Mobility in Cities: Comparative Analysis of Mobility Models Using Geo-tagged Tweets in Australia

Saeed Farooq Khan, Neil Bergmann
School of Information Technology & Electrical Engineering
University of Queensland
Brisbane, Australia
e-mail: s.khan@uq.edu.au, n.bergmann@itee.uq.edu.au

Raja Jurdak, Branislav Kusy, Mark Cameron
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Australia
e-mail: {firstname.lastname}@csiro.au

Abstract—**Modeling human movement has recently received significant research interest due to its applications in disaster management, transportation planning, communication networks and epidemic modeling and prediction. Most of the prior human mobility studies use data which is coarse-grained, proprietary or limited in size. In contrast, this study utilizes a large amount of publicly available data from Twitter. Geo-data associated with Twitter posts have been analyzed to characterize and model movement patterns within Australia. Gravity and Radiation models have been used to analyze human movements within major Australian cities. The Gravity models show better performance compared to the Radiation model in estimating flows between places. We further find that fitted parameters vary across cities, highlighting the need for city-specific models to accurately represent movement flows.**

Keywords-mobility; Twitter; gravity; estimation

I. INTRODUCTION

The study of human mobility focuses on discovering the underlying dynamics and intrinsic characteristics of people's movements: what drives these movements, what is the pattern of these movements and how to technically define them. The applications of this understanding span domains such as disaster management, transportation network planning, infrastructure management, communication networks and epidemic modeling and prediction.

The study of human mobility requires real mobility data which can be acquired from various sources such as a population census [1], cell phone records [2], RFID [3] and Wi-Fi [4], all with various limitations. Census data typically involves one snapshot every 5 years; RFID and Wi-Fi have limited sensing range around readers or access points. Mobile phone call records have privacy issues and low spatial resolution, and are not publically available. A preferable data source should offer higher resolution, real-time spatiotemporal data which is publically and easily available at large scale and volume.

In recent years, Twitter has become the subject of many studies. Twitter is primarily a social networking platform, and there has been research on content analysis of tweets [5]. Analysis of Twitter data has also been used to help manage crisis situations and emergency services [6] and other domains such as crime prediction [7].

Since Twitter addresses many of the limitations of other mobility data sources, it has been chosen as the human mobility proxy for this research. While previous work from our team has focused on understanding human mobility from Twitter [8] from a statistical-physics' perspective, it is still unclear how well Twitter captures intra-city mobility and whether its patterns match existing mobility models.

This paper analyzes mobility flows in major cities of Australia using data from geotagged tweets, with two main contributions: (1) to identify the most representative mobility model; and (2) to investigate the heterogeneity of mobility dynamics across cities.

In a previous paper we have shown that Twitter mobility data is best described by the gravity mobility model at national, state, and city scale using the 20 most populous entities at each scale [9]. The scope of the current analysis focuses on intra-city movements, and includes a larger number of cities at a uniform and standard geographic scale defined by the Australian Bureau of Statistics [10]. Three versions of the Gravity model (Gravity-4, Gravity-2, Exponential Gravity) plus the Radiation model are compared. The accuracy of a single nationwide model is compared with city-specific models to explore the degree of heterogeneity in mobility dynamics across the country. Comparisons of Twitter users and census population for different regions is also analyzed to identify locations where the Twitter data over-represents the actual population.

II. RELATED WORK

This section first discusses a few key mobility models that will be used in the subsequent analysis. This is followed by a review of data sources for mobility proxies.

Theory of Intervening Opportunities: According to this theory, "the number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities" [11]. This theory was applied on a census data set which described the migration of families in a U.S city. It looked at the mobility problem from the sociologists' perspective who at that time were keen on finding a theory that could explain human movement [12].

Gravity Model: The Gravity model suggests that mobility between two places, namely an origin and a destination is

978-1-5090-3619-6/17/$31.00 ©2017 IEEE

proportional to the product of population of these two places, and inversely proportional to the distance between them [13]. It is also used for estimating trade flows between countries as well as communication volume between cities. There are several different variants of this model:

Gravity-4 Model: $$F_{ij} = C\frac{m_i^{\alpha} n_j^{\beta}}{d^{\gamma}} \quad (1)$$

Gravity-2 Model: $$F_{ij} = C\frac{m_i n_j}{d^{\gamma}} \quad (2)$$

Exponential-Gravity Model: $$F_{ij} = C\frac{m_i^{\alpha} n_j^{\beta}}{e^{\gamma d}} \quad (3)$$

where m_i and n_j are the masses (populations) of source i and destination j respectively, d_{ij} is the distance between them, F_{ij} is the number of trips between them, and α, β, γ and C are the model's parameters.

