

Article

Outlier Detection in Time-Series Receive Signal Strength Observation Using Z-Score Method with S_n Scale Estimator for Indoor Localization

Abdulmalik Shehu Yaro ^{1,2,*} , Filip Maly ¹ and Pavel Prazak ¹ 

¹ Department of Informatics and Quantitative Methods, Faculty of Informatics and Management, University of Hradec Kralove, 500 03 Hradec Kralove, Czech Republic

² Department of Electronics and Telecommunications Engineering, Ahmadu Bello University, Zaria 810106, Nigeria

* Correspondence: abdulmalik.yaro@uhk.cz

Abstract: Collecting time-series receive signal strength (RSS) observations and averaging them is a common method for dealing with RSS fluctuation. However, outliers in the time-series observations affect the averaging process, making this method less efficient. The Z-score method based on the median absolute deviation (MAD) scale estimator has been used to detect outliers, but it is only efficient with symmetrically distributed observations. Experimental analysis has shown that time-series RSS observations can have a symmetric or asymmetric distribution depending on the nature of the environment in which the measurement was taken. Hence, the use of the Z-score method with the MAD scale estimator will not be efficient. In this paper, the S_n scale estimator is proposed as an alternative to MAD to be used with the Z-score method in detecting outliers in time-series RSS observations. Performance comparison using an online RSS dataset shows that the Z-score with MAD and S_n as scale estimators falsely detected about 50% and 13%, respectively, of the RSS observations as outliers. Furthermore, the average absolute RSS median deviations between raw and outlier-free observations are 3 dB and 0.25 dB, respectively, for the MAD and S_n scale estimators, corresponding to a range error of about 2 m and 0.5 m.

Keywords: RSS fluctuation; MAD; S_n scale-estimator; outlier



Citation: Yaro, A.S.; Maly, F.; Prazak, P. Outlier Detection in Time-Series Receive Signal Strength Observation Using Z-Score Method with S_n Scale Estimator for Indoor Localization. *Appl. Sci.* **2023**, *13*, 3900. <https://doi.org/10.3390/app13063900>

Academic Editors: Sven Casteleyn, Aleksandr Ometov and Joaquín Torres-Sospedra

Received: 15 February 2023

Revised: 6 March 2023

Accepted: 16 March 2023

Published: 19 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Indoor wireless localization involves locating an indoor radio frequency (RF) transmitter using a position-dependent signal parameter (PDSP) obtained from the detected RF signal with a localization algorithm [1–3]. The received signal strength (RSS) is the most used PDSP for indoor localization because it can be derived from a variety of wireless technologies with comparable localization performance. The RSS represents the average power of the received RF signal and can be used with two different kinds of localization algorithms, namely fingerprinting and trilateration [3,4]. The trilateration approach makes use of path-loss propagation models to establish a relationship between the RSS measurements obtained by the receiver and the actual separations between the RF transmitters and receiver [4]. The fingerprinting approach involves two phases. The first phase, known as the offline phase, involves the generation of a radio map using RSS measurements obtained from several RF transmissions using a single receiver at fixed reference locations. The second phase is referred to as the “online phase”. It involves locating an indoor user by searching the radio map for the corresponding reference location to the indoor user’s instantaneously acquired RSS measurements [3]. There are several wireless technologies used for indoor localization, but the most used are Wi-Fi, Bluetooth, Zigbee, and ultra-wideband [3,5].

The localization accuracy of the indoor wireless localization system depends on several factors. One of these factors is the accuracy of the RSS measurement, which is affected

by RSS fluctuation [6,7]. RSS fluctuates due to the presence of walls, furniture, crowds, ambient and temporal conditions, and differences in RSS measuring device hardware and configurations [6,8]. Several research papers [6,7,9–21] have proposed various techniques for dealing with RSS fluctuations. Taking multiple RSS observations for each reference location is recommended in most research works. The measurements are then smoothed using a mean-averaging filter, a Kalman filter, a moving average filter, or a median filter to produce a single RSS representative of the observation [6,13,19–21]. Other, less commonly used techniques are based on RSS measurement diversity. This diversity is achieved by using Wi-Fi or Bluetooth anchors with multiple transmitter antennae. Alternatively, in the case of BLE, it can be accomplished by using different broadcast channels (channels 37, 38, and 39), and then aggregating the RSS measurements to obtain a single RSS measurement [11,22,23]. Table 1 summarizes recent research on how to deal with RSS fluctuation.

Table 1. A summary of techniques for dealing with RSS fluctuation.

Research Work	Summary of Technique
Koubâa et al. (2012) [20]	At each reference location, several RSS observations were collected and a smoothing technique was used to find a mean RSS representative.
Cheng et al. (2013) [11]	Instead of measuring aggregated RSS, the RSS ratio coefficient was used. Utilized a 3-antenna receiver system where the RSS ratio coefficient was computed using the log-normal shadowing model.
Ibrahim et al. (2018) [24]	The RSS measurements were normalized using the Z-score after the time-series RSS matrix was generated (concatenation of multiple observations of the RSS fingerprint).
Huang et al. (2019) [22]	Instead of the aggregated RSS obtained by the smartphone device, RSS measurements from each BLE advertisement channel were used.
Polak et al. (2021) [23]	The mean RSS was calculated by averaging two consecutive observations from a wireless AP with multiple advertising antennas.
Zhou et al. (2021) [7]	Instead of using direct RSS subtraction, the authors transformed RSS to Q-based RSS and used a Q-based RSS subtraction.
Balaji et al. (2021) [21]	Each AP had N RSS measurements taken, which were then arranged in descending order. The RSS representative is determined by taking the median of RSS measurements above the -90 dBm threshold.
Zhou et al. (2022) [19]	A non-stationary, non-linear RSS sequence was generated, and then smoothed using the empirical mode decomposition threshold smoothing (EMDT) technique to obtain the RSS representative of the observation.
Flueratoru et al. (2022) [6]	Utilizing aggregated RSS measurements from multiple observations, RSS representative was obtained.
Wenqing et al. (2022) [17]	Ten RSS measurements were obtained using ten different parameters from the power–distance equation, which were then aggregated.
Nabati et al. (2023) [25]	At each reference location, multiple RSS measurements from each AP were obtained and averaged.

When there are outliers in the time-series RSS observation, most of the techniques presented in Table 1 are less effective. For instance, smoothing techniques such as exponential smoothing and moving averages can help reduce the variability in time-series RSS observations, making it easier to identify outliers. However, this technique is ineffective in reducing the impact of outliers on the observation in the presence of highly influential outliers. RSS outliers are RSS observations that appear to be significantly different from the rest of the time-series RSS observations. They have a disproportionately large impact on the RSS mean calculation. In this paper, a method to detect outliers in time-series RSS observations is proposed.

2. Overview of Time-Series RSS Observation Outlier Detection Methods

The best outlier detection method for a univariate time-series observation depends on several factors. Some of these factors include the distribution of the observation, the size of the observation, and the desired level of sensitivity to outliers [26]. Some common methods for detecting outliers in univariate time-series observations are presented in the following subsections.

2.1. Z-Score Method

This method calculates the difference between each RSS observation and the mean of the time-series RSS observations. Then, the result is divided by the standard deviation (SD) of the observation [27]. RSS observations that have a Z-score greater than a pre-determined threshold are considered outliers.

Given the time-series RSS observations shown in Equation (1) obtained at the i -th reference location from the j -th wireless access point (AP):

$$\mathbf{rss}_{i,j} = [rss(1), rss(2), \dots, rss(N)] \quad \text{for } 1 \leq n \leq N \quad (1)$$

where $rss(n)$ is the individual RSS observations.

The Z-score of an RSS observation, $rss(n)$ is calculated as follows [26,28]

$$Z_{score} = \frac{rss(n) - \mu}{\sigma} \quad (2)$$

where “ μ ” and “ σ ” are the mean and SD, respectively, of the time-series RSS observation.

A Z-score of 0 indicates that the RSS observation is equal to the time-series observation's mean. A positive Z-score and a negative Z-score show that the RSS observation is above and below the mean, respectively. An RSS observation is considered an outlier if its Z-score value exceeds a predefined threshold. The most common threshold value for detecting outliers is ± 3 . This means that when the Z-score value of an RSS observation is greater than ± 3 , it is an outlier.

2.2. Interquartile Range Method

This method calculates the interquartile range (IQR) of the time-series RSS observation, which is the difference between the 75th and 25th percentiles. RSS observations that are outside of the range defined by the IQR are considered outliers [28]. To use the IQR method for outlier detection, the following steps are followed:

- Calculate the first quartile (Q1), the third quartile (Q3), and the interquartile range ($IQR = Q3 - Q1$) of the data.
- Identify the lower and upper bounds, which are defined as $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively.
- Any RSS observation that falls below the lower bound or above the upper bound is considered an outlier.

The 1.5 multiplier used in the upper and lower bound calculations corresponds to approximately 3 SD from the median. This covers 99.7% of the data in a normal distribution. The value 1.5 was chosen at random and can be adjusted to be strict in detecting outliers. It is important to note that the IQR method assumes that the data is approximately normally distributed. Therefore, it may not be appropriate for observations with non-normal distributions or highly skewed observations [26]. Alternative methods, such as the Z-score method or the modified Z-score method, may be more appropriate.

