IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 8, 2017, accepted October 16, 2017, date of publication October 30, 2017,
date of current version November 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2766076

Predictability Analysis of Spectrum State Evolution: Performance Bounds and Real-World Data Analytics

JIACHEN SUN[1], LIANG SHEN[1], GUORU DING[1,2], (Senior Member, IEEE), RONGPENG LI[3], AND QIHUI WU[4], (Senior Member, IEEE)

[1]College of Communications Engineering, Army Engineering University, Nanjing 210007, China
[2]National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
[3]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[4]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Corresponding authors: Liang Shen (shenliang671104@sina.com); Guoru Ding (dr.guoru.ding@ieee.org)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501510 and Grant 61631020, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20150717, in part by the China Postdoctoral Science Foundation Project under Grant 2016M590398, and in part by the Jiangsu Planned Projects for Postdoctoral Research Funds under Grant 1501009A.

ABSTRACT Predictability in spectrum prediction refers to the degree to which a correct prediction of the radio spectrum state (RSS) can be made quantitatively. It is obvious that the possibility that the future RSS is accurately predicted will be different when using different spectrum prediction algorithms. However, the fundamental limits on the accuracy of various spectrum prediction algorithms should exist and be worthwhile to be paid attention to. In this paper, we define these fundamental limits as the performance bounds of predictability, which can be the important indexes when evaluating the performance of different spectrum prediction algorithms. Real-world spectrum data is involved to present comprehensive and profound analysis of the predictability. We first transform large amount of spectrum data into symbol sequences by sampling and quantization, to calculate the entropy of the symbol sequence, which represents the randomness of the RSS evolution. Then, we derive the upper bound and the lower bound of the predictability mainly from entropies of the symbol sequences. Further, we conduct the detailed analysis on the performance bounds of the predictability of the RSS. Based on real-world data analytics, the key insights among others include: 1) entropies almost have no relationship with selection of sampling intervals in the data preprocessing; 2) the upper and the lower bounds of the predictability will both decrease as the quantization level rises and tend to be stable around a value at last; and 3) two kinds of lower bounds of the predictability are proposed, and one of the lower bounds, the regularity R, can reveal the tidal effect of the evolution of the RSS.

INDEX TERMS Predictability, spectrum state, entropy rate, real-world spectrum data, data analytics

I. INTRODUCTION

A range of applications in cognitive radio networks, from adaptive spectrum sensing to predictive spectrum mobility and dynamic spectrum access, depend on our ability to foresee the state evolution of radio spectrum [1]–[4]. A number of spectrum prediction techniques have been proposed, such as time series-based prediction, autoregressive model-based prediction, hidden Markov model-based prediction, neural networks-based prediction, and Bayesian inference-based prediction, etc. (see e.g. the surveys in [2] and [5], and the references therein). Just as Shannon capacity gives the upper bound of various modulation and coding schemes [6], there should be fundamental performance bounds, in terms of predictability, of various spectrum prediction algorithms.

Predictability is the degree to which a correct prediction or forecast of a system's state can be made either qualitatively or quantitatively [8]. When it comes to spectrum prediction of interest in this paper, predictability refers to the degree to which a correct prediction of the radio spectrum state (RSS) can be made quantitatively. Seeing that in Shannon's theorem [6], the channel capacity is defined as the upper bound on the rate at which information can be reliably transmitted over a communication channel when the length of the source symbol string goes to infinity, that means there

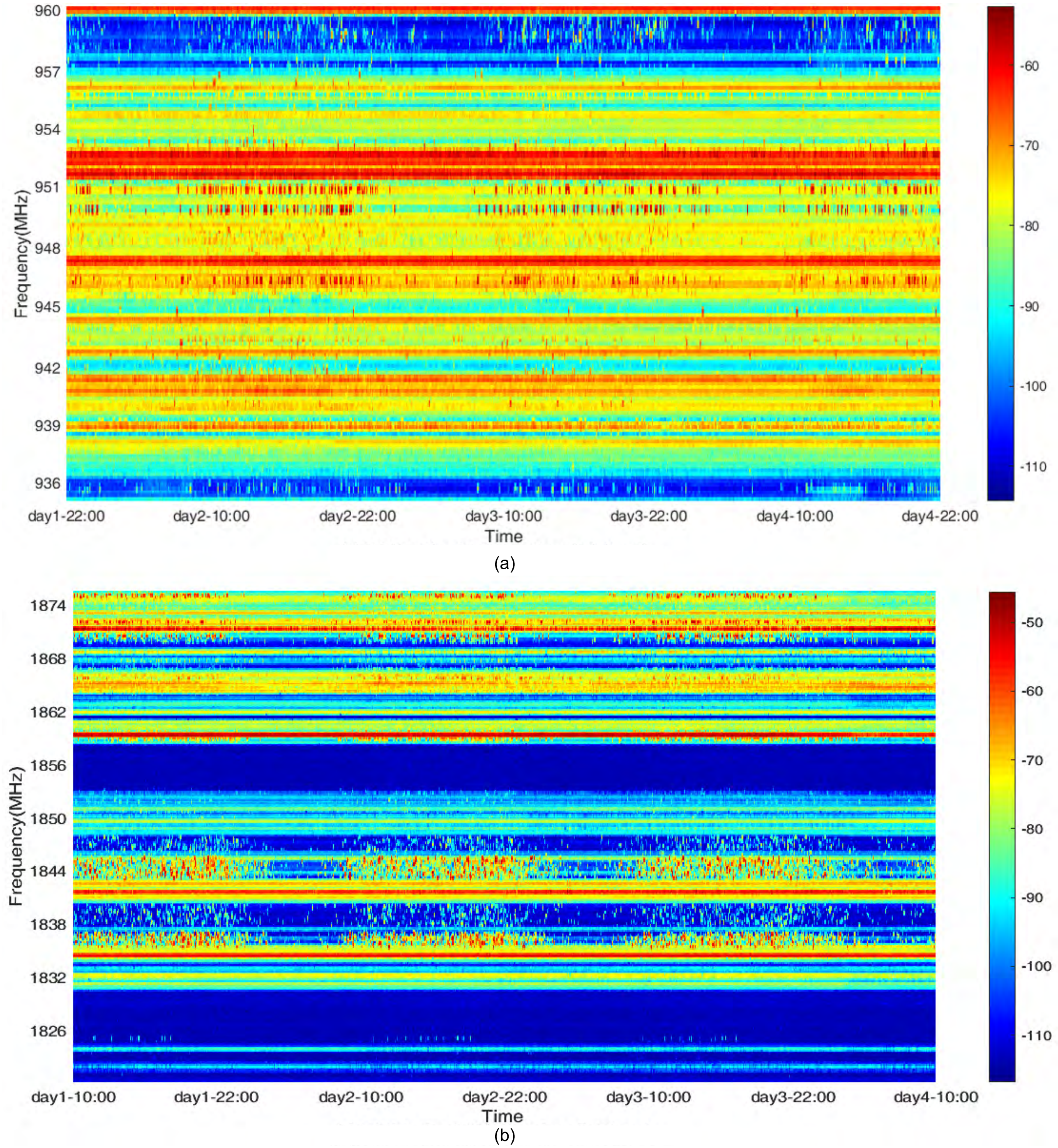


FIGURE 1. The evolution trajectories of RSS in the GSM900 downlink bands (935MHz ~ 960MHz) and the GSM1800 downlink bands (1820MHz ~ 1875MHz) during the 3-day measurement. (a) RSS in the GSM900 downlink bands. (b) RSS in the GSM1800 downlink bands.

must be an encoding method to make the rate of the error-free information transmission over a communication channel be infinitely close to the channel capacity. Similarly, in this paper, the upper bound of predictability in spectrum prediction can be defined as the upper bound on the possibility that the future RSS predicted by a certain spectrum prediction algorithm agrees with the actual RSS at the next moment when the length of the history RSS sequence goes to infinity, that means there must be an appropriate spectrum prediction algorithm to make the possibility be infinitely close to the upper bound of predictability. Meanwhile, there should also exist a lower bound of predictability. This means the possibility that one simplest spectrum prediction algorithm, whose prediction is only based on the frequencies of the RSSs in the history sequence, can predict the future RSS accurately. We can assess the reliability of the spectrum prediction algorithm by making comparison between the prediction accuracy and the bounds of predictability on the same spectrum dataset. So, predictability in RSS dynamics can be an important index when evaluating the performance of different spectrum prediction algorithms. In this paper, we will pay attention to the fundamental limits on the accuracy of spectrum prediction algorithms, instead of constraining ourselves to discuss the performance of a particular spectrum prediction algorithm.

Real-world spectrum data can provide evidence for the research of performance bounds of predictability on spectrum prediction. As an example, Fig. 1 shows the evolution trajectories of the 3-day real-world RSS in the GSM900 downlink bands and the GSM1800 downlink bands. The measured power spectrum density (PSD) values,[1] which are used to represent the RSS. The larger the PSD value is, the higher the signal strength is and the busier the service of the corresponding frequency band is. As can be seen from both Fig. 1(a) and (b), the RSS of part of measurement points has the variation trend like tidal effect, which means the frequency bands are relatively clear in the rest time, like deep night and early morning, but relatively busy in the working time. The RSS of minority measurement points remains unchanged and the RSS of other measurement points evolves randomly and disorderly.

Accordingly, we can observe that regularity and randomness coexist in the evolution of the RSS of bands over the time. The future RSS can be always predicted accurately to some extent. Certainly, it is obvious that the possibility that the future RSS is accurately predicted will be different when using different spectrum prediction algorithms. Even if we use an entirely accurate spectrum prediction algorithm, predictability may have an upper bound due to the inherent randomness in the RSS dynamics and the limited amount of history data.