Radiation Model: This model is based on the particle diffusion theory in which particles emitted at a source have certain probability of being absorbed by a destination location [14]. It states that the absorption probability depends on the origin and destination population and is also determined by the population within a circle centered at the origin, with the radius equal to the distance between the origin and the destination (excluding the origin and destination).

$$T_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \quad (4)$$

where m_i and n_j represent populations of source and destination respectively, T_i is the total number of trips emanating from location i, s_{ij} is defined as the total population within radius d from the center of the source area with the source and destination areas excluded, and T_{ij} is the estimated flux from i to j.

It is difficult to directly track the movement of humans, so instead various proxies have been used, often some sort of article carried by the studied subjects whose location is more easily checked.

Bank Note: In [15], the authors analyzed the circulation of bank notes in the United States as a proxy for mobility. They found that distribution of travelling distances decays as a power law and they suggested that human travel on geographical scale is a random process.

Mobile Phone: In [16], authors studied the mobility of mobile phone users by tracking their positions when making calls. They found that in contrast with the random walk model proposed in [15], human trajectories show a high degree of spatial and temporal regularity. They suggested that there is a high probability of users returning to a few frequently visited locations. However mobile phone data is not publically available.

Twitter: In [17], authors made use of geo-tagged tweets to examine mobility profiles of different countries. They compared countries based on mobility rate, spatial spread of movement, diversity of destinations and balance of inflows and outflows. They also discovered global patterns of increased mobility at the end of the year. This study, however, does not take into account finer spatial scales such as movement within cities.

In previous work in our lab [9], geo-tagged tweets posted within Australia have been analyzed and they have been found to capture features of human mobility such as diversity of movement orbits of people and mobility within and between cities. This paper focuses on more comprehensive intra-city movement modeling to characterize and understand human movements within major cities in Australia.

III. METHODOLOGY

The analyzed data set consists of over 6 million geo-tagged tweets collected within Australia using Twitter Streaming API [18] between September 2013 and April 2014. Each tweet includes latitude and longitude. Using an R-tree method [19], every tweet is then mapped to its corresponding Statistical Area Level 2 (SA2) location as defined by the Australian Bureau of Statistics [10]. An SA2 area roughly corresponds to a suburb.

Nine major cities of Australia have been chosen for mobility analysis namely Sydney, Melbourne, Brisbane, Adelaide, Perth, Canberra, Gold Coast, Darwin and Hobart. For each city, the 20 largest SA2 areas (Twitter user-wise) have been selected for analysis and the number of tweets and census population have been recorded for each SA2 area. The only exceptions are Canberra, Gold Coast, Darwin and Hobart for which 19, 18, 10 and 9 SA2 areas respectively have been chosen due to their relatively smaller geographical sizes. To calculate Twitter fluxes or trips between SA2 areas, the number of pairs of consecutive tweets appearing first at source m_i at time t_i and then at destination n_j at time t_j are counted. Gravity-4, Gravity-2, Exponential Gravity and Radiation models have been used to estimate the fluxes. For Gravity models, training has been done on those pairs of SA2 areas which have at least ten Twitter fluxes between them and linear regression has been used to fit the data. The testing is then done on the complete data set.

IV. DISTANCE DISTRIBUTION

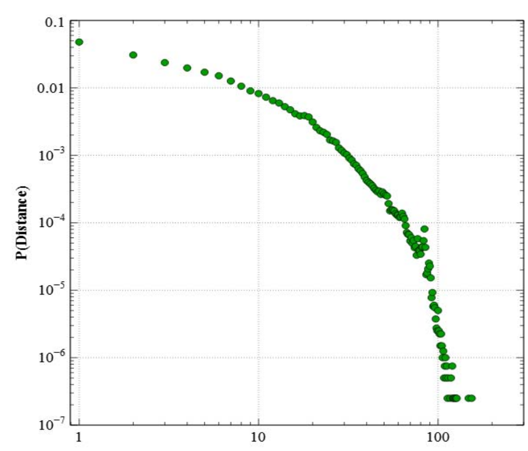


Figure 1. Distribution of distances.