2.3. Moving Average and Moving Median Method

This method replaces each RSS observation with the average or median of surrounding RSS observations over a sliding window. Outliers are detected by comparing the original RSS observations to the moving average or median. Using the time-series RSS observation

in Equation (1), the moving average of RSS observations within a window size “ m ” is calculated. Mathematically, the moving average at time “ t ” can be calculated as follows [26]:

$$MA(t) = \frac{rss(t) + rss(t-1) + \dots + rss(t-m+1)}{m} \quad (3)$$

where $rss(t), \dots, rss(t-m+1)$ are the m consecutive RSS observations at time t .

Like the moving average, the moving median takes the median value of the set of RSS observations as opposed to averaging the RSS observations. The value that divides the observation into its upper and lower halves is known as the median. Since the moving median is unaffected by outliers or skewness, it can be used to smooth out time-series RSS observations that contains these elements.

Given the RSS observations in Equation (1), the moving median at time “ t ” for a window size “ m ” is calculated as follows [26]:

- Sort the m consecutive RSS observations $rss(t), rss(t-1), \dots, rss(t-m+1)$ in ascending or descending order.
- Take the middle RSS observation as the moving median if m is odd.
- Take the average of the two middle RSS observations as the moving median if m is even.

Both moving average and moving median methods are used to smooth out time-series RSS observations and make it easier to identify trends and patterns. However, the moving median is typically more robust to outliers and skewness in the observations compared to the moving average.

2.4. Density-Based Method

This method uses a density function to calculate the likelihood of a data point being an outlier. Data points that have a low likelihood of being part of the data distribution are considered outliers. The local outlier factor (LOF) is a common density-based method for outlier detection. The LOF computes an RSS observation’s density in relation to its neighbours and uses this information to identify outliers. Using the RSS observation in Equation (1), the LOF of an RSS observation, $rss(n)$ is calculated as follows [26,27]:

- For each RSS observation, $rss(n)$, find the “ k ” nearest neighbours using a distance metric such as Euclidean distance.
- Calculate the reachability distance for each RSS observation, $rss(w)$ in the “ k ” nearest neighbours of, $rss(n)$, which is the maximum of the Euclidean distance between $rss(n)$ and $rss(w)$ and the distance between $rss(n)$ and its “ $k-1$ ” nearest neighbour.
- Calculate the local reachability density (LRD) of each RSS observations in the “ k ” nearest neighbours of $rss(n)$, which is the inverse of the average reachability distance of $rss(w)$ to its “ k ” nearest neighbours.
- Calculate the LOF of each RSS observation, $rss(n)$ as the ratio of the average LRD of its “ k ” nearest neighbours to its own LRD. Mathematically, the LOF is:

$$LOF(rss(n)) = \frac{1}{k} \sum \left(\frac{LRD(rss(l))}{LRD(rss(n))} \right) \quad (4)$$

- Assign a threshold value to the LOF, typically 1, above which an RSS observation is considered an outlier.

The LOF measures how unusual an RSS observation is in comparison to its neighbours. A high LOF value indicates that the RSS observation is an outlier due to its low local reachability density in comparison to its neighbours [27].

As mentioned earlier, there is no best outlier detection technique for any univariant time-series observation. It depends on several factors, including the size and distribution of observations. However, when dealing with a small observation sample size, outlier detection using the modified Z-score method based on the median absolute deviation (MAD)

scale estimator is considered to be the fastest, simplest, and most effective method [29]. As a result, the focus of this paper will be on the application of the modified Z-score method to the detection of outliers in time-series RSS observations.

Several research works have used Z-score based on the MAD scale estimator to detect outliers [30–32]. However, this method is only effective for symmetrically distributed univariant time-series observations. Experimental analyses have shown that time-series RSS observations measured at a fixed reference location do not always follow a normal distribution [33–35]. The distribution of RSS observations over time can be affected by a variety of factors. These factors include the environment, transmit power, receiver sensitivity, and the distance between the transmitter and receiver, among others. The time-series RSS observations may have a Gaussian distribution if the environment is relatively stable and there are no significant obstacles or changes in the environment. Also, it may not have a non-Gaussian distribution if the environment is dynamic and there are significant changes in the environment. Such changes could include the presence of obstacles, the movement of people or vehicles, or changes in weather conditions. This demonstrates that the modified Z-score method based on the MAD scale estimator will not be efficient in detecting outliers in time-series RSS observations. A robust scale estimator to be used with the Z-score method as an alternative to the MAD is proposed and presented in the next section. This scale estimator is highly efficient for both normally and skewedly distributed time-series RSS observations.

3. Z-Score Method Using the Proposed Scale Estimator

This section, as previously stated, presents an alternative scale estimator for use with the Z-score method for detecting outliers in both symmetric and asymmetrically distributed observations. The conventional modified Z-score method based on the MAD scale estimator is presented first, followed by the Z-score method using the proposed scale estimator.

3.1. Detecting Outliers with the MAD Scale Estimator

The conventional Z-score method based on Equation (2) is highly influenced by outliers. An outlier can skew the results of an observation, causing the mean to no longer be representative of the observation. The approach to detecting outliers using the modified Z-score is the same as using the conventional Z-score method; however, the only difference is that the mean of the distribution is replaced by the median of the distribution, and the SD is replaced with the MAD.

Given the time-series RSS observations shown in Equation (1), the modified Z-score value of an $\text{rss}(n)$ is obtained using Equation (5).

$$z_{MAD}(\text{rss}(n)) = \frac{|\text{rss}(n) - \text{MED}_{\text{rss}_{i,j}}|}{\alpha \text{MAD}(\text{rss}(n))} \quad (5)$$

where $\alpha = 1.4285$ is called the correction factor and is used to make the MAD scale estimator consistent with the distribution of the time-series RSS observation, $\text{MED}_{\text{rss}_{i,j}}$ is the median on the RSS observation in Equation (1) and $\text{MAD}(\text{rss}(n))$ is the MAD scale estimator mathematically obtained using Equation (6) [28,36].

$$\text{MAD}(\text{rss}(n)) = \text{median}\left\{|\text{rss}(n) - \text{MED}_{\text{rss}_{i,j}}|\right\} \quad (6)$$

An RSS observation is considered an outlier if its Z-score value obtained using Equation (5) is above a predefined threshold value “ γ ”. The values of γ range from ± 0.5 to ± 5 . That is, for a given RSS observation, in Equation (1), it is considered an RSS outlier if:

$$|z_{MAD}(\text{rss}(n))| > |\gamma| \quad (7)$$

The MAD scale estimator based on Equation (6) has several drawbacks, including the following: (1) It has only 37% Gaussian efficiency; and (2) the approach to symmetric dispersion, which is based on giving equal weight to positive and negative deviations from the median value, contradicts the general theory of M-estimators [29]. Furthermore, MAD has a high gross error sensitivity, which means that it is more likely to mistake valid observations for outliers. Other well-known scale estimators, such as the least median square (LMS), S_n , and Q_n scale estimators, are also used to identify outliers in univariate time-series observations and share the same 50% breakdown point as the MAD scale estimator [29]. Although the LMS is less computationally complex, it has the same Gaussian efficiency as the MAD, which is lower than that of the S_n and Q_n scale estimators. The Q_n scale estimator is computationally complex but has the lowest gross error sensitivity. The S_n scale estimator is thought to have moderate computational complexity (lower than Q_n but higher than LMS and MAD) and moderate efficiency (higher than LMS and MAD but lower than Q_n) among the four scale estimators [29]. Thus, as an alternative to MAD, the S_n scale estimator is proposed to be used in determining the Z-score value of an RSS observation. Outlier detection using the Z-score based on the S_n scale estimator is presented in the next subsection.

3.2. Proposed S_n Scale Estimator for Outlier Detection

The proposed S_n scale estimator as an alternative to MAD scale estimators has several advantages, one of which is lower gross error sensitivity. This is because the S_n scale estimator considers the relationship that is the distance between the RSS observations, making it less sensitive to outliers [29]. Furthermore, the S_n scale estimator has a higher Gaussian efficiency of 58% and is very effective at detecting outliers in both symmetric and asymmetrically distributed observations, which has been extensively validated theoretically and empirically in earlier research works [29,31,37].

Mathematically, the S_n scale estimator is presented in Equation (8) [29].

$$S_n = c \times \text{median}_m \{ \text{median}_n \{ |rss(m) - rss(n)| \} \} \quad (8)$$

where $m \in [1, 2, \dots, N]$, $n \in [1, 2, \dots, N]$, $m \neq n$, $c = 1.1926$

The Z-score value based on the S_n scale estimator is presented in Equation (9).

$$Z_{S_n}(rss(n)) = \frac{|rss(n) - \text{MED}_{rss,i,j}|}{\alpha \times S_n} \quad (9)$$

For each $rss(n)$ in Equation (1), the Z-score value is calculated using Equation (9) with the scale estimator obtained using Equation (8) and it is an outlier if it satisfies Equation (10).

$$|Z_{S_n}(rss(n))| > |\gamma| \quad (10)$$

where “ γ ” is called the rejection criterion above which an RSS observation is considered an outlier.

