Mining Spectrum Usage Data: A Large-Scale Spectrum Measurement Study

Sixing Yin, Dawei Chen, Qian Zhang, Mingyan Liu, Shufang Li

Presented by Xiao MA

Outline for this presentation

- Introduction
- Methodology
- Statistic and correlation
- FPM

Introduction I -- background

- This paper includes a measurement experiment from 20MHz to 3 GHz band in Guangdong Province, China.
- 2 parameters (will be introduced in the next section) are measured and used for further observation.
- Measurements taken in 4 nearby spaces for a variety of services.

Introduction I Background

- Dynamic Spectrum Access is a scheme to utilize unlicensed band in a proper scheme so that it does not interfere with licensed usage.
- In DSA, sensing spectrum hole is an important step for the “Proper scheme”
Introduction II-- Aim

- This paper aims to find spectrum usage relationships according to service types and locations.
- This relationship can be used to infer spectrum holes which is an important step for DSA/CR application.

Methodology I-- Terms

- Channel vacancy duration (CVD).
- Service Congestion Rate (SCR). This can be used to predict degree of congestion for a service
- Channel State (CS): a binary function of time and channel $CS(t,c)=0/1$ means channel c at time t is idle/busy.

Introduction III-- innovation

- Different approach to model channel parameters.
- Strong and obvious correlation found for channel occupancy data.

Methodology II-- measurement

- 4 locations in Guangdong (Table 1)
- 1500hrs Feb 16 2009~1500hrs Feb 23 2009
- Equipment: R&S EM 550 VHF/UHF Digital Wideband Receiver, 20MHz-3.6GHz, resolution 1 per 0.2MHz, 14900 frequency data collected every timeslot.
- Timeslot about 75s
- Overall 8058 timeslots in 7 days measurement.
Methodology III– data output

- Each channel considered as 200KHz bandwidth and indexed by sequence.
- Service of a particular channel can be found in Table 2.
- This measurement returns raw data of Channel States for each time slot, which are used to calculate other parameters.

Methodology IV– detecting CS

- Detection Threshold issue: anything more than 3dB of min energy considered as signal presence.

Methodology V– determine CVD

- CVD can be observed from related CS data.
- Direct observation from CS results in integer data, not desired.
- By measuring threshold crossing time, real number data are obtained via this traditional method
- The devices for this traditional method may not be mentioned in this paper.

Methodology VI-- SCR

- Service Congestion Rate (SCR)
- Defined as the ratio between the number of busy channels in $S$ at time $t$ and the total number of channels in $S$.
- $S$ represents the set of channels for a particular service
- Can be derived from CS

$$SCR(t, S) = \sum_{c \in S} CS(t, c)/n,$$
**Methodology VII -- MSU**

- Mobile service Utilization (MSU) defined as the time varying difference in total received energy.
- In this paper it is further assumed that each user introduce same energy in the channel, reducing the complexity for analysis.

**Statistic I -- CVD**

- The cumulative density distribution can be approximated and modeled as $y = a + b e^{c x}$, an exponential like distribution.
- The regression parameters is calculated by:

  $$r^2 = 1 - \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - f_i)^2},$$

  Pre-defined parameter

**Statistics I -- CVD**

- CVD shows exponential-like PDF

Note: the CVD, albeit can be obtained from CS as positive integers, are measured using the threshold trespassing record, in order to get more precise real number results.

