Estimating Micro-Populations through Social Media Analytics

Richard O. Sinnott, Wei Wang

Abstract— Estimation of crowd sizes or the occupancy of buildings and skyscrapers can often be essential. However, traditional ways of estimation through manual counting, image processing or in the case of skyscrapers, through total water usage are awkward, inefficient and often inaccurate. Social media has developed rapidly in the last decade. In this work, we provide novel solutions to estimate the population of suburbs and skyscrapers – so called micro-populations, through the use of social media. We develop a big data solution leveraging large-scale harvesting and analysis of Twitter data. By harvesting real-time tweets and clustering tweets within suburbs and skyscrapers, we show how micro-populations can be calculated. To validate this we construct linear and spatial models for the suburbs in four cities of Australia using census data and geospatial data models (shapefiles). Our prediction of micro-population shows that Twitter can indeed be used for population prediction with a high degree of accuracy.

Index Terms—E.0.e Knowledge and data engineering tools and techniques, C.5.1.a Super (very large) computers**,** G.3.b Correlation and regression analysis**.**

◆

1 INTRODUCTION

Analysis of crowds and the number of individuals in a given area is extremely important [1]. In sport matches, concerts and public events and demonstrations, analysis of the crowd is needed for management strategies, e.g. in case of unexpected events where evacuation procedures are required [6]. It is the case that there are no scalable solutions to tackle this. In skyscrapers and apartment blocks, it is challenging to truly determine the occupancy rates in a given building. Knowing the number of purchased apartments and bedrooms is only a loose guideline to understanding how many people might be living in a giving building at any given time. Over-occupancy and under-occupancy are daily challenges facing health and safety planners as well as city planners in deciding whether further apartment blocks should be established to meet estimations of the population size. The typical population profile of countries is done through a periodic Census – in Australia this is conducted every five years by the Australian Bureau of Statistics (www.abs.gov.au), however this is often a poor approximation with inaccurate estimates since it does not consider students, tourists, legal and illegal immigrants. It also does not deal with individual the number of individuals at any given time in a suburb or inhabiting a skyscraper. To address this, approximate measures are often adopted, e.g. energy or water consumption for skyscrapers is used as a loose approximation for how many residents might live there at any given time. However this has several disadvantages: (i) inaccessibility: the data for energy and water companies are held by different energy/water companies and is not typically available; (ii) diversity: energy and water utilisation in a given building can be for residential and commercial purposes with different usage profiles; (iii) variance: the amount of energy/water usage for different residents can vary greatly depending on individual habits, e.g. number of baths taken.

To support this, we recognize that there are different types of crowd. Crowds can form where people are gathered together for a specific purpose, such as participating in a political rally or watching a football match. Alternatively crowds or collections of people in a given area more generally can arise. This might be the actual number of individuals living in a suburb at a given time or the inhabitants of a skyscraper. We call these *micro-populations*. But how can we estimate the size or a micro-population at any given time? For sport matches and concerts, we can often obtain the exact number of people through the official attendances, e.g. the people that bought a ticket and went through the turnstyle. For public rallies, we can obtain get a rough estimate given the public area and average area per person and then make use of a grid and the density of occupation if the grid [7, 8]. However, the above two estimation strategies are not generalizable and are only effective in certain situations, e.g. where people are visible. This model does not work with the population of skyscrapers or indeed with the number of inhabitants in a given suburb at a given time. It is quite possible to conduct surveys, but this is time-consuming and prone to human nature/error, e.g. a landlord renting out an apartment in a building to too many students is unlikely to declare the true occupancy of the apartment.

In the past decade, mobile phone and social media have swept the globe and are used extensively by populations as a whole. In 2016, it has been estimated that there will be 2 billion smartphone users. There has been an associated explosion in Internet traffic with social

---

- *Richard O. Sinnott is with the Department of Computing and Information Systems, University of Melbourne, Melbourne, Victoria 3010, Australia. E-mail: rsinnott@unimelb.edu.au.*
- *Wei Wang is with the Department of Computing and Information Systems, University of Melbourne, Melbourne, Victoria 3010, Australia. E-mail: weiw9@student.unimelb.edu.au.*

networking now an indispensable part in the daily life of many people. Platforms such as Foursquare, Facebook and Twitter are used by vast swathes of the population. According to the official Twitter and Facebook sites, there are approximately 320 million monthly active users of Twitter, and 1.04 billion daily active users of Facebook (March 2016). There are over 2.9million users of Twitter in Australia from a population of 24.1million. Given this, a key question we focus on here is can social media be used as a model for population estimation and especially focusing on *micro-populations*. To answer this we leverage the fact that many social media platforms include geospatial information. If a mobile device has the location-based service activated, then the precise location of the Tweeter can be ascertained. Through collecting large amounts of data, we wish to assess whether social media use can be used to accurately assess the size of micro-populations. We focus specifically on suburbs and skyscrapers around the cities of Australia, but the work is generalizable to other scenarios and in other contexts.

The rest of the paper is structured as follows. Section 2 covers related work. Section 3 focuses on the architecture of the systems used to explore mico-population estimation and the associated data. Section 4 focuses on the methodology that was adopted for the analysis. Section 5 focuses on the results and discussions, and finally section 6 conclusions on the work as a whole and outline areas of future work.

2 RELATED WORK

Knowing the size of crowds is important but has hitherto often been based on guesswork [7]. Guesses often exaggerate the crowd size considerably and manual counting is time-consuming, often impossible and is typically unreliable. Although individual head counts are the most accurate model this depends on crowd visibility. It is also especially difficult with moving crowds. Approximations are thus adopted or using sampling techniques [8]. One example is to record when a crowd moves forward passed a fixed point. By counting the number of individuals that pass through at that time and measuring the time for the rest of the crowd to go passed that point, a broad brush estimate of the crowd size can be achieved. This is notoriously error prone since participants may join or leave the parade before or after inspection point. The 'double count and spot-check' approach proposed by Yip et al. [6] solves this. Instead of one inspection point, two points are used to achieve the count. Although this approach adds extra count costs, it shows increased immediacy and efficiency.

An alternative is the 'grid/density' approach. Five steps are needed to support estimation of crowd sizes [9]. First, a best position to observe or photograph the whole crowd is used. Second, a symmetric grid is used to cover the space that the crowd occupies. Third, given a grid cell, the number of people within it is calculated. Fourth, the number of grids, which are filled with people, is calculated. Finally, the crowd size is calculated as the product of the number of grid cells and people within one grid cell. There are several strengths of this method compared to head counts. Head counts are impracticable for a large crowd, and some people are too small to be counted as they are hidden in the crowd. Furthermore, as the crowd density varies among different grid cells of the crowd, this approach can introduce large errors. To allow for density variation, the nebulous boundary of crowds is also not suitable for grid/density-based approaches.

Image processing technology and image recognition has also been explored to estimate crowd sizes. As the most discriminating part of the human body, many researchers use the face as a feature for crowd estimation solutions. Swets et al. [2] make use of a genetic algorithm for object (e.g. face-based) localization through image segmentation. However, the result is highly related to the training set and is limited to only work well when faces are similar sizes and orientations. Li et al. [3] propose a pyramid-detector to detect a multi-view face to tackle this problem. Jones et al. [4] offer another approach by building various detectors based on the rotated face and facial profile views. In order to determine the viewpoint class, a decision tree is trained. Furthermore, a strategy for detecting heads in the crowd has also been proposed [5]. This consists of three steps: get the likelihood information based on the image obtained from the camera; filter and remove unnecessary parts of the body, and finally using a likelihood map and a mean-shift algorithm, the heads and hence head-count can be iteratively extracted. These approaches belong to a local approach as described by Ryan et al. [10] since they utilize detectors and/or local image features within local regions of an image.

