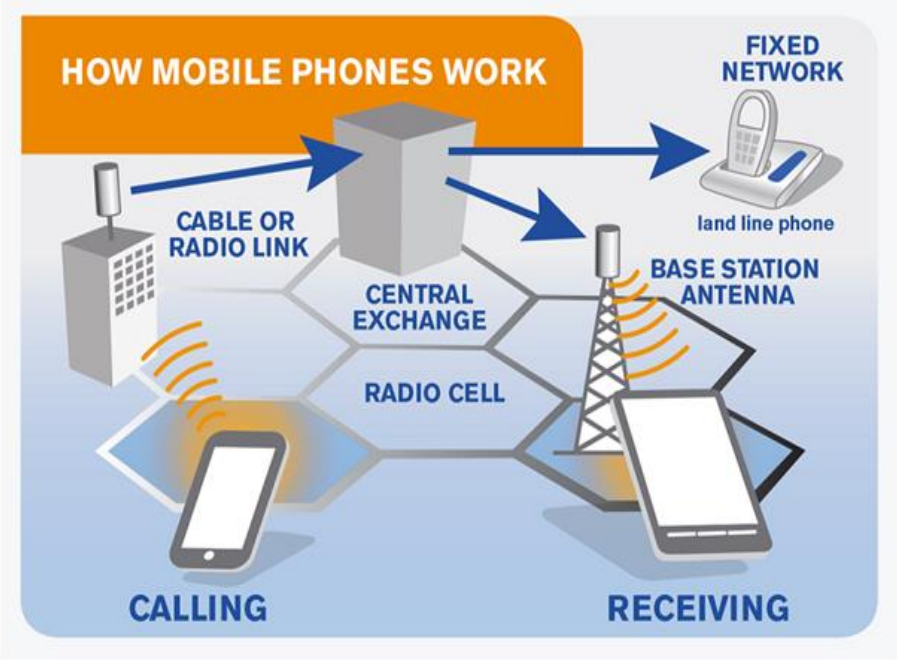
Predicting population using mobile device counts

International Conference on Big Data and Official Statistics
20-22 October, 2015

by S Tam, Chief Methodologist
Australian Bureau of Statistics



Reference – ITU EMF Guide 2014

Outline

- The statistical question?
  - Can we use mobile device counts from base stations to estimate population counts?
- Quality of mobile device counts
- Simulated population and mobile device counts
- Model fitting and results
- Conclusion

Source - http://barnraisersllc.com/2015/06/25-examples-of-companies-doing-something-with-big-data/

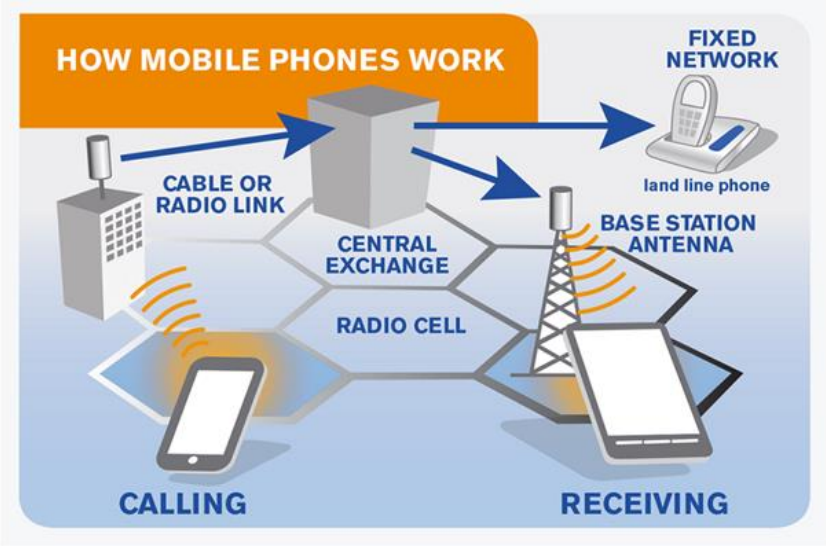
Does size matter?