The next step after detecting outliers is to decide what to do with them. The approach to dealing with outliers is presented in the section that follows.

4. Managing RSS Outliers and Selecting RSS Representatives

After outliers have been identified in an observation, the next step is to treat them. How an outlier is treated depends on several factors, one of which is the size of the observation. There are several techniques to treat outliers, and the most commonly used are as follows [37,38].

- a. Deletion: outliers can simply be removed from the observation. This approach is often used when the number of outliers is small, and the dataset is large enough that removing a few observations will not significantly impact the analysis.

- b. Winsorization: involves replacing the extreme values (outliers) with values that are closer to the mean or median of the observation. This method can preserve the observation size and reduce the influence of outliers on the results, but it can also alter the distribution of the observation and introduce bias.
- c. Trimming: involves removing a fixed percentage of the outliers from the observation. This method is similar to deletion, but it can reduce the impact of outliers on the results while preserving a larger portion of the data. However, trimming can also introduce bias and alter the distribution of the data.
- d. Transformation: involves changing the scale of the observation to reduce the influence of outliers. Common transformations include logarithmic or reciprocal transformations, which can reduce the impact of outliers by transforming the observation into a more symmetrical distribution.
- e. Median or mean imputation: Outliers can be replaced with the median or mean of the dataset. This approach is useful when the outliers are affecting the analysis and removal is not an option.

All the techniques mentioned above have their advantages and disadvantages; however, for a limited number of observations, the imputation method is considered to be simple and efficient [39]. In this paper, the median imputation technique will be used to treat outliers because the mean, as previously mentioned, is heavily influenced by outliers.

Let $MED_{RSS_{i,j}}$ be the median of the RSS observations in Equation (1), the process for the detection and treatment of outliers is as follows: Given an RSS observation, $rss(n)$, if:

$$||z_{score}(rss(n))|| > |\gamma| \text{ for } z_{score} \in [Z_{sn}(n), Z_{MAD}(n)] \quad \text{and } 1 \leq n \leq N \quad (11)$$

then

$$rss(n) = rss(n) \quad \text{for } 1 \leq n \leq N \quad (12)$$

else

$$rss(n) = MED_{RSS_{i,j}} \quad \text{for } 1 \leq n \leq N \quad (13)$$

The next step after detecting and treating the RSS outlier is the determination of the RSS observation representative. The median approach is considered to be the optimum approach to determine RSS representative of a time-series RSS observation. Let the RSS observation vector in Equation (14) be the cleaned RSS observations without the outliers.

$$\mathbf{rss}_{i,j}^{clean} = [rss_{clean}(1), rss_{clean}(2), \dots, rss_{clean}(N)] \quad (14)$$

where $rss_{clean}(n)$ is a non-outlier RSS observation.

Using the median approach, the median RSS that represents the observation in Equation (14) is obtained as follows:

$$rss_{median} = median\{\mathbf{rss}_{i,j}^{clean}\} \quad (15)$$

The performance of the RSS outlier detection techniques presented in Section 2 is compared in the following section.

5. Simulation Results and Discussion

The performance of the outlier detection methods presented in Section 3 is determined and compared using an online RSS dataset available in [40]. The dataset is made up of LTE RSS power measurements taken with an LTE-enabled mobile device at a fixed reference location. Measurements were taken between the hours of 7:30 a.m. and 11:00 p.m. at 30 min intervals for 30 days, resulting in approximately 32 RSS observations per day and 960 RSS observations for the entire months. Only eight RSS datasets from eight different days have 100% RSS observations, i.e., no missing measurements, and these eight RSS datasets will be used for the analysis. Figure 1 shows the time-series RSS distribution for each dataset.

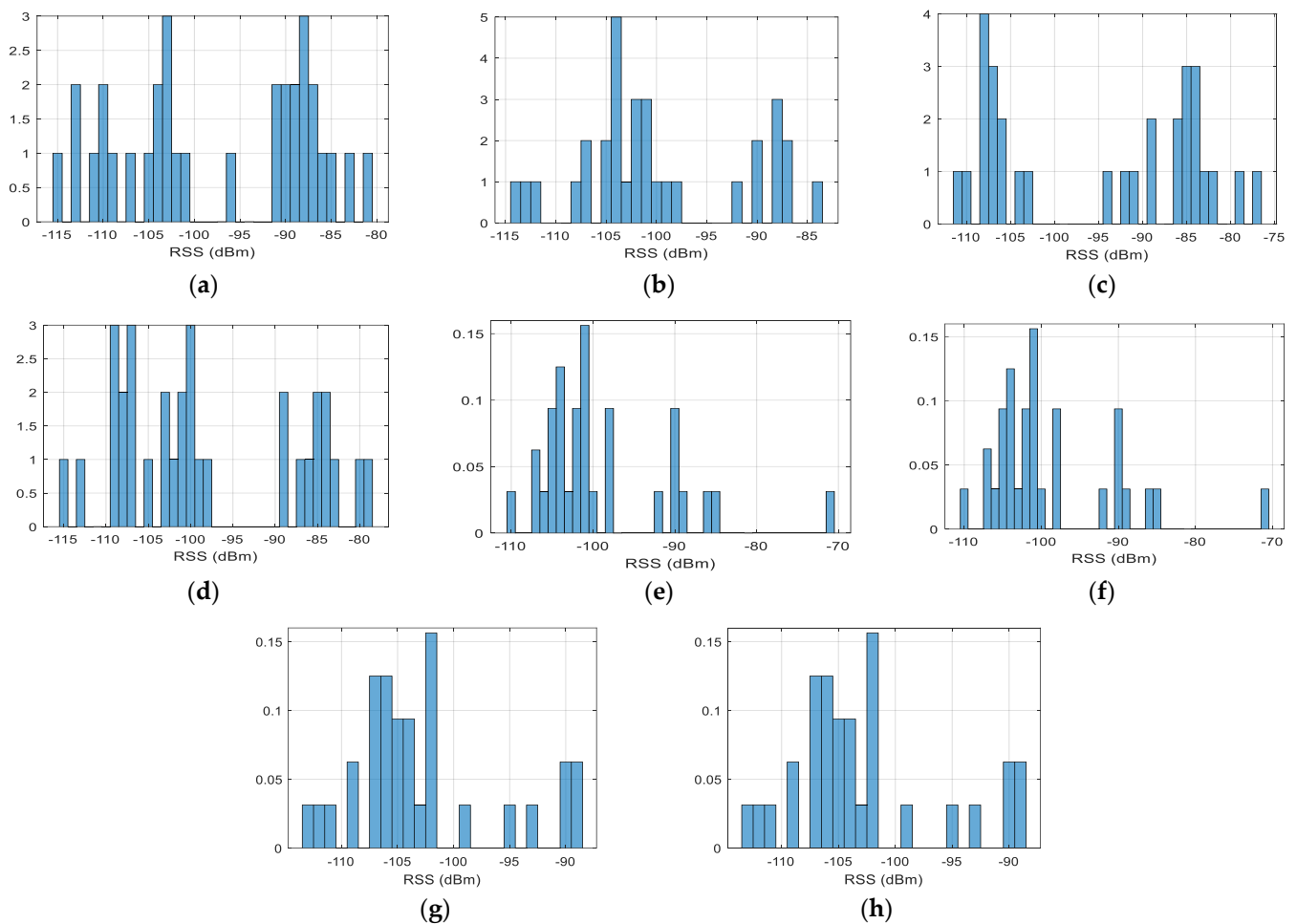


Figure 1. RSS Dataset distribution: (a) RSS observation 1, (b) RSS observation 2, (c) RSS observation 3, (d) RSS observation 4, (e) RSS observation 5, (f) RSS observation 6, (g) RSS observation 7, and (h) RSS observation 8.

Looking at the distribution of each RSS observation in Figure 1, we can see that the time-series RSS observations lack a Gaussian distribution and are thus asymmetrical. This means that detecting outliers in these datasets using the Z-score method with MAD as a scale estimator will be inefficient.

The γ , as mentioned earlier, is the Z-score threshold value above which an RSS observation is considered to be an outlier. The total number of RSS observations in each dataset considered an outlier is determined and presented in Figure 2 using each of the techniques presented in Section 3 and by varying γ from 0.5 to 5 at a step of 0.5.

Figure 2 shows that for all eight RSS datasets, the Z-score method based on the S_n scale estimator detects fewer outliers than the Z-score method based on the MAD scale estimator. Detecting more or fewer outliers is insufficient to determine an outlier detection method's performance. A comparison of the gross error sensitivity of the S_n and MAD scale estimators, on the other hand, reveals that the MAD scale estimator has a higher gross error sensitivity, making it more likely to detect an outlier falsely. This could be the reason why the Z-score method based on the MAD scale estimator detected a higher number of outliers.

