MINING SPECTRUM USAGE DATA: A LARGE-SCALE SPECTRUM MEASUREMENT STUDY

Sixing Yin[†], Dawei Chen[‡], Qian Zhang[‡], Mingyan Liu[∨] and Shufang Li[†]

[†] Beijing University of Posts and Telecommunications
yinsixing@sina.com, lisf@bupt.edu.cn
[‡] Hong Kong University of Science and Technology
{dwchen,qianzh}@cse.ust.hk
[∨] University of Michigan
mingyan@eecs.umich.edu

Abstract

Dynamic spectrum access has been a hot topic for extensive study in recent years. The increasing volumes of literatures calls for a deeper understanding of the characteristics of current spectrum utilization. In this paper we present a detailed spectrum measurement study, with data collected in the 20MHz to 3GHz spectrum band and at four locations concurrently in Guangdong province of China. We examine the statistics of the collected data, including channel vacancy statistics, channel utilization within each individual wireless service, and the spectral and spatial correlation of these measures. Main findings include that the channel vacancy durations follow an exponential-like distribution, but are not independently distributed over time, and that significant spectral and spatial correlations are found between channels of the same service. We then exploit such spectrum correlation to develop a 2-dimensional frequent pattern mining algorithm that can predict channel availability based on past observations with considerable accuracy.

Index Terms

Spectrum Measurement, Channel Vacancy Duration, Service Congestion Rate, Spectrum Usage Prediction, Frequent Pattern Mining, FPM-2D, Spectral Correlation, Spatial Correlation

I. INTRODUCTION

Recent advances in software defined radio (SDR) [10] and cognitive radio (CR) [6], [19] combined with ever-increasing demand for wireless spectrum resources have led to the notion of dynamic spectrum access; wireless devices with the ability to detect spectrum availability and the flexibility to adjust operating frequencies can opportunistically access under-utilized spectrum. This type of access is aimed at significantly improving spectrum efficiency in light of evidence that there exists abundant spectrum availability [12], [13] in the current allocation. This has also led to the notion of open access, whereby unlicensed users or devices are encouraged to access to licensed spectrum bands such that spectrum opportunity can be fully exploited[1].

These concepts have motivated extensive studies on both technical and policy issues related to dynamic spectrum access. With this comes the need for better and quantitative understanding of current spectrum utilization, beyond the qualitative knowledge of the existence of ample spectrum opportunities. With such understanding one can help (1) validate spectrum/channel models often used in analysis without questioning, (2) provide grounds for more realistic channel models with better predictive competence, and ultimately, (3) allow devices to adaptively make more effective dynamic spectrum access decisions.

[1]For instance, the FCC on November 4, 2008 approved unlicensed wireless devices that operate in the empty white space between TV channels, after four years of effort.

To achieve these goals, we recently conducted a comprehensive spectrum measurement study in the 20MHz to 3GHz spectrum band in Guangdong province, China. This paper reports our methodology and findings from this study. There has been a number of spectrum measurement studies published in recent years, like [12], [13] conducted in the US, one in Singapore [8], one in New Zealand [2], and one in Germany [18]. A common finding among these studies is that spectrum resources is indeed heavily underutilized at the moment.

Compared to the prior work, the salient features of the data sets we collected are:

- Our measurements are carried out in four locations concurrently;
- Our measurement locations are specifically selected (2 urban and 2 suburban locations) in order to study the potential spatial correlation of spectrum usage between similar and different types of locations.

There are two parts in this study. In the first part, we examine statistics of the collected data and use a variety of models to fit the data. These include (1) channel vacancy statistics (precisely defined later), over time, across channels, and for different wireless services (that group multiple channels), (2) service congestion rate that reveals how well channels assigned to a particular wireless service are utilized, and (3) the use of a subscriber model to explain the dynamics in channel utilization for specific services, (4) spectral and spatial correlation of spectrum usage. In the second part of the study we exploit the spectrum correlations to develop a 2-dimensional frequent pattern mining algorithm that can predict channel availability based on past observations with considerable accuracy.

The key findings and contributions of this work are summarized as follows:

1) The channel vacancy duration (CVD) distribution is shown to have an exponential tail (but not exactly an exponential distribution; this is empirically obtained but with very high statistical significance), in all channels and locations we studied. This evidence to a certain extent supports some widely used channel models (e.g., the 0-1 Gilbert-Eliot model) under which such vacancy durations are exponentially distributed. On the other hand, statistical analysis on our data reveals that these vacancy durations are *not* independently distributed over time, as is commonly assumed. This finding suggests that spectrum usage is inherently more predictable than current models imply, and that better and more sophisticated models may allow us to exploit such predictability.

2) The service congestion rate (SCR), the spectrum utilization within a specific wireless service, can well fitted by autoregressive model, which makes possible the high-precision prediction for SCR such that degree of congestion for a service can be known about in advance.
3) Spectral correlation of spectrum utilization is significant between channels within the same service, and quite insignificant otherwise. This reflects the difference in the nature of these services, and the resulting different usage patterns.
4) There is very significant spatial correlation between the SCRs of the same service (e.g. GSM900 uplink) at different locations. The spatial correlation is even higher when the two locations are of the same type (both in urban or both in suburban areas). This suggests that usage patterns are heavily influenced by the nature/type of the wireless service, rather than the location. There is a population factor (high vs. low density), but overall similarities in collective usage pattern of the same service are significant in different regions.
5) Motivated by the strong correlation in spectral and spatial dimensions, we propose an effective 2D frequent pattern mining algorithm, which can predict spectrum usage with the accuracy exceeding 95%.

We hope that these findings will lead to more discussions on how to better model current spectrum utilization, i.e., the behavior of primary users. This in turn can help us design better and more efficient spectrum sensing and access schemes for secondary users.

The remainder of the paper is organized as follows. Section II presents how our data is collected and processed. We then present a comprehensive statistical analysis on the measurement data including CVD, SCR, and subscriber model in Sections III. Spectrum correlation (spectral and spatial) results are presented in Sections IV. In Section V we develop in detail a 2D frequent pattern mining technique to predict spectrum availability. We discuss how these results can be useful in spectrum sensing and access in Section VI. Related work is presented in Section VII, and Section VIII concludes the paper.

II. Data Collection and Preprocess

A. Data Collection

The results presented in this paper are based on the analysis of 4 sets of measurement data, which were collected at four different locations in Guangdong province, China, from 15:00 Feb

TABLE I

LOCATION OVERVIEW

Location	Type	Coordinate
1.Trade Center,Guangzhou	Downtown	E 113°15′25″ N 23°08′01”
2.Canadian Garden,Guangzhou	Downtown	E 113°21′45″ N 23°08′20”
3.Jiangmen	Suburban	E 113°7′59.9″ N 22°22′46.9”
4.Zhongshan	Suburban	E 113°27′24.8″ N 22°25′32.5”

TABLE II

SPECTRUM ALLOCATION OF POPULAR SERVICES

Services	Band
CDMA uplink	825MHz - 835MHz
CDMA downlink	870MHz - 880MHz
GSM900 uplink	885MHz - 915MHz
GSM900 downlink	925MHz - 960MHz
GSM1800 uplink	1710.0MHz - 1785.0MHz
GSM1800 downlink	1805.0MHz - 1880.0MHz
Broadcasting TV1	48.5 - 92MHz
Broadcasting TV2	167 - 233MHz
Broadcasting TV3	470MHz - 566MHz
Broadcasting TV4	606 - 870MHz
ISM	2400 - 2500MHz

16, 2009 to 15:00 Feb 23, 2009. Locations 1 and 2 are in the downtown area of Guangzhou, the main metropolis of Guangdong province. These two locations are roughly 10 kilometers apart. Locations 3 and 4 are in suburban areas of two under-developed cities and are roughly 45 kilometers apart. These locations are listed in Table I.

We are primarily interested in spectrum usage of the frequency band between 20MHz and 3GHz. Within this range, the list of wireless services along with their spectrum assignment in the local region are provided in Table II.

The measurement equipment we used is an R&S EM550 VHF / UHF Digital Wideband Receiver. EM550 is a superheterodyne receiver that covers a wide frequency range, from 20 MHz to 3.6 GHz. The measurement resolution is one per 0.2MHz, resulting in a total of 14,900 frequency readings per time slot (or sweep time), roughly 75 seconds. There are 8,058 (7 days/ 75s) time slots. As a result, there are 14,900$\times$8,058 data points in the data set (roughly 2GB in size) per location.