Ryan et al. [10] also present two other holistic and intermediate approaches for crowd estimation. In the holistic approach, global image features extracted from frames of a video sequence are utilized and the feature space is mapped to the estimated crowd size through a regression model or classifier. This mapping-based approach needs features, such as edges [12], textures [13] and foreground pixels [11] and then models such as linear regression and neural network are constructed that map relationships between those features for different sizes of crowd. Davies et al. [11] obtained an approximately linear relationship between the crowd size and the number of pixels in foreground or edges. The intermediate approach proposed by Kong et al [14] uses a combination of local and holistic approaches, and makes use of blob size histograms that are reflected at a holistic level. The blob size histogram features are obtained by detecting image edges and subtracting the background. Through supervised learning, they obtain a relationship between the crowd size and the feature histograms.

The previous approaches are based upon visibility of the crowd. This is often not possible. Mobile phone and social media use represents an alternative model, especially when location based services are activated. If we know the locations of users in a crowd, we may be able to estimate the crowd size directly. There are two main methods to obtain location-based information: positioning systems and geotagged data, e.g. social media check-ins. A hybrid positioning system provides a more advanced approach, which combines several positioning technologies, such as cell tower signals, IP addresses, network environment data etc [15].

Botta et al. [28] quantify the crowd size through mobile phone data. They considered whether data such as calls and SMS etc. from mobile phones, could be utilized to estimate crowd sizes. They considered a time interval from 1 November 2013 to 31 December 2013 for the city of Milan. They investigated two special locations: San Siro football stadium and Linate Airport, in order to adjust their model. In both cases the crowd sizes were known based on the attendance at the stadium and to a lesser extent, based on the flight schedules and passengers arriving/departing at the airport.

However, the crowd size given by mobile data or social media is a statistical estimation. It is essential to consider the spatial-temporal and statistical properties of individuals and crowds to accurately estimate crowd sizes. Gonzalez et al. [16] tracked hundreds of thousands of anonymous mobile phone users. They found people conform to simple reproducible patterns, as different from the random trajectories predicted by the Lévy flight [29]. This pattern is predictable according to the findings by Song et al [17]. They give a 93% potential predictability of individual mobility, by analysing the entropy from the trajectory of mobile phone users. Their predictability is lower than for stationary people, but still above 80%. Calabrese et al. [18] also analyse human mobility when attending special events based on cell phones. They found a strong relationship between an event and the people who live nearby. This relationship is directly beneficial to decision makers who often want to manage events and mitigate potential congestion issues.

One of the most popular research areas nowadays, is to make insights through the social media. However, an estimation of the crowd size or furthermore, modelling the crowd, is quite challenging. Firstly, user data in social media is often incomplete, and not all people in a crowd will make posts on social media. Secondly, the size of a crowd usually changes rapidly and has a short lifetime. To tackle these limitations, Liang et al. [19] propose a time-evolving model, where people might join and leave a crowd at any time and location. They make use of the model to predict traffic flows and events in New York based on 22million geotagged check-ins and 120,000 event-related tweets. The result shows that the proposed model can be reasonably effective.

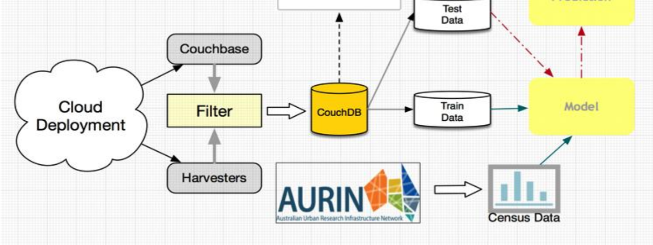
There are two major ways that individuals in the crowd can provide their location: explicitly (event-driven) or implicitly (location-driven). For the former, people officially express their location, e.g. in a tweet text (I am @theMCG) or through check-in on FourSquare for example. For the latter, this can often be captured as metadata based on their phone settings, i.e. they may/may not explicitly wish to provide the exact location of their tweet but this information is sent as a metadata as part of the tweet. This gives rise to a range of privacy issues.

In order to analyse event patterns in three cities, London, Chicago and New York, Georgiev et al. [20] utilize check-in data from Foursquare. They show that there are various forces, including spatio-temporal and social phenomenon that drive one human to attend certain events. Based on these forces, he estimated the potential crowd size for given certain events. An exploration of millions of check-in data from Foursquare by Cheng et al. [21] also unravelled the spatio-temporal and social forces and their relationship focusing especially on user mobility. Scellato et al. [22] studied the socio-spatial properties of social networks through three different online location-based social media such as Brightkite, Foursquare and Gowalla.

One of the most predominant social media resources, Twitter is used globally and has been analysed by many researchers. Real-time tweets with geotagged data benefits geographically grounded situational awareness [23], which is of great importance to many decision-making domains, such as crisis management [25], e.g. earthquakes [24], congestion [39], road accidents [41], health events [40] and pandemics [27]. Through geotagged tweets, events can be detected, e.g. based on the irregular patterns in the number of tweets in certain areas [26]. Sinnott et al. [38] estimate the crowd size at public events through large-scale collections of Twitter data, however this typically involves major events such as major soccer matches where the actual (definitive) number of attendees is known. Dealing with smaller areas and micro-populations remains a challenge and demands big data analytics with advanced spatial capabilities.

3 SYSTEM DATA AND TECHNOLOGIES

Establishing the number of individuals in a micro-population through social media such as Twitter requires the geo-location of the tweeter. A tweet comprises many fields, such as the username, the language, device type, the profile location (e.g. Melbourne), the tweet text and occasionally the actual tweet location. Twitter provides programmatic APIs for data collection: a Search API and a Streaming API. Real-time tweets are accessible through the Streaming API. A returned tweet is in JSON format. In the platform developed (Fig. 1), we support two document-oriented NoSQL databases, Apache CouchBase and CouchDB [37], both of which store tweet data in JSON format. Different from the traditional relational databases, a CouchDB database is a collection of independent documents. CouchDB employs views that can be realized through a variety of languages, of which JavaScript is the default one. The computation can be



done in parallel through MapReduce and leverage data analytics offered through Cloud capabilities.

Fig. 1. Architecture of Implementation

In order to get the tweets within a suburb or a skyscraper, it is necessary to filter the tweets by their coordinates, i.e. whether the coordinates of a tweet fall within a particular polygon. For micro-populations these polygons are suburbs or skyscrapers. The polygons themselves are available as Shapefiles. In this work we utilized the Australian Urban Research Infrastructure Network (AURIN) to access the suburb geospatial information (polygons) encoded as Shapefiles [42].

By utilizing the Twitter APIs, we obtain a subset of the sent tweets in a given area or on a given topic depending on the queries sent. We have developed and deployed twitter harvesters to collect tweets from four major cities in Australia: Melbourne, Sydney, Perth and Brisbane. We combine and process the collected data on the Australia-wide openStack-based research Cloud offered by the National eResearch Collaborations Tools and Resources (NeCTAR – www.nectar.org.au) Research Cloud. In total over 65million tweets were collected over a 6-month period as shown in Table 1.