![Histogram of CVD distribution for GSM900 uplink at Location 2.](image)

![Approximated PDF of CVD for GSM900 uplink at Location 2.](image)

**Table 3**

<table>
<thead>
<tr>
<th>Service</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM900 uplink</td>
<td>0.014917</td>
<td>7.052599</td>
<td>1.266810</td>
<td>0.950487</td>
</tr>
<tr>
<td>GSM900 downlink</td>
<td>0.03809</td>
<td>0.710870</td>
<td>0.483754</td>
<td>0.986417</td>
</tr>
<tr>
<td>GSM1800 uplink</td>
<td>0.062653</td>
<td>0.846733</td>
<td>0.060588</td>
<td>0.985644</td>
</tr>
<tr>
<td>GSM1800 downlink</td>
<td>0.031259</td>
<td>1.058021</td>
<td>0.691488</td>
<td>0.993801</td>
</tr>
<tr>
<td>CDMA uplink</td>
<td>-0.016388</td>
<td>0.250563</td>
<td>0.140147</td>
<td>0.980726</td>
</tr>
<tr>
<td>CDMA downlink</td>
<td>0.022279</td>
<td>1.251644</td>
<td>0.732049</td>
<td>0.98638</td>
</tr>
<tr>
<td>TV1</td>
<td>0.027716</td>
<td>1.021406</td>
<td>0.599504</td>
<td>0.988016</td>
</tr>
<tr>
<td>TV2</td>
<td>0.039241</td>
<td>0.714230</td>
<td>0.518904</td>
<td>0.988958</td>
</tr>
<tr>
<td>TV3</td>
<td>0.042968</td>
<td>0.725349</td>
<td>0.595759</td>
<td>0.986116</td>
</tr>
<tr>
<td>TV4</td>
<td>0.046346</td>
<td>0.736806</td>
<td>0.601420</td>
<td>0.943959</td>
</tr>
</tbody>
</table>

Regression equation: $y = a + bc^x$. 
Statistics I CVD

- The distribution does not satisfy the memory less property in Markov model. Recall: \( p(t) = \lambda e^{-\lambda t} \) while in this paper \( p(t) = a + be^{-ct} \)
- Statistics from CS series also indicates:
  \[
  \Pr[CS(t + 1, c) = 0 | CS(t, c) = 0] = 0.91895, \\
  \Pr[CS(t + 1, c) = 0 | CS(t, c) = 0, CS(t - 1, c) = 1] = 0.5545.
  \]

Statistics II SCR

- SCR in the same location seems to have similar appearance.

Fig. 10. SCR series at Location 2: GSM1800 uplink versus GSM900 uplink.

Statistics II SCR

- Auto-correlation function and partial correlation function produces different statistics

Fig. 11. Autocorrelation coefficient of SCR of GSM1800 uplink at Location 2.
Fig. 12. Partial autocorrelation coefficient of SCR of GSM1800 uplink at Location 2.

Statistics I CVD

- 75 second time slot issue
- Verified by increasing the sampling interval to see how regression fits
Statistics II SCR

- It can be inferred that from partial correlation that SCR can be modeled as auto regressive process

$$SCR(t, S) = \sum_{m=1}^{\infty} e_m SCR(t - m, S) + n(t).$$

![Fig. 13: SCR prediction with the third-autoregressive model.](image)

Statistics III MSU

- Defined to evaluate energy difference in time sequence in a channel.
- Assume each user introduces same amount of energy:

$$P(X(t) = m) = \begin{cases} e^{-(\lambda_1 + \lambda_2)} \lambda_1^m \sum_{k=1}^{\infty} \frac{\lambda_2^k \lambda_1^{m-k}}{k!} & m \geq 0 \\ -e^{-(\lambda_1 + \lambda_2)} \lambda_2^m \sum_{k=1}^{\infty} \frac{\lambda_1^k \lambda_2^{m-k}}{k!} & m < 0. \end{cases}$$