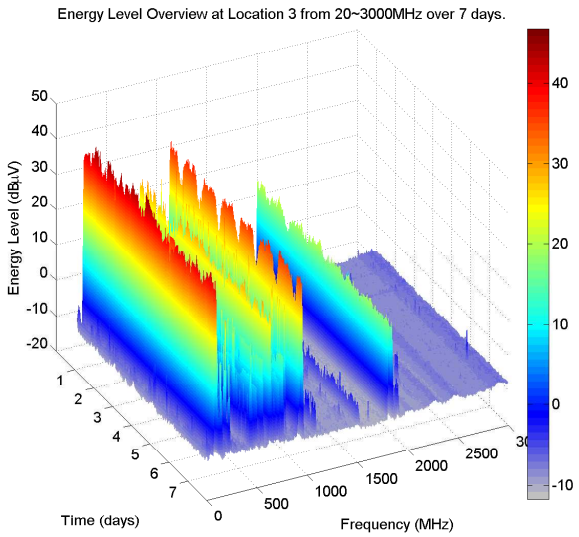

Fig. 1. 3-D view of the energy level over all bands.

Here we would like to briefly compare our data sets with those reported on the Shared Spectrum Company (SSC) website[2]. Judging by the published reports, our data sets are of a similar nature and have been collected in a similar way, e.g., the antennas are placed outdoor on the roof of a building while the receivers are placed indoor.

For illustration purpose, Figure 1 shows a 3-D depiction of the set of data at Location 3. The color coding (energy level from low to high on a scale from blue to bright red) on the figure is an attempt to make the figure easier to visualize, but does not provide extra information, as the vertical dimension already shows the energy reading (in dBμV).

These data sets provide us with a fairly rich set of measurements, based on which spectrum utilization and patterns can be identified and analyzed as we show in subsequent sections.

B. Preprocessing

To conduct the sequence of analysis presented in later sections, we will first convert the above measurement data (in absolute energy reading) into a binary sequence of 0s and 1s, through a thresholding process, with 0 denoting a channel being unused, idle or available, and 1 denoting the opposite (i.e., used, busy or unavailable).

[2]http://www.sharedspectrum.com/measurements, one taken in Maine, one in Chicago, and one in Ireland.

We begin by defining the following terms.

- Channel: a channel is an interval of radio frequency of bandwidth 200KHz. Channels are indexed sequentially; Channel X is the frequency interval $[(20 + 0.2(X - 1))$MHz, $(20 + 0.2X)$MHz$], X > 0$. Since 200KHz is the resolution of our measurement devices, a channel is the smallest unit at which we can distinguish energy.
- Service: a service is a set of channels that have been assigned to the same application/service, as listed in Table 2. Without ambiguity we will use this term to mean both the service and the set of channels assigned to the service.
- Channel state (CS): this is a function of time and channel. $CS(t, c) = 0$ indicates that Channel c is idle at time slot t, and $CS(t, c) = 1$ otherwise.
- Channel vacancy: this is the period in which a channel remains idle.

Converting energy level to CS (0 or 1) is essentially a binary hypothesis testing process. For lack of a priori knowledge on channel statistics, it is done in a simplistic way here through a thresholding procedure: for channel c, a threshold is set to be 3dB higher than the minimum value seen in this channel over the entire duration of the trace collected. At time slot t if the energy level is lower than this threshold, then $CS(t, c) = 0$; otherwise $CS(t, c) = 1$. The reason for this thresholding scheme is the following: Figure 2 shows the maximum, minimum and average energy level of some noise channels (those higher than 2GHz but below the ISM band) at Location 2. They are called noise channels because they are currently assigned to satellite-to-satellite communications (the signal does not reach the ground and thus the only energy present on the ground is due to noise). We see that for these channels, the maximum and minimum power levels are all within a 3dB range. Assuming that noise behaves similarly across channels (which is not exactly true, but probably close), anything more than 3dB above the minimum power level suggests the presence of signal, hence the above thresholding is reasonable. Decreasing this threshold will improve signal detection probability, but the false positive will become higher, while lowering the threshold increases false negative. [3] While an important subject in its own, calibrating the error in such a process is out of the scope of the present paper.

The result of this process is a sequence of CSs (0s and 1s) for each spectrum channel of

[3]We did try increasing this threshold from 3dB to 4.5 dB and found the resulting 0-1 sequence to be nearly the same. The same thresholding process was used in a measurement study conducted in Aachen, Germany [18].

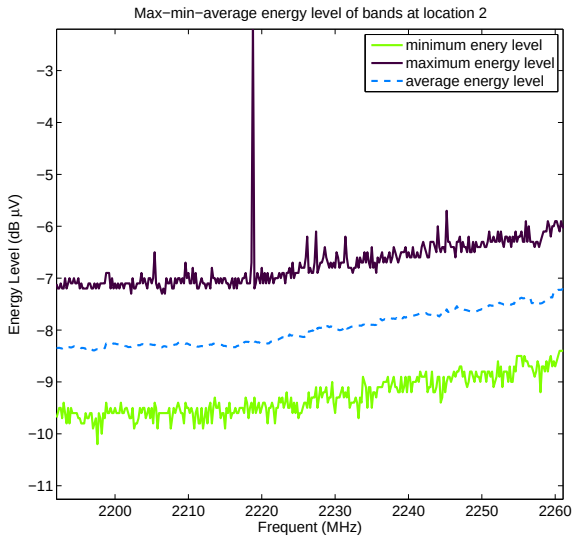

Fig. 2. The maximum, minimum and average energy levels of noise channels at Location 2.

200KHz wide, representing its availability over a time resolution of approximately 75 seconds. This is shown in Figure 3 where black dots indicate busy channels. In subsequent sections we will try to uncover the properties inherent in these sequences.

III. Statistical Properties of The Measurement Data

In this section, we perform comprehensive statistical analysis on the measurement data including channel vacancy durations (CVD), service congestion rate (SCR) series, as well as a subscriber model to explain channel utilization dynamics.

A. channel vacancy duration (CVD) distribution

To make better spectrum access decision, we are often interested in knowing how long a channel will remain idle. CVD is defined to capture this feature. In this section we show that our measurement data suggest that it has an exponential tail, but is not exactly an exponential distribution, nor is it independently distributed over time.

As defined earlier, channel vacancy is the period in which a channel remains idle. If we use the CS time series of a given channel, then the channel vacancy duration will always be a nonnegative integer (i.e., the number of consecutive 0's) since the CS is defined for discrete time slots. In reality, however, the channel state does not in general change at slot boundaries. In

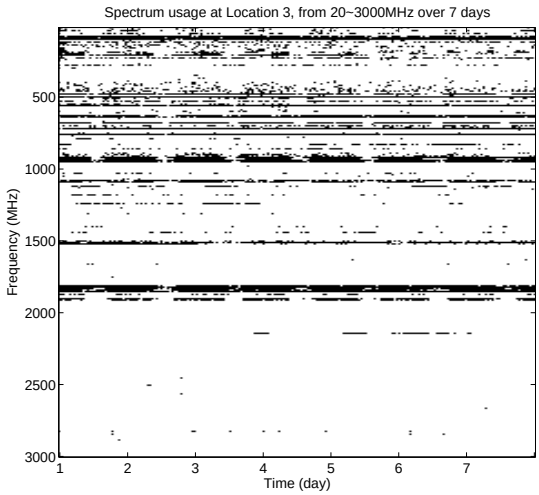

Fig. 3. CS map at Location 3

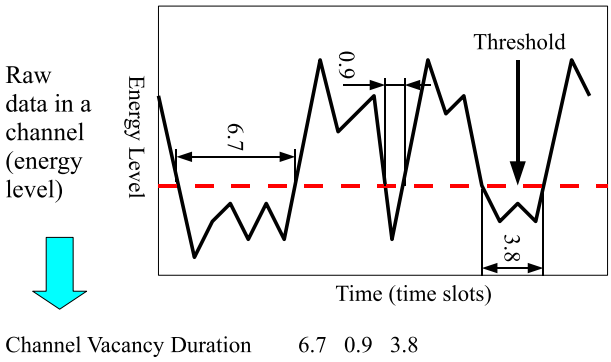

Fig. 4. Extract channel vacancy durations from raw data.

order to obtain a better, fractional estimate of the CVD, we use the original energy measurement data: By treating it as a continuous time signal and determining the threshold-crossing times, we can obtain CVD estimates in real numbers rather than integers. This is illustrated in Figure 4.