In our previous work [7], we introduced the concept of the predictability into spectrum prediction and explored the limits of the predictability in radio spectrum state (RSS) dynamics by studying the RSS evolution patterns in spectrum bands of several popular services based on the real-world spectrum measurements, including TV bands, ISM bands, and Cellular bands. We investigated the measured power spectral density values (PSD), instead of the binary spectrum occupancy (BSO), to analyze the predictability mainly for the reasons that the PSD is the original raw data, also the BSO highly depends on the selection of the detection threshold inevitably and introduces detection or sensing errors obtained by comparing with the detection threshold. We conducted the entropy analysis by taking the real-world measurements in TV bands as an example and quantizing the PSD values into 8 levels. With the obtained actual entropy, we calculated the upper bound $\Pi^{\max}$ of the predictability. Furthermore, we illustrated the cumulative distribution functions for the predictability of various services from a statistical perspective. We can derive some conclusions from the above work. On the one hand, the predictability in the real-world RSS dynamics can reach up to 90% despite the apparent randomness. On the other hand, the predictability of various services and its distribution would be different due to humans' spectrum usage.

To present more comprehensive and profound analysis of the predictability, we continue our research on the basis of the previous work. Specifically, the new contributions of this paper are summarized as follows:

- We obtain two kinds of lower bounds, Π^{unc} and R, of the predictability.
- We analyse the impact of the size of data, including sampling intervals and the size of original data, on calculating entropies.
- We analyse the impact of different quantization levels on entropies, the upper bound $\Pi^{\max}$ and the lower bound Π^{unc} of the predictability.

The key insights of real-world spectrum data analytics in this paper include: i) entropies almost have no relationship with selection of sampling intervals in the data preprocessing; ii) the upper bound $\Pi^{\max}$ and the lower bound Π^{unc} of the predictability will both decrease as the quantization level rises and tend to be stable around a value at last; iii) the other lower bound, the regularity R, can reveal the tidal effect of the evolution of the RSS.

The remainder of this paper is organized as follows. Section II will introduce the procedure of data preprocessing. Section III will talk about computing the entropy of a symbol sequence with finite length. The upper bound and the lower bounds of the predictability will be presented in details in Section IV. Section V will show the results and discussions about the predictability. Conclusions are drawn in Section VI.

II. DATA PREPROCESSING

The real-world spectrum dataset of the RWTH Aachen University spectrum measurement campaign[2] is made up of the measured PSD values in chronological order of a number of measurement points in the bands. We can try to regard the sequence of measured PSD values as the information sequence. The measured PSD values are represented in the decimal form and data preprocessing can help us transform the sequence of measured PSD values into the symbol sequence in the appropriate form which is convenient

[1] All measured data used to present the figure are from the open source real-world spectrum dataset of the well-known RWTH Aachen University spectrum measurement campaign [9].

[2] The researchers have conducted a strict and complete spectrum measurement from December 2006 to July 2007 at two locations in Aachen, Germany, and one location in Maastricht, Netherlands. The measured bands ranges from 20MHz to 6GHz and are composed of four subbands of each 1.5GHz bandwidth. A resolution bandwidth of 200 kHz is chosen as a compromise between frequency resolution and the maximum supported span. So, each subband of 1.5GHz bandwidth includes 8192 measurement points, which will cause a small overlap between adjacent measurement channels. The inter-sample time is about 1.8s that means there will be about 1000 measured PSD values obtained in 30 minutes for each measurement point by the measurement system. In this paper, almost all the bands carrying the popular services will be included in our analysis.

for analysis. Data preprocessing can be divided into two parts: sampling and quantization.

A. SAMPLING

The original inter-sample time in the RWTH Aachen University spectrum measurement campaign is about 1.8s, which results in about 48000 measured PSD values one day for each individual spectrum point. It is hard for us to process such huge data with a PC in time. We may obtain a new spectrum dataset by taking one PSD value as a sample every several consecutive PSD values to facilitate the presentation and analysis. We can call the procedure ‘sampling’. The interval time between the two adjacent samples based on the original measurements can be defined as the sampling interval, which may have effect on the entropy of the sequence and then on the predictability. It is obvious that the details of evolution trajectories will be lost more or less after sampling. The longer the sampling interval is, the more details are lost. How much impact sampling has on the predictability will be discussed in Section V-C.

B. QUANTIZATION

The measured PSD with continuous values can be transformed into the symbols via quantization to facilitate further processing. convenient for analysis. We take the maximum value and the minimum value of the original dataset as the upper and lower bound respectively and quantize the PSD values between the upper and lower bound into Q levels equally ($Q \geq 3$). The PSD values inside each quantization interval will be represented by each corresponding symbol.

There is still a special case taken into consideration. When $Q = 2$, it is inappropriate to quantize the PSD values into 2 levels equally for the FCC’s final rules. The sensitive threshold is specified as −114dBm/200kHz in the rule [9], which means if the PSD with continuous values is bigger than −114dBm, the corresponding channel should be considered as busy and the RSS can be symbolized with the symbol ‘1’; otherwise, the RSS will be symbolized with the symbol ‘0’.

III. ENTROPY ANALYSIS

Given a symbol sequence with finite length, how can we depict the randomness of the discrete source transmitting a symbol? In the Shannon’s information theory [6], the information entropy is the general information measure of the source which reflects the randomness of the source.

For a source of the known probability space composed by the prior probabilities that the source transmits different symbols with, we can easily calculate its information entropy according to the formula of source entropy

$$S\left(X\right) = -\sum_{i=1}^{m} p\left(a_i\right) \log p\left(a_i\right), \tag{1}$$

where m represents the total kinds of source symbols and $p\left(a_i\right)$ represents the frequency that the symbol a_i is transmitted [10]. Only the symbol sequence with finite length is given in our problem. Maybe we can estimate the probability distribution of all symbols appearing in the sequence, which will help us obtain the information entropy. The length of the symbol sequence will have a significant influence on the accuracy of the estimated probability distribution. The longer the sequence is, the more accurate the estimation is with the higher computation cost. The temporal correlation of adjacent symbols should be taken into account, too.

Let $X = \{X_i\}$ be the symbol sequence, where X_i represents a symbol at the sampling time slot i, which can be position i mentioned below.

The entropy of a stationary ergodic process X is defined by

$$S(X) = \lim_{n\to\infty} \frac{1}{n} S\left(X_1, X_2, \ldots, X_n\right) \tag{2}$$

and the limit must exist. This is the per symbol entropy of the n-length sequence. Another related quantity for entropy rate is defined by

$$S'(X) = \lim_{n\to\infty} S\left(X_n \left| X_{n-1}, X_{n-2}, \ldots, X_1\right.\right) \tag{3}$$

which represents the conditional entropy of the last symbol given the past sequence with $n-1$ length. According to [10], for a stationary stochastic process X, we have

$$S(X) = S'(X). \tag{4}$$

$S\left(X_n \left| X_{n-1}, \ldots, X_1\right.\right)$ is nonincreasing in n and has a limit $S'(X)$.

So, the entropy of $X = \{X_i\}$ can be written as

$$\begin{aligned} S &\triangleq \lim_{n\to\infty} \frac{1}{n} S\left(X_1, X_2, \ldots, X_n\right) \\ &= \lim_{n\to\infty} S\left(X_n \left| X_{n-1}, X_{n-2}, \ldots, X_1\right.\right) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(X_i \left| h_{i-1}\right.\right) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(i\right), \end{aligned} \tag{5}$$

where $h_{i-1} = \{X_{i-1}, X_{i-2}, \ldots, X_1\}$ and $S\left(n\right) \triangleq S\left(X_n \left| h_{n-1}\right.\right)$ is defined as the conditional entropy of the symbol at position n.

The definition of the entropy of a stationary ergodic process has been given in (2), which holds under the condition that the length of the sequence approaches infinity. In practice, some straightforward methods for estimating the entropy rate of an unknown source would be to run a universal coding algorithm on a finite long sequence of the source output. If the symbol sequence is long enough for the algorithm to converge, the compression ratio is a good estimate for the source entropy. Here, we invoke the Lempel-Ziv algorithm [11] to address this issue. The algorithm reveals deep connections between the entropy rate of a stationary ergodic process and the longest match-length of the subsequence in the process of encoding compression.

Specifically, for $i \leq j$, X_i^j denotes the subsequence $\{X_i, \ldots, X_j\}$. For $i \geq 1$, $\Lambda_i = k$ represents the length of the shortest subsequence X_i^{i+k-1} starting at position i that does not appear as a contiguous subsequence of the previous i-length symbol subsequence X_0^{i-1}. The length of the sequence is denoted by n, namely the window size. Then, the entropy of the sequence can be defined as

$$S^{est} = \frac{\log_2 n}{n}\left(\sum_{i=1}^{n} \Lambda_i\right)^{-1}. \tag{6}$$

This method for estimating entropy using Lempel-Ziv algorithm is actually an exhaustive-searching process with heavy computation. When n approaches infinity, S^{est} converges to the actual entropy [11].

If the time correlation between two adjacent symbols isn't taken into account, the entropy of the symbol sequence will be much easier to be computed. The problem degenerates into only considering the frequency of each symbol. The time-uncorrelated entropy can be defined as $-\sum_{i=1}^{M} p_i \log_2 p_i$, denoted as S^{unc}, where M represents the total kinds of all different symbols and p_i represents the appearance frequency of each symbol. Note that when the measured PSD values are transformed into a symbol sequence by sampling and quantization, the relationship between the kinds of symbols M and the quantization level Q is $M \leq Q$. For the non-uniform distribution of the original measured PSD values, there may be no values inside some quantization intervals, leading to that some corresponding symbols don't exist in the symbol sequence.

Furthermore, the problem can degenerate by ignoring the different appearance frequencies of symbols, only thinking about that symbols appear with equal probability. The random entropy can be defined as $\log_2 M$, denoted as S^{rand}.