Fig. 1 shows the distribution of distance between consecutive tweets for all SA2 areas of all nine cities, divided into bins of size 1km. Only tweets where consecutive tweets are located within the same city are plotted. According to this, the probability of having smaller distance between two consecutive tweets is higher, and, as the distance increases, there seems to be a lesser number of tweets between two areas. This suggests people undertake short trips in their usual daily routine and confirms similar findings in previous mobility studies.

V. MOBILITY ANALYSIS

This section compares how well the mobility models of section II estimate the fluxes between pairs of SA2 areas for all nine cities. For each pair, we extract the movement data from the tweets for the 20 SA2 areas with largest Twitter populations. Because for n SA2 areas we get n^2 fluxes, using a larger number makes the diagram unwieldy. For all models, m_i is the Twitter population (unique Twitter users) of source SA2 area, n_j is the Twitter population of destination SA2 area and d_{ij} is the distance between centers of source and destination areas. Model parameters are fitted using Linear Regression in a log-log space for each city. For training parameters only fluxes equal to or greater than 10 are used to avoid model fitting being skewed by many small fluxes. For testing, all non-zero fluxes are used.

Fig. 2(a), (b) show the comparison of the mobility models for Australia's largest two cities Sydney and Melbourne respectively. Here, the x-axis represents the fluxes calculated using mobility models and labelled as 'Estimated Flux' whereas the y-axis represents historic movements extracted from the tweets and labelled as 'Twitter Flux'. The dots represent pairs of estimated and actual Twitter fluxes, and the green line shows the relationship given by the fitted model parameters. Each plot contains the relevant Pearson correlation coefficient (R) and the mean absolute percentage error (MAPE) for comparison purposes.

$$\mathrm{MAPE} = \frac{100}{n}\sum_{k=1}^{n} \frac{|A_k - F_k|}{|A_k|} \tag{5}$$

where A_k is the actual value and F_k is the forecast value.

A visual comparison of the plots for both cities shows that Gravity models in general perform better than the Radiation model. The points in the Gravity plots are closely packed around the green line indicating a better fit, and this is mathematically evidenced by the higher values of R and lower values of MAPE than the Radiation model. Among gravity models, Gravity-2 performs best, confirming the effectiveness of these models over short (intra-city) distances [9]. A closer examination suggests that the estimation error is roughly bounded by one decade.

Table I lists the R and MAPE values for all four models as well as fitting parameters for the three gravity models for all nine cities. The best model for each city is highlighted in bold. For Darwin, however, the R value for all the models is much lower than for any of the other cities (see Table 1) and so none of the models fit that city well – we still need to investigate the reasons further. In all cases Gravity-2 is the best or second-best model, and so we concentrate on that model for further experiments.

The values of fitting parameters differ substantially among cities, and it appears that every city has unique mobility model parameters reflecting unique movement dynamics. Notably, Perth, Canberra and Brisbane have significantly lower γ values than all other cities (except Darwin), indicating that movement between suburbs in these cities is less sensitive to distance. Sydney and Melbourne both have the highest Pearson correlation and the lowest MAPE, suggesting that Twitter movement fits the gravity-2 model best in larger cities.

To further explore the heterogeneity in observed Twitter mobility dynamics, we fit all the Twitter mobility data to a single model for all cities. All 2196 SA2 areas of Australia are used and mobility is calculated between every pair. The Gravity-2 model parameters are fitted to this complete data (without setting a limit of minimum 10 flows between SA2 areas). Using these new parameters, flows between top SA2 areas for each major city are estimated using the single national model. Fig. 3(a), (b) show plots of actual and estimated fluxes for Sydney and Melbourne.

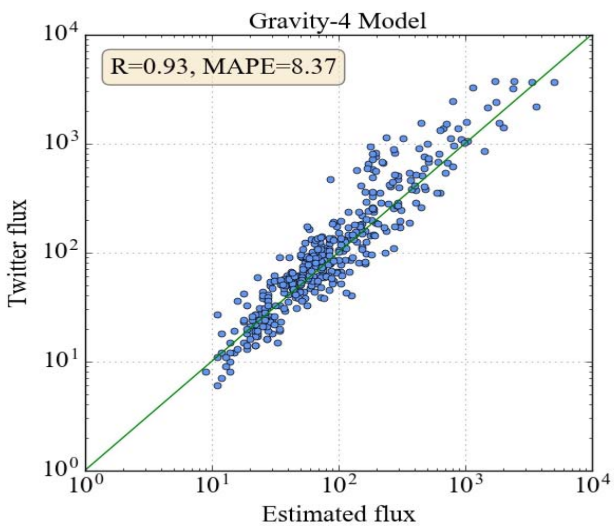
For each city, Table II shows the R and MAPE values using the new generalized gravity-2 parameters and the city-specific gravity-2 parameters. In all cases, the R value for each city is much lower, in most cases significantly lower. The MAPE is higher for most cities except Darwin which, as explained above, is not modelled well with any of the models. In the case of Sydney, Melbourne and Hobart the MAPE is much higher with the generalized model. This suggests that city-specific model parameters are necessary for representative mobility estimates.