Further analysis of the outlier detection performance is performed with $\gamma = 1.28$, which is considered the optimum Z-score threshold value to detect an outlier as presented in [28]. Figures 3–10 compare the outlier detection performance of the two scale estimators at $\gamma = 1.28$.

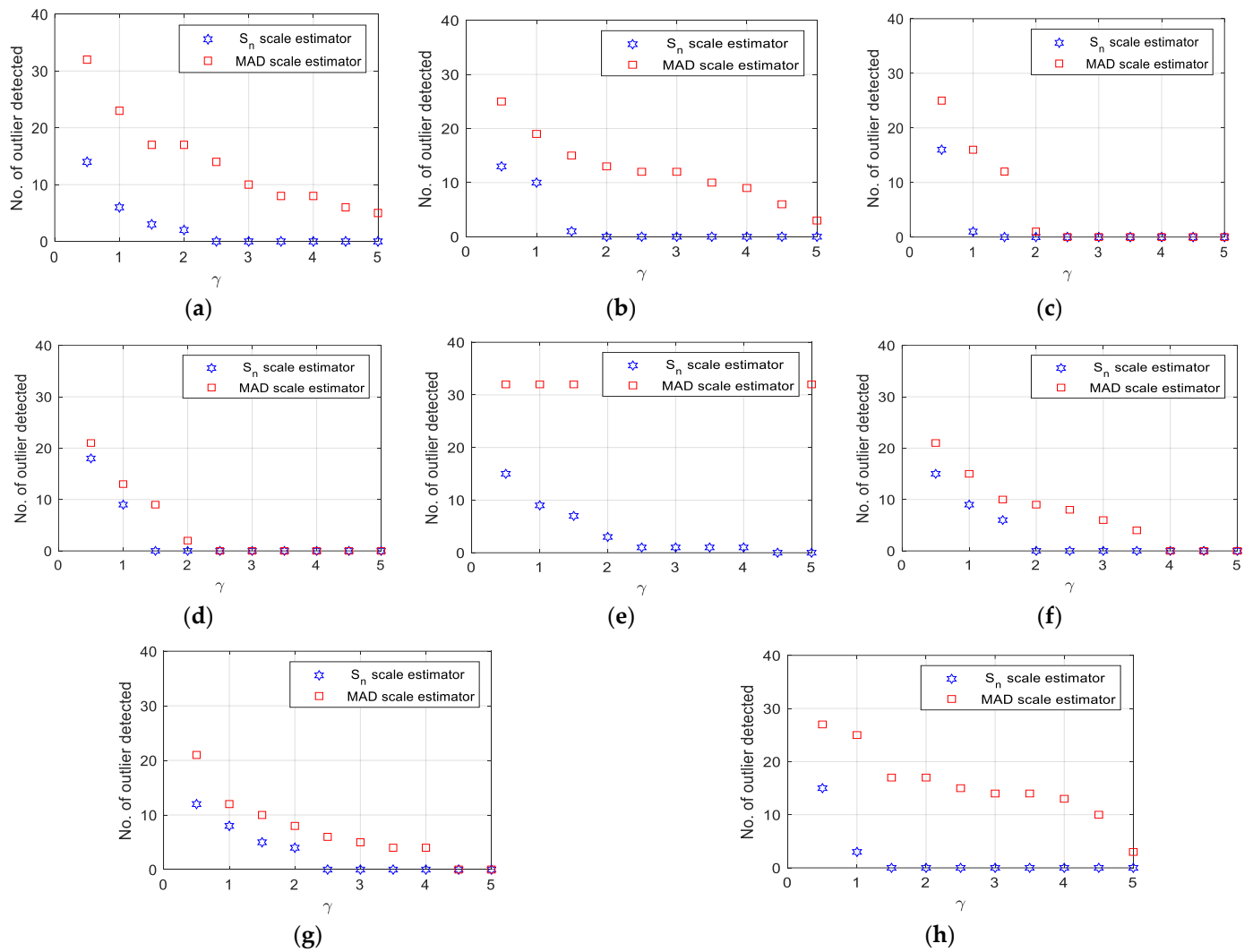


Figure 2. Number of detected RSS outliers for different γ values: (a) RSS observation 1, (b) RSS observation 2, (c) RSS observation 3, (d) RSS observation 4, (e) RSS observation 5, (f) RSS observation 6, (g) RSS observation 7, and (h) RSS observation 8.

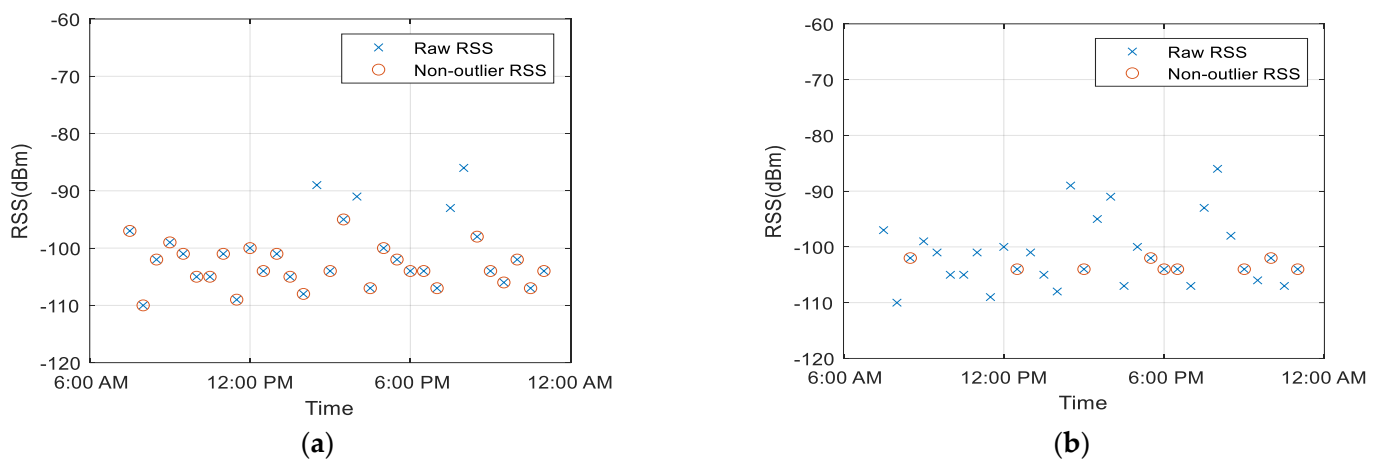


Figure 3. RSS outlier detection comparison for RSS Dataset 1. (a) S_n scale estimator, and (b) MAD scale estimator.

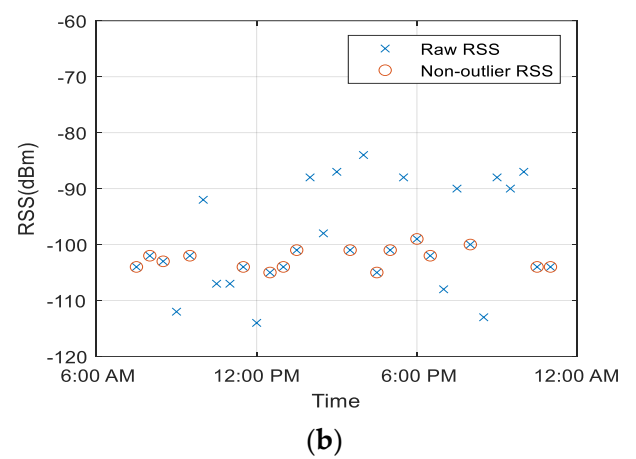
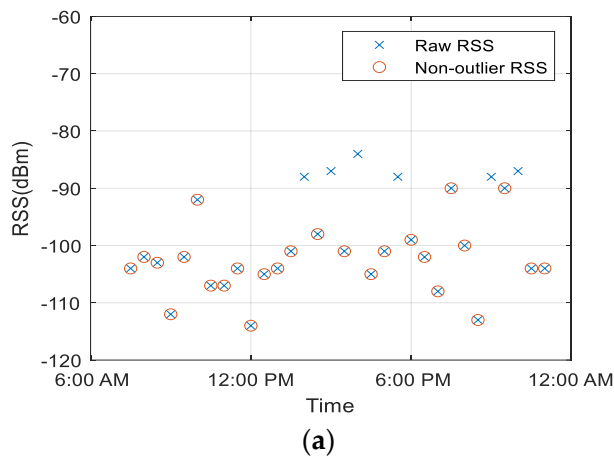


Figure 4. RSS outlier detection comparison for RSS Dataset 2. (a) S_n scale estimator, and (b) MAD scale estimator.