There is no doubt that for any symbol sequence, $S^{\text{actual}} \leq S^{\text{unc}} \leq S^{\text{rand}}$. Except the actual entropy, the other two entropies are both defined in the case of ignoring some statistical property of the symbol sequence. Entropy represents the nondeterminacy that the source transmits the symbol, also the randomness of the source. The bigger the entropy is, the bigger the randomness of the source is. When the source is regarded as transmitting each symbol with equal probability, the random entropy must be the biggest one among three entropies.

IV. THE UPPER BOUND AND THE LOWER BOUND OF PREDICTABILITY

As mentioned above, entropy represents the nondeterminacy that the source transmits the symbol. When the entropy is equal to 0, it means that there is not any nondeterminacy that the source transmits the symbol. The symbol transmitted by the source at the next moment is determined by the history symbol sequence. In this case, predictability, the possibility that an appropriate prediction algorithm can predict the symbol transmitted at the next moment accurately [8]. When it comes to the random entropy $S^{\text{rand}} = \log_2 M$, it means the probability of transmitting each symbol is equal, then the predictability will not exceed $1/M$ accordingly. So, the relationship between the entropy and the bounds of the predictability can be established.

A. DEFINITION OF PREDICTABILITY

Let $h_{n-1} = \{X_{n-1}, X_{n-2}, \ldots, X_1\}$ denote a history symbol sequence of a measurement point from position 1 to position $n-1$. Let $\Pr[X_n = \hat{X}_n | h_{n-1}]$ be the probability that the estimated next symbol $\hat{X}_n$ agrees with the actual next symbol X_n based on the history sequence h_{n-1}. Let $\pi(h_{n-1})$ be the probability the next symbol will agree with the most likely next symbol x_{ML} based on the history sequence h_{n-1}. Thus

$$\pi(h_{n-1}) = \sup_x \{\Pr[X_n = x | h_{n-1}]\}, \tag{7}$$

where $\Pr[X_n = x | h_{n-1}]$ is the probability that the next symbol X_n is x based on the history sequence h_{n-1}. That is, $\pi(h_{n-1})$ contains the full predictive power including the potential long-range correlations present in the data.

Let $P_a(\hat{X}_n | h_{n-1})$ be the distribution generated by an arbitrary spectrum prediction algorithm $a*$ over the next possible symbol $\hat{X}_n$. Let $P(X_n | h_{n-1})$ be the true distribution over the next symbol. Thus, the probability of correctly predicting the next symbol of the sequence is $\Pr_a\{X_n = \hat{X}_n | h_{n-1}\} = \sum_x P(x | h_{n-1}) P_a(x | h_{n-1})$. Since $\pi(h_{n-1}) \geq P(x | h_{n-1})$ for any x, we have

$$\begin{aligned}\Pr_a\left\{X_n = \hat{X}_n | h_{n-1}\right\} &= \sum_x P(x | h_{n-1}) P_a(x | h_{n-1}) \\ &\leq \sum_x \pi(h_{n-1}) P_a(x | h_{n-1}) \\ &= \pi(h_{n-1}).\end{aligned} \tag{8}$$

In other words, any prediction based on the history sequence h_{n-1} cannot do better than the one that the source transmits the most likely symbol at the next position.

For a hypothetical prediction algorithm $a*$, the maximal predictability is theoretically achievable which can be denoted as

$$P_{a*}(x | h_{n-1}) = \begin{cases} 1 & x = x_{ML} \\ 0 & x \neq x_{ML}. \end{cases} \tag{9}$$

That is to say that $a*$ can always choose the most likely symbol as its prediction of the symbol transmitted at the next position. Then, we have

$$\begin{aligned}\Pr_{a*}\left\{X_n = \hat{X}_n | h_{n-1}\right\} &= \sum_x P(x | h_{n-1}) P_{a*}(x | h_{n-1}) \\ &= \pi(h_{n-1}).\end{aligned} \tag{10}$$

Therefore, $\pi(h_{n-1})$ is not only an upper bound, but is in principle attainable by an appropriate algorithm.

Next, we define the predictability $\Pi(n)$ for the given history sequence whose length is $n-1$. Let $P(h_{n-1})$ be the probability of obtaining a particular history sequence h_{n-1}. Then, the predictability is given by

$$\Pi\left(n\right)\equiv\sum_{h_{n-1}}P\left(h_{n-1}\right)\pi\left(h_{n-1}\right),\tag{11}$$

denoting that all possible history sequences are summed. Taking the limit, we define the overall predictability Π as

$$\Pi\equiv\lim_{n\rightarrow\infty}\frac{1}{n}\sum_{i}^{n}\Pi\left(i\right).\tag{12}$$

Since $\Pi\left(n\right)$ is the best success rate to predict the symbol at the position n, Π may be viewed as the time averaged predictability.

B. UPPER BOUND OF PREDICTABILITY

The conditional entropy is defined in (5). Next, we will relate the conditional entropy to the predictability.

Assuming that $\pi\left(h_{n-1}\right)=p$, we can obtain the inequation about the conditional entropy

$$S\left(X_n\left|h_{n-1}\right.\right)\le S_F\left(\pi\left(h_{n-1}\right)\right),\tag{13}$$

by invoking the Fano's inequality $H\left(X\left|Y\right.\right)\le H\left(P_e\right)+P_e\log\left(m-1\right)$. The derivation is as follows

$$\begin{aligned}S\left(X_n\left|h_{n-1}\right.\right)&\le H\left(P_e\right)+P_e\log\left(m-1\right)\\&=H\left(1-\pi\left(h_{n-1}\right)\right)+\left(1-\pi\left(h_{n-1}\right)\right)\log\left(m-1\right)\\&=-\left[p\log_2 p+\left(1-p\right)\log_2\left(1-p\right)\right]\\&\quad+\left(1-p\right)\log_2\left(M-1\right)\\&\triangleq S_F\left(p\right)=S_F\left(\pi\left(h_{n-1}\right)\right),\end{aligned}\tag{14}$$

where M represents the total kinds of all different symbols, S_F is the notation of the newly defined function in the equations above and $S_F\left(p\right)$ is concave and monotonically decreases with p. Then, we can establish the relationship between the conditional entropy and the predictability by using Jensen's inequality as follows [8]

$$\begin{aligned}S\left(n\right)&=\sum_{h_{n-1}}P\left(h_{n-1}\right)S\left(X_n\left|h_{n-1}\right.\right)\\&\le\sum_{h_{n-1}}P\left(h_{n-1}\right)S_F\left(\pi\left(h_{n-1}\right)\right)\\&\le S_F\left(\sum_{h_{n-1}}P\left(h_{n-1}\right)\pi\left(h_{n-1}\right)\right)\\&=S_F\left(\Pi\left(n\right)\right).\end{aligned}\tag{15}$$

Similarly, we can obtain the relationship between S and Π as

$$\begin{aligned}S&=\lim_{n\rightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}S\left(i\right)\\&\le\lim_{n\rightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}S_F\left(\Pi\left(i\right)\right)\\&\le S_F\left(\lim_{n\rightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}\Pi\left(i\right)\right)\\&=S_F\left(\Pi\right).\end{aligned}\tag{16}$$

For $S_F\left(\Pi\right)\ge S$ and that S_F is a monotone decreasing concave function, we can obtain the minimum value of S_F, namely $S=S^{\mathrm{actual}}$, when Π is set as its maximum value $\Pi^{\max}$. We can describe this relationship by the equations as follows

$$\begin{aligned}S^{\mathrm{actual}}&=S_F\left(\Pi^{\max}\right)\\&=-\left[\Pi^{\max}\log_2\Pi^{\max}+\left(1-\Pi^{\max}\right)\log_2\left(1-\Pi^{\max}\right)\right]\\&\quad+\left(1-\Pi^{\max}\right)\log_2\left(M-1\right).\end{aligned}\tag{17}$$

Thus, $\Pi^{\max}$ is an upper bound of the predictability Π.

C. LOWER BOUND OF PREDICTABILITY

As mentioned above, $S_F\left(\Pi\right)$ is a monotone decreasing concave function in Π. We can refer to the details of entropy analysis in Section III for the value of S_F. The relationship among three entropies is $S^{\mathrm{actual}}\le S^{\mathrm{unc}}\le S^{\mathrm{rand}}$. When S_F reaches its minimum value S^{actual}, the value of variable Π can be regarded as the upper bound of the predictability, denoted as $\Pi^{\max}$. When the temporal correlation isn't taken into account, namely $S_F=S^{\mathrm{unc}}$, we can also regard the value of variable Π as the lower bound of the predictability, denoted as Π^{unc}. The defination of this lower bound of the predictability can be described as