Statistics II SCR

### Table 4

<table>
<thead>
<tr>
<th>Service</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_3$</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM9000 uplink</td>
<td>0.580122</td>
<td>0.512329</td>
<td>0.305352</td>
<td>0.960674</td>
</tr>
<tr>
<td>GSM9000 downlink</td>
<td>0.356719</td>
<td>0.322443</td>
<td>0.322548</td>
<td>0.970801</td>
</tr>
<tr>
<td>GSM1800 uplink</td>
<td>0.592203</td>
<td>0.319233</td>
<td>0.286660</td>
<td>0.974579</td>
</tr>
<tr>
<td>GSM1800 downlink</td>
<td>0.487140</td>
<td>0.312867</td>
<td>0.279747</td>
<td>0.973845</td>
</tr>
<tr>
<td>TV1</td>
<td>0.476766</td>
<td>0.285754</td>
<td>0.235663</td>
<td>0.906869</td>
</tr>
<tr>
<td>TV2</td>
<td>0.424089</td>
<td>0.302902</td>
<td>0.270873</td>
<td>0.861388</td>
</tr>
<tr>
<td>TV3</td>
<td>0.457287</td>
<td>0.322248</td>
<td>0.208510</td>
<td>0.963755</td>
</tr>
<tr>
<td>TV4</td>
<td>0.409256</td>
<td>0.336229</td>
<td>0.254212</td>
<td>0.931405</td>
</tr>
</tbody>
</table>

Statistics III MSU

- This model fits observation
- Observation does not change significantly according to time and channel, and location
This model fits well when the sampling interval is less than 10 time slots (750s).

Most CSs in GSM900 uplinks in Location 4 are highly correlated.

Spectral correlation of SCR at Location 1

<table>
<thead>
<tr>
<th>GSM900</th>
<th>GSM900</th>
<th>GSM1800</th>
<th>GSM1800</th>
<th>TV1</th>
<th>TV2</th>
<th>TV3</th>
<th>TV4</th>
</tr>
</thead>
<tbody>
<tr>
<td>uplink</td>
<td>downlink</td>
<td>uplink</td>
<td>downlink</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>0.873</td>
<td>0.832</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.952</td>
<td>0.955</td>
<td>0.817</td>
<td>0.827</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.674</td>
<td>0.666</td>
<td>0.713</td>
<td>0.690</td>
<td>0.634</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.730</td>
<td>0.749</td>
<td>0.769</td>
<td>0.710</td>
<td>0.833</td>
<td>0.711</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.588</td>
<td>0.581</td>
<td>0.597</td>
<td>0.565</td>
<td>0.560</td>
<td>0.721</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

L1 and L2 are similar, up to a constant shift. So do L3 and L4.
All share general trend since subscribers at different locations share common behavioral.
Correlation II– Spatial SCR

<table>
<thead>
<tr>
<th>Location</th>
<th>Loc. 1</th>
<th>Loc. 2</th>
<th>Loc. 3</th>
<th>Loc. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loc. 1</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loc. 2</td>
<td>0.833</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loc. 3</td>
<td>0.858</td>
<td>0.846</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Loc. 4</td>
<td>0.854</td>
<td>0.880</td>
<td>0.909</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Correlation III– FPM and prediction

- Some same pattern in the CS series in all the channels occur frequently. These pattern can be used for prediction.
- This is a 2D prediction, that means to predict multiple channel CSs based on known multiple channel CSs.
- 2 issues: find pattern, and find associations among the patterns

FPM– Find pattern

- Need to find maximum possible patterns
- Trivial search is needed and hard
- An iterative-like decision process reduces the complexity: e.g. (x,y) block can be judged as negative if one of its (x-1,y) sub block is negative

FPM algorithm terms

- Use hash tables $T_{\downarrow x \times y}$ to save positive recognized block patterns
- For each block pattern in $T_{\downarrow x \times y}$, a variable is dedicated recording the repeating time for the pattern.
- Record in $T_{\downarrow(x+1) \times y}$ and $T_{\downarrow x \times (y+1)}$
FPM algorithm overview

1. N=2 till infinite
2. For all x, y, x+y=N, establish new hash table \( T_{x \times y} \)
3. Check each possible block \( B_{x \times y} \) in training data, eliminate impossible patterns.
4. Save all positive \( B_{x \times y} \)'s and their repetitions respectively in hash table \( T_{x \times y} \).
5. Delete the insufficient repeated blocks in \( T_{x \times y} \).
6. If nothing found in T, direct break out and end this program, all previous saved T are the results learned in this process.