Table 1. Harvested Twitter Dataset for Cities of Australia

	Melbourne	Sydney	Perth	Brisbane
Original	20,465,251	36,568,680	4,223,247	4,252,048
Geo-tagged	1,433,429	1,903,487	482,049	423,563
Ratio (G/0)	7.0%	5.2%	11.4%	10.0%

The system was designed explicitly to be extensible and all of the software is deployed to support elastic scaling through use of the Boto and Ansible scripting languages. The system that was deployed comprised 8 Virtual Machines (VM) with 32Gb RAM and with 100Gb volume storage attached to each of them. The harvesters themselves were designed to overcome the rate-limiting issues in accessing data from Twitter. This included support for multiple concurrent requests from separate harvesters (on different VMs with different IP addresses). The harvesters themselves used bounding boxes targeted to the areas of interest (the CBDs of the major cities of Australia). The system avoids duplicate tweets through indexing on the tweet Id, and the fact that CouchDB is a version controlled data solution where duplicates are prohibited.

As can be observed in Table 1, the majority of tweets do not include the precise geo-location of the tweeter. Nevertheless a significant amount of geo-located data has been amassed for micro-population estimation. Tweets with geotagged data include coordinates as latitudes and longitudes that need to be mapped to real locations where micro-population estimation is required. Specifically, we need to identify whether a tweet occurred within a given polygon. It is important to note that AURIN provides *live* access to over 2500 data sets from over 70 major and typically definitive data providers from Australia and can be seen as the measure of truth for statically collected official data. This includes official population statistics for a rich variety of geospatial aggregation levels from organisations such as the ABS, e.g. population statistics from Statistical Area Levels (SA4-SA1) amongst many others.

Given the concentration of the population and skyscrapers in the cities, we focus primarily on the central business districts (CBD) and their associated diasaggregation levels and related suburbs. Table 2 shows an overview of the suburbs of the CBDs.

Table 2. Spatial Decomposition of Australian Cities (SA4-SA2)

City	SA4	SA3 (#)	SA2 (#)
Brisbane	Brisbane Inner City	4	33
Perth	Perth - Inner	2	13
Sydney	Sydney – City and Inner South Sydney – Inner South West Sydney – Inner West	10	57
Melbourne	Melbourne – Inner Melbourne – Inner East Melbourne – Inner South	13	87

In addition to the suburbs, we require the building footprints of the major skyscrapers of the cities of Australia. This information is not directly accessible within AURIN but was obtained instead from BBBike (http://extract.bbbike.org), which provides data (shapefiles) for OpenStreetMap. It is noted that the live access to official and definitive data related to building footprints is challenging especially in cities like Melbourne with the very rapid expansion of the city skyline. For each skyscraper we identify the purposes (residential vs office); the height; the number of floors; the number of apartments and the gross floor area (GFA) as shown in Table 3. Using the building footprint, we calculate the number of tweets that have occurred inside the building using standard point-polygon algorithms.

Table 3. Skyscrapers of Melbourne & Sydney and Tweet Counts

City	Skyscraper	Purpose	Ht (m)	Floors	Apartment (#)	GFA (m2)	Tweet (#)
Sydney	Suncrop Place	O	193	48	-	49,954	118
	ANZ Tower	O	195	43	-	55,000	114
	MLC Centre	O	228	60	-	100,000	94
	Chifley Tower	O	216	50	-	90,000	96
	Deutsche Bank Place	O	160	39	-	67,370	38
	Governor Phillip Tower	O	227	54	-	55,000	31
	RBS Tower (Aurora Place)	O	188	41	-	46,500	17
Melbourne	Eureka Tower	R	297	91	560	123,000	859
	Prima Pearl	R	254	72	661	102,000	282
	Melbourne Central Tower	O	211	51	-	123,000	281
	Freshwater Place	R	205	60	536	-	212
	Rialto Towers	O	251	63	-	84,000	195
	101 Collins St	O	195	50	-	83,000	180
	Telstra Corp. Centre	O	193	47	-	63,158	171
	Bourke Place	O	223	49	-	62,407	128
	120 Collins St	O	222	52	-	65,000	58
	568 Collins St	R	224	69	568	-	26

To attempt to validate the estimation of the number of individuals in a given skyscraper, it is necessary to ensure that *known* populations can be accurately predicted. To achieve this we focus on the official Census data from the ABS at the SA2 geographical aggregation level. This equates to a suburb and we develop models to predict the population of a polygon (suburb) based on the number of residential tweeters in that suburb.

It is obvious that a larger population base in a suburb will likely have more twitter users and hence make more tweets. However other factors can impact on the volume of tweets. The age and language distribution within the population also plays an important role in social media analysis. Younger people tend to be more engaged in social media. To address this, we consider the age distribution of suburbs in three groups, 15~29yrs, 30~44yrs and 45~59yrs for each suburb. We also consider the top ten languages spoken at home according to the Census data, which we divide into three groups: English, which is the official language for Australia; Asian language groups, including Mandarin, Cantonese, Arabic, Vietnamese, Hindi and Tagalog, and European language groups, including Italian, Greek and Spanish. We also consider factors such as the Gini coefficient, poverty rate, and median disposable income synthetic estimates. These estimates are produced by NATSEM's Spatial Microsimulation model, which is described further in Tanton et al. [30]. The Gini coefficient is the most commonly used measure of inequality. In order to evaluate the socio-economic conditions, we utilize the poverty rate. This is based on the median disposable income for a given area (suburb). Finally, we consider the influence of unemployment rates on use of social media, from three age groups, 15~24yrs, 25~44yrs and 45~64yrs. All of these criteria can have a direct impact on the estimates of micro-populations and should be factored in to the analysis.

4 METHODOLOGY

Our ultimate aim is to construct a model that can be used to estimate the number of individuals in a micro-population, e.g. individuals living in a skyscraper. This demands statistical modeling and of particular relevance here, it must handle spatial regression.

Regression analysis is a common and useful way to discovery the relationships among a *dependent* variable *Y*, and one or more explanatory or *independent* variables *X1, X2, X3,…* also called *predictors*. Here the independent variables of interest includes the population size, the poverty rate, the Gini coefficient, and the ratio of individuals that only speak English at home, whilst the dependent variable is the number of tweets. All of these independent variables are accessible through the AURIN platform for the suburbs of Australia. After the regression model has been built, it can be used to estimate micro-population sizes.

As the model is based on Twitter data from the suburbs of inner cities and many suburbs are contiguous, we have to consider effects of the spatial correlation when building these regression models. Once the model based and validated on data from suburbs, we are able to apply it to estimate the population of skyscrapers or indeed other micro-populations.

There are many possible regression models. Here we initially apply a linear model to explore the relationship between the suburb population and the number of geotagged tweets from users. In doing this, it is essential to factor in potential errors or noise in the model. More formally, given a suburb population P, a number of tweeters U, and a given noise term ε , an estimate α of the relationship between these variables is given by:

$$U=\alpha P+\varepsilon \qquad (1a)$$

Through ordinary least squares (OLS), we can calculate the estimate $\hat{\alpha}$, also known as the best linear unbiased estimator. We have two assumptions about the variables: (i) the dependant variable should be normally distributed; (ii) independence should exist in observations and residuals. A few standard assumptions about the error term ε also need to be made: (i) the expected value of the error is zero, namely, there is no systematic bias in equation (1a); (ii) the random error is also independent, i.e. it is not correlated, and its variance should be constant (homoscedasticity), and (iii) the random error is normally distributed.