On average each channel CS time sequence (more than 8000 time slots) contains on the order of hundreds of channel vacancies. This sample size turns out to be too small to derive the CVD distribution. To increase this sample size we collect CVDs across all channels within the same service. This is done based on the observation that spectrum usages of channels within the same service are statistically very similar (shown in Section V.B). For example, spectrum

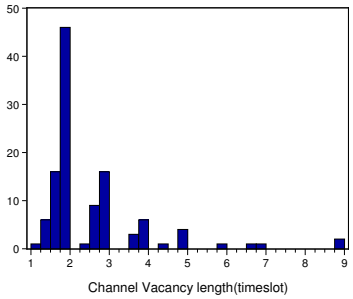

Fig. 5. Histogram of CVD distribution for GSM900 uplink at Location 2.

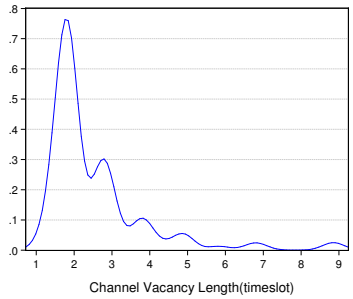

Fig. 6. Approximated PDF of CVD for GSM900 uplink at Location 2.

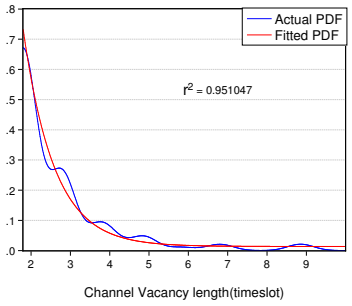

Fig. 7. Fitting curve of CVD for GSM900 uplink at Location 2

usages in channels within GSM900 uplink service (885-915MHz) are nearly the same in terms of occupancy, periodicity, average energy level, etc. This gives us enough samples to obtain the empirical distribution of channel vacancies.

Figure 5 shows the statistical histogram of the CVD in GSM900 uplink service at Location 2. Then we plot the Gaussian-kernelled density graph such that the curve of its probability density function (PDF) can be approximately restored as shown in Figure 6. Here we only focus on the falling part of the curve Figure 6, where CVD samples that exceeds two timeslot, because a longer vacancy period can provide access opportunities for a variety of services, no matter real-time and delay-sensitive services such as push-to-talk and video teleconference, or best-effort services such as file transferring. We then apply the least-square regression analysis to falling part of the curve in Figure 6. The significance of the fitting is measured by the coefficient of determination r^2, defined as:

$$r^2 \equiv 1 - \frac{\sum_i (y_i - \bar{y})^2}{\sum_i (y_i - f_i)^2} \tag{1}$$

where y_i is the sample value with mean $\bar{y}$ and f_i is the modelled/fitted value.

As shown in Figure 7, the CVD distribution is very well approximated ($r^2 > 0.95$) by an exponential-like distribution $y = a + be^{-cx}$. We repeated this exercise in all the services (GSM900 / 1800 uplink / downlink, broadcasting TV, CDMA, ISM) at all 4 locations and obtained similar results. The regression results at Location 2 are showed in Table III, where y denotes $\Pr[CVD = x]$ and $x = 1, 2, 3, ...C$ time slot(s), where C is a constant integer that indicates the maximal value of the CVD obtained from the data.

TABLE III

CHANNEL VACANCY DURATION DISTRIBUTION REGRESSION RESULTS AT LOCATION 2. REGRESSION EQUATION:

$$y = a + be^{-cx}$$

Service	a	b	c	r^2
GSM900 uplink	0.0134917	7.052599	1.268810	0.951047
GSM900 downlink	0.023809	0.710870	0.485794	0.986417
GSM1800 uplink	0.042653	0.884733	0.669833	0.983644
GSM1800 downlink	0.033259	1.058021	0.691488	0.993801
CDMA uplink	-0.016088	0.250563	0.140147	0.930726
CDMA downlink	0.022279	1.253164	0.723049	0.986838
TV1	0.027716	1.021404	0.599504	0.988016
TV2	0.039241	0.714230	0.548904	0.989958
TV3	0.042968	0.725089	0.599759	0.959616
TV4	0.046246	0.756806	0.631420	0.943959

It should be noted that $y = a + be^{-cx}$ has an exponential tail, but is *not* an exponential distribution. A direct consequence of this is that it does not have the memoryless property, i.e., how long a channel is going to remain in a certain state is a function of its history, rather than being independent of it. This latter independence assumption has been commonly used in channel access studies, see for example [15], [17], [20]. More precisely, these studies assume a two-state Markov chain model for the channel (i.e., an Eliot-Gilbert model). This channel model implies that the duration of channel vacancy are geometrically distributed (the discrete equivalent of an exponential distribution), and that these durations are independently distributed. Our results here indicate that such an assumption is inaccurate, and there is significantly more memory in the channel state information.

The above observations indicate that we will not be able to describe the CS series using a first order Markov model, whereby future state is independent of history given present state. To illustrate, we empirically obtain the following conditional probabilities for the GSM1800 uplink band at Location 2:

$$\Pr[CS(t+1,c) = 0 | CS(t,c) = 0] = 0.918953$$

while

$$\Pr[CS(t+1,c) = 0 | CS(t,c) = 0, CS(t-1,c) = 1] = 0.55454$$

Clearly the channel state is highly history dependent, a feature this type of Markov model fails to capture. We tried higher order Markov models, by defining a higher-dimensional state space (a higher-dimension state consists of a sequence of channel states, which results in an increase the state space), without much success. Besides, since CVD in our case is a statistic for the whole service rather than a single channel such that the whole set of CVD samples can not constitute a time series, we did not analyze CVD samples in any time-series way, such as autoregressive model fitting.

All the above indicate that the channel state information possess some far richer properties. Technically, any discrete system can be modeled as a Markov chain provided we embed sufficient memory into the state, but the resulting expansion in the state space is in general computationally prohibitive. In Section V we use a frequent pattern mining technique to get around this problem. This technique exploits the potential correlation in CSs and provides accurate prediction.

B. Service Congestion Rate (SCR)

The channel state $CS(t, c)$ is a microscopic level measure of the channel utilization – indeed a single time slot together with a single channel is the finest resolution we can obtain from the measurement data. Examining it over time for each channel results in CVD, a measure we examined in the previous section. In this section we will examine it across channels within the same service for a given time slot, a measure captured in the service congestion rate (SCR). We will then show how this measure evolves over time.

SCR for service S at time t, denoted by $SCR(t, S)$ is defined as the ratio between the number of busy channels in S at time t and the total number of channels in S. Thus $SCR(t, S) = \sum_{c \in S} CS(t, c)/n$, where n is number of channels in service S. SCR is thus a measure of the level of congestion in a service; the larger the SCR of a service is, the fewer idle channels there are. This is a value ranging from 0 to 1.

Figure 8 shows the SCR series of service GSM1800 uplink and GSM900 uplink at Location 2. We can see that the SCR series is cyclic in its outline with a period of one day, as expected. Also note that the SCR series of the two services are highly correlated: the rises and drops are very much in synchrony.

Of particular interest is the high SCR regions in Figure 8, i.e., those “plateau” regions. Within these regions, the SCR as a random process appears to be stationary. This turns out to be true: it

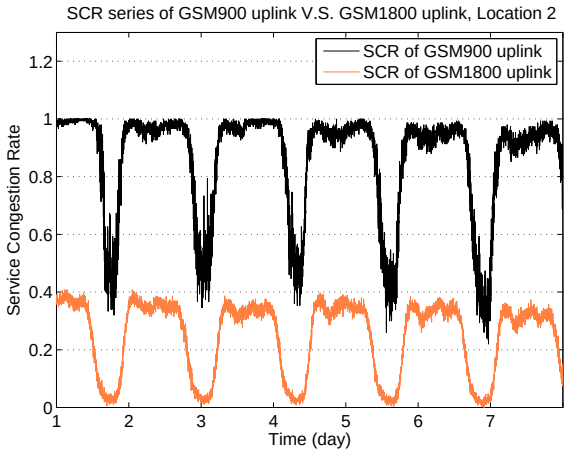

Fig. 8. SCR series at Location 2. GSM1800 uplink vs. GSM900 uplink.

passes augmented Dickey-Fuller (ADF) test with significance 0.01, verifying its stationarity [11] (This test was performed for each of the plateau regions). We further calculate the correlogram of the SCR series of service GSM900 uplink at Location 2 and results are shown in Figure 9 and 10. Here correlogram refers to autocorrelation function (ACF), denoted by $\rho_{t,s}$, and partial autocorrelation function (PACF), denoted by $\phi_{t,s}$ of a time series Y, which are respectively given by

$$\rho_{t,s} = Corr(Y_t, Y_{t-s})$$

$$\phi_{t,s} = Corr(Y_t, Y_{t-s}|Y_{t-1}, Y_{t-2}, ..., Y_{t-s+1})$$

where $Corr$ refers to the correlation coefficient.