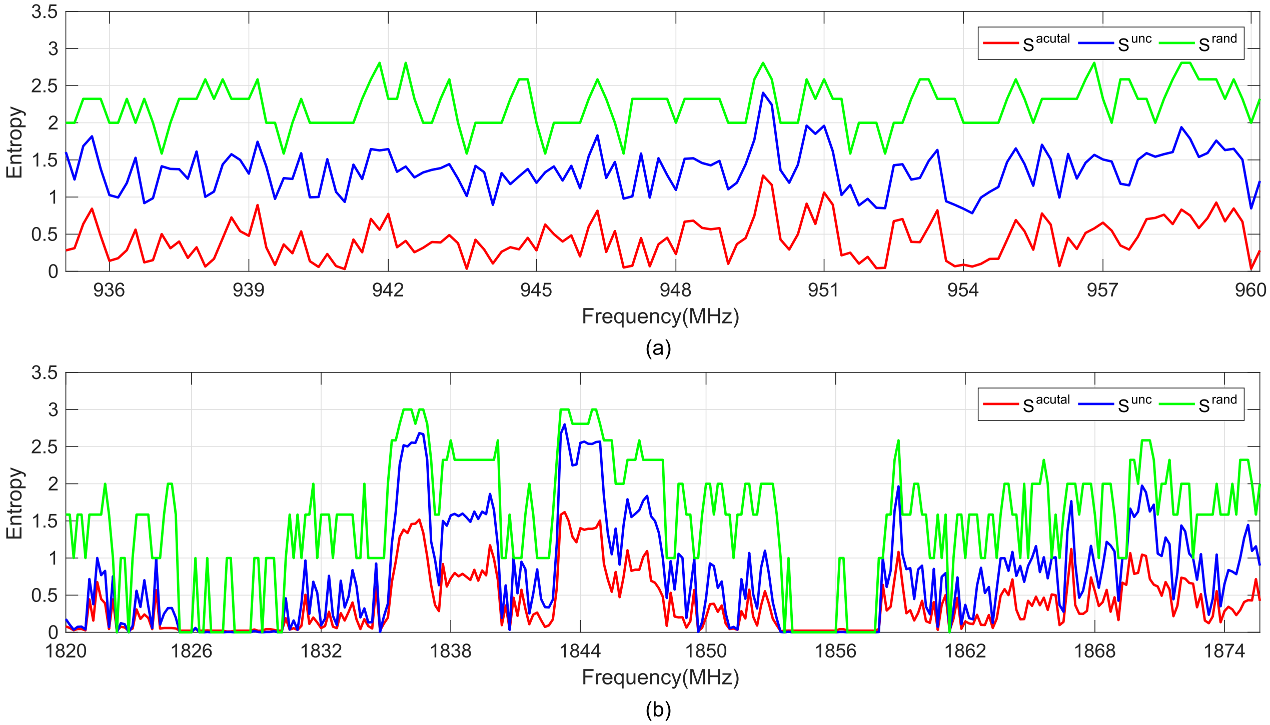
$$\begin{aligned}S^{\mathrm{unc}}&=S_F\left(\Pi^{\mathrm{unc}}\right)\\&=-\left[\Pi^{\mathrm{unc}}\log_2\Pi^{\mathrm{unc}}+\left(1-\Pi^{\mathrm{unc}}\right)\log_2\left(1-\Pi^{\mathrm{unc}}\right)\right]\\&\quad+\left(1-\Pi^{\mathrm{unc}}\right)\log_2\left(M-1\right).\end{aligned}\tag{18}$$

Here, we leave S^{rand} out of consideration, since S^{rand} represents the entropy that each symbol shares equal probability which is too loose to serve as a lower bound of the predictability.

From another angle, we can separate the sequence into several segments to measure the respective regularity,[3] since the long sequence seems random. The regularity of the long sequence is another thought of the lower bound.

Let h'_{n-1} be one segment of the long sequence and $\pi\left(h'_{n-1}\right)\equiv P\left(x'_{ML}\left|h'_{n-1}\right.\right)$. Let $P\left(h'_{n-1}\right)$ be the probability of

[3] Regularity can be considered as the probability of the symbol at the next moment being consistent with the symbol with the highest frequency within a certain period of time [8]. The regularity is denoted as R and also ignores the temporal correlations among symbols, similar to Π^{unc}.



(a)

(b)

FIGURE 2. Entropies of the RSS dynamics in the GSM900 downlink bands (935MHz ∼ 960MHz) and the GSM1800 downlink bands (1820MHz ∼ 1875MHz) during the 6-day measurements ($Q = 8$, sampling interval = 3 min). (a) The entropy of the RSS dynamics in the GSM900 downlink bands. (b) The entropy of the RSS dynamics in the GSM1800 downlink bands.

obtaining the particular segment. Regularity can be denoted as follows

$$R\left(n\right)=\sum_{h'_{n-1}\in h_{n-1}} P\left(h'_{n-1}\right)\pi\left(h'_{n-1}\right). \tag{19}$$

We can prove that $R\left(n\right)$ represents a lower bound for $\Pi\left(n\right)$.

$$\begin{aligned}\Pi\left(n\right) &\equiv \sum_{h_{n-1}} P\left(h_{n-1}\right)\pi\left(h_{n-1}\right)\\ &= \sum_{h_{n-1}}\left(\sum_{h'_{n-1}\in h_{n-1}} P\left(h'_{n-1}\right)P\left(h_{n-1}\left|h'_{n-1}\right.\right)\right)\pi\left(h_{n-1}\right)\\ &= \sum_{h'_{n-1}\in h_{n-1}} P\left(h'_{n-1}\right)\left(\sum_{h_{n-1}} P\left(h_{n-1}\left|h'_{n-1}\right.\right)\pi\left(h_{n-1}\right)\right)\\ &\geq \sum_{h'_{n-1}\in h_{n-1}} P\left(h'_{n-1}\right)\left(\sum_{h_{n-1}} P\left(h_{n-1}\left|h'_{n-1}\right.\right)P\left(x\left|h_{n-1}\right.\right)\right)\\ &= \sum_{h'_{n-1}\in h_{n-1}} P\left(h'_{n-1}\right)P\left(x\left|h'_{n-1}\right.\right)\\ &= \sum_{h'_{n-1}\in h_{n-1}} P\left(h'_{n-1}\right)\pi\left(h'_{n-1}\right)\\ &= R\left(n\right)\end{aligned} \tag{20}$$

Then, the overall regularity is defined as

$$R\equiv\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}R\left(i\right)\leq\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Pi\left(i\right)=\Pi. \tag{21}$$

Combining this result with the upper bound, the predictability Π satisfies $R\leq\Pi\leq\Pi^{\max}$. Thus R is one lower bound of the predictability.

To summarize this section on entropy analysis, we propose three bounds of the predictability in spectrum prediction in all. The upper bound $\Pi^{\max}$ and the lower bound Π^{unc} are considering the grobal predictability from the whole long sequence of a specified long period of time. But the lower bound R is separating the long sequence into sub-segments to consider the respective predictability of every sub-period of time. The former one $\Pi^{\max}$ is derived from searching for the huge data space, bringing heavy computation; the latter two, Π^{unc} and R, both ignore the temporal correlations among symbols of the sequence, with less computation cost. In practice, to make the accuracy of the prediction algorithm as close as possible to the upper bound of the predictability and as higher as possible than the lower bound of the predictability is our pursuit when designing the spectrum prediction algorithms.

V. RESULTS AND DISCUSSION

In this section, we conduct extensive data analysis on the measured PSD values of the first two subbands from 20MHz to 3GHz which were collected on the 3rd floor balcony

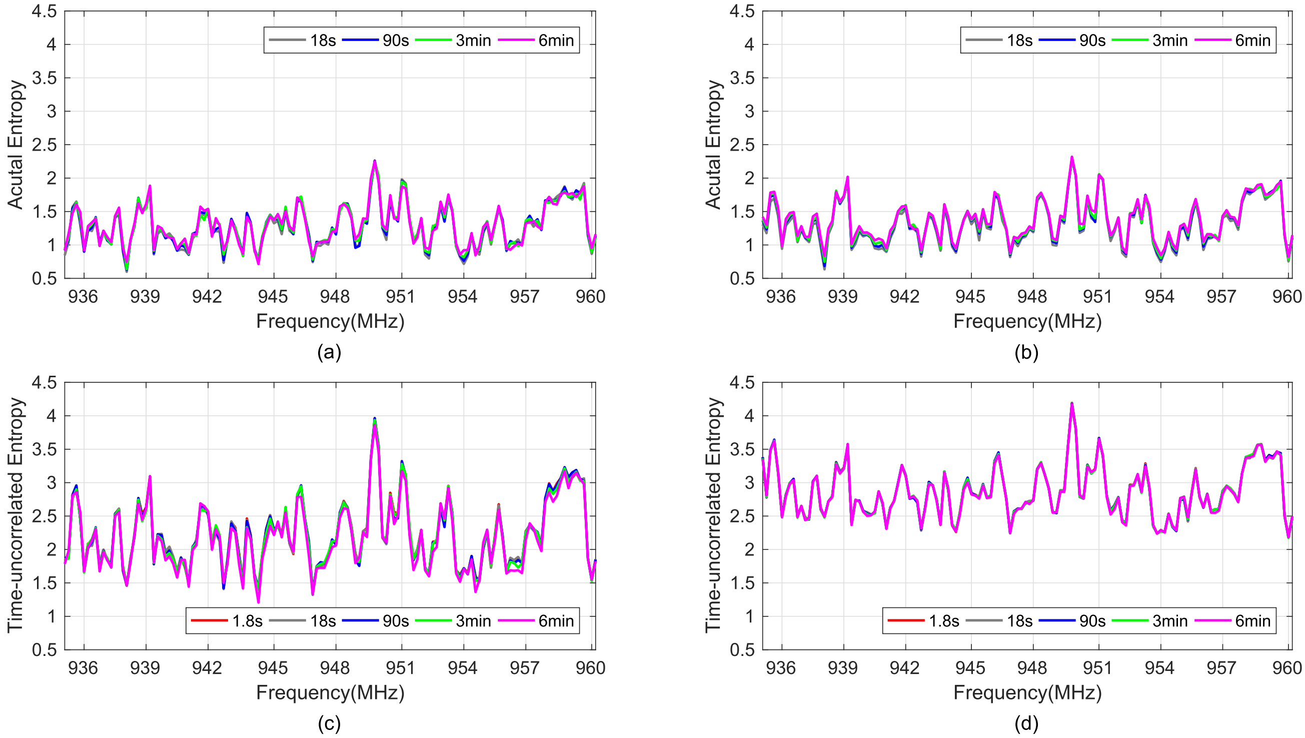


FIGURE 3. The actual entropy and the time-uncorrelated entropy of the RSS dynamics with different sampling intervals in the GSM900 downlink bands (935MHz∼960MHz) during the 1-day measurements and the 6-day measurements respectively ($Q = 32$). (a) During the 1-day measurement ($Q = 32$). (b) During the 6-day measurement ($Q = 32$). (c) During the 1-day measurement ($Q = 32$). (d) During the 6-day measurement ($Q = 32$).

in a residential area in Aachen, Germany. Discussions and analysis on the entropies and bounds of the predictability of the popular services will be specified in the following.