In order to stabilize the variance, we make a logarithmic transformation to the above equation. This is the most widely used approach to achieve normality when the variable is highly skewed as shown in equation 1b:

$$logU=log\alpha+logP+log\varepsilon \qquad (1b)$$

which we simplify to:

$$logU = \beta_0 + \beta_1 logP + \varepsilon' \qquad (1c)$$

From equation (1c), if $\hat{\beta}_1 = 1$, then we can say, U is linear to P with a constant proportion $\exp(\hat{\beta}_0)$. Meanwhile, another extreme is the null hypothesis $H_0 : \beta_1 = 0$, which means there is no relationship between the population and the number of tweet users, i.e. the tweeters are just a random sampling from the whole population. In case that the value of P and U is zero and we make a modification to the equation (1c):

$$log(1+U) = \beta_0 + \beta_1 log\,(1+P) + \varepsilon' \qquad (1d)$$

We also attempt to make similar models regarding the relationship between the tweeters and the Census data respectively, including the Gini coefficient, the poverty rate, the English-speaking rate, the age and the unemployment rate distribution of different age groups. Finally, we get equation (1e), where $X_1, X_2 \ldots$ are the explanatory variables from the Census.

$$\begin{aligned} log(1+U) = \beta_0 + \beta_1\, &log(1+X_1) \\ +\beta_2\, &log(1+X_2) + \cdots \varepsilon\,' \qquad (1e) \end{aligned}$$

For equation (1e), another important assumption is required: predictors should not be strongly correlated to each other, i.e. there is no multi-collinearity. With this linear regression for a predictor, we assume that the observations are independent of each other. For example, for the two suburbs in Melbourne, Carlton and Collingwood, the unemployment rate for the age group 25~44 is 9.74% and 11.66% respectively. However the fact is that these two suburbs are adjacent geographically and this can bias the analysis, e.g. unemployment rates for two adjacent suburbs are unlikely to be independent (highly affluent areas are rarely adjacent to poorer areas)? Addressing this requires spatial correlations and associated analytics to be considered.

Spatial effects can occur for two reasons: one is the collected data from areas do not reflect the actual properties of those areas, and the other is that the spatial autocorrelation itself plays an important role in modelling. This relationship can be positive, i.e. the observational values are more similar in nearby locations, or negative, i.e. more dissimilar values appear in nearby locations. To tackle this we need to establish a weights matrix, which stores the relationships between spatial units. The weights matrix can be specified in many ways:

- *Fixed distance*: all spatial units within a specified distance of each unit are included, and the units outside the critical distance are excluded;
- *Inverse distance*: the impact of one spatial unit on another one decreases with distance;
- *K-nearest neighbors*: the closest k units are included;
- *Queen contiguity*: polygon units that share a boundary and/or share a node are neighbors;
- *Rook Contiguity*: polygon units that share a boundary are neighbors;
- *Delaunay triangulation*: a mesh of non-overlapping triangles is created from unit centroids, and units associated with triangle nodes that share edges are neighbors;
- *Convert table*: Spatial relationships are defined in a table.

We consider three weights matrices: k-nearest neighbors, queen and rook contiguity. The weights matrices are constructed from polygon boundaries (shapefiles). For the k-nearest neighbors weights, given a spatial unit i, and other units j, we can get centroid distance d: d_{ij}. With these distances, and given parameter k, we can get the k nearest units for each unit i. Finally we obtain the matrix W, with spatial weights of the form:

$$w_{ij} = \begin{cases} 1, & j \in N_k(i) \\ 0, & otherwise \end{cases} \tag{2a}$$

Alternatively, we can consider a symmetric version of equation (2a):

$$w_{ij} = \begin{cases} 1, & j \in N_k(i) \; or \; i \in N_k(j) \\ 0, & otherwise \end{cases} \tag{2b}$$

Consider the example shown in Fig. 2 showing the relationship of the four-nearest neighbors among Inner Melbourne suburbs.

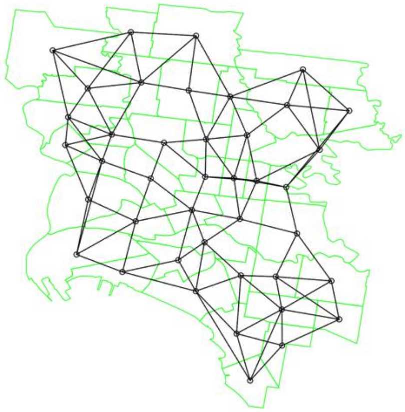


Fig. 2 Four-nearest neighbors for Inner Melbourne suburbs

For *queen contiguity weights*, if the set of boundary nodes of unit i are denoted by $bnd(i)$, then the weights are defined by:

$$w_{ij} = \begin{cases} 1, & bnd(i) \cap bnd(j) \neq \emptyset \\ 0, & bnd(i) \cap bnd(j) = \emptyset \end{cases} \tag{2c}$$

However, (2c) considers spatial units that share only one single boundary point, which is not a strong condition. The *rook contiguity weights* solved this inadequacy, where l_{ij} is defined to denote the length of the shared boundary, $bnd(i) \cap bnd(j)$, between unit i and j. This is given as:

$$w_{ij} = \begin{cases} 1, & l_{ij} > 0 \\ 0, & l_{ij} = 0 \end{cases} \tag{2d}$$

There are many spatial autocorrelation measurements, with the same origin based on a general statistical cross-product [31, 32] of the following form:

$$\Gamma = \sum_i \sum_j w_{ij} p_{ij} \tag{3a}$$

where w_{ij} is the spatial weights matrix mentioned above, and p_{ij} is used to describe the neighborhood relationship from other aspects, e.g. the Manhattan distance. Two common measures derived from (3a) are Moran's I and Geary's C. In the process of studying stochastic phenomena, Moran [33] defined the Moran's I measurement, which has since been applied to several spatial autocorrelation problems. The formula is given as:

$$I = \frac{n}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3b}$$

Similarly, w_{ij} is the spatial weights matrix, and $(x_i - \bar{x})(x_j - \bar{x})$ depicts the proximity p_{ij}. The value of I ranges from -1 to 1. Geary's C [34] is defined as (3c), with values varying within [0,2]. Table 4 lists the positive-, negative- or non-autocorrelation, given the values of Moran's I and Geary's C.

$$C = \frac{(n-1)}{2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3c}$$

In most situations, both Moran's I and Geary's C are praticable, and suitable for global and local conditions respectively. However, Cliff and Ord [35] identified that to obtain consistent progress, Moran's I is preferred over Geary's C. In this work, we apply both approaches.

Table 4. Moran's I and Geary's C

Test	**Negative**	**Non**	**Positive**
Moran's I	[-1, -1/(n-1))	-1/(n-1)	(-1/n-1, 1]
Geary's C	(1, 2)	1	[0, 1)

There are three different kinds of spatial autocorrelation typically existing within spatial data. The first one happens within the dependent variable, which is often termed the spatial lag of Y. The second is the autocorrelation of the predictor, namely the spatial lag of s. The last is the spatial error, which exists within the residuals of the model, $Y{\sim}X$. Considering the spatial error, our assumptions about the error in standard regression identified previously are not all tenable, and given spatial lags, the observations are also not independent of one another. That is to say, if the spatial autocorrelation exists (it does!) the result of the standard regression will be inaccurate. Therefore, a spatial model needs to be constructed. This in turn requires autocorrelation tests.

Our tests use the R package *spdep* and include Moran's I, Geary's C and Lagrange Multiplier tests. We make three tests on the OLS regression model. Both Moran's I and Geary's C test, implemented through *moran.test and geary.test* can be used to test the spatial autocorrelation of dependent variables and predictors. Residual autocorrelation can be tested by *lm.morantest.* Lagrange Multiplier test (*lm.LMtests*) can be used to detect both spatial error and spatial lag. The Lagrange Multiplier test also suggests a distinction of the underlying spatial models.

In addition to spatial autocorrelation, another spatial effect is spatial heterogeneity. Spatial autocorrelation, or spatial dependence, is related to the notion of relative space or location, where neighbouring spatial units are more alike than faraway ones. Spatial heterogeneity shows the varying relationships across space. There are two types of heterogeneity, structural instability and heteroscedasticity. In structural instability, parameters are different for the same model for different spatial units. For heteroscedasticity, various error variances exist across space. However, as we build models for different regions of the cities, we do not consider spatial heterogeneity.