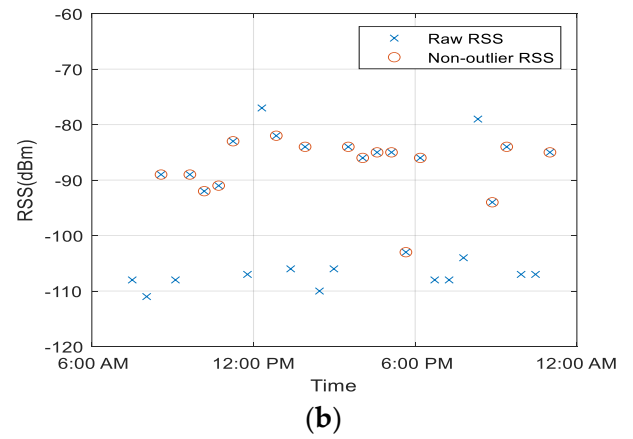
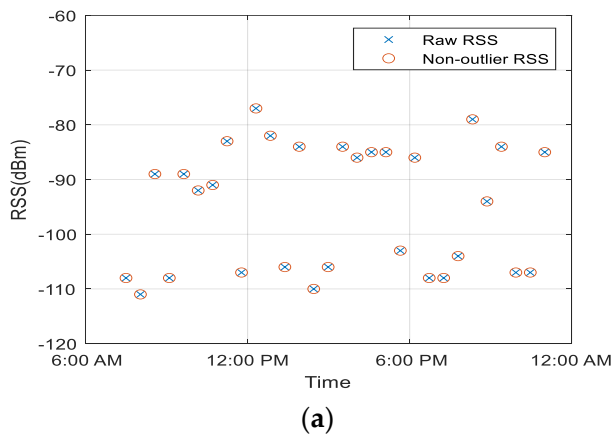


Figure 5. RSS outlier detection comparison for RSS Dataset 3. (a) S_n scale estimator, and (b) MAD scale estimator.

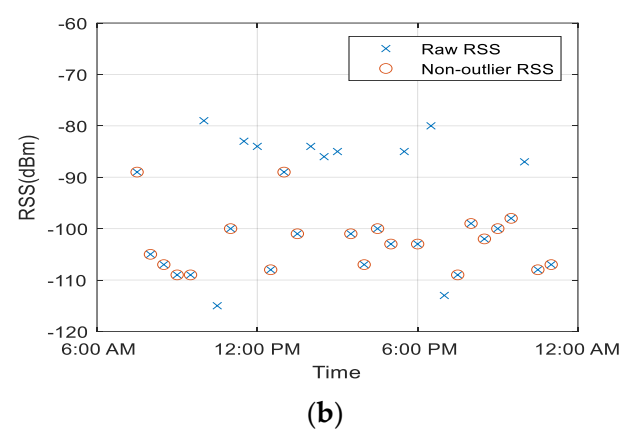
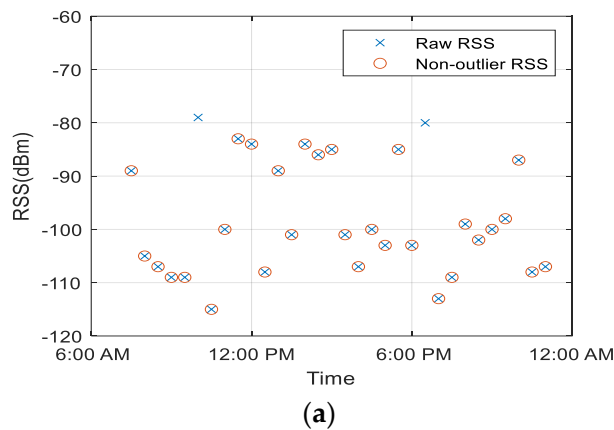
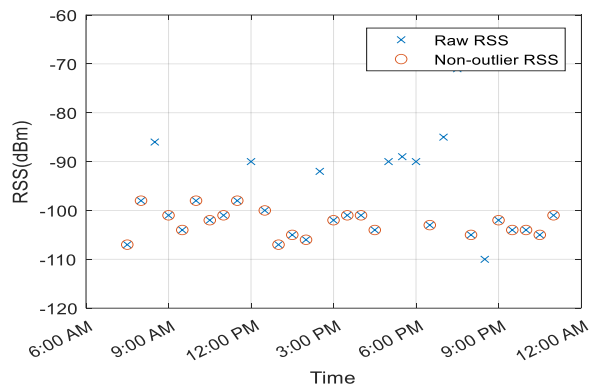
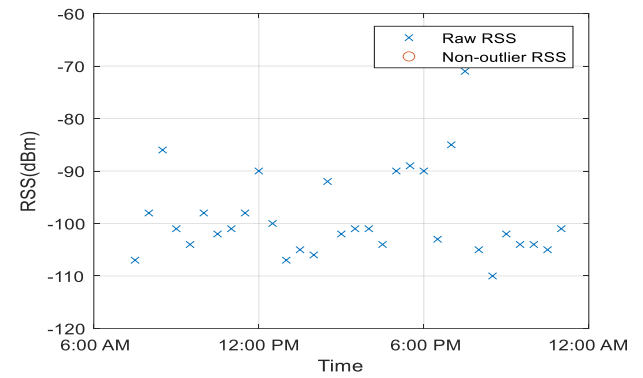


Figure 6. RSS outlier detection comparison for RSS Dataset 4. (a) S_n scale estimator, and (b) MAD scale estimator.

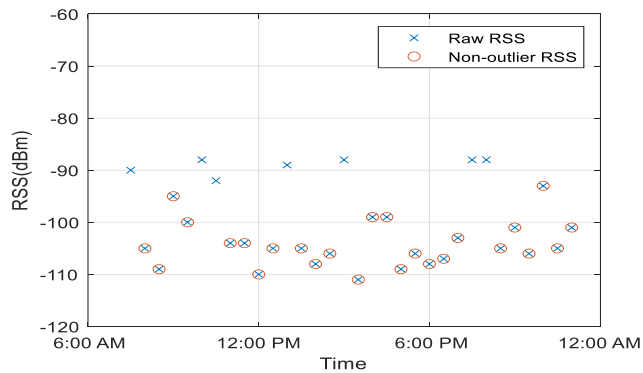


(a)

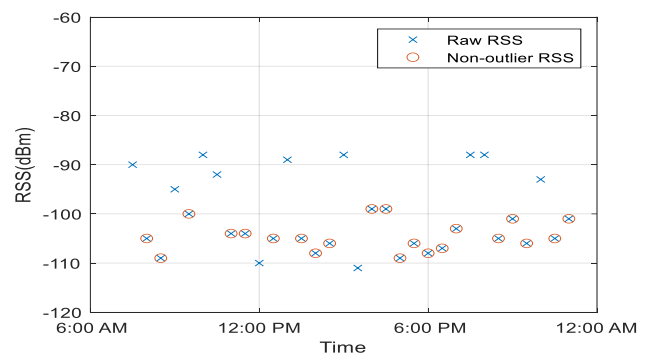


(b)

Figure 7. RSS outlier detection comparison for RSS Dataset 5. (a) S_n scale estimator, and (b) MAD scale estimator.

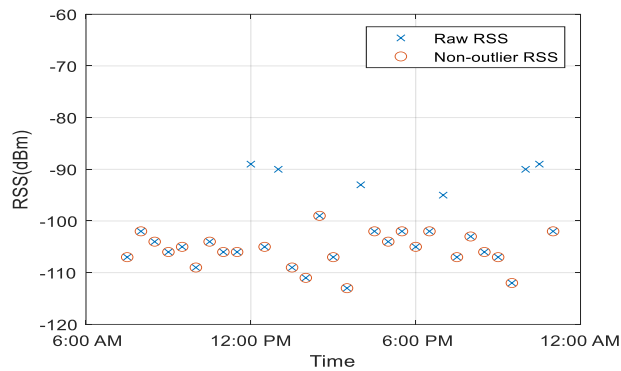


(a)

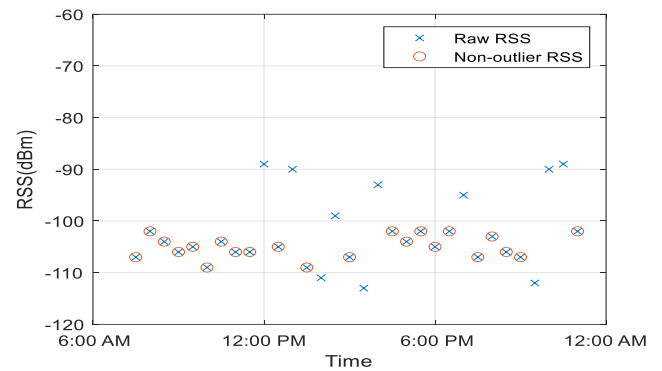


(b)

Figure 8. RSS outlier detection comparison for RSS Dataset 6. (a) S_n scale estimator, and (b) MAD scale estimator.



(a)



(b)

Figure 9. RSS outlier detection comparison for RSS Dataset 7. (a) S_n scale estimator, and (b) MAD scale estimator.