A. EXPERIMENTAL SETUP

Firstly, we take the datasets of the PSD values in the GSM900 downlink bands and the GSM1800 downlink bands as examples to calculate entropies of the RSS. The followed discussion and analysis are mainly based on these datasets. Then, we discuss the impacts of the size of data and quantization levels on entropies.

Secondly, we derive the upper bound $\Pi^{\max}$ and the lower bound Π^{unc} of the predictability based on the computation of entropies. Then, the impact of quantization levels on predictability is discussed in details. With the appropriate selection of quantization levels and the size of data, more spectrum data are involved in our research and predictability of different bands will be illustrated.

The bounds above are both considering the global predictability, while the lower bound R focuses on the time-segmented predictability which will be included at last.

B. ENTROPIES OF BANDS

Entropies of the RSS in the GSM900 downlink bands and the GSM1800 downlink bands have been shown in Fig. 2. Here, the random entropy S^{rand}, the time-uncorrelated entropy S^{unc} and the actual entropy S^{actual} of the bands are plotted in the same figure to make comparison. The size of the PSD values is 6-day measurements; the quantization level Q is 8 and the sampling interval is 3min. The relationship between three entropies is confirmed exactly $S^{\text{actual}} \leq S^{\text{unc}} \leq S^{\text{rand}}$, consistent with the theory.

Actual entropies of measurement points in the GSM900 downlink bands are uniformly distributed in the range of 0 to 1, which means that the RSSs of some measurement points relatively evolve randomly while those remain regular relatively. Actual entropies of some measurement points in the GSM1800 downlink bands equal to 0, which means the RSS of them keep almost unchanged. Meanwhile, the RSS of the rest of measurement points in the GSM1800 downlink bands evolves randomly, consistent with the colorful strips in the Fig. 1.

C. IMPACT OF THE SIZE OF DATA ON ENTROPIES

The size of data have great impact on the computation of entropies, especially the actual entropy. When calculating the actual entropy, the entire sequence is searched for ‘the shortest subsequence’ at each position. The longer the sequence is, the more time the process of calculating the actual entropy costs. Thus, it is important for us to reduce the data amount involved in our analysis to provide convenience.

A direct method to reduce the data amount is to involve less days’ measurements in the analysis. The subgraphs in the left column in Fig. 3 use 1-day measurements and the subgraphs in the right column use 6-day measurements. Curve clusters of the subgraphs in the same row approximately have the same trend. The more data is involved, the more states of the radio spectrum there will occur, thus, the bigger the entropies

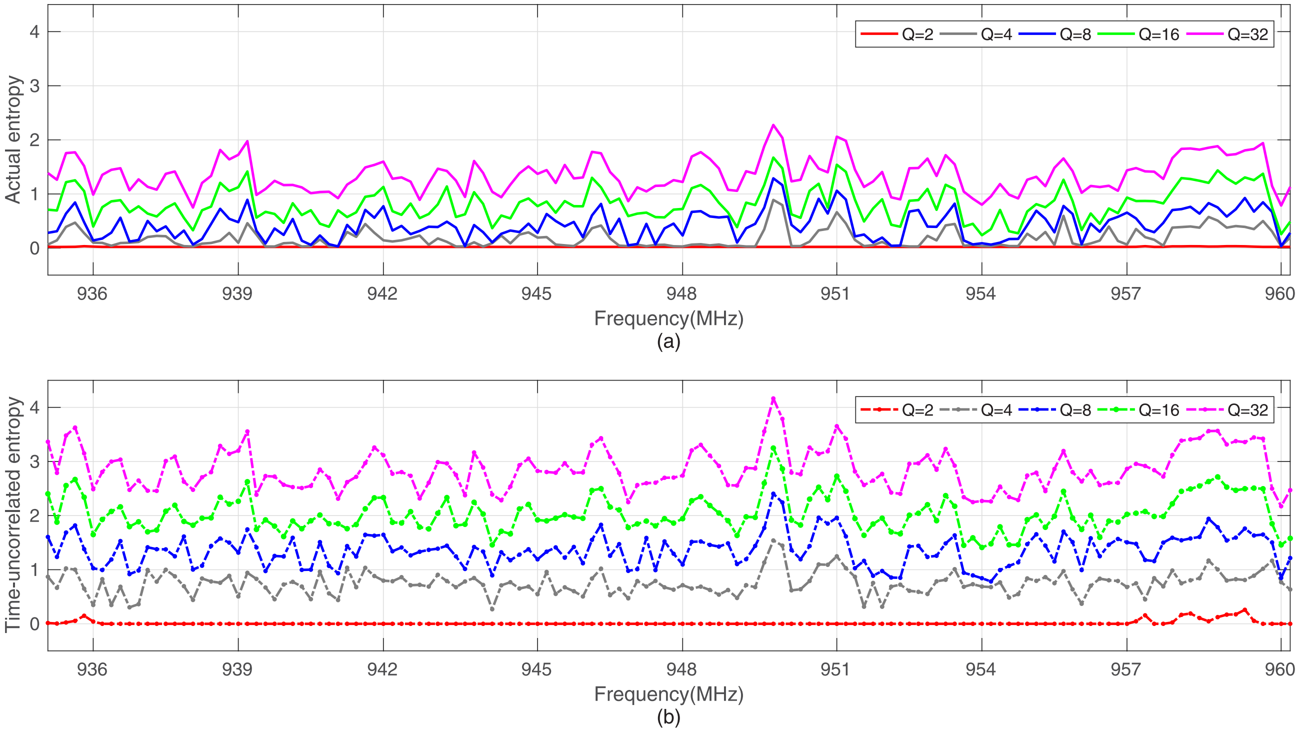


FIGURE 4. Entropies of the RSS dynamics with different quantization levels in the GSM900 downlink bands (935MHz ∼ 960MHz) during the 6-day measurements (sampling interval = 3min). (a) The actual entropy. (b) The time-uncorrelated entropy.

are, especially for the time-uncorrelated entropies which are directly proportional to the kinds of symbols M.

Another method to reduce data is sampling. Prolonging the sampling interval is to reduce the data amount when keeping days of measurements unchanged. Subgraphs in Fig. 3 also compare the entropies when selecting different sampling intervals. Whatever days of measurements are and whatever quantization levels are, curve clusters of actual entropies and time-uncorrelated entropies almost overlap respectively. As a result, we can find that the sampling interval has ignorable effect on the entropies. When it comes to processing large amount of data, like several-day measurements, we can prolong the sampling interval to make the processing procedure as short as ensuring the abundant sampling data.

D. IMPACT OF QUANTIZATION LEVELS ON ENTROPIES

Quantization, as an important part of data preprocessing, helps us transform the measured PSD The measured PSD with continuous values into the symbols. It is convenient for us to calculate entropies or do further analysis on the symbol sequence. When the quantization level Q equals to 2, there are only two kinds of symbols in the symbol sequence and the entropy of the symbol sequence will be so small that no more than 1. The larger the quantization level is, the more kinds of symbols there are in the symbol sequence and the bigger the entropies are.

The above conclusions obtained from theoretical analysis are reinforced in Fig. 4. Entropies will increase as the quantization levels rise and the time-uncorrelated entropies rise with the larger magnitudes. Moreover, the selection of quantization levels will affect the computation speed of entropies. More quantization levels can bring the shorter data processing procedure.

If we continue to predict the future RSS on the basis of the history symbol sequence, the prediction result of the future RSS must be the symbols instead of the numerical PSD values. When the quantization level Q equals to 2, the possibility of predicting the future RSS accurately is at least 50%. The larger the quantization level is, the more kinds of symbols there will occur in the future and the smaller the possibility of predicting the future RSS accurately is. Whether this conclusion is correct will be discussed in the following section.

E. PREDICTABILITY OF BANDS

On the basis of the above analysis and discussion, we investigate the upper bound and the lower bound of the predictability of the RSS dynamics in the GSM900 downlink bands and the GSM1800 downlink bands, where the lower bound of the predictability here refers to Π^{unc}. The size of the PSD values is 6-day measurements; the quantization level Q is 8 and the sampling interval is 3min.

Fig. 5(a) shows the predictability of the RSS dynamics in the GSM900 downlink bands and the upper bounds of the predictability of all measurement points are up to 80%. The lower bounds of the measurement points are above 60% except a few measurement points. Although it seems that high randomness exists in the evolution trajectories, there have a