We see that the correlation coefficient decreases at an negligible rate, while the partial correlation coefficient tails off rapidly. This suggests that the SCR series may be potentially well modeled as an autoregressive process [16]. The autoregressive model can be written as

$$SCR(t, S) = \sum_{m=1}^{O} c_m SCR(t - m, S) + n(t) \tag{2}$$

where O and $n(t)$ are the order and the residual of the model, respectively.

A third-order autoregressive model of the above form is applied to the SCR series of all

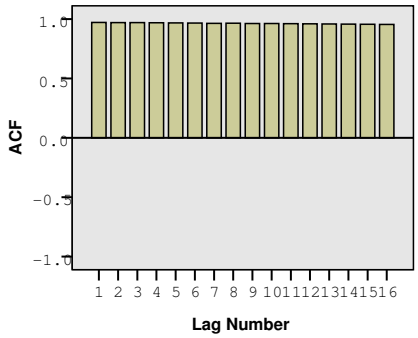

Fig. 9. Autocoorelation coefficient of SCR of GSM900 uplink at Location 2.

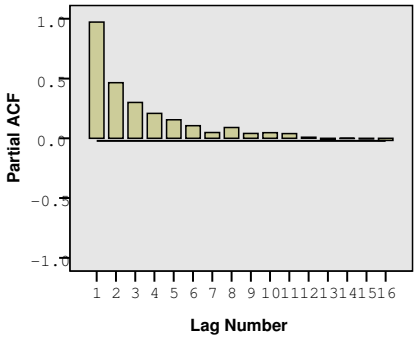

Fig. 10. Partial autocoorelation coefficient of SCR of GSM900 uplink at Location 2.

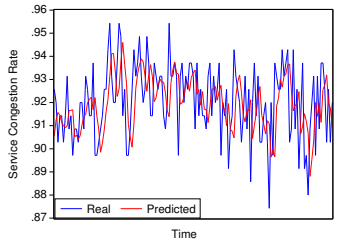

Fig. 11. SCR prediction with 3rd-autoregressive model.

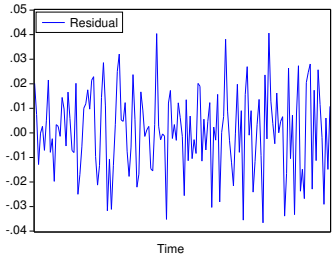

Fig. 12. SCR prediction residual with 3rd-order autoregressive model.

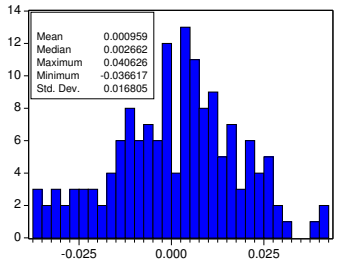

Fig. 13. Histogram of prediction residual with 3rd-autoregressive model.

services at Location 2[4]. The regression results are shown in Table IV. We did not include the regression results of CDMA service as the existence of CDMA signal cannot be verified simply using energy detection [4]. We see that in all cases the 3rd-order autoregressive model achieved very high accuracy, as indicated by the coefficients of determination shown in 1.

Since autoregressive model can approximate the SCR series quite well, we can leverage it into temporal prediction for SCR series, that is, to use its past to predict its future. Therefore, again we use 3rd-order autoregressive model to make such prediction. For illustration purpose, we only show the prediction results for the service of GSM900 uplink at Location 2 as shown in Figure 11, results for other services are quite similar. We also show the residual of the prediction

[4]Regarding selecting the right order for this model: a higher order generally results in better approximate; on the other hand, the order cannot be arbitrily high based on Akaike Information Criterion, in which a high order will improve regression quality but will also lead to the reduction of degree of freedom [1]. The choice of 3rd-order in our experiment seems to be appropriate; in particular we were not able to obtain significant fitting improvement by increasing the order beyond 3.

TABLE IV

3RD-ORDER AUTOGRESSIVE MODEL REGRESSION RESULTS AT LOCATION 2.

Service	c_1	c_2	c_3	r^2
GSM900 uplink	0.380212	0.313529	0.305352	0.960674
GSM900 downlink	0.354719	0.322443	0.322548	0.970801
GSM1800 uplink	0.392203	0.319233	0.286660	0.974579
GSM1800 downlink	0.407140	0.312867	0.279747	0.973845
TV1	0.476766	0.285754	0.233663	0.906869
TV2	0.424009	0.302502	0.270873	0.861388
TV3	0.457287	0.332248	0.208510	0.963755
TV4	0.409256	0.336329	0.254212	0.921405

model and its statistical histogram as shown in Figure 12 and 13 respectively. The probability distribution of the residual passes the Anderson-Darling hypo-test with confidence level 0.1 such that it is testified to comply with a normal distribution with zero mean, which indicates that the autoregressive prediction model is quite adequate for the SCR series [16].

C. Dynamic Utilization of GSM Services

Of all the services listed in Table II, mobile services have the special feature of very high dynamic utilization. This is because they are subscriber based and the level of energy present in these channels are driven by the arrivals and departures of subscribers (or rather, their calls). Understanding how the channel utilization changes over time as a function of the subscriber dynamics can be very helpful in marketing and operating strategies including pricing and promotions. In this section we show that the channel utilization given by the measurement data can be well described using a simple queuing model. Due to the difficulty in energy-detecting CDMA signal as mentioned before, we will only consider GSM services in this section.

To this end, we note that the received energy at a measurement location is the summation of all transmitted signals, whose strength gets attenuated over the propagation distance. While this means that different users/callers will have a different contribution to this sum depending on where they are located, if we assume that they are uniformly distributed in space, one would expect that the sum of received energy changes proportional to the change in the active caller population within an applicable region (i.e., the region in which a transmitted signal results in non-negligible reception at the measurement location). Below we will specifically consider the

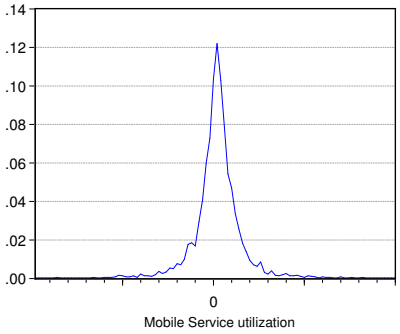

Fig. 14. Distribution curve of MSU of GSM900 uplink at Location 3.

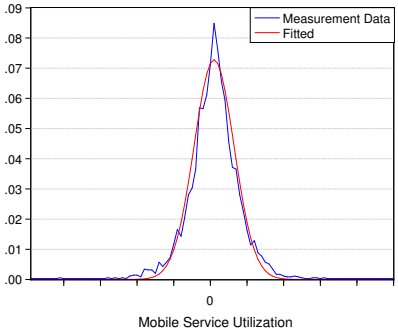

Fig. 15. Fitting result of MSU of GSM900 uplink at Location 4.

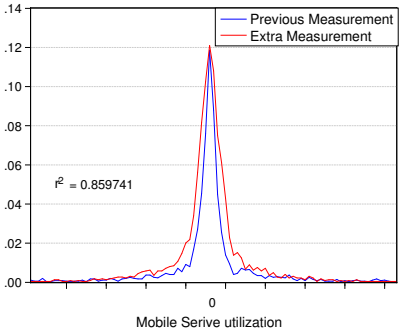

Fig. 16. Comparison of MSU of GSM900 uplink at Location 1 between two measurements.

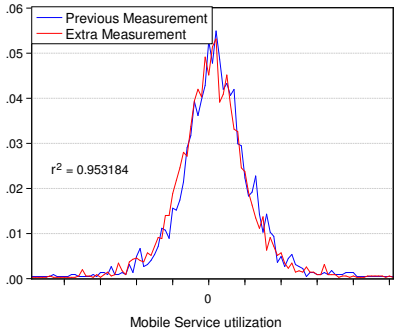

Fig. 17. Comparison of MSU of GSM1800 uplink at Location 1 between two measurements.