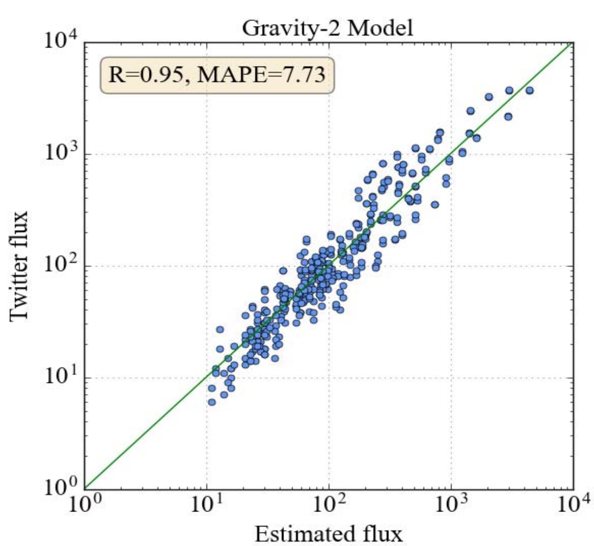
VI. TWITTER VS CENSUS POPULATION

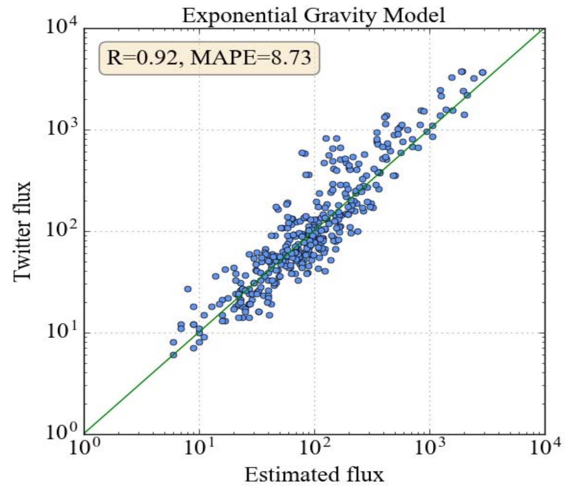
In this section, the relation between Twitter and census population is explored. Fig. 4 shows a scatter plot of the census versus Twitter population for all 271 SA2 areas of Sydney. There is no clear relationship between the two. Twitter population does not always increase with the higher census population and vice versa. Similar results apply to other cities. The SA2 areas of cities for which we have done mobility analysis, Table III shows the ratio of aggregate Twitter and census populations as well as the ratio of number of tweets to these populations. Central Business Districts (CBDs) show the highest tweets to census population ratio. These variations in tweeting dynamics and population indicate specific attributes relevant to each city.

Next, we attempt to identify the reason for variations in Twitter population relative to the census population for Brisbane. Our conjecture is that areas with high Twitter population include a larger proportion of (non-resident) visitors. Fig. 5 shows the heat map of ratio of Twitter to census population and identifies some key SA2 areas where Twitter users are relatively over-represented such as inner suburbs around the CBD, airport and Mt. Coot-tha tourist area. This intuitively makes sense as central areas of the city have a larger number of visitors that appear to be highly represented in their Twitter volume, compared to the actual census population of these areas. The airport appears to have very high Twitter over-representation as it is an area with a very small population, yet has a large number of active Twitter users travelling in and out of the city. Mt. Coot-tha is a tourist area in Brisbane that includes city lookouts, botanic gardens, and hiking trails, all of which attract visitors though its census population is quite small. This confirms our conjecture that the varying degrees of representation of Twitter users relative to census populations is due to the visitor tweets.

Next, we turn our attention to the areas with absolute high volumes of tweets, rather than high volume relative to census populations. Fig. 6 shows the heat map of number of tweets for biggest SA2 areas (Twitter-user wise) of Brisbane and identifies hot spots such as shopping centers, markets, universities, hospital and central business district of the city.