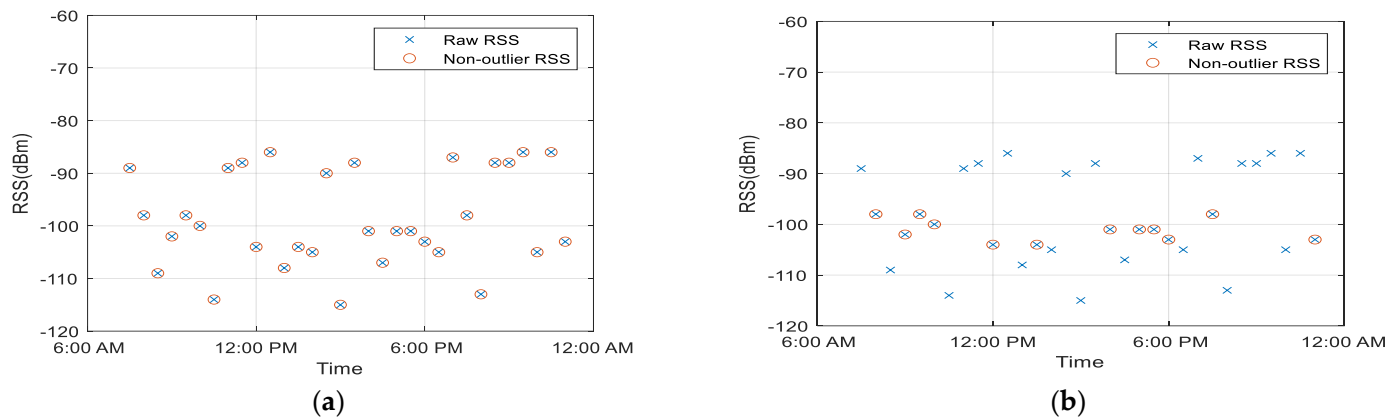


Figure 10. RSS outlier detection comparison for RSS Dataset 8. (a) S_n scale estimator, and (b) MAD scale estimator.

Looking at the distribution of RSS observations for each dataset, it is possible that very few, if any, of the RSS observations could be considered outliers. Both methods identified some RSS observations as outliers. As shown in Table 2, the Z-score method using the MAD as a scale estimator identified at least 50% of the observations in each dataset as outliers. On the other hand, the Z-score method with the S_n scale estimator identified, on average, 13% of the observations in each dataset as outliers. This implies that the Z-score method based on the MAD scale estimator considered several valid RSS observations as outliers, demonstrating its inefficiency when applied to time-series RSS observations. A number of factors could contribute to such inefficiency. The first one includes the previously mentioned high gross error sensitivity. The second is its symmetrical approach to detecting outliers from observation medians, which is inefficient when applied to an asymmetrical observation, such as time-series RSS observations.

Table 2. Comparison of the number of detected outliers.

RSS Dataset	S_n Scale Estimator		MAD Scale Estimator	
	No. of Outlier Detected	% of Total Observation	No. of Outlier Detected	% of Total Observation
1	4	13	23	72
2	6	19	16	50
3	0	0	14	44
4	2	6	11	34
5	8	25	32	100
6	7	22	11	34
7	6	19	10	31
8	0	0	20	63

There were no clearly visible outliers in the datasets used to evaluate the performance of the outlier detection methods earlier. To further demonstrate the S_n scale estimator's outlier detection capability, artificial outliers are deliberately inserted into two of the datasets, namely, datasets 2 and 4. This is achieved by replacing some of the valid RSS observations with artificial outliers. Figures 11 and 12, respectively, show the RSS datasets 2 and 4 with and without the artificial outliers. The artificial outliers are marked with the colour red. In Figure 11, all the artificial outliers are inserted in such a way that they are visually alien to the observations. In Figure 12, one of the artificial outliers is placed very close to an RSS observation. If the outlier detection method considers the RSS observation to be a non-outlier observation, it is expected that the artificial outlier will also be considered a non-outlier observation due to its close proximity to the RSS observation.

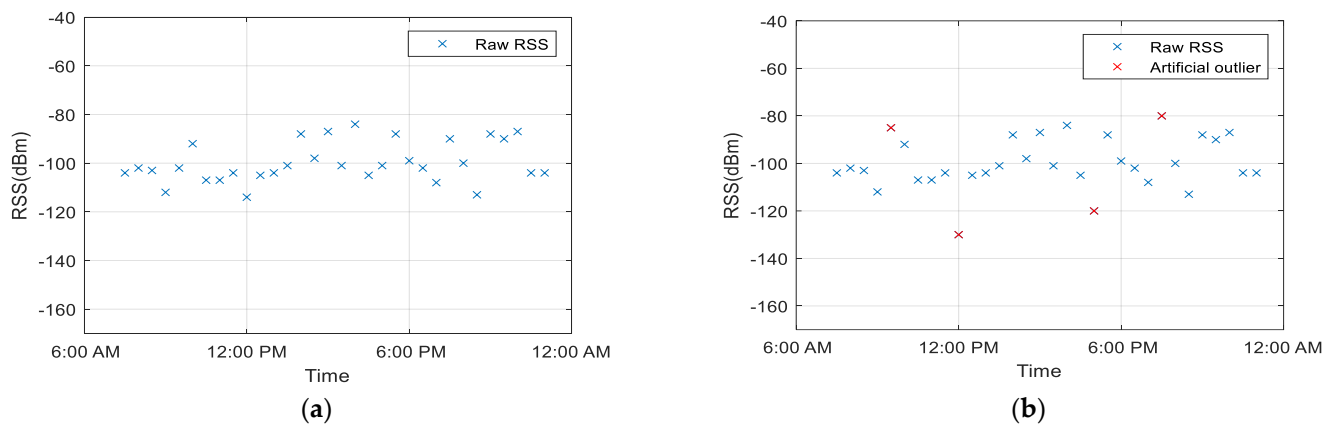


Figure 11. RSS Dataset 2 with and without the artificial outliers. (a) Without artificial outlier (b) With artificial outliers.

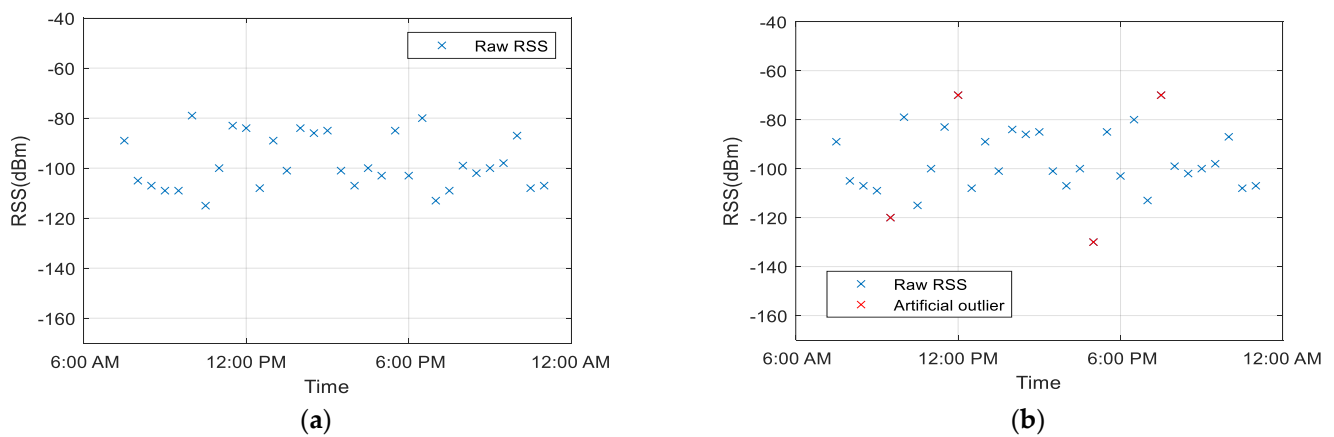


Figure 12. RSS Dataset 4 with and without the artificial outliers. (a) Without artificial outlier (b) With artificial outliers.

Figures 13 and 14 show, respectively, the outlier detection performance comparisons of the Z-score method with MAD and S_n as scale estimators for the RSS datasets taken into consideration with and without the artificial outliers. From Figure 13, both the Z-score methods using the S_n and MAD scale estimators did not consider the artificial outliers as valid RSS observations, which is what is expected. However, in Figure 14, the Z-score method with the S_n scale estimator considered one of the artificial outliers that is close to the valid RSS observation as a non-outlier observation. This is expected, as the distance between the artificial outlier and the valid RSS observation is not large enough to consider it an outlier. It can be recalled that the S_n scale estimator uses the distance between pairs of observations to classify them as outliers. This is different from the MAD scale estimator, which uses the distance between an observation and the median of the observations to detect outliers. Hence, it considered the artificial outlier, which was close to the valid RSS observation, an outlier as it was very far from the median of the observation.

Even though outliers have been removed from the observations, it is possible that the observations are skewed. As a result, it is recommended that the median approach be used to determine the RSS observation representative. The median RSSs obtained using Equation (15) for the outlier-free observations generated using the MAD and S_n scale estimators are presented in Table 3 and compared with the median RSS of the raw RSS observations of each dataset.