Prediction after FPM

- Prediction accuracy and missing rate VS size of training

Prediction after FPM

Example: We already observed that [0010] occurs 1000 times and [00101] occurs 990 times.
So we can make a decision that if next time 0010 is observed, the next bit is 1 with probability of 0.99
Unmatched blocks considered as missing in prediction (No prediction for these blocks)

### Prediction Compartments

<table>
<thead>
<tr>
<th>Service</th>
<th>Occupancy</th>
<th>1st order Markov Prediction Accuracy</th>
<th>Miss rate</th>
<th>FPM-2D Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM1800 downlink</td>
<td>85.1%</td>
<td>85.2%</td>
<td>11.8%</td>
<td>96.9%</td>
</tr>
<tr>
<td>GSM1800 uplink</td>
<td>60.3%</td>
<td>77.4%</td>
<td>24.8%</td>
<td>95.1%</td>
</tr>
<tr>
<td>GSM1800 downlink</td>
<td>50.2%</td>
<td>83.7%</td>
<td>16.2%</td>
<td>96.6%</td>
</tr>
<tr>
<td>TV2</td>
<td>92.1%</td>
<td>92.6%</td>
<td>5.4%</td>
<td>96.9%</td>
</tr>
<tr>
<td>TV3</td>
<td>44.5%</td>
<td>75.0%</td>
<td>4.2%</td>
<td>97.8%</td>
</tr>
<tr>
<td>TV4</td>
<td>41.9%</td>
<td>74.5%</td>
<td>6.3%</td>
<td>97.7%</td>
</tr>
</tbody>
</table>

The training/testing sets are from the same service.
Cross service Prediction

TABLE 9
The Spectrum Usage Prediction Results at Location 1

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Testing Set</th>
<th>Miss Rate</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM900 downlink</td>
<td>TV1</td>
<td>66.1%</td>
<td>79.2%</td>
</tr>
<tr>
<td>TV1</td>
<td>GSM900 uplink</td>
<td>75.3%</td>
<td>80.4%</td>
</tr>
<tr>
<td>GSM900 uplink</td>
<td>GSM900 downlink</td>
<td>35.2%</td>
<td>86.4%</td>
</tr>
<tr>
<td>GSM900 downlink</td>
<td>GSM900 uplink</td>
<td>31.8%</td>
<td>87.4%</td>
</tr>
</tbody>
</table>

The training/testing sets are from different services.

Cross location prediction

TABLE 10
The Cross-Location Spectrum Usage Prediction Results

<table>
<thead>
<tr>
<th>Service</th>
<th>Training Set</th>
<th>Testing Set</th>
<th>Miss Rate</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV1</td>
<td>Location 1</td>
<td>Location 2</td>
<td>5.7%</td>
<td>95.3%</td>
</tr>
<tr>
<td>TV1</td>
<td>Location 3</td>
<td>Location 4</td>
<td>7.3%</td>
<td>97.4%</td>
</tr>
<tr>
<td>TV1</td>
<td>Location 1</td>
<td>Location 3</td>
<td>6.5%</td>
<td>96.7%</td>
</tr>
<tr>
<td>TV1</td>
<td>Location 3</td>
<td>Location 1</td>
<td>7.7%</td>
<td>95.8%</td>
</tr>
</tbody>
</table>

Prediction summary

- Self service self location: more than 95%. Missing rate 4%~25%
- Cross service prediction 60%~80%, missing rate 30%~70%
- Self service cross locate predict has similar performance to self service self location
- 3h training time is sufficient

Problems--model

- Every estimation based on CS, CS based on sensing channel. Threshold dilemma, unable for CDMA detection
- Assumption: each user introduces same energy.
Problem--Algorithm

- Are the blocks include arbitrary combination of locations (rows)?
- Minimum repetition requirement may end up mess.
- Enhanced for multiple time slot prediction with some probabilistic reasoning.