- No, value creation does - Create big value from big (and small) datasets!
- Yes, (big) garbage in, (big) garbage out

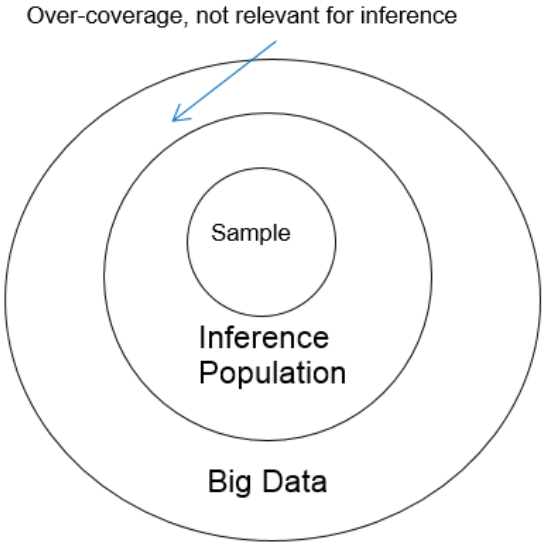
Mobile device data and population mobility

Base Stations generate Call Data Records

- **No under-coverage issues**
- **Measurement error issues**



Reference – ITU EMF Guide 2014

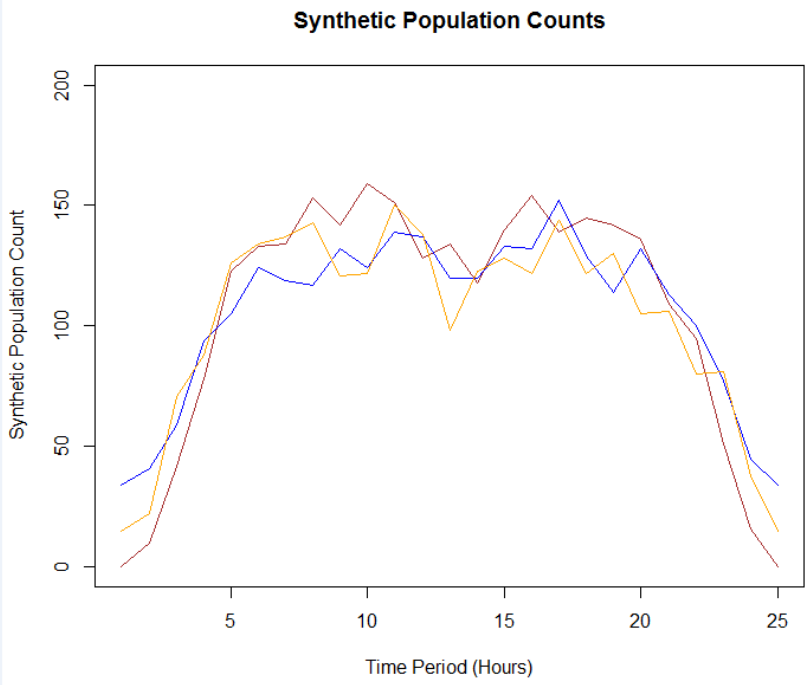


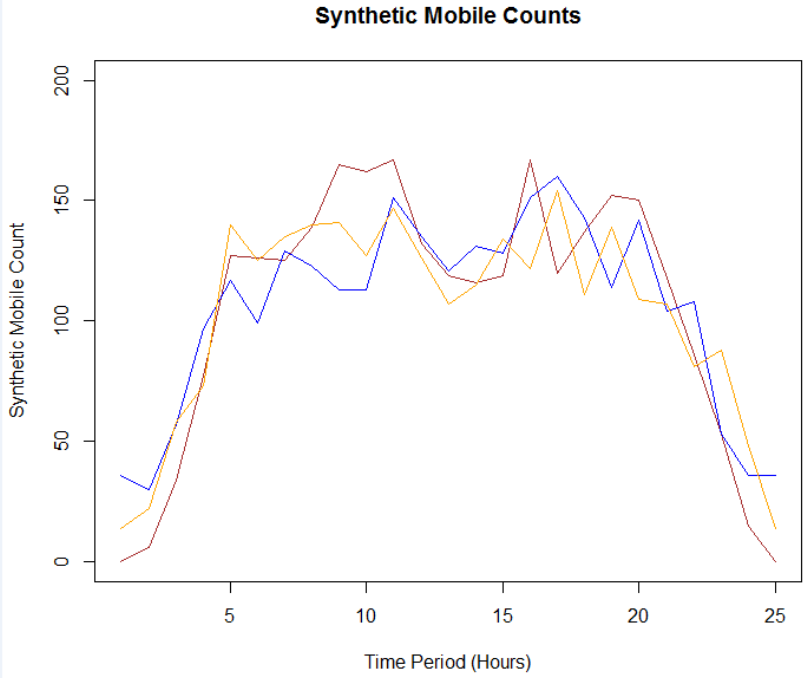
- Mobile phones comprise a transmitter and receiver
- Call made/received via Base stations
- CDR can be used to estimate pop movement