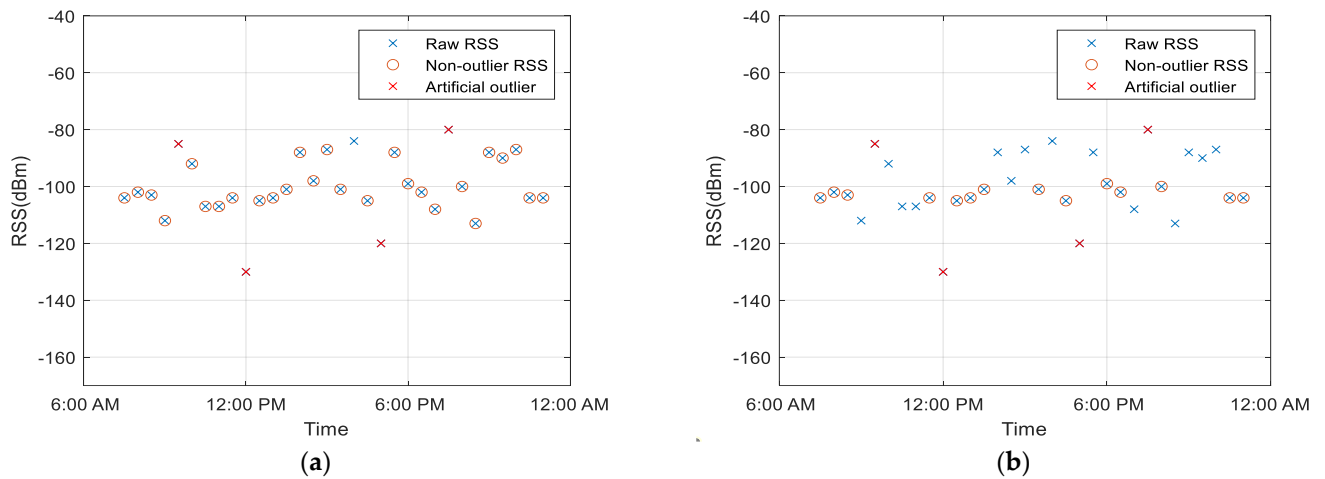


Figure 13. Outlier detection comparison of RSS Dataset 2 with and without artificial outliers for: (a) S_n scale estimator (b) MAD scale estimator.

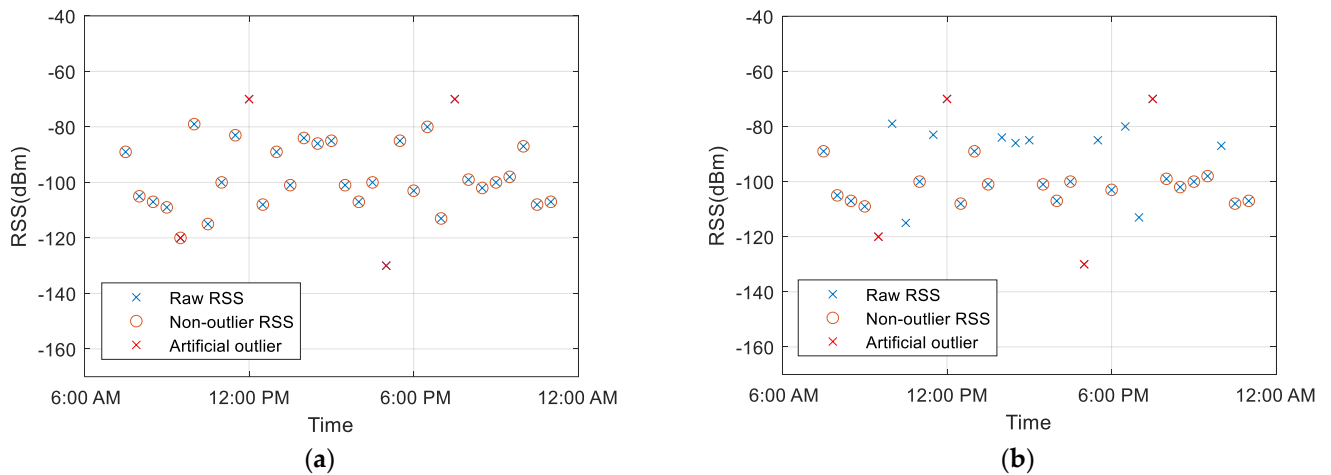


Figure 14. Outlier detection comparison of RSS Dataset 4 with and without artificial outliers for: (a) S_n scale estimator (b) MAD scale estimator.

Table 3. RSS observation representative determination.

RSS Dataset	Median RSS (dBm)		
	Raw Data	S_n	MAD
1	−103	−104	−98
2	−102	−102	−99
3	−92	−92	−95
4	−101	−101	−99
5	−101	−101	nil
6	−104	−104	−102
7	−105	−105	−103
8	−101	−101	−99

All the median RSSs of the eight raw observations and that obtained for the S_n scale estimator's observations are approximately the same, with an average absolute deviation of about 0.25 dB. For the MAD scale estimator, the median RSS absolute deviation from the raw observation is around 3 dB. A power error of 0.25 dB and 3 dB will translate to an error in distance of approximately 0.5 m and 2 m, respectively. This is under the assumption

of a free-space propagation model with no environmental effect and using a Wi-Fi AP operating at a frequency of 2.4 GHz. A distance error of 2 m translates to a very large localization error. As such, when dealing with RSS time-series observations that are known to be asymmetric in nature, it is recommended to use the S_n scale estimator for outlier detection and the median approach to determine the aggregated RSS representative.

One of the limitations of using the S_n scale estimator when compared to MAD is the computational time. Figure 15 shows a computational time comparison between the MAD and S_n scale estimators for varying observation sizes.

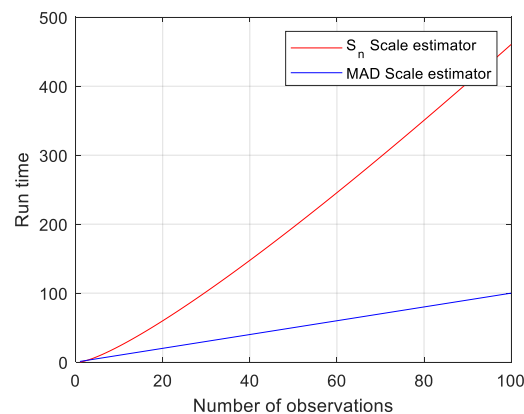


Figure 15. Computational time comparison.

The run time of the MAD is $O(n)$, while that of the S_n scale estimator is $O(n \log n)$, making it the slowest. The S_n scale estimator becomes exponentially slower as the number of RSS observations increases. For example, for an observation size of 32, the S_n scale estimator is approximately 3.5 times slower than the MAD. However, based on our previous analysis, the MAD has approximately 50% average false detection of an outlier.

In summary, both estimators can be used to detect outliers in time-series RSS observations. However, if the observation is known to be skewed, that is, asymmetrical, or the presence of outliers is expected, then the S_n scale estimator may be a more appropriate choice. On the other hand, if computational efficiency is a concern, then MAD may be a better choice. However, if computational efficiency is not a concern and the observation is known to be skewed, a hybrid of the two techniques can be used, and the S_n scale estimator should be applied first, followed by the MAD scale estimator.

6. Conclusions

In this paper, a S_n scale estimator is proposed as an alternative to the MAD scale estimator in detecting outliers in RSS time-series observations. The scale estimator is more efficient, has a lower gross error sensitivity, and performs well with both symmetrically and asymmetrically distributed observations. This is in contrast to the MAD scale estimator, which has a high gross error sensitivity and works best on only symmetrical data. The performance of the proposed scale estimator is determined and compared with the MAD scale estimator using an online RSS dataset. Results show that the S_n scale estimator performs better at detecting RSS outliers compared with the MAD scale estimator. This is as a result of the S_n scale estimator's better performance when dealing with asymmetrically distributed observations, which the RSS time-series observations are known to be. Future research will combine the S_n scale estimator with smoothing techniques such as the Kalman filter and moving average filter to further enhance the reliability of the RSS measurements.