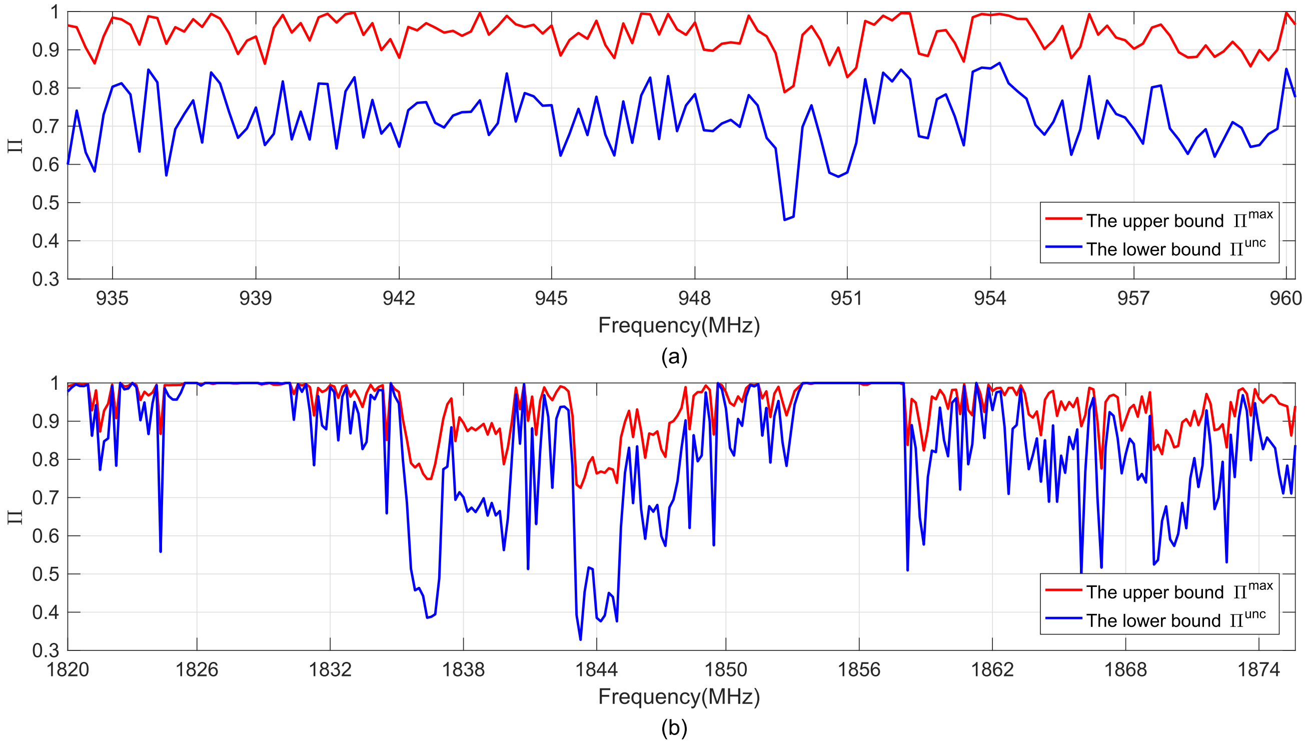


FIGURE 5. Predictability of the RSS dynamics in the GSM900 downlink bands (935MHz∼960MHz) and the GSM1800 downlink bands (1820MHz∼1875MHz) during the 6-day measurements ($Q = 8$, sampling interval = 3min). (a) Predictability of the RSS dynamics in the GSM900 downlink bands. (b) Predictability of the RSS dynamics in the GSM1800 downlink bands.

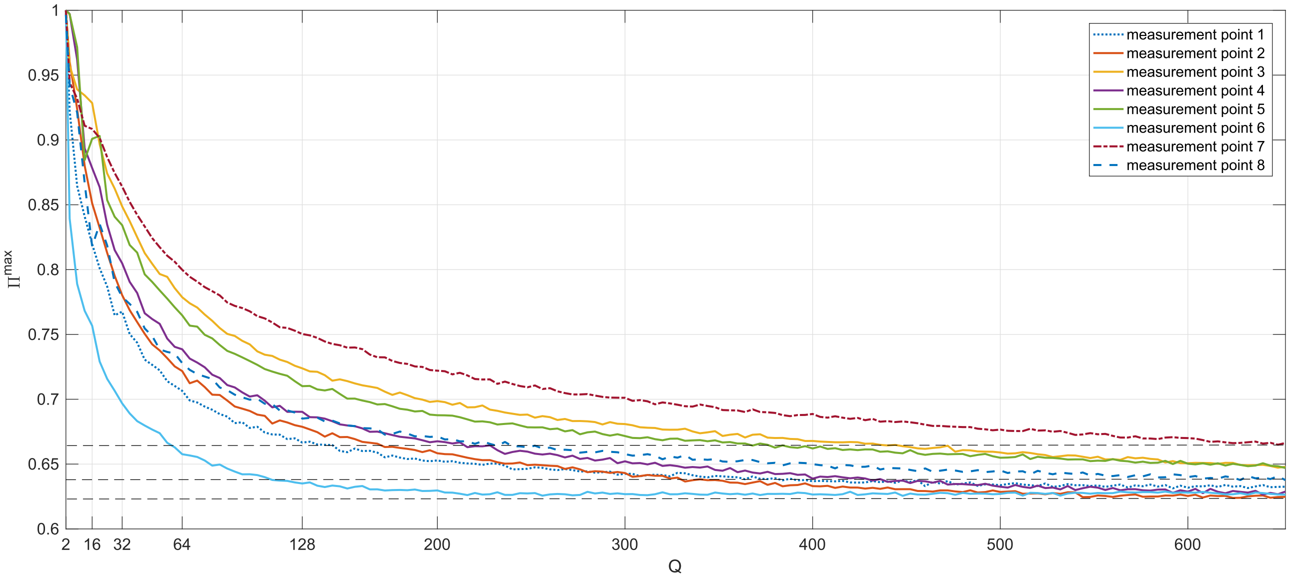


FIGURE 6. The upper bound of the predictability with different quantization levels of several measurement points in the GSM900 downlink bands (935MHz∼960MHz) during the 6-day measurements (sampling interval = 3min).

good performance on the predictability. Similar observations can be found on the predictability of the RSS dynamics in the GSM1800 downlink bands in Fig. 5(b).

F. IMPACT OF QUANTIZATION LEVELS ON PREDICTABILITY

To explore whether the quantization level has an impact on the predictability of the RSS, the upper bounds and the lower bounds of the predictability when selecting different quantization levels are compared with each other. To facilitate the presentation, we randomly select 8 measurement points among all measurement points of the GSM900 downlink bands to plot their bounds of the predictability of the RSS for comparison in Fig. 6 and Fig. 7.

It is consistent that the upper bounds and the low bounds of the predictability will both decrease as the quantization

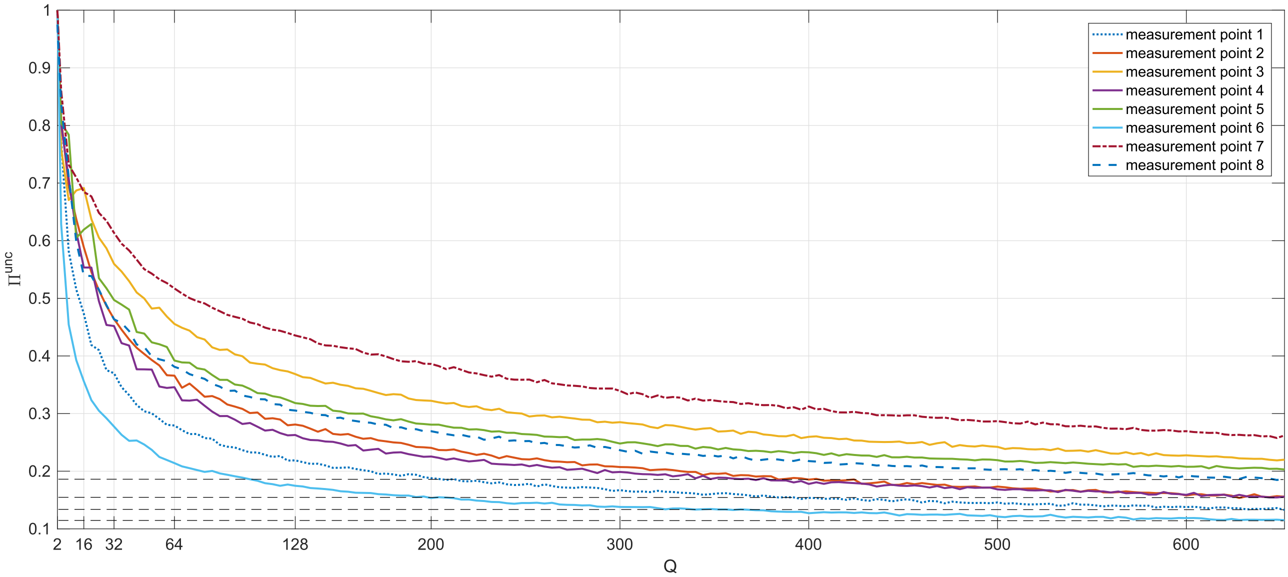


FIGURE 7. The lower bound Π^{unc} of the predictability with different quantization levels of several measurement points in the GSM900 downlink bands (935MHz~960MHz) during the 6-day measurements (sampling interval = 3min).

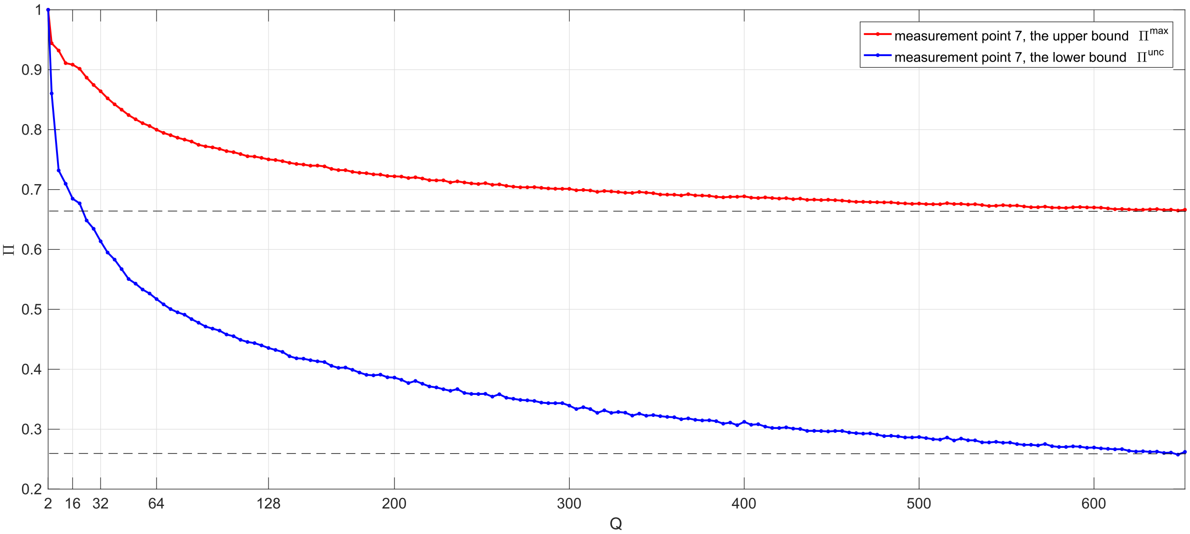


FIGURE 8. The upper bound Π^{max} and the lower bound Π^{unc} of the predictability of one measurement point in the GSM900 downlink bands (935MHz~960MHz) during the 6-day measurements (sampling interval = 3min).

level Q rises, and then decrease slowly, keeping stable around a value at last. The difference between the maximum value and the minimum value of the original dataset we used is about 65dBm. When the quantization level Q is greater than 650, which means the corresponding quantization interval is less than 0.1dBm, the corresponding resolution of prediction is high enough to neglect the effect of quantization on the original data.