If spatial autocorrelation occurs in the OLS regression model, it is necessary to re-estimate it through a spatial regression model. There are three types of spatial regression models needed for our work: (i) the spatial lag of X model (SLX model) [36], (ii) the spatial autoregressive model (SAR model), and (iii) the spatial error model (SEM model).

The SLX model presumes that the dependant variable Y is influenced by the explanatory variables $X_1, X_2, \ldots, X_k$ and their spatial lags $LX_1, LX_2, \ldots, LX_k$. That is to say, Y is not only determined by predictors in the same spatial unit, but also by predictors in nearby units. This is given as:

$$Y = \alpha\iota_N + \beta_1 X_1 + \cdots \beta_j X_j + \cdots + \beta_K X_K + \theta_1 WX_1 + \cdots + \theta_j WX_j \ldots + \theta_K WX_K + \varepsilon \tag{4a}$$

where Y is the dependent variable, an $N \times 1$ vector, while X_j is an $N \times 1$ vector of the explanatory variable. ι_N is an $N \times 1$ vectors of 1s, that represents the constant parameter α. W is an $N \times N$ spatial weights matrix mentioned above, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^T$ is a vector of errors, where ε_i is assumed to meet the standard assumptions for a linear regression model discussed previously. β_j and θ_j ($j = 1,2, \ldots K$) are regression coefficients of the exogenous variables and spatial lags. A more compact form of (4a) is:

$$Y = \alpha\iota_N + X\beta + WX\theta + \varepsilon \tag{4b}$$

Actually, the SLX model is a special case of Spatial Durbin Model (SDM), with $\rho = 0$ in the formula (4c):

$$Y = \rho WY + \alpha\iota_N + X\beta + WX\theta + \varepsilon \tag{4c}$$

A spatial autoregressive process is defined by:

$$Y = \rho WY + \alpha\iota_N + \varepsilon \tag{5a}$$

In a pure SAR process, it is assumed that spatial correlation is described by the spatial lagged variable WY. But in empirical work, a pure SAR process seldom exists. In regional science models, several geo-referenced variables are usually incorporated. In this case, the basic SAR model (5a) can be defined as:

$$Y = \rho WY + \alpha\iota_N + X\beta + \varepsilon \tag{5b}$$

Again, let $\theta = 0$, we can obtain (5b) from (4c). In R, we carry out SAR model, or spatial lag model, through maximum likelihood estimation, with the *lagsarlm()* function. By solving (5b) for ε, we get:

$$\varepsilon = (I - \rho W) \cdot Y - X\beta - \alpha\iota_N \tag{5c}$$

The key to calculate the maximum likelihood is to calculate the value of the Jacobian:

$$J = \left|\frac{\partial\varepsilon}{\partial Y}\right| = |I - \rho W| \tag{5d}$$

In a standard multiple regression model, the formula is given by (6a), similar to (1a) but in a matrix form:

$$Y = \alpha\iota_N + X\beta + u \tag{6a}$$

Here, u is an $N \times 1$ vector of errors, but different from the standard regression model (1a), the errors are not assumed to be identically distributed. In contrast, u follows a spatially autocorrelated process, and is defined by:

$$u = \lambda Wu + \varepsilon \tag{6b}$$

But when using maximum likelihood to estimate the spatial error model, the error term ε must follow normal assumptions. From (5a) and (5d), we get the final formula for the SEM model:

$$Y = \alpha\iota_N + X\beta + (I - \lambda W)^{-1}\varepsilon \tag{6c}$$

This is still a special case of the Spatial Durbin Model, with $\theta = -\rho\beta \; and \; \lambda = \rho$. By solving (5c) for ε, we get $\varepsilon = (I - \lambda W)(Y - \alpha\iota_N - X\beta)$, where again we can obtain the Jacobian (5d).

Having introduced all relevant basic mathematical theory for constructing models the micro-population estimation is based on three steps:

- Run a standard linear regression;
- Test OLS Residuals for spatial correlation, including Moran's I, Geary's C and Lagrange Multiplier tests, and
- Estimate the spatial model via the maximum likelihood, including the SAR model, the SEM model, the SDM model, and a mixture of the three models.

5 RESULTS AND DISCUSSION

The analysis and result consists of three parts: observational results from intuitive observation of raw data and simple processed data, e.g. graphs and scatter plots; models established through linear and spatial regression; and finally, predictions derived from the selected models.

5.1 Observations from Raw Data

Observations are based on two types of data, tweet data and official Census data. For tweet data, we identify unique users with focus on residential users versus passers-by or tourists. As noted, preprocessing of the data provides geotagged tweets that are aggregated to the suburbs of inner cities (as listed in Table 2) through polygon filters. However in the prediction models, some suburbs cannot be taken into consideration due to their inherent noise. Table 5 lists examples of some of the noisy suburbs that were found in the cities of Australia. These suburbs include industrial zones or public places, either with no population, or only a small population in comparison to the volume of tweets. For example, Kings Park (WA) has an official population of 34, but it has 2293 tweets. Obvious these tweets predominantly arise from people visiting the park and not from residents. Similarly, Sydney Airport has excessive amouns of tweets from people who are non-residents (at least they don't live at the airport!). Thus, in establishing the model part we exclude these suburbs.

Table 5. Noisy Suburbs

City	Suburb	Tweet (#)	Pop. (#)
Perth	Kings Park (WA)	2293	34
Sydney	Sydney Airport	27842	15
	Banksmeadow	247	14
	Port Botany Industrial	122	0
Melbourne	Flemington Racecourse	4268	86
	Port Melbourne Industrial	1796	15
	West Melbourne	1700	0

We also cannot use *all* tweets directly since users of Twitter consist of numerous different entities: people, organisations, businesses and advertisers. Furthermore, many twitter users don't tweet very often. This is the so-called long tail phenomenon. Fig. 3 shows the logarithmic plot of the number of tweets for given users for inner city Melbourne. The same phenomenon exists for all of the selected cities.

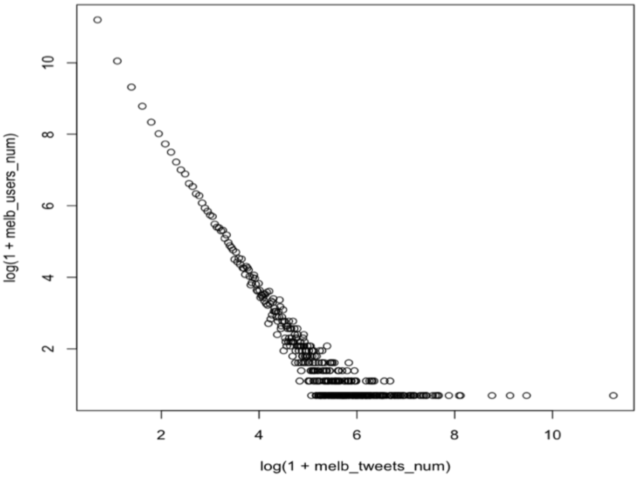


Fig. 3 Number of Melbourne Tweets per Tweeter

The long tail phenomenon suggests that utilizing the number of tweets directly is inadvisable. Table 6 lists a few abnormal tweeters, i.e. they post far more than other tweeters. These tweets are often sent from official Twitter accounts. Such tweeters also have to be removed from the total tweeter base to ensure that the model is not biased.