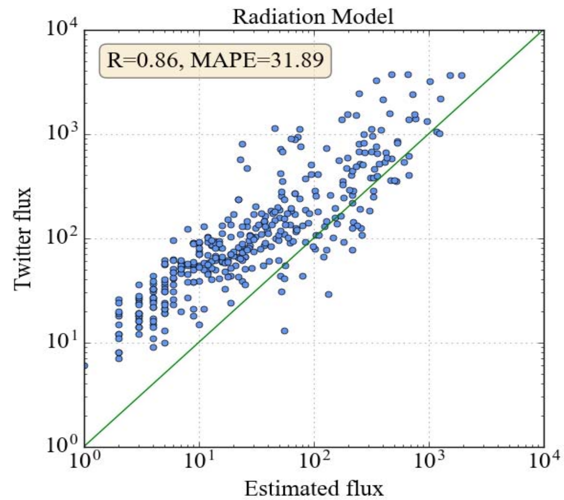
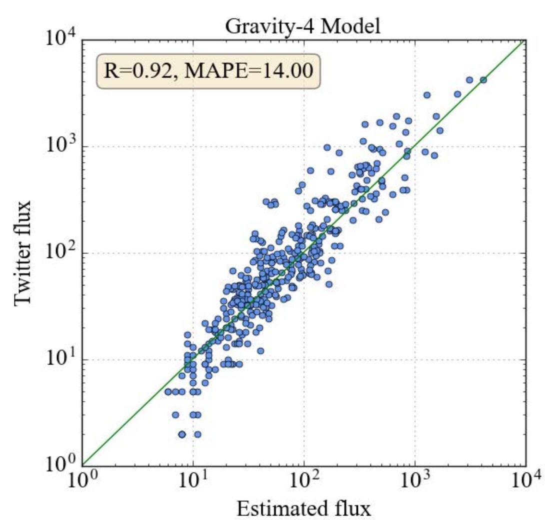


Figure 2. (a) Sydney.



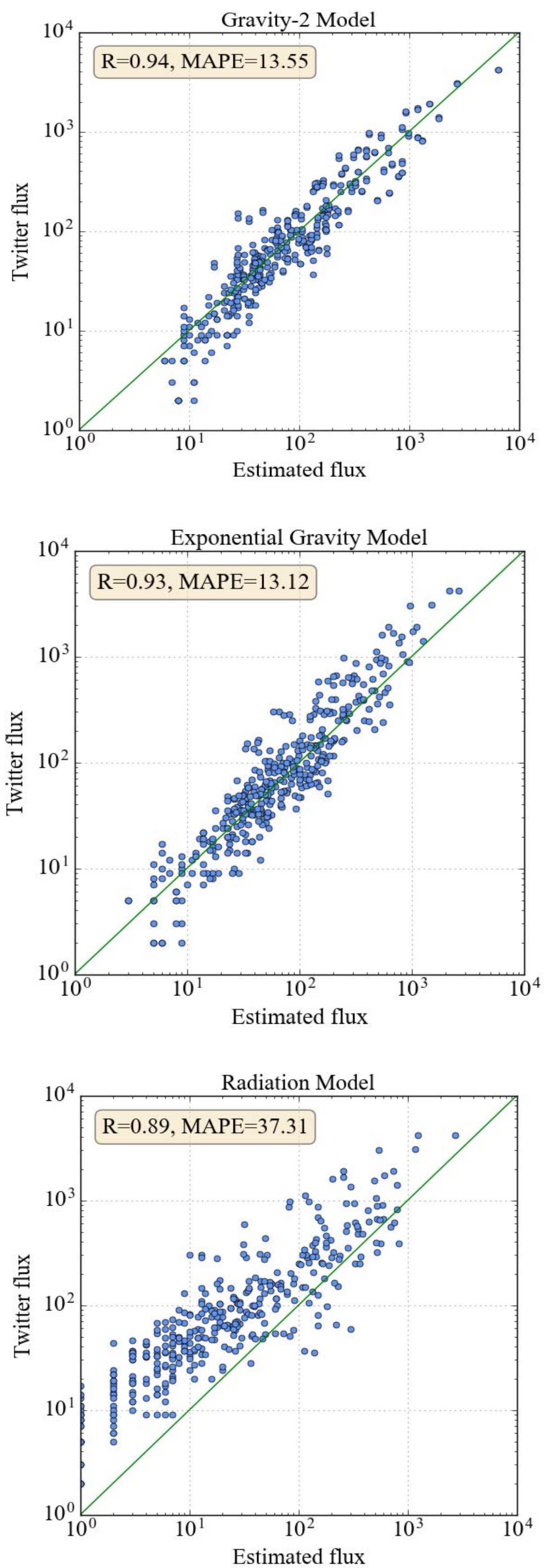


Figure 2. (b) Melbourne.