We can also pay attention to one of the 8 measurement points to make comparison between the upper bound Π^{max} and the lower bound Π^{unc} of the predictability. The upper bound of the predictability of the RSS of measurement-point-7 tends to be around 0.67 and the lower bound tends to be around 0.265 in Fig. 8. Without quantization, the maximum possibility to predict the future RSS accurately with the appropriate prediction algorithm based on the original dataset is still up to 67%. Certainly, the bound will vary with each measurement point.

G. PREDICTABILITY OF DIFFERENT BANDS

The upper bound Π^{max} and the lower bound Π^{unc} of the predictability are related to the actual entropy S^{actual} or the time-uncorrelated entropy S^{unc} and the number of the symbols M in (17) and (18). So, the upper bound Π^{max} and the lower bound Π^{unc} will be influenced by the quantization level Q and the dataset itself involved in the data processing.

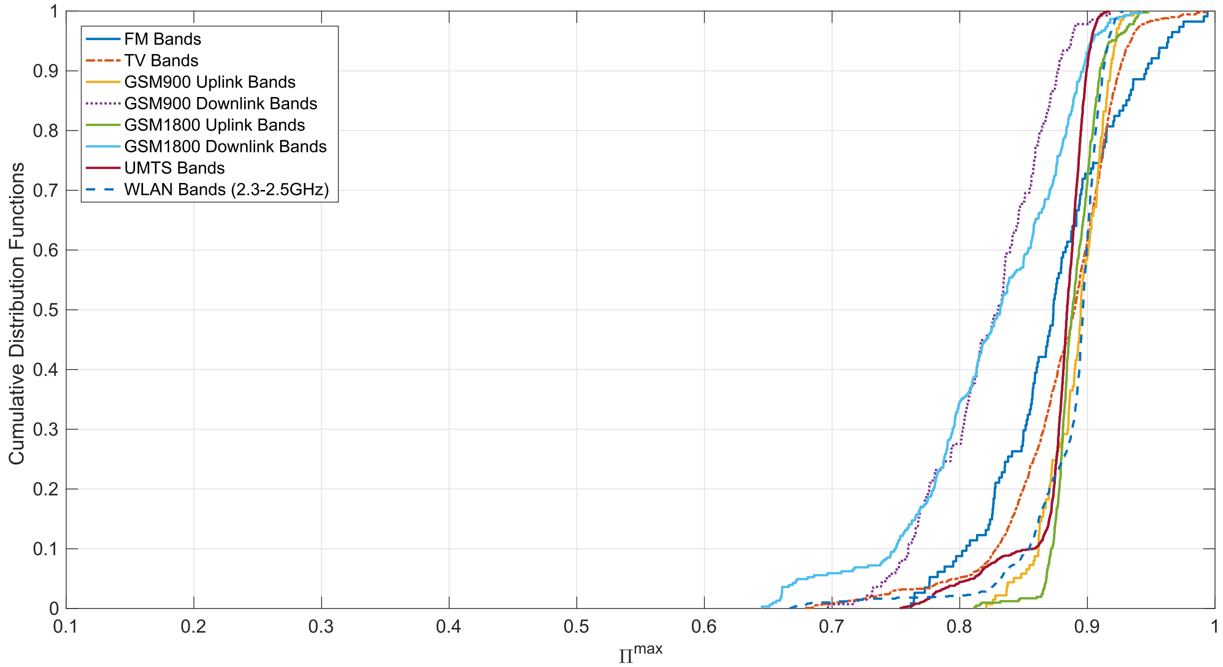


FIGURE 9. The cumulative density functions (CDFs) for the upper bound of the predictability in different bands during the 6-day measurements ($Q = 32$, sampling interval = 3min).

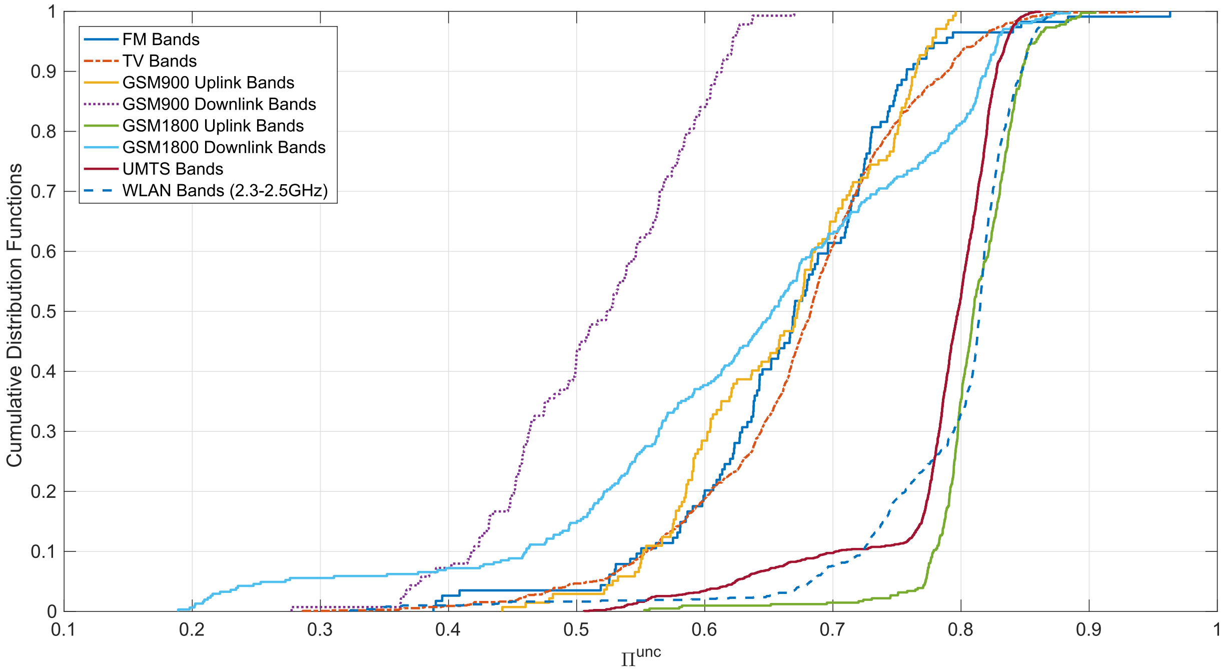
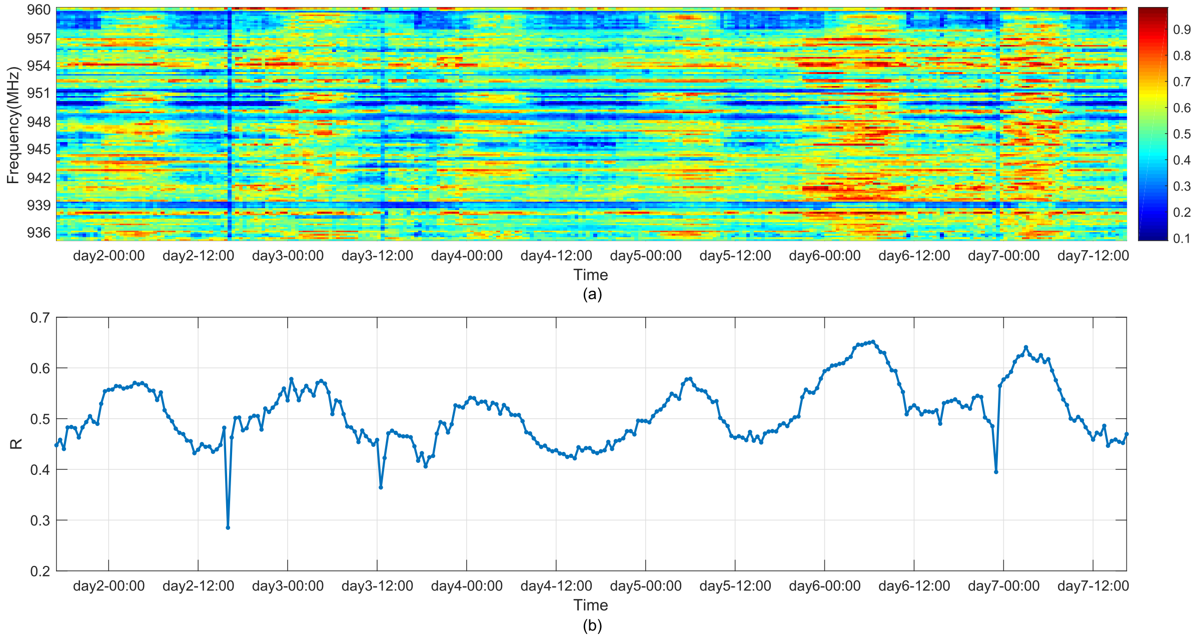


FIGURE 10. The cumulative density functions (CDFs) for the lower bound Π^{unc} of the predictability in different bands during the 6-day measurements ($Q = 32$, sampling interval = 3min).

