Communications (WPNC)*, Oct 2018, pp. 1–6.
- [9] D. Plets, N. Podevijn, J. Trogh, L. Martens, and W. Joseph, "Experimental performance evaluation of outdoor tdoa and rss positioning in a public lora network," in *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Sep. 2018, pp. 1–8.
- [10] N. Podevijn, D. Plets, J. Trogh, L. Martens, P. Suanet, K. Hendrikse, and W. Joseph, "Tdoa-based outdoor positioning with tracking algorithm in a public lora network," *WIRELESS COMMUNICATIONS & MOBILE COMPUTING*, p. 9, 2018. [Online]. Available: <http://dx.doi.org/10.1155/2018/1864209>
- [11] W. Choi, Y.-S. Chang, Y. Jung, and J. Song, "Low-power lora signal-based outdoor positioning using fingerprint algorithm," *ISPRS International Journal of Geo-Information*, vol. 7, no. 11, 2018. [Online]. Available: <http://www.mdpi.com/2220-9964/7/11/440>
- [12] P. Gotthard and T. Jankech, "Low-cost car park localization using rssi in supervised lora mesh networks," in *2018 15th Workshop on Positioning, Navigation and Communications (WPNC)*, Oct 2018, pp. 1–6.
- [13] T. Janssen, M. Aernouts, R. Berkvens, and M. Weyn, "Outdoor fingerprinting localization using sigfox," in *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Sep. 2018, pp. 1–6.
- [14] J. Torres-Sospedra, R. Montoliu, S. Trilles, scar Belmonte, and J. Huerta, "Comprehensive analysis of distance and similarity measures for wi-fi fingerprinting indoor positioning systems," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9263 – 9278, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0957417415005527>