An example based on simulated data

1. Aim: Use model device counts to predict population counts
2. Mobility of 100,000 persons was simulated
3. Each person wanders between ‘home’ base station (BS), ‘destination’ BS and ‘home’ BS throughout a 24 hr period; throughout journey, each person reaches an intermediary BS each hour. All BS are randomly assigned
4. Each person has a 65% chance of being picked up by a BS; and each person’s number of mobile device is governed by a Poison distribution with mean = 1.5.
5. A total of 1,000 Base Station pairs of mobile device and population counts were simulated
6. Dynamic Linear Model fitted for random 100 Base Station pairs
7. The fitted Model was used to predict the other 900 Base Station population counts
8. Relative prediction error was calculated for each of the 900 Base Station pop counts
9. Steps 6, 7 and 8 were repeated for another random sample of 100 Base Stations
10. Step 9 was repeated 200 times.

Sample plots of Population and MD counts over the 24-hour period





Modelling base station data using dynamic linear models

"English" version

- Population (Pop) counts assumed to be stochastically related to the mobile device (MD) counts through a "Pop to MD" ratio
- The ratio is allowed to change over time
- Ratio is estimated using a "EM" algorithm
- The estimated ratio is used for Pop counts prediction

"Greek" version

$$\begin{bmatrix} \mathbf{Y}_{ot} \\ \mathbf{Y}_{rt} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{ot} \\ \mathbf{Z}_{rt} \end{bmatrix} \beta_t + \begin{bmatrix} \mathbf{e}_{ot} \\ \mathbf{e}_{rt} \end{bmatrix}$$

$$\beta_t = \beta_{t-1} + \boldsymbol{\varepsilon}_t \ , \ \beta_t \perp \mathbf{Z}_t$$