Author Contributions: Conceptualization, A.S.Y. and F.M.; writing—original draft preparation, A.S.Y.; writing—review and editing, A.S.Y. and F.M.; supervision, F.M.; project administration, P.P.; funding acquisition, F.M. and P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the SPEV project 2023 at the Faculty of Informatics and Management, University of Hradec Kralove, the Czech Republic. The technical support of Ing. Kruncik is kindly acknowledged.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yaro, A.S.; Sha'ameri, A.Z. Development of an Association Technique for a 3-Dimensional Minimum Configuration Multilateration System. *Int. J. Integr. Eng.* **2020**, *12*, 59–71.
2. Kriz, P.; Maly, F.; Kozel, T. Improving Indoor Localization Using Bluetooth Low Energy Beacons. *Mob. Inf. Syst.* **2016**, *2016*, 2083094. [[CrossRef](#)]
3. Yaro, A.S.; Maly, F.; Prazak, P. A Survey of the Performance-Limiting Factors of a 2-Dimensional RSS Fingerprinting-Based Indoor Wireless Localization System. *Sensors* **2023**, *23*, 2545. [[CrossRef](#)] [[PubMed](#)]
4. Asaad, S.M.; Maghddid, H.S. A Comprehensive Review of Indoor/Outdoor Localization Solutions in IoT era: Research Challenges and Future Perspectives. *Comput. Netw.* **2022**, *212*, 109041. [[CrossRef](#)]
5. Maly, F.; Kriz, P.; Adamec, M. Pervasive Game Utilizing WiFi Fingerprinting-Based Localization. In Proceedings of the Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection, Nicosia, Cyprus, 31 October–5 November 2016; pp. 836–846.
6. Flueraoru, L.; Shubina, V.; Niculescu, D.; Lohan, E.S. On the High Fluctuations of Received Signal Strength Measurements with BLE Signals for Contact Tracing and Proximity Detection. *IEEE Sens. J.* **2022**, *22*, 5086–5100. [[CrossRef](#)]
7. Zhou, R.; Yang, Y.; Chen, P. An Rss Transform—Based Wknn for Indoor Positioning. *Sensors* **2021**, *21*, 5685. [[CrossRef](#)]
8. Roy, P.; Chowdhury, C. A Survey on Ubiquitous WiFi-Based Indoor Localization System for Smartphone Users from Implementation Perspectives. *CCF Trans. Pervasive Comput. Interact.* **2022**, *4*, 298–318. [[CrossRef](#)]
9. Chen, Y.C.; Sun, W.C.; Juang, J.C. Outlier Detection Technique for RSS-Based Localization Problems in Wireless Sensor Networks. In Proceedings of the SICE Annual Conference 2010, Taipei, Taiwan, 18–21 August 2010; pp. 657–662.
10. Ye, Q.; Fan, X.; Fang, G.; Bie, H. Exploiting Temporal Dependency of RSS Data with Deep for IoT-Oriented Wireless Indoor Localization. *Internet Technol. Lett.* **2022**, e366. [[CrossRef](#)]
11. Cheng, W.; Tan, K.; Omwando, V.; Zhu, J.; Mohapatra, P. RSS-Ratio for Enhancing Performance of RSS-Based Applications. In Proceedings of the 2013 Proceedings IEEE INFOCOM, Turin, Italy, 14–19 April 2013; pp. 3075–3083.
12. Fang, S.-H.; Lin, T.-N. Accurate WLAN Indoor Localization Based on RSS, Fluctuations Modeling. In Proceedings of the 2009 IEEE International Symposium on Intelligent Signal Processing, Budapest, Hungary, 26–28 August 2009; pp. 27–30.
13. Zhu, H.; Tsang, K.-F.; Liu, Y.; Wei, Y.; Wang, H.; Wu, C.K.; Chi, H.R. Extreme RSS Based Indoor Localization for LoRaWAN with Boundary Autocorrelation. *IEEE Trans. Ind. Inform.* **2020**, *17*, 4458–4468. [[CrossRef](#)]
14. Rozum, S.; Sebesta, J. SIMO RSS Measurement in Bluetooth Low Power Indoor Positioning System. In Proceedings of the 2018 28th International Conference Radioelektronika (RADIOELEKTRONIKA), Prague, Czech Republic, 19–20 April 2018; pp. 1–5.
15. Xin-Di, L.; He, W.; Tian, Z.S. The Improvement of Rss-Based Location Fingerprint Technology for Cellular Networks. In Proceedings of the 2012 International Conference on Computer Science and Service System, Nanjing, China, 11–13 August 2012; pp. 1267–1270.
16. Yu, F.; Jiang, M.; Liang, J.; Qin, X.; Hu, M.; Peng, T.; Hu, X. Expansion RSS-Based Indoor Localization Using 5G WiFi Signal. In Proceedings of the 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, India, 14–16 November 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 510–514.
17. Ji, W.; Zhao, K.; Zheng, Z.; Yu, C.; Huang, S. Multivariable Fingerprints With Random Forest Variable Selection for Indoor Positioning System. *IEEE Sens. J.* **2022**, *22*, 5398–5406. [[CrossRef](#)]
18. Fronckova, K.; Prazak, P. Possibilities of Using Kalman Filters in Indoor Localization. *Mathematics* **2020**, *8*, 1564. [[CrossRef](#)]
19. Zhou, R.; Meng, F.; Zhou, J.; Teng, J. A Wi-Fi Indoor Positioning Method Based on an Integration of EMDT and WKNN. *Sensors* **2022**, *22*, 5411. [[CrossRef](#)]
20. Koubaa, A.; ben Jamaa, M.; AlHaqbani, A. An Empirical Analysis of the Impact of RSS to Distance Mapping on Localization in WSNs. In Proceedings of the Third International Conference on Communications and Networking, Hammamet, Tunisia, 29 March–1 April 2012; IEEE: Piscataway, NJ, USA, 2014; pp. 1–7.
21. Ezhumalai, B.; Song, M.; Park, K. An Efficient Indoor Positioning Method Based on Wi-fi Rss Fingerprint and Classification Algorithm. *Sensors* **2021**, *21*, 3418. [[CrossRef](#)] [[PubMed](#)]
22. Huang, B.; Liu, J.; Sun, W.; Yang, F. A Robust Indoor Positioning Method Based on Bluetooth Low Energy with Separate Channel Information. *Sensors* **2019**, *19*, 3487. [[CrossRef](#)]

23. Polak, L.; Rozum, S.; Slanina, M.; Bravenec, T.; Fryza, T.; Pikrakis, A. Received Signal Strength Fingerprinting-Based Indoor Location Estimation Employing Machine Learning. *Sensors* **2021**, *21*, 4605. [\[CrossRef\]](#)
24. Ibrahim, M.; Torki, M.; ElNainay, M. CNN Based Indoor Localization Using RSS Time-Series. In Proceedings of the 2018 IEEE Symposium on Computers and Communications (ISCC), Natal, Brazil, 25–28 June 2018; pp. 01044–01049.
25. Nabati, M.; Ghorashi, S.A. A Real-Time Fingerprint-Based Indoor Positioning Using Deep Learning and Preceding States. *Expert Syst. Appl.* **2023**, *213*, 118889. [\[CrossRef\]](#)
26. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection. *ACM Comput. Surv.* **2009**, *41*, 1–58. [\[CrossRef\]](#)
27. Ian, H.W.; Eibe, F.; Mark, A.H. *Data Mining: Practical Machine Learning Tools and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011; ISBN 9780123748560.
28. Wilcox, R.R. 3—SUMMARIZING DATA. In *Applying Contemporary Statistical Techniques*; Wilcox, R.R., Ed.; Academic Press: Burlington, NJ, USA, 2003; pp. 55–91, ISBN 978-0-12-751541-0.
29. Rousseeuw, P.J.; Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [\[CrossRef\]](#)
30. Bae, I.; Ji, U. Outlier Detection and Smoothing Process for Water Level Data Measured by Ultrasonic Sensor in Stream Flows. *Water* **2019**, *11*, 951. [\[CrossRef\]](#)
31. Rousseeuw, P.J.; Hubert, M. Robust Statistics for Outlier Detection. *WIREs Data Min. Knowl. Discov.* **2011**, *1*, 73–79. [\[CrossRef\]](#)
32. Kulanuwat, L.; Chantrapornchai, C.; Maleewong, M.; Wongchaisuwat, P.; Wimala, S.; Sarinnapakorn, K.; Boonya-aroonnet, S. Anomaly Detection Using a Sliding Window Technique and Data Imputation with Machine Learning for Hydrological Time Series. *Water* **2021**, *13*, 1862. [\[CrossRef\]](#)
33. Lin, M.; Chen, B.; Zhang, W.; Yang, J. Characteristic Analysis of Wireless Local Area Network's Received Signal Strength Indication in Indoor Positioning. *IET Commun.* **2020**, *14*, 497–504. [\[CrossRef\]](#)
34. Belmonte-Fernández, Ó. Modeling the Received Signal Strength Intensity of Wi-Fi Signal Using Hidden Markov Models. *Expert Syst. Appl.* **2021**, *174*, 114726. [\[CrossRef\]](#)
35. Kaemarungsi, K.; Krishnamurthy, P. Properties of Indoor Received Signal Strength for WLAN Location Fingerprinting. In Proceedings of the First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, MOBIQUITOUS 2004, Boston, MA, USA, 26 August 2004; pp. 14–23.
36. Pearson, R.K.; Neuvo, Y.; Astola, J.; Gabbouj, M. Generalized Hampel Filters. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 87. [\[CrossRef\]](#)
37. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods (with R)*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2019; ISBN 9780470010921.
38. Krämer, W. Andrew Gelman and Jennifer Hill: Data Analysis Using Regression and Multilevel/Hierarchical Models. *Stat. Pap.* **2011**, *52*, 741–742. [\[CrossRef\]](#)
39. Seliem, M.M. Handling Outlier Data as Missing Values by Imputation Methods: Application of Machine Learning Algorithms. *Turk. J. Comput. Math. Educ. TURCOMAT* **2022**, *13*, 273–286.
40. Karanja, H.S.; Atayero, A. Cellular Received Signal Strength Indicator Dataset. *IEEE Dataport*. 2020. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.