Table 6. Non-residents (atypical) Tweeters

City	Suburb	Tweet Id	Tweet Name	Tweet (#)	Purpose
Brisbane	Brisbane City	2439997368	@bikewatchbne	12899	service
	Brisbane City	132038515	@trendsbrisbane	1894	topic
Perth	Perth City	132035314	@trendsperth	1573	topic
	Perth City	256410521	@iamMariza	1541	ad
Sydney	Sydney Haymarket - The Rocks	123791748	@trendssydney	13376	topic
	Sydney Haymarket -The Rocks	211362581	@trendsaustralia	8393	topic
Melbourne	North Melbourne	317680173	@will_i_ammg	76512	weather
	Kensington	2836002698	@3031weather	12964	weather

In order to tackle this long tail problem, we have to not only consider the total number of tweets, but also the number of tweets from *unique* twitter users. Table 7 lists the number of unique users from the suburbs of inner city Sydney. The column tweets (<3) is the number of unique users in each suburb that have posted less than three tweets.

Our hypothesis here is that many of these people are likely to be tourists or people passing through that area and not residents of the suburb. However given people can tweet just a few times, we do not simply count the (<3) column as the potential tweeter residents. Instead we calculated the potential residential tweeters by the unique column minus half of the tweets (<3) column.

The value 3 was selected based on a reasonable assumption of how many tweets could remove background noise, e.g. tweets from tourists at a given location.

Table 7. Unique Tweeters of Inner City Sydney

Suburbs	Tweet (#)	Unique	Twitter (<3)	Filtered
Botany	4146	317	224	205
Darlinghurst	23758	4892	3473	3155
Erskineville - Alexandria	16422	3765	2802	2364
Glebe - Forest Lodge	14884	2343	1687	1499
Marrickville	10822	2045	1446	1322
Mascot - Eastlakes	8564	1277	969	792
Newtown-Camperdown-Darlington	34222	6238	4245	4115
Pagewood - Hillsdale - Daceyville	2810	502	380	312
Petersham - Stanmore	11772	1953	1442	1232
Potts Point - Woolloomooloo	23927	4657	3272	3021
Pyrmont - Ultimo	36022	6654	4617	4345
Redfern - Chippendale	36770	6364	4335	4196
Surry Hills	31868	7031	4826	4618
Sydenham - Tempe - St Peters	5165	1448	1128	884
Sydney - Haymarket - The Rocks	246442	32066	18422	22855
Waterloo - Beaconsfield	17917	2461	1701	1610

In establishing the prediction model, we assume that the dependent variable, i.e. the unique residential tweeters, is normally distributed. However, this is not the case. It follows a highly skewed distribution. To tackle this, we make a logarithmic transformation as shown in Fig. 4.

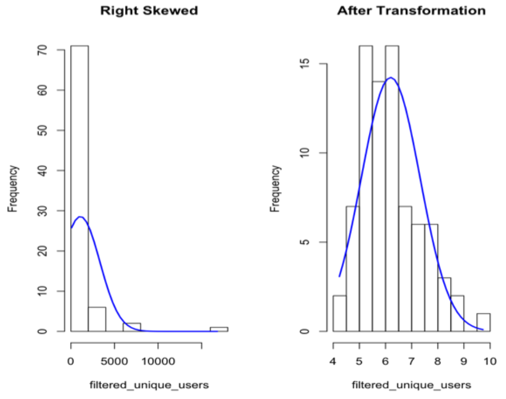


Fig. 4 Normalising the Distribution of Unique Tweeters

We also need to consider the potential mechanisms in selecting the Census data that can influence tweet numbers. To aid in this process we assess the relationship between the Census data and the unique twitter users. Fig. 5 and Fig. 6 show scatterplots of Melbourne data (80 data points) for each inner city suburb including, the unique twitter users (*U*), the population (*PP*), the poverty rate (*PR*), the Gini coefficient (*GC*), the median disposable income (*MDI*), the three language groups (English group (*ENG*), European group (*ERG*) and Asian group (*ASG*)), three age groups (age group youth (*AGY*), age group middle (*AGM*), age group old (*AGO*)), and the three age distributed unemployment groups: *UEY*, *UEM* and *UEO*. It should be noted that all the variables are in the form of $\log(1 + variable)$. The plots themselves are based on a linear regression.

From the first column, we see positive relationships between *U* and *PR*, *GC*, *ASG* respectively, and a negative relationship between *U* and *ENG*, *ERG* respectively, and no apparent relationship between *U* and *MDI*, *PP* respectively. We would expect that larger populations produce more unique twitter users, but this seems not to be the case. Inequality and poverty also impacts on this. In contrast, lower English language speaking rates represent more diversity, and further bring more unique twitter users. From the Census data, we can see that lower *ENG* exists in various suburbs, such as 47.6% in Brisbane City, 50.1% in Perth City, 25.7% in Sydney – Haymarket – The Rocks, and 29.0% in Melbourne. The four suburbs are the locations of the CBD. Just considering the English-speaking rate, we know that Sydney is the most advanced city, and Perth the least. Similar to the English language group, the European group shows exactly the same phenomenon. However, for the Asian group, the situation is reversed. An increase in *ASG* results in a larger number of twitter users.

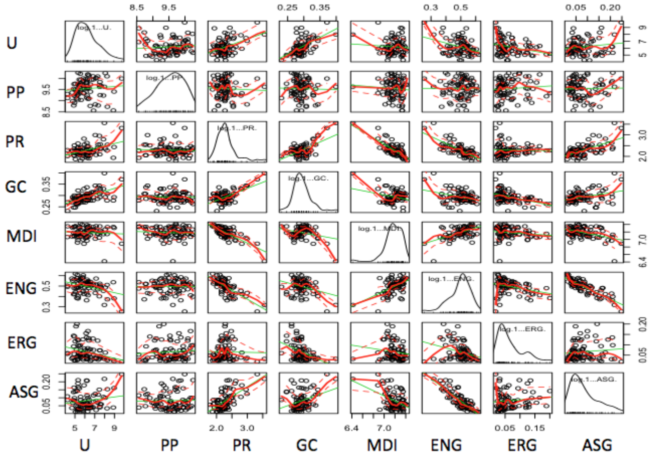


Fig. 5 Scatterplot of Melbourne Data Part I

From the second column, we observe no apparent relationships between *PP* and the other variables. Actually, the size of the population doesn't have much effect on inequality, poverty and language distribution. Hence we can conclude that they are independent from the population. Comparing poverty and language distribution, we see a strong relationship, that is to say, if we want to build a linear model, we should not consider all of the variables since one variable can be linearly represented by other variables. Thus, we need to handle these variables carefully when constructing the combined model.

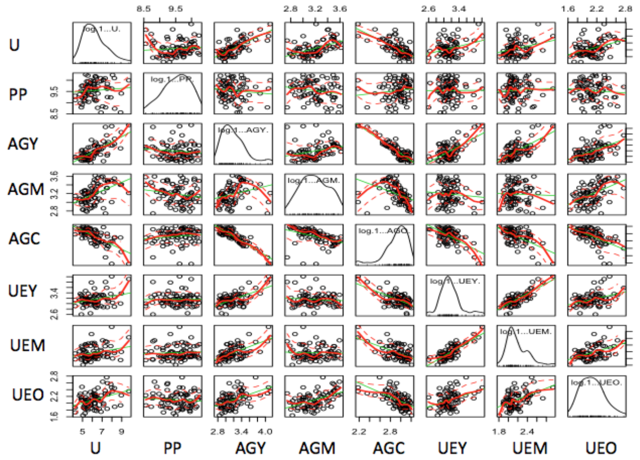


Fig. 6 Scatterplot of Melbourne Data Part II

Figure 6 gives the scatterplot of Melbourne data, including U, PP, three age groups (AGY, AGM, AGO), and three unemployment groups (UEY, UEM, UEO). Positive relationships exist between U and AGY, AGM, and the three unemployment groups respectively, whilst negative relationships exist between U and AGO. This means that, young and middle aged people are more devoted to Twitter, while old are less so (as expected). For the three unemployment rate groups, there is a general indication that areas with larger unemployment rates have larger numbers of Twitter users. Furthermore, we also see a strong negative relationship between AGY and AGO, so if we construct a linear model, we only need to consider one of these variables. Since there are no obvious relationships between population and the age groups, we can treat them as independent variables. Similar results can be obtained for other three cities.