$$\beta_1 \sim N(\beta_0, Q)$$

$$\mathbf{e}_t \sim \text{independent } N(\mathbf{0}, \Sigma)$$

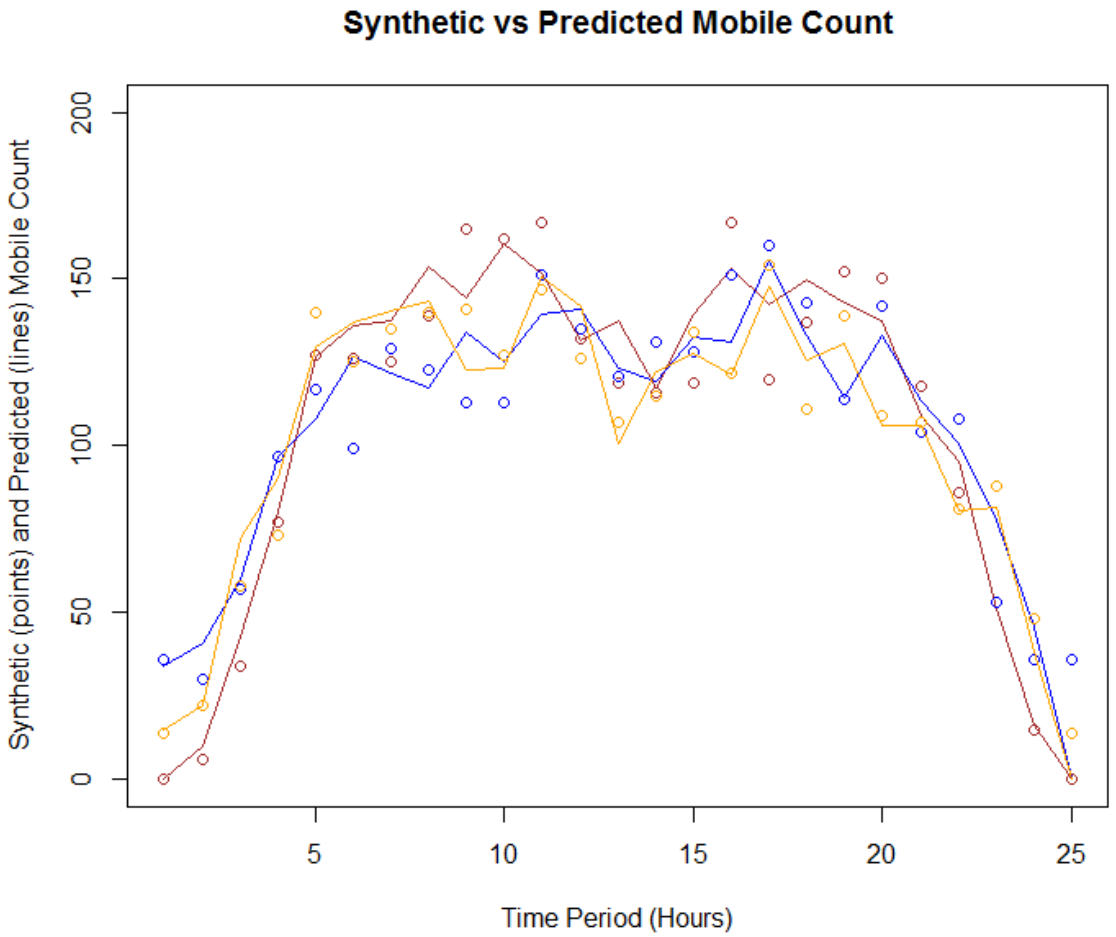
$$\boldsymbol{\varepsilon}_t \sim \text{independent } N(0, Q) \ , \ \boldsymbol{\varepsilon}_t \perp \mathbf{D}^{(t)}$$

where β_t, Σ and Q are unknown for $t = 0,\ldots,T$. Bayesian Hierachical Modelling (BHM) requires priors to be specified for Σ and Q. In this example, we use Empircal Hierachical Modelling (EHM) by plugging MLEs of Σ and Q, using the EM algorithms, into the updating equations.