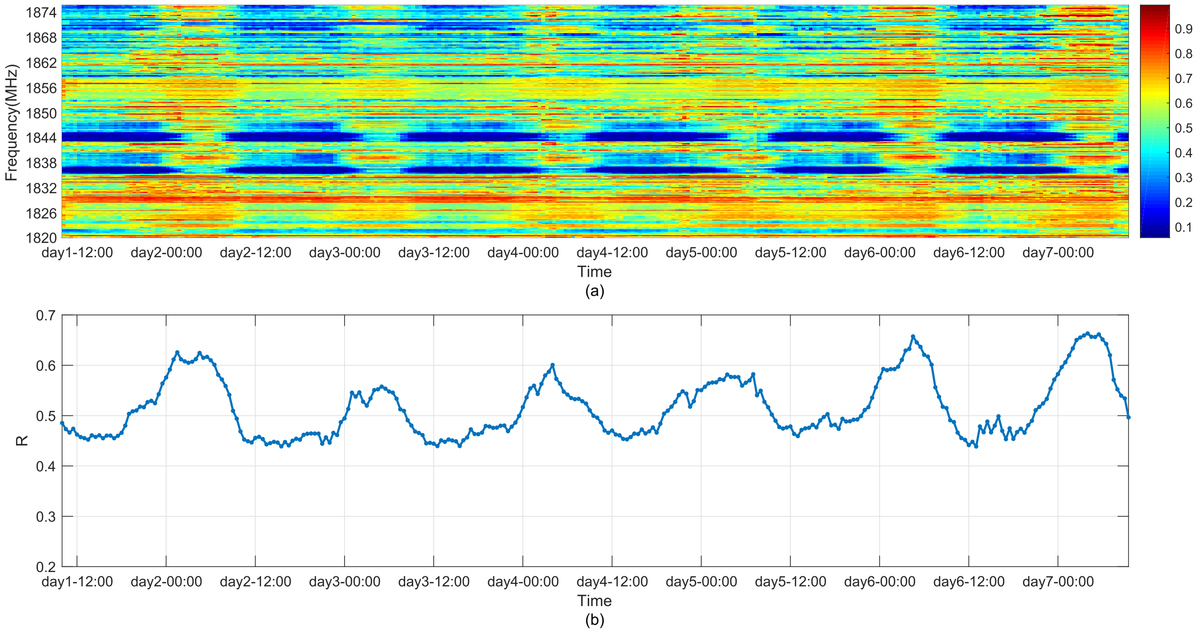
Predictability of the RSS in the bands of all popular services is illustrated in Fig. 9 and Fig. 10. The sizes of the PSD values of different services are all 6-day measurements; the quantization level Q is 32 and the sampling interval is 3min. It follows that the upper bounds of the predictability of most popular services are around 90% and the upper bounds of the predictability of the GSM900 downlink bands and the GSM1800 downlink bands have a relatively poor performance due to the influence of the randomness of human activities. The lower bounds Π^{unc} of the predictability range from 40% to 80%. Some bands have high lower bounds of the predictability, which represents the future RSS of these



(a)

(b)

FIGURE 11. The regularity of the GSM900 downlink bands during the 6-day measurements ($Q = 32$, sampling interval = 1.8s). (a) The regularity of measurement points in the GSM900 downlink bands within each half-hourly interval. (b) The average regularity of the GSM900 downlink bands within each half-hourly interval.



(a)

(b)

FIGURE 12. The regularity of the GSM1800 downlink bands during the 6-day measurements ($Q = 32$, sampling interval = 1.8s). (a) The regularity of measurement points in the GSM1800 downlink bands within each half-hourly interval. (b) The average regularity of the GSM1800 downlink bands within each half-hourly interval.

bands can be easily predicted accurately. When it comes to predicting the future RSS of these bands, simple and efficient algorithms without taking the time correlation into account are enough to predict the future RSS accurately.

H. THE LOWER BOUND R OF PREDICTABILITY

Another lower bound of the predictability R is to consider the probability of the symbol at the next moment being consistent with the symbol with the highest frequency within a certain period of time. To let each half-hour measurements as one segment of the symbol sequence, we calculate the regularity of each segment of the symbol sequence of all measurement points in the bands, then averaging the regularity of all measurement points in each half-hour period as the average regularity R of the bands in each half-hour period. The curves of R with time varying of the GSM900 downlink bands and the GSM1800 downlink bands are respectively plotted in Fig. 11(b) and Fig. 12(b). The lower bounds R of the predictability mainly range from 40% to 60% and both evolve clearly like tidal effect. It accords with the speciality of the bands of the two popular service that clear in the rest time and busy in the working time.

VI. CONCLUSION

This paper studies the predictability of the RSS in the popular bands with the real-world spectrum data. The first contribution is to find that entropies almost have no relationship with selection of sampling intervals in the data preprocessing. We can prolong the sampling interval to make computation quicker to facilitate the analysis and research. The second contribution is to find that the upper bounds and the lower bounds of the predictability will both decrease as the quantization level rises and tend to be stable around a value at last, which means there exists the fundamental limits of the predictability of the RSS. The third contribution is to propose two kinds of lower bounds of the predictability, and one of the lower bounds, the regularity R, can reveal the tidal effect of the evolution of the RSS.

REFERENCES

[1] E. Axell, G. Leus, E. G. Larsson, and H. V. Poor, "Spectrum sensing for cognitive radio: State-of-the-art and recent advances," *IEEE Signal Process. Mag.*, vol. 29, no. 3, pp. 101–116, May 2012.

[2] X. Xing, T. Jing, W. Cheng, Y. Huo, and X. Cheng, "Spectrum prediction in cognitive radio networks," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 90–96, Apr. 2013.

[3] G. Ning and P. Nintanavongsa, "Time prediction based spectrum usage detection in centralized cognitive radio networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 300–305.

[4] S. Yin, D. Chen, Q. Zhang, and S. Li, "Prediction-based throughput optimization for dynamic spectrum access," *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, pp. 1284–1289, Mar. 2011.

[5] G. Ding *et al.*, "Spectrum inference in cognitive radio networks: Algorithms and applications," *IEEE Commun. Surveys Tuts.*, to be published.

[6] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[7] G. Ding *et al.*, "On the limits of predictability in real-world radio spectrum state dynamics: From entropy theory to 5G spectrum sharing," *IEEE Commun. Mag.*, vol. 53, no. 7, pp. 178–183, Jul. 2015.

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[9] M. Wellens, "Empirical modelling of spectrum use and evaluation of adaptive spectrum sensing in dynamic spectrum access networks," Ph.D. dissertation, RWTH Aachen Univ., Aachen, Germany, May 2010.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[11] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1319–1327, May 1998.

JIACHEN SUN received the B.S. degree in information engineering from Southeast University, Nanjing, China, in 2016. She is currently pursuing the M.S. degree with the College of Communications Engineering, Army Engineering University, Nanjing, China. Her research interests include data analytics, wireless communications, and cognitive radio networks.

LIANG SHEN received the B.S. degree in communications engineering and the M.S. degree in communications and information system from the Institute of Communications Engineering, Nanjing, China, in 1988 and 1991, respectively. He is currently a Professor with the Army Engineering University, China. His current research interests are information theory and digital signal processing, and wireless networking.

GUORU DING (S'10–M'14–SM'16) received the B.S. degree (Hons.) in electrical engineering from Xidian University, Xi'an, China, in 2008, and the Ph.D. degree (Hons.) in communications and information systems from the College of Communications Engineering, Nanjing, China, in 2014. Since 2014, he has been an Assistant Professor with the College of Communications Engineering and a Research Fellow with the National High Frequency Communications Research Center of China. Since 2015, he has been a Post-Doctoral Research Associate with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. His research interests include cognitive radio networks, massive MIMO, machine learning, and big data analytics over wireless networks.

He has acted as a Technical Program Committee member for a number of international conferences, including the IEEE Global Communications Conference, the IEEE International Conference on Communications, and the IEEE Vehicular Technology Conference. He is also a Voting Member of the IEEE 1900.6 Standard Association Working Group. He was a recipient of the Best Paper Awards from EAI MLICOM 2016, IEEE VTC 2014-Fall, and the IEEE WCSP 2009. He also received the Alexander von Humboldt Fellowship in 2017 and the Excellent Doctoral Thesis Award of China Institute of Communications in 2016. He has served as a Guest Editor of the IEEE Journal on Selected Areas in Communications (Special issue on spectrum sharing and aggregation in future wireless networks). He is currently an Associate Editor of the *Journal of Communications and Information Networks*, the *KSII Transactions on Internet and Information Systems*, and the *AEU-International Journal of Electronics and Communications*

RONGPENG LI received the B.E. degree (Hons.) from Xidian University, Xi'an, China, in 2010, and the Ph.D. degree (Hons.) from Zhejiang University, Hangzhou, China, in 2015. He was a Visiting Doctoral Student with Supélec, Rennes, France, in 2013, and an Intern Researcher with the China Mobile Research Institute, Beijing, China, in 2014. From 2015 to 2016, he was a Researcher with the Wireless Communication Laboratory, Huawei Technologies Company, Ltd., Shanghai, China. He is currently a Post-Doctoral Researcher with the College of Computer Science and Technologies, Zhejiang University. He has authored/co-authored over 35 papers in the related fields. His research interests currently focus on resource allocation of cellular networks (especially full duplex networks), applications of reinforcement learning, and analysis of cellular network data. He was granted by the National Post-Doctoral Program for Innovative Talents, which has a grant ratio of 13% in 2016. He served as the Web Design Chair of the IEEE OnlineGreenComm 2015 and the Web and System Chair of the IEEE ISCIT 2011. He serves as an Editor of *China Communications*.

QIHUI WU (SM'13) received the B.S. degree in communications engineering, the M.S. and Ph.D. degrees in communications and information systems from the Institute of Communications Engineering, Nanjing, China, in 1994, 1997, and 2000, respectively. From 2003 to 2005, he was a Post-Doctoral Research Associate with Southeast University, Nanjing, China. From 2005 to 2007, he was an Associate Professor with the College of Communications Engineering, PLA University of Science and Technology, Nanjing, China, where he served as a Full Professor from 2008 to 2016. In 2011, he was an Advanced Visiting Scholar with the Stevens Institute of Technology, Hoboken, USA. Since 2016, he has been a Full Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests span the areas of wireless communications and statistical signal processing, with emphasis on system design of software defined radio, cognitive radio, and smart radio.