5.2 Linear and Spatial Regression Models

Table 8 lists the relationships between twitter users (U) and other provided variables for all four cities. We can conclude that, there are positive relationships between U and PP, PR, AGY respectively, and negative relationships between U and ENG, AGO respectively, and no relationship between U and ERG.

Table 8. Relationships between Tweeters (U) and Other Variables

City	Positive	Negative	Neutral
Brisbane	*PP, PR, GC, MDI, AGY*	*ER, AGO*	*AGM, ERG*
Perth	*PP, PR, AGY, AGM*	*ER, MDI, AGO*	*GC, ERG*
Sydney	*PP, PR, GC, AGY, AGO*	*ER, AGO*	*MDI, ERG*
Melbourne	*PP, PR, GC, AGY*	*ER, AGO*	*MDI, AGM, ERG*

In the following model part, we study these relationships from a statistical aspect and use the models for estimating the micro-population of the inhabitants of skyscrapers.

5.3 Predictions from Spatial Models

We consider the skyscrapers from the four cities all together. Consider first the heights of skyscrapers versus the number of tweets within them as shown in Fig. 7.

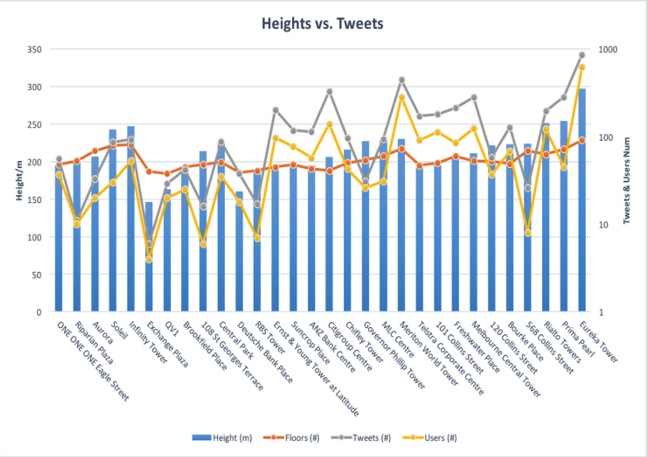


Fig. 7 Correlation between Skyscraper Height and Tweeters

Fig. 7 contains all thirty skyscrapers with from left to right: five for Brisbane, five for Perth, ten for Sydney and ten for Melbourne. Within each city group, the height increases from left to right. In almost every city, the number of floors is (as expected!) directly proportional to the height of the skyscraper. The variation in trend of the number of users keeps pace with the number of tweets within each skyscraper. As each city has its own character and spatial heterogeneity, we only make comparisons of skyscrapers within that city.

In general, the number of tweets is directly proportional to the height of the skyscraper, with some exceptions: Riparian Plaza in Brisbane, 108 St Georges Terrace in Perth, Governor Phillip Tower in Sydney, 120 Collins Street and 568 Collins Street in Melbourne. There are various reasons for these exceptions that we explain through Fig. 8 and Fig. 9.

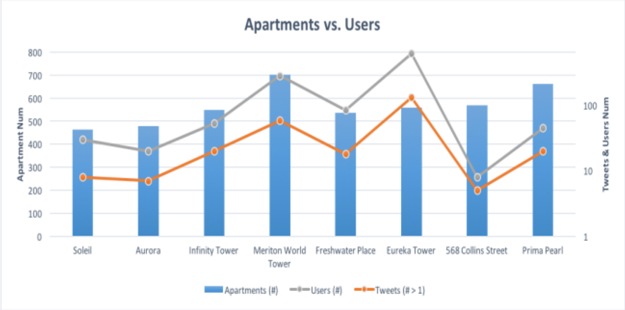


Fig. 8 Correlation between #Skyscraper Apartments and Tweeters

Fig. 8 visualizes the relationships between the number of apartments in a skyscraper and the Twitter users in those skyscrapers. From left to right, the first three skyscrapers are in Brisbane, Meriton World Tower in Sydney, and the rest are in Melbourne. It should be noted that no residential skyscrapers are in Perth. As a residential skyscraper, we only consider the users that post tweets more than once as residents. As seen, the number of users in a skyscraper is directly proportional to the number of apartments within it. The exceptions to this are the Melbourne skyscrapers: Eureka Tower and 568 Collins Street. For the Eureka Tower, the reason lies in that it is the highest skyscraper in Melbourne and an attraction with a skydeck that attracts a large number of people and tourists. For 568 Collins Street, it was completed in 2015.

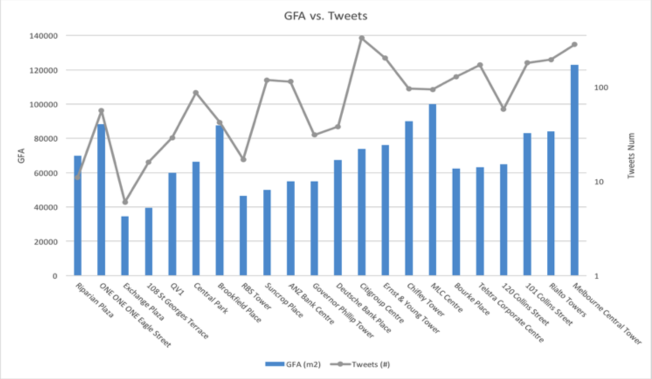


Fig. 9 Correlation between GFA and Tweeters

Fig. 9 visualizes the relationship between the gross floor area (GFA) and the number of tweets. From left to right, the first two skyscrapers are in Brisbane, followed by five skyscrapers in Perth, nine skyscrapers in Sydney, and then finally six skyscrapers in Melbourne. In Brisbane, Perth and Melbourne, a directly proportional relationship exists between GFA and the number of tweets, with the exception of 120 Collins Street. One reason for this anomaly is that this skyscraper is home to a lot of high-profile commercial tenants like the Rio Tinto Group, Bank of America Merrill Lynch etc, i.e. it is not just residential.

From Table 8, we have a glimpse of the possible relationships between tweeters (U) and explanatory variables, however a more rigorous statistical analysis is required. Analysis of spatial autocorrelation on the above models needs to be handled to discover potential spatial effects. To address this we build spatial models based on the linear models and provide spatial autocorrelation findings. Fig. 10 shows the relationship between the suburb population and the number of twitter users for inner city Melbourne.

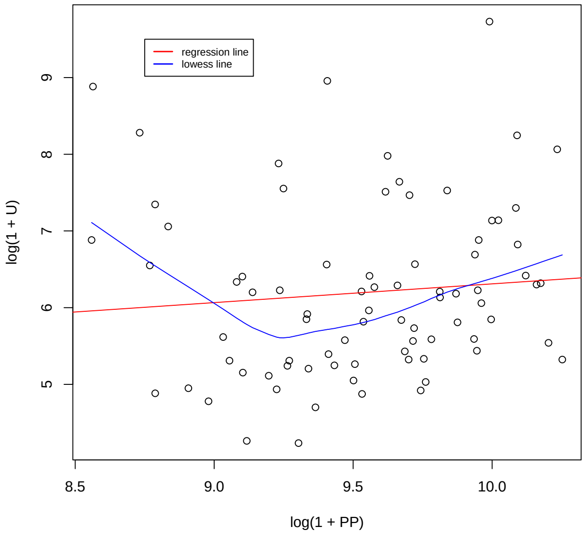


Fig. 10 Melbourne Suburb Population vs. Twitter Users

With the fitted regression line, we see no obvious linear regression, and indeed this model with a p-value of 0.4174 means that the model is not trustworthy. Actually, we can see an obvious line after excluding the points in the top left part of the figure, which is shown as the left segment of the locally weighted scatterplot smoothing (lowess) line. The right segment of the lowess line shows a very good regression.