TABLE I. COMPARISON OF MOBILITY MODELS

City	Grav-4	Grav-2	Exp-Grav	Radiation
Syd	α = 1.224 β = 0.7308 γ = 0.9305 C = 6.97e-05 R = 0.93 MAPE = 8.37%	**γ = 0.9482** **C = 5.82e-05** **R = 0.95** **MAPE = 7.72%**	α = 1.217 β = 0.926 γ = 0.105 C = 6.92e-06 R = 0.92 MAPE = 8.72%	R = 0.86 MAPE = 31.88%
Mel	α = 1.025 β = 0.564 γ = 1.115 C = 0.0019 R = 0.92 MAPE = 14%	**γ = 0.9499** **C = 7.29e-05** **R = 0.94** **MAPE = 13.55%**	α = 1.011 β = 0.689 γ = 0.175 C = 0.0004 R = 0.93 MAPE = 13.12%	R = 0.89 MAPE = 37.31%
Bri	α = 0.863 β = 0.1164 γ = 0.882 C = 0.149 R = 0.79 MAPE = 45.77%	**γ = 0.675** **C = 0.0001** **R = 0.86** **MAPE = 42.75%**	α = 0.873 β = 0.236 γ = 0.139 C = 0.035 R = 0.79 MAPE = 41.89%	R = 0.80 MAPE = 42.2%
Ade	α = 0.945 β = 0.381 γ = 0.964 C = 0.0335 R = 0.87 MAPE = 34.72%	**γ = 0.880** **C = 0.0004** **R = 0.90** **MAPE = 33.26%**	α = 0.955 β = 0.395 γ = 0.127 C = 0.0121 R = 0.86 MAPE = 32.17%	R = 0.87 MAPE = 39.37%
Per	**α = 1.103** **β = 1.041** **γ = 0.655** **C = 7.43e-05** **R = 0.89** **MAPE = 28.16%**	γ = 0.647 C = 0.0002 R = 0.89 MAPE = 29.94%	α = 1.109 β = 1.142 γ = 0.051 C = 1.597 R = 0.87 MAPE = 30.66%	R = 0.80 MAPE = 37.18%
Can	α = 0.918 β = 0.457 γ = 0.537 C = 0.0128 R = 0.81 MAPE = 39.8%	**γ = 0.486** **C = 0.0003** **R = 0.84** **MAPE = 37.12%**	α = 0.957 β = 0.414 γ = 0.078 C = 0.0087 R = 0.76 MAPE = 39.5%	R = 0.76 MAPE = 39.72%
GC	α = 0.512 β = 0.964 γ = 0.863 C = 0.0207 R = 0.87 MAPE = 39.13%	**γ = 0.90** **C = 0.0005** **R = 0.89** **MAPE = 26.78%**	α = 0.518 β = 0.92 γ = 0.061 C = 0.0077 R = 0.85 MAPE = 39.19%	R = 0.85 MAPE = 37.36%
Dar	α = 0.691 β = 1.130 γ = 0.468 C = 0.0076 R = 0.53 MAPE = 122.6%	γ = 0.517 C = 0.0026 R = 0.55 MAPE = 105.63%	α = 0.694 β = 1.122 γ = 0.109 C = 0.0071 R = 0.51 MAPE = 125.9%	**R = 0.64** **MAPE = 65.91%**
Hob	**α = 1.052** **β = 1.001** **γ = 0.946** **C = 0.002** **R = 0.94** **MAPE = 30.61%**	γ = 0.932 C = 0.0027 R = 0.94 MAPE = 32.35%	α = 1.003 β = 0.995 γ = 0.131 C = 0.0013 R = 0.91 MAPE = 35.37%	R = 0.72 MAPE = 38.56%