GSM uplink service while excluding downlink services as the GSM downlink traffic contains not only user traffic but also that of signaling applications.

We define the mobile service utilization (MSU) of a GSM service S, denoted by $MSU(t, S)$, as the time-varying difference in the total received energy, expressed as

$$MSU(t, S) = \sum_{c \in S} e(t, c) - \sum_{c \in S} e(t-1, c).$$

Figure 14 shows the probability distribution of MSU series of GSM900 uplink at Location 3 over a week. For proprietary reasons we do not show the actual value of $MSU(t, S)$.

We next show that this distribution can be well described using a simple queuing model. Suppose that in each time slot the arrival and departure of calls each follows a Poisson distribution with rate λ_a and λ_h, respectively. Assume further that they are independently distributed (note

that this is equivalent of saying that the call durations are exponentially distributed). Denote the number of active callers at the beginning of time slot t by $N(t)$. Then we have $N(t+1) = N(t) + A(t) - D(t)$, where $A(t)$ ($D(t)$) is the arrival (departure) within time slot t. Then the change in the number of active callers in the system, denoted by $X(t) = N(t+1) - N(t)$, is given by

$$P(X(t)=m) = \begin{cases} e^{-(\lambda_a+\lambda_h)}\lambda_a^m \sum_{k=1}^{\infty} \frac{\lambda_a^k \lambda_h^k}{k!(k+m)!} & m \geq 0 \\ e^{-(\lambda_a+\lambda_h)}\lambda_h^{-m} \sum_{k=1}^{\infty} \frac{\lambda_a^k \lambda_h^k}{k!(k-m)!} & m < 0 \ . \end{cases} \tag{3}$$

The probability distribution curve in Figure 14 is centered around zero, suggesting that in steady state λ_a and λ_h are approximately equal. Equation 3 can therefore be simplified to

$$P(X=m) = \begin{cases} e^{-2\lambda}\lambda^m \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{k!(k+m)!} & m \geq 0 \\ e^{-2\lambda}\lambda^{-m} \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{k!(k-m)!} & m < 0 \end{cases} \tag{4}$$

We then compare Eqn (4) with the distribution curves of MSU of the GSM uplink services at four locations. For brevity, we only show the result of GSM900 uplink at Location 4 as shown in Figure 15 while noting the other results are similar. As shown in Figure 15, the two are very close, and the fitting coefficient of determination is over 0.97.

We also examined whether the above utilization dynamics change significantly over time. Figure 16 and 17 show the MSU distributions obtained using measurement data collected over two different weeks, for the GSM900 uplink service and the GSM1800 uplink service, respectively. In both cases we see that this distribution stay roughly the same from week to week, indicating a rather steady collective calling pattern in terms of arrival and departure statistics.

IV. Spectral and Spatial Correlations

The more we know about spectrum usage characteristics (of the primary users), the better we can predict spectrum opportunity, and the better we can make dynamic spectrum sensing and access decisions. Much of this predictive power lies in the spectral and spatial dependence of spectrum usage. For instance, if everything is independently distributed, then knowing the past does not offer information for the future. On the other hand, Figure 8 shown in the previous section suggests that measuring/sensing channels in GSM900 uplink provides ample information about channel availability in GSM1800 uplink. We thus set out to take a more in-depth look at

the dependence characteristics of our data sets in this section. Specifically, we will analyze the spectral, and spatial correlation of the CS series and the SCR series, respectively.

We will use the following two measures of correlation, the first one defined for two random variables and the second one defined for two 0-1 random sequences. The correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with sample mean μ_X and μ_Y and sample standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} \tag{5}$$

where cov is the covariance operator. This coefficient ranges between -1 and 1, extreme values indicating X and Y are (inversely) fully correlated. In general correlation is considered high (i.e., one random variable proving a lot of information about the other) when the absolute value of the coefficient is closed to 1.

The correlation between two discrete-time 0-1 series $X(t)$ and $Y(t)$ are defined as follows:

$$Corr_{X(t),Y(t)} = \frac{\sum_t I\{X(t) = Y(t)\} - \sum_t I\{X(t) \neq Y(t)\}}{\sum_t I\{X(t) = Y(t)\} + \sum_t I\{X(t) \neq Y(t)\}} \tag{6}$$

where $I\{A\}$ is the indicator function: $I\{A\} = 1$ if A is true and 0 otherwise. The two summations in the above equation are the total number of positions that the two sequences coincide and differ, respectively. This is commonly used for evaluating the correlation between two binary sequences.

A. Spectral Correlation

In order for investigation on spectral correlation, we take the SCR and CS series and cross correlate them with their counterparts from a different service and different channel within the same service, using Eqn (5) and Eqn (6), respectively.

Figure 18 shows the CS correlation coefficients between every two channels in the GSM900 uplink at Location 4. We see that these coefficients are extremely high for almost every two channels within the same service. This is because in most cases the channels are either all busy or all idle. There are also cases where the spectral correlation coefficients are closed to -1. This is because some channels are always idle while some others are always busy. For instance, there are channels in GSM that are kept idle to avoid inter-channel interference. These results are representative of what we found in other services.

Table V shows the spectral correlation coefficients between two SCR series at Location 1; these results are also representative of what we found at the other locations.

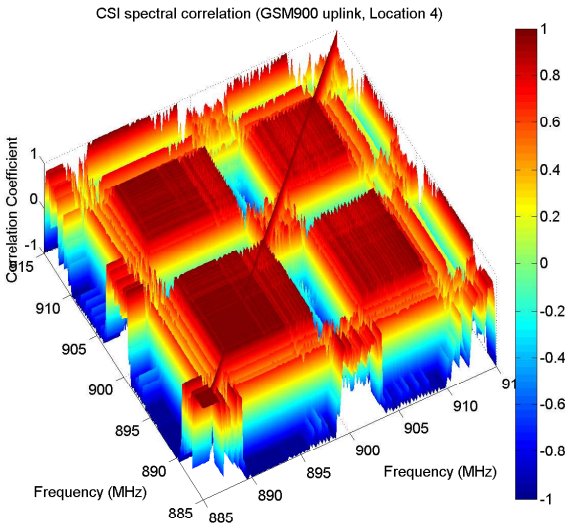

Fig. 18. Spectral Correlation Coefficients of CS series within GSM900 uplink at Location 4.

From the results shown in Table V, we see that there is significant correlation among these services, for none of the coefficients falls below 0.55, even when between a broadcasting TV service and a GSM service. In addition, services of the same type are particularly correlated, as high as 0.952 in the case between GSM900 uplink and GSM1800 uplink.

B. Spatial Correlation

Figure 19 shows 4 SCR series of the same service (GSM900 downlink) at all four locations. At a high level, these series appear all correlated with each other; they share common changes in value. In particular, Locations 1 and 2 are very similar, up to a constant shift, and Locations 3 and 4 are very similar, also up to a constant shift. This suggests spatial correlation across different locations. We thus cross correlate SCR / CS series from the same service / channel at different locations.

Table VI shows the spatial correlation coefficients of the SCR series in GSM900 downlink service among four locations. These are very high and none of the values falls below 0.8. It appears that within the same service, the spectrum utilization is highly correlated across different locations and different types of locations.

For other services the results are quite similar and are thus not presented separately. For instance between Locations 1 and 2, the spatial correlation coefficient of SCR series is as high

TABLE V

SPECTRAL CORRELATION COEFFICIENTS OF SCR AT LOCATION 1.

	GSM900 uplink	GSM900 downlink	GSM1800 uplink	GSM1800 downlink	TV1	TV2	TV3	TV4
GSM900 uplink	1.000							
GSM900 downlink	0.873	1.000						
GSM1800 uplink	0.952	0.832	1.000					
GSM1800 downlink	0.855	0.747	0.827	1.000				
TV1	0.674	0.616	0.713	0.636	1.000			
TV2	0.730	0.669	0.700	0.690	0.634	1.000		
TV3	0.789	0.742	0.809	0.710	0.833	0.711	1.000	
TV4	0.588	0.581	0.557	0.566	0.655	0.567	0.721	1.000

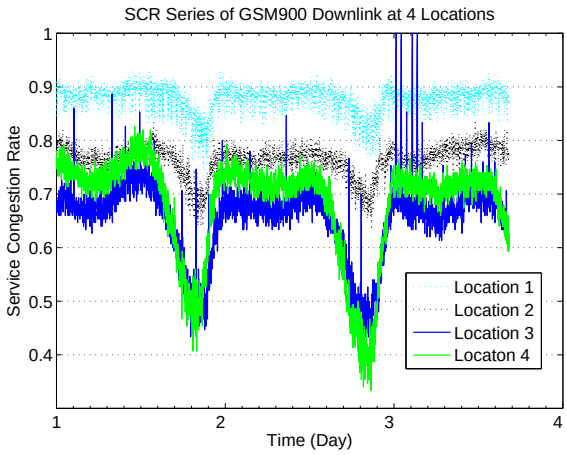

Fig. 19. SCR Series of GSM900 Downlink at 4 Locations.

as 0.962.