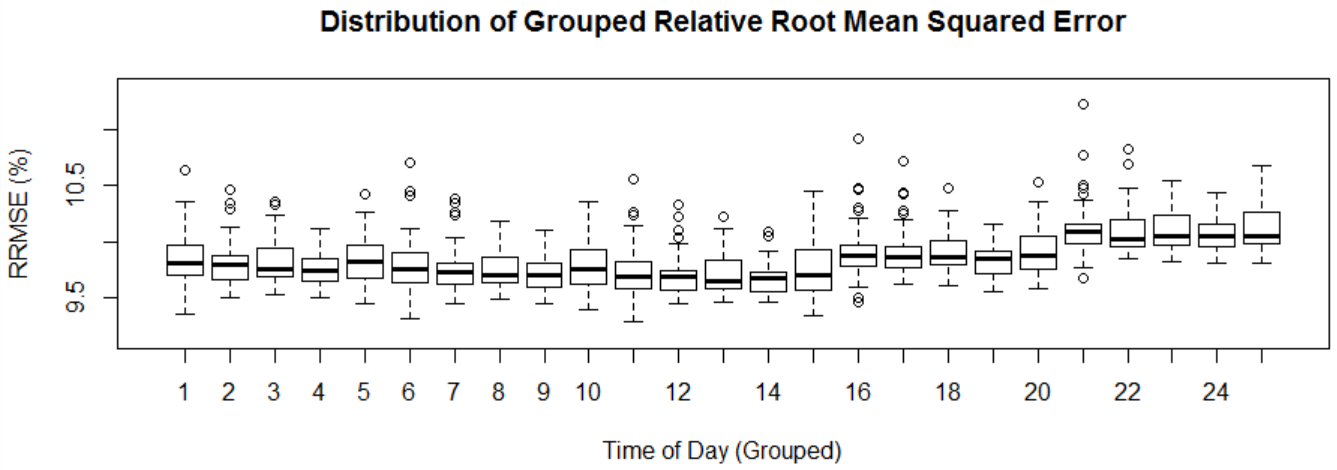
EM Algorithm – Latent variable are β_t

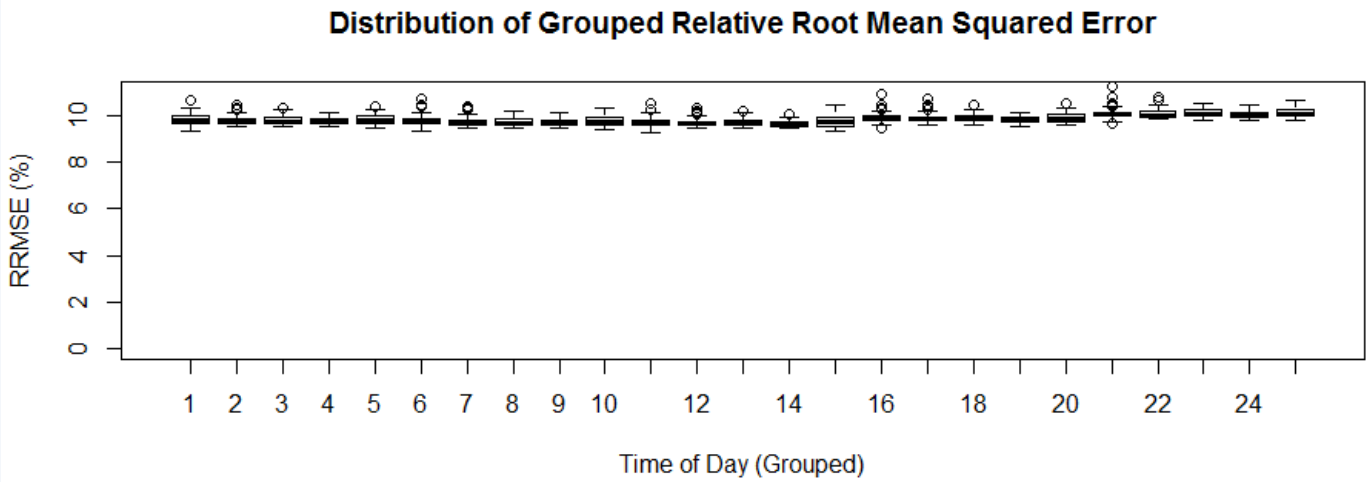
i. Compute "maximisers" from the log likelihood - (M-step)
ii. Compute Expectation of the maximisers based on guesses of Q and Σ - (E-step):
   - Basically Expected values of $\beta_t, \beta_{t-1}\beta_t', \beta_t\beta_t'$, for $t = 1,\ldots T$ given the data $D^{(T)}$ i.e. Kalman smoothers.
iii. Update parameters using the maximisers
iv. Repeat (i) and (ii) until convergence

Predicted vis BS population counts



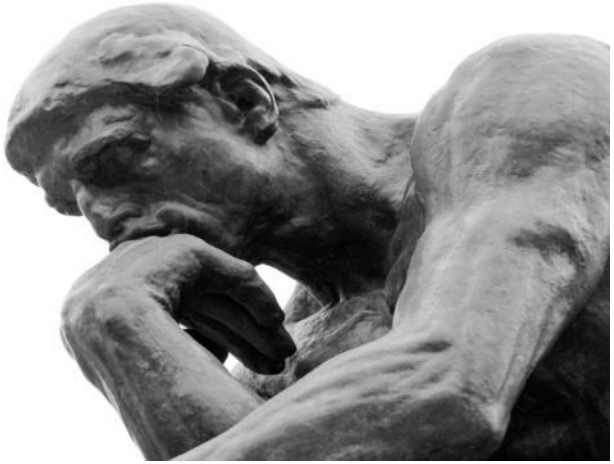
Relative root mean square prediction errors - Leave 900 out CV procedure





Take-home messages

- Relatively accurate population counts can be predicted using mobile device counts by employing a Dynamic Linear Model
  - Modelling requires
    - Availability of mobile device counts for all base stations
      - from all telecommunication service providers
    - Ground truths available from a random sample of base stations
      - To estimate the “Pop to MD” ratio
      - Accuracy of prediction will improve over time if this ratio is allowed to change over time



siu-ming.tam@abs.gov.au