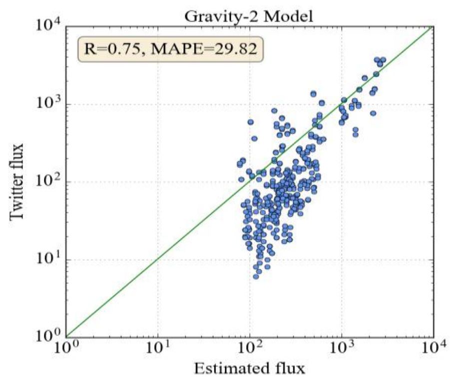


Figure 3. (a) Sydney: Model fit with generalized parameters

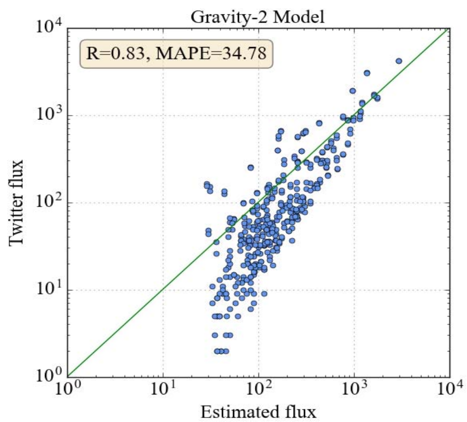


Figure 3. (b) Melbourne: Model fit with generalized parameters

TABLE II. GENERALIZED VS CITY-SPECIFIC G-2 MODELS

City	Generalized G-2		City-Specific G-2	
	R	*MAPE*	*R*	*MAPE*
Syd	0.75	29.82%	0.95	7.72%
Mel	0.83	34.78%	0.94	13.55%
Bri	0.72	45.65%	0.86	42.75%
Ade	0.76	34.32%	0.90	33.26%
Per	0.77	29.52%	0.89	29.94%
Can	0.74	40.18%	0.84	37.12%
GC	0.66	33.26%	0.89	26.78%
Dar	0.51	93.08%	0.55	105.63%
Hob	0.77	80.6%	0.94	32.35%

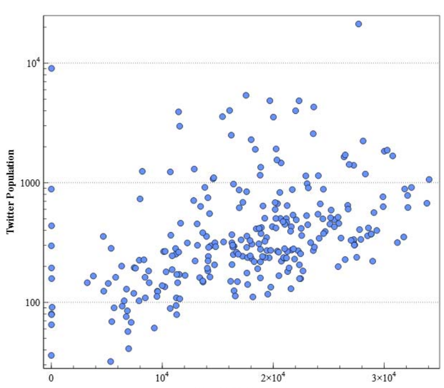


Figure 4. Census vs Twitter population.

TABLE III. POPULATION AND TWEET RATIOS

City	Census/Twitter Population	Tweets/Twitter Population	Tweets/Census Population
Syd	5.46	7.54	1.38
Mel	6.61	7.75	1.17
Bri	12.47	8.36	0.67
Ade	21.07	10.46	0.50
Per	14.99	9.09	0.61
Can	8.70	6.58	0.76
GC	19.04	8.84	0.46
Dar	25.50	9.90	0.39
Hob	23.77	11.50	0.48
CBD	2.62	10.32	3.94

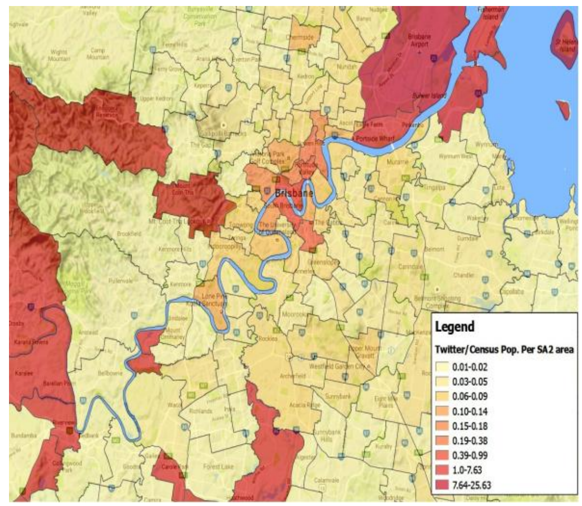


Figure 5. Twitter/Census Population heat map.