It's easy to understand the reason behind such high correlation: for the same service, such as GSM voice calls, subscribers at different locations share common behavioral patterns (e.g., same peak calling hours of the day), even though the actual SCR values are different.

V. PREDICTION USING FREQUENT PATTERN MINING (FPM)

The correlation structure presented in the previous section suggests that it can be exploited to help us better predict channel state or spectrum opportunity based on measurements made in

the past, in adjacent channels, or at similar locations. More precisely, we are interested in the question of whether one could accurately predict the value of $CS(t,c)$ based on the knowledge of $CS(t-k,c'), k>0$, and if so how many past observations (what values of k) and over what set of channels (what values of c') are needed.

There are different approaches one could take to model such correlation. For instance, as pointed out in Section III, the memory in channel state information could conceptually be captured by a sufficiently high-ordered Markov chain (and one can use measurement data to collect statistics on the transition probabilities of this chain), but with significant technical difficulty due to the exponential increase in the state space.

To overcome this difficulty, in this section we present a technique based on frequent pattern mining (FPM) ([3], [14] and a survey [5]) through an efficient pattern identification process over the spectrum data. This technique generates predictions of future channel state based on a collection of past observation in a set of channels. It uses both temporal and spectral correlations examined in earlier sections. A unique feature of this approach is that it automatically adjusts the algorithm to the appropriate size of past observations (both in time and in channels) based on the data set. This method along with our experimental results are detailed in the remainder of this section.

A. FPM and Prediction

We begin by illustrating how FPM can be used for spectrum usage prediction and the challenges in doing so. Suppose we know the CSs of Channels $c, c+1$ of previous 8000 time slots (from time slot $t-7999$ to t) and we would like to predict the CS of the next time slot of Channel c and $c+1$ (i.e., $CS(t+1,c)$ and $CS(t+1,c+1)$). The known CSs can be written into a single matrix shown below:

$$\begin{bmatrix} a_{t-7999,c} & a_{t-7998,c} & \dots & a_{t,c} \\ a_{t-7999,c+1} & a_{t-7998,c+1} & \dots & a_{t,c+1} \end{bmatrix},$$

where $a_{i,j} = CS(i,j)$.

Define a submatrix as a pattern if it appears no less than 200 times throughout the CS series of these channels. For instance, if the submatrix $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ appears 1000 times, it's considered a

TABLE VI

SCR SPATIAL CORRELATION COEFFICIENTS FOR GSM900 DOWNLINK AMONG 4 LOCATIONS

	Loc 1	Loc 2	Loc 3	Loc 4
Loc 1	1.000			
Loc 2	0.833	1.000		
Loc 3	0.858	0.846	1.000	
Loc 4	0.854	0.880	0.909	1.000

pattern. We may find another pattern $\begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$ which appears 990 times. Comparing these two patterns we can predict that $CS(t+1,c)=1$ and $CS(t+1,c+1)=0$ with probability 99% (990/1000) if $CS(t-2,c)=1, CS(t-1,c)=0, CS(t,c)=1, CS(t-2,c+1)=1, CS(t-1,c+1)=1$, and $CS(t,c+1)=0$.

Clearly this prediction method needs to successfully handle two issues: the first is to find frequent patterns, referred to as frequent pattern mining. The second issue is to find associations among these patterns, referred to as pattern association rules mining.

In our setting the dimension (number of rows and columns) of the patterns of interest are not fixed in advance, i.e., we do not know in advance how much history and how many neighboring channels are needed in order to have accurate prediction. Rather, this has to be learned during the mining process. This 2-D (of patterns) learning element is a unique challenge in our mining process, compared to existing FPM literature, e.g., [3], [5], [14]. In addition, in all these studies the patterns are 1-D and can be written in a row, although the data sets can be in multiple rows. In our problem, the patterns are in 2-D, which is another unique challenge. To summarize, both the width and the height of the patterns are variable, and their appropriate sizes need to be automatically identified in this process.

In the following we will refer to our problem as a FPM-2D problem.

B. FPM-2D Problem Definition

The goal of the FPM-2D problem is to find all relevant 2D patterns. Once this is achieved, it is fairly easy to compute the probabilities of future channel state (spectrum prediction). Table VII contains a list of terminologies used in FPM-2D.

Formally, the FPM-2D problem is stated as follows: Given the input set $\mathbb{M}$, min_area, and min_rep, find all valid block patterns and the corresponding match numbers.

TABLE VII

NOTATIONS / CONCEPTS IN FPM-2D

Notation / Concept	Definition
$\mathbf{\Gamma}$	The literals set. $\mathbf{\Gamma} = \{0, 1\}$
$\mathbb{M}$	An input matrix $\mathbb{M}_{m\times n} = \begin{bmatrix} a_{1,1} & a_{1,2} & ... & a_{1,n} \\ a_{2,1} & a_{2,1} & .. & a_{2,n} \\ ... & & & \\ a_{m,1} & a_{m,2} & .. & a_{m,n} \end{bmatrix}$ where $a_{i,j} \in \mathbf{\Gamma}$, m: the number of adjacent channels, n: the number of consecutive time slots
block	a submatrix of $\mathbb{M}$
subblock	a submatrix of a block
block area	the number of elements in a block
block pattern	a block whose area $\geq min\_area$. "pattern" and "block pattern" are used interchangeably.
subpattern	submatrix of a pattern
matches (supports)	If a pattern $\mathbf{P}$ and a block $\mathbf{A}$ are identical, we say $\mathbf{A}$ is a match (support) of $\mathbf{P}$, or $\mathbf{A}$ matches (supports) $\mathbf{P}$
match number	the number of ALL matches of a pattern
valid	a pattern is valid if its match number is no less than min_rep

C. Proposed Algorithm

We start by scanning the input matrix $\mathbb{M}$ from left to right, top to bottom to find all the blocks with size $x \times y$. This is done for all x and y such that $x \times y \geq min\_area$. We use a hash table to store the blocks for efficiently search, since it takes $O(1)$ time to search an item in a hash table. A potential problem is the number of blocks might be too large; it is $2^{x\times y}$ in the worst case. This problem is addressed by the following simple property.

Consider a block $\mathbf{A}_{x\times y}$. We say block $\mathbf{B}$ is $\mathbf{A}_{x\times y}$'s *parent block* if: a) $\mathbf{B}$ is $\mathbf{A}_{x\times y}$'s subblock, b) $\mathbf{B}$'s size is $(x-1)\times y$ or $x \times (y-1)$, and c) $\mathbf{B}$'s area is not less than min_area. A simple yet key property concerning a valid pattern is that for any block $\mathbf{A}_{x\times y}$, it has at most 4 parent blocks and it is a valid pattern only if all its parent blocks are valid patterns.

Thus, if any block has a parent block that is not valid, then we can simply skip this block

Algorithm 1 FPM-2D

```
for N=2 to ∞:
    flag ← false
    for each (x, y) s.t. (x + y = N, x × y ≥ min_area):
        T_{x,y} ← new empty hash table
        if T_{x−1,y} and T_{x,y−1} are both empty:
            continue;
        for each block B_{x×y} in M:
            if one of B_{x×y}'s parent blocks
is not a valid pattern:
                continue
            if B_{x×y} is in T_{x,y}:
                T_{x,y}[B_{x×y}] ← T_{x,y}[B_{x×y}] + 1
            else:
                T_{x,y}[B_{x×y}] ← 1
        for each block P_{x×y} in T_{x,y}:
            if T_{x,y}[P_{x×y}] < min_rep:
                remove P_{x×y} from T_{x,y}
        if T_{x,y} is not empty:
            flag ← true
            output the valid patterns in T_{x,y} and
            the corresponding match numbers
    if flag == false:
        break
```

because itself cannot be valid. By checking parent blocks, a large number of blocks can be ignored, which significantly reduces the memory consumption and improves the performance.