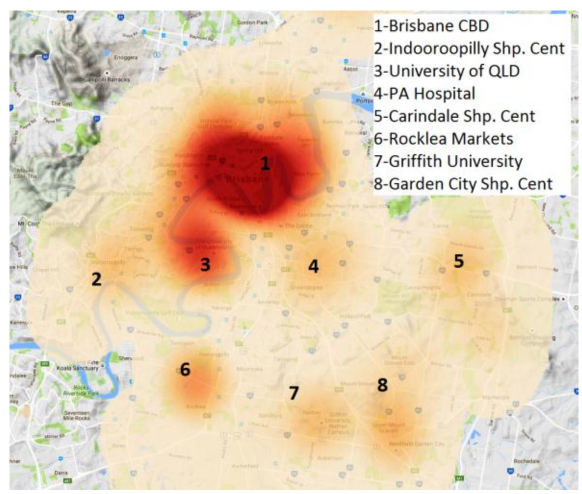


Figure 6. Tweets heat map.

VII. CONCLUSIONS

Various mobility models were examined for Australian cities. A Gravity-2 model with city-specific parameters provides the best results. One city (Darwin) does not fit any model well. Comparing Twitter and census populations has been demonstrated to be useful for identifying mobility hotspots for visitors. Future work is to look at spatiotemporal patterns, i.e. how Twitter mobility varies over space and time rather than the purely spatial patterns examined here.

REFERENCES

[1] D. Balcan, V. Colizza, B. Gonalves, H. Hu, J. J. Ramasco, and A. Vespignani. "Multiscale mobility networks and the spatial spreading of infectious diseases." *Proc. National Academy of Sciences,* vol. 106, no. 51, pp. 21484–21489, 2009.

[2] S. Jiang et al. "A review of urban computing for mobile phone traces: current methods, challenges and opportunities." *Proc. 2nd ACM SIGKDD International Workshop on Urban Computing*, ACM 2013, pp. 21-29

[3] K. Zhao, J. Stehl´e, G. Bianconi, and A. Barrat, "Social network dynamics of face-to-face interactions," *Physical Review E*, vol. 83, no. 5, pp. 056109, 2011.

[4] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, June 2007.

[5] S. A. Golder, and M. W. Macy. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Day length Across Diverse Cultures." *Science,* vol. 333, no. 6051, pp. 1878-1881.

[6] A. M. MacEachren, A. C. Robinson, A. Jaiswal, S. Pezanowski, A. Savelyev, J. Blanford, and P. Mitra. "Geo-Twitter Analytics: Applications in Crisis Management." *Proc. 25th International Cartographic Conference,* pp. 3-8, 2011.

[7] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems,* vol. 61, pp. 115-125, 2014

[8] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. "Understanding human mobility from Twitter." *PloS One*, 10(7): e0131469.

[9] Liu, J. et al. (2015) Multi-scale population and mobility estimation with geo-tagged Tweets. *31st IEEE International Conference on Data Engineering Workshop, BioBAD 2015*

[10] "Australian Bureau of Statistics". [Online]: http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/88F6A0EDEB8879C0CA257801000C64D9, [Sep. 10, 2016].

[11] S. A. Stouffer, "Intervening Opportunities: A Theory Relating Mobility and Distance," *American Sociological Review*, vol. 5, no. 6, pp. 845-867, 1940

[12] H. Makower, J. Marschak and H. W. Robinson, "Studies in Mobility of Labour: Analysis for Great Britain, Part I," *Oxford Economic Papers*, 1939.

[13] G. K. Zipf, "The P_1 P_2/D Hypothesis: On the Intercity Movement of Persons," *American Sociological Review*. vol. 11, no. 6, pp. 677-686, 1946

[14] F. Simini, M. C. Gonzalez, A. Maritan, and A. L. Barabasi, "A Universal Model for Mobility and Migration Patterns," *Nature*, vol. 484, no. 7392, pp. 96-100, 2012.

[15] D. Brockmann, L. Hufnagel, and T. Geisel. "The scaling laws of human travel." *Nature*, vol. 439, no. 7075, pp. 462-465, 2006

[16] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi. "Understanding individual human mobility patterns." *Nature*, vol. 453, no. 7196, pp. 779-782, 2008

[17] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located Twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science,* vol. 41, no. 3, pp. 260–271, 2014.

[18] "Streaming APIs". [Online]: https://dev.Twitter.com/streaming/overview, [Jun. 21, 2013]

[19] A. Guttman. "R-Trees: A Dynamic Index Structure for Spatial Searching". *Proc. 1984 ACM SIGMOD International Conference on Management of Data – SIGMOD*, pp. 47-57, 1984