The pseudo code of the proposed algorithm is given in Algorithm 1. In Algorithm 1, $\mathbf{T}_{x,y}$ is the hash table to store patterns with size $x \times y$.

After all valid patterns have been identified, the prediction rules are extracted as follows. A prediction rule is defined as $\mathbf{P}_1 \rightarrow \mathbf{P}_2$, where $\mathbf{P}_1$ and $\mathbf{P}_2$ are all valid patterns. $\mathbf{P}_2$ has the form $[\mathbf{P}_1\ \mathbf{V}]$, where $\mathbf{V}$ is a column vector $(v_1, v_2, v_3, ...v_k)^T, v_i \in \mathbf{\Gamma}$. Thus $\mathbf{P}_1$ is one of the parent blocks of $\mathbf{P}_2$. Let $M(\mathbf{P}_1)$ denote the match number of $\mathbf{P}_1$, then the *transferring rate* $R(\mathbf{P}_1 \rightarrow \mathbf{P}_2) = M(\mathbf{P}_2)/M(\mathbf{P}_1)$ What this rule says is that if the current CS appears to match $\mathbf{P}_1$, then the CS in the next time slot will match $\mathbf{V}$ with probability $R(\mathbf{P}_1 \rightarrow \mathbf{P}_2)$. Clearly, a similar procedure can be used to predict the CS over multiple future slots, by simply considering $\mathbf{V}$ as multiple column vectors.

D. Experiment Result

To test our algorithm we split the measurement data into two part, one as a training set on which we run Algorithm 1 and find the prediction rules, while the other as the testing set on which we apply the prediction rules and perform spectrum usage prediction. We set $min\_rep = 200, min\_area = 4$.

In our experiment we adopted the following prediction methods: a) We only consider those rules whose transferring rates are larger than 0.9. b) If the current CSs appears to match the pattern $\mathbf{P}1$ in a rule $\mathbf{P}1 \rightarrow \mathbf{P}2$, we predict the CSs matches $\mathbf{P}2$ in the next time slot. c) If the current CSs do not match any patterns, we do not predict and count it as a miss.

We are interested in answering the following questions:

1) what is the prediction accuracy if the training set and corresponding testing set are from the same service (self-service prediction)?
2) what is the missing rate of the prediction?
3) can the mining result in one service be used to predict the CSs in another service (cross-service prediction)?
4) can the mining result in a service at one location be used to predict the CSs in the same service at a different location (cross-location prediction)?
5) how large the training set needs to be for accurate prediction?

We define prediction accuracy to be the ratio between the number of correctly predicted CSs and the total number of predicted CSs, and define missing rate as the ratio between the number of CSs that cannot be predicted and the total number of CSs.

We first study the case where the training set and the corresponding testing set are from the same service at Location 1, i.e., both from TV1 or both from GSM900 uplink. The training sets are CS series over durations from 1 hour to 3 days. The testing sets are CSs of the last 4 days. The results are shown in Figure 20. We observe that:

- The prediction accuracy is larger than 0.95, a very encouraging sign.
- The missing rate is around 5% for TV1 service, which is very low. It is higher, around 15% for GSM900 uplink. The reason why the missing rate on GSM900 uplink is higher than TV1 is that TV service is a pure broadcast service, while GSM service is an interactive service whose patterns are more complicated to match.

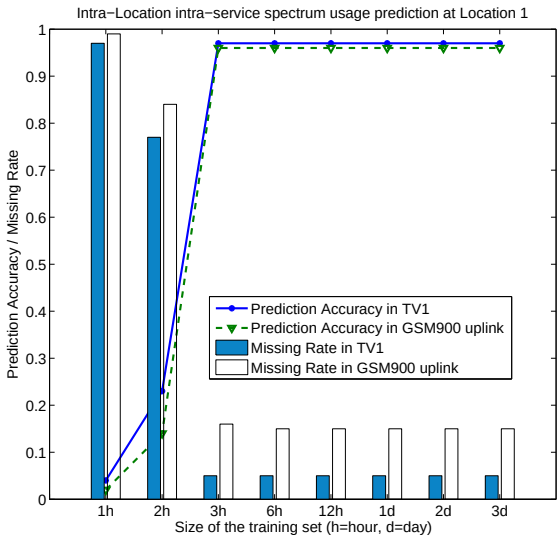

Fig. 20. The experiment results of intra-location intra-service spectrum usage prediction at Location 1. The prediction accuracy is larger than 0.95 if the training set size is no less than 3 hours.

- The training set cannot be less than 2 hours. Otherwise FPM-2D cannot find sufficient prediction rules.
- A 3-hour training set appears sufficient and the performance of the algorithm saturates at this level. Beyond this threshold, more training data does not seem to help improve the prediction accuracy or reduce the missing rate.

For other services of Location 1, the results are similar except for the missing rate, which is listed in Table VIII. We see that the prediction accuracy is consistently high but the missing rate varies from 4% up to 25%. We also give the overall occupancy of the services as a reference. If the overall occupancy is a, then the prediction does not help if its accuracy is lower than $\max\{a, 1-a\}$. This is because we can always achieve this accuracy simply by guessing.

For comparison, we have also shown the prediction accuracy using the 1st-order Markov Chain model (1st MC), which is one of the most commonly used models in the existing papers [15], [17], [20]. We could see that the prediction accuracy of the 1st MC is only around 80%, which is much lower than that of FPM-2D, except for those services whose occupancy is high. In these latter cases the prediction accuracy can be naturally high even by guessing.

The results of the other locations are similar, and thus not repeated.

TABLE VIII

THE SPECTRUM USAGE PREDICTION RESULTS AT LOCATION 1. THE TRAINING / TESTING SETS ARE FROM THE SAME SERVICE.

Service	occu-pancy	1-st order Markov Prediction Accuracy	Miss rate	FPM-2D Prediction Accuracy
GSM900 downlink	85.1%	85.2%	11.8%	96.9%
GSM1800 uplink	60.3%	77.4%	24.8%	95.1%
GSM1800 downlink	30.2%	83.7%	16.2%	96.6%
TV2	92.1%	92.6%	5.4%	96.9%
TV3	44.5%	75.0%	4.2%	97.8%
TV4	41.9%	74.5%	6.3%	97.7%

TABLE IX

THE SPECTRUM USAGE PREDICTION RESULTS AT LOCATION 1. THE TRAINING / TESTING SETS ARE FROM DIFFERENT SERVICES.

Training Set	Testing Set	Miss Rate	Prediction Accuracy
GSM900 downlink	TV1	66.1%	79.2%
TV1	GSM900 uplink	75.3%	80.4%
GSM900 uplink	GSM900 downlink	35.2%	86.4%
GSM900 downlink	GSM900 uplink	31.8%	87.4%

We next study the case where the training set (CSs of 3 days) and the testing set (CSs of 4 days) are from different services. The results are shown in Table IX.

We see that the accuracy of cross-service prediction is much lower than that of self-service prediction, and the missing rate is quite high. This is because the patterns in different services collide, i.e. patterns and prediction rules found in one service might lead to wrong prediction results in other services. This is another manifestation of the conclusion drawn earlier that the spectral correlation is high for channels within the same service but low across different services.

Consistent with earlier correlation result, we also observe that the prediction accuracy is high

TABLE X

THE CROSS-LOCATION SPECTRUM USAGE PREDICTION RESULTS.

Service	Training Set	Testing Set	Miss Rate	Prediction Accuracy
TV1	Location 1	Location 2	5.7%	95.3%
TV1	Location 3	Location 4	7.3%	97.4%
TV1	Location 1	Location 3	6.5%	96.7%
TV1	Location 3	Location 1	7.7%	95.8%

if we use the prediction rules at one location to predict the CSs of the same service at another location, due to the high spatial correlation of channels within the same service. Table X Shows the experiment results of this cross-location self-service prediction.

To summarize this section, we conclude that:

1) The self-service self-location spectrum usage prediction accuracy is higher than 95%, which is significantly larger than that of the commonly used 1st-order Markov model.
2) The missing rate varies from 4% to 25%, an overall acceptable range.
3) The cross-service prediction accuracy ranges from 60% to 80%, much lower than the self-service prediction. The corresponding missing rate is above 30%, sometimes over 70%, which is too high.
4) The accuracy and missing rate of cross-location self-service prediction are nearly as high as that of self-location self-service prediction.
5) CSs of 3 hours appear sufficient for training purpose.

VI. DISCUSSIONS

In this section we briefly summarize how findings and results presented in previous sections may be used toward both the theory and practice in spectrum sensing and access within the context of cognitive radio networks.

Broadly speaking, these results contribute to two aspects of dynamic spectrum access: (1) the construction of better channel models (as a way of describing the spectrum usage of the primary users), that may be more generally applicable in other environments, and (2) the prediction of channel availability in a similar wireless spectrum environment.

Channel models are an essential component in an array of spectrum access studies, especially theoretical analysis. Our study has shown certain weaknesses of existing channel models (e.g.,

insufficient capture of history-dependence, inability to describe spectral and spatial dependence). Our results here can help build better channel models that more accurately reflect the primary users' activity. In particular, the statistics we collected on channel occupancy/vacancy, its rich dependency property, as well as the statistics on the CS series may motivate the construction of certain type of discrete event models (e.g., a Petri net) to describe the channel behavior. This may allow us to model the memory structure as well as the spatial and spectral correlation while avoiding a large state space.

The frequent pattern mining technique introduced here can be a very powerful tool in analyzing spectrum usage data. For specific wireless environments where such data are available for training, we have shown that using this technique can generate very accurate predictions on channel availability (especially in the TV broadcast channels in our study). This allows a secondary user to make far better channel sensing and access decisions, and exploit much more effectively under-utilized spectrum opportunity.

VII. Related Works

The Shared Spectrum Company (SSC) performed extensive spectrum measurements at 7 locations in the US and one outside the US between 2004 to 2007 [12], [13]. These measurements are all wide-band and over long periods of time. For instance, the measurement in Chicago scanned the energy level from 30MHz to 2900MHz and lasted 46 hours. The goal of these measurements is to gain a better understanding of the actual utilization of spectrum in rural and urban environments. To achieve this, the authors set two fixed thresholds for channels in higher and lower frequency bands, respectively, and considered a channel busy if the power level is above the corresponding threshold, and idle otherwise. According to their reports, among those 7 locations in the US, the lowest average occupancy is 1% at Greenbank, West Virginia while the highest is 17.4% at Chicago, Illinois.

In addition to SSC, there have been a few similar measurement studies recently. In 2008, Institute for Infocomm Research (I^2R) published their spectrum measurement results in Singapore [8]. They scanned from 80MHz to 5850MHz for 12 weekdays. They found the average occupancy to be 4.54% and most of the allocated frequencies were heavily underutilized except the TV broadcast channels and cell phone channels. Similar results were reported from Auckland, New Zealand according to [2].

The above cited work primarily focused on collecting statistics on the average utilization of wireless spectrum, and they all confirmed that it is indeed heavily underutilized. Correlations in the temporal, spectral and spatial dimensions were not a focus in these studies.

There has also been work in exploring correlations. A spectrum measurement was carried out during the 2006 Football World Cup in Germany, in the cities of Kaiserslautern and Dortmund [7]. They found that the change of spectrum usage (energy level) was clearly related to specific events (football match). Moreover they investigated the autocorrelation structure of changes in energy levels. Later in 2007 another measurement was conducted in the US on the public safety band (around 800MHz) [9]. The authors collected data concurrently at two locations, with a total of 5 pairs of locations with distance ranging from 5 meters to a few kilometers. They investigated the adjacent channel interference and spatial correlation. They revealed that very different energy detection results were obtained at closely located detection stations (5 meters apart); this was attributed to the difference in sensitivity in the sensing devices used.

Compared to this set of studies, our analysis also explored spectral correlation, both within the same service and across services. The service congestion rate (SCR) is a unique notion that allows us to examine spectrum usage service by service. In addition, our measurement involves the most concurrent locations (4), is over a fairly long duration (7 days), and scans from 20MHz to 3GHz. This allowed us to conduct a very detailed analysis on both the first and second order statistics of these data sets.

VIII. Conclusions

In this paper we carried out a set of spectrum measurements in the 20MHz to 3GHz spectrum band at 4 locations concurrently in Guangdong province of China. Using these data sets we conducted a set of detailed analysis of the first and second order statistics of the collected data, including channel occupancy / vacancy statistics, channel utilization within each individual wireless service, also spectral and spatial correlation of these measures. Moreover, we also utilized such spectrum correlation to develop a 2-dimensional frequent pattern mining algorithm that can accurately predict channel availability based on past observations.

We believe our findings will spur more discussions on how to better model current spectrum utilization and help us design more efficient spectrum sensing and accessing schemes for secondary users.

IX. Acknowledgments

The research was support in part by grants from RGC under the contracts CERG 622508 and N_HKUST609/07, the grant from Huawei-HKUST joint lab, and the NSFC oversea Young Investigator grant under Grant No. 60629203, and the National Natural Science Foundation of China under Grant No. 60632030. Meanwhile, we would like to show our gratitude to the Radio Monitoring Center of Guangzhou, Jiangmen and Zhongshan for providing us measurement devices and helping us collect the measurement data.

References

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974.

[2] R. Chiang, G. Rowe, and K. Sowerby. A Quantitative Analysis of Spectral Occupancy Measurements for Cognitive Radio. *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, pages 3016–3020, April 2007.

[3] G. Cong, K.-L. Tan, A. Tung, and F. Pan. Mining frequent closed patterns in microarray data. *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, pages 363–366, Nov. 2004.

[4] W. Gardner and S. CA. *Cyclostationarity in communications and signal processing*. IEEE press New York, 1994.

[5] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[6] S. Haykin. Cognitive Radio: Brain-Empowered Wireless Communications. *IEEE Journal on Selected Areas of Communications (JSAC)*, 23(2):201–220, February 2005.

[7] O. Holland, P. Cordier, M. Muck, L. Mazet, C. Klock, and T. Renk. Spectrum Power Measurements in 2G and 3G Cellular Phone Bands During the 2006 Football World Cup in Germany. *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, pages 575–578, April 2007.

[8] M. Islam, C. Koh, S. Oh, X. Qing, Y. Lai, C. Wang, Y.-C. Liang, B. Toh, F. Chin, G. Tan, and W. Toh. Spectrum Survey in Singapore: Occupancy Measurements and Analyses. *Cognitive Radio Oriented Wireless Networks and Communications, 2008. CrownCom 2008. 3rd International Conference on*, pages 1–7, May 2008.

[9] S. Jones, E. Jung, X. Liu, N. Merheb, and I.-J. Wang. Characterization of Spectrum Activities in the U.S. Public Safety Band for Opportunistic Spectrum Access. *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, pages 137–146, April 2007.

[10] J. Kennedy and M. Sullivan. Direction Finding and "Smart Antennas" Using Software Radio Architechtures. *IEEE Communications Magazine*, pages 62–68, May 1995.

[11] P. S. Mann. *Introductory Statistics*. John WileySons, 2003.

[12] M. A. McHenry. NSF spectrum occupancy measurements project summary. In *Shared Spectrum Company Report*, August 2005.

[13] M. A. McHenry, P. A. Tenhula, D. McCloskey, D. A. Roberson, and C. S. Hood. Chicago spectrum occupancy measurements & analysis and a long-term studies proposal. In *The first international workshop on Technology and policy for accessing spectrum*. ACM Press New York, NY, USA, 2006.

[14] A. Ng and A. Fu. Mining frequent episodes for relating financial events and stock trends. In *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, 2003.

[15] H. Su and X. Zhang. Cross-Layer Based Opportunistic MAC Protocols for QoS Provisionings Over Cognitive Radio Wireless Networks. *Selected Areas in Communications, IEEE Journal on*, 26(1):118–129, Jan. 2008.

[16] R. S. Tsay. *Analysis of financial time series*, pages 28–42. John Wiley Sins, Inc., 2001.

[17] M. Wellens, A. de Baynast, and P. Mahonen. Exploiting Historical Spectrum Occupancy Information for Adaptive Spectrum Sensing. *Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE*, pages 717–722, 31 2008-April 3 2008.

[18] M. Wellens, J. Wu, and P. Mahonen. Evaluation of Spectrum Occupancy in Indoor and Outdoor Scenario in the Context of Cognitive Radio. *Cognitive Radio Oriented Wireless Networks and Communications, 2007. CrownCom 2007. 2nd International Conference on*, pages 420–427, Aug. 2007.

[19] Q. Zhang, J. Jia, and J. Zhang. Cooperative relay to improve diversity in cognitive radio networks. *Communications Magazine, IEEE*, 47(2):111–117, February 2009.

[20] Q. Zhao, L. Tong, and A. Swami. Decentralized cognitive mac for dynamic spectrum access. *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, pages 224–232, Nov. 2005.