As anomalies can exist in the population data, we rarely see a linear relationship between suburb populations and twitter users. Hence we further include potential predictors into the model. We first consider all of the available predictors and get a p-value less than 2.2e-16, i.e. the predictor models fit very well to the data. Table 9 lists the significance of the different predictors and the dependent variables as $\log(1+U)$.

Table 9. Linear Regression for all Predictors

Predictor	**Standard Error**	**Pr(>t)**
(Intercept)	7.9082	0.03575 *
$\log(1+PP)$	0.1612	7.18e-05 **
$\log(1+PR)$	0.7007	0.49579
$\log(1+GC)$	4.5670	0.00525 **
$\log(1+MDI)$	0.8967	0.02925 *
$\log(1+ENG)$	3.1477	0.04723 *
$\log(1+ERG)$	2.3890	0.17172
$\log(1+ASG)$	3.2824	0.22140
$\log(1+AGY)$	0.7192	0.13016
$\log(1+AGM)$	0.5022	0.00234 **
$\log(1+AGO)$	1.0649	0.24605
$\log(1+UEY)$	0.5228	0.59319
$\log(1+UEM)$	0.8141	0.30990
$\log(1+UEO)$	0.4350	0.18116

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5754 on 66 DF
Multiple R-squared: 0.78,
Adjusted R-squared: 0.7366
F-statistic: 18 on 13 and 66 DF
p-value: < 2.2e-16

All predictors are inadequate as seen from the significance codes. There are only three predictors that show any significance: the population, Gini coefficient and middle-aged people. Taking all predictors is thus unwise. On the one hand, this model is too complex with many variables, whilst on the other hand some of the variables are dependent on one another. For example, the strong negative relationship between young people and old people mentioned previously. However, compared to (1c), this model is a big step forward, as we not only find that the population makes an effect together with the occurrence of other variables, but we also reveal some significant variables such as the importance of the Gini coefficient and middle-aged people.

Models with one or many predictors are not ideal. Thus we need to find the best combination of predictors in order to get the best performance from a statistical sense. Our approach is based on AIC (Akaike Information Criterion) [43], as a means of model selection that is constructed from information theory. Given a data set and several models, it estimates the information lost by each model, making a compromise between goodness of fit and complexity for that model. There are three choices to make a model selection: backward, forward or both. As we start from a full model, backward and both are optional, and the result is the same. Table 10 lists the whole process of model selection from backward AIC.

Table 10. Backward AIC

Step	Model	AIC
0	Full model	-75.81
1	- log(1+***UEY***)	- 77.466
2	- log(1+***PR***)	- 79.063
3	- log(1+***AGO***)	- 79.644
4	- log(1+***UEO***)	- 80.511

Four predictors are filtered, *UEY*, *UEO*, *PR* and *AGO*. Intuitively, this can be explained as follows: the unemployment rate of young and old people doesn't greatly influence the number of twitter users, since they are at either the start or the end of their career. Second, as we consider the inner city of Melbourne, the poverty rate doesn't change a lot excluding only a few anomalous suburbs. Finally, it is known that old aged people have little uptake of social media compared to young and middle-aged people. However the predictor is still too high. Considering the F-test and T-test[t] shown in Table 11, we further filter four predictors, *MDI*, *ERG*, *ASG* and *UEM*, which are of little importance. One question is why are the European and Asian language groups not that important compared to the English group? We believe that these two language groups are a subset of languages spoken at home, and many other languages belonging to these groups are not considered. Even adding all languages into the same group, they cannot be compared to English, which changes considerably across suburbs and hence is a good predictor.

Table 11. F-test and T-test of Best AIC Model

Predictor	F value	Pr(>F)	T-value	Pr(>|t|)
log(1+***PP***)	2.5794	0.1127652	4.146	9.33e05 ***
log(1+***GC***)	101.1495	3.150e-15 ***	3.711	0.000411 ***
log(1+***MDI***)	2.9484	0.0903818 .	2.216	0.0299-11 *
log(1+***ENG***)	20.0806	2.826e-05 ***	-2.734	0.007910 **
log(1+***ERG***)	8.3041	0.0052477 **	-1.453	0.150557
log(1+***ASG***)	52.0538	5.061e-10 ***	-1.662	0.100994
log(1+***AGY***)	29.4758	7.708e-07 ***	4.243	6.65e-05 ***
log(1+***AGM***)	15.3892	0.0002019 ***	3.813	0.000292 ***
log(1+***UEM***)	3.2903	0.0739748 .	-1.814	0.073975 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Finally we have five predictors: *PP*, *GC, ENG*, *AGY* and *AGM*, reflecting population, inequality, language and age as key influential factors in use of social media. Therefore, our final linear regression model is a form of formula (1e). Table 12 lists the final linear regression formulas for the four cities.

Table 12. Final Linear Regression Formula

City	Formula
Melbourne	*U~PP+GC+ENG+AGY+AGM*
Sydney	*U~PP+AGY+AGM+UEM*
Brisbane	*U~PP+ASG+AGM*
Perth	*U~PP+GC+AGM*

Before constructing the spatial model, we need to determine whether spatial autocorrelation exists within the dataset. As mentioned, this requires construction of a spatial weights matrix. For every type of weights matrix, many coding schemes, basic binary, row standardised, global standardised exist. Taking the basic binary weights for example, the weight is one or zero, in which two spatial units are listed as neighbors or not. Given these weights matrix, we can make tests of spatial autocorrelation on a given variable using Moran's I and Geary's C values. This variable can include dependent variables, predictors, or even residuals of a linear regression model.

Our focus is on the variables whose Moran's I value is bigger than 0.5, and Geary's C value less than 0.5. As all our linear models are associated with the dependent variable $\log(1+U)$ (the Tweeters), which is actually strongly positive spatial autocorrelated, there would be a large error in the linear models. Therefore, we must take the spatial autocorrelation into the models that is to say we have to add the spatially lagged variables, including dependent and explanatory ones, into our model, which becomes consequently a SDM or a SAR model.

In Fig. 10, we see no obvious linear regression, because the residuals of $\log(1+U) \sim \log(1+PP)$ are strongly spatially correlated, thus a simple linear regression is far from sufficient. In the linear model, we see little relationship between unique tweet users and language groups, or with unemployment groups, however we see strong spatial correlation in the residuals. In comparison, no obvious spatial correlation exists for the age groups. Thus, adding spatially correlated error terms into our model is meaningful and indispensable, which is exactly what the SEM model achieves.

Up to now, we have illustrated potential spatially autocorrelated lagged variables and residuals of the model. With those autocorrelations, we can choose variables to construct appropriate spatial models, which can eliminate the effects of autocorrelation. We also will employ the Lagrange Multiplier test statistics, which allows a distinction between spatial error models and spatial lag models.

We observe spatial autocorrelation in dependent variables for the SAR models; in explanatory variables for the SLX models, and the SDM models for both of them. We also observe spatial autocorrelation in residuals for the SEM model.

Considering the basic linear regression model, $\log(1+U) \sim \log(1+PP)$ we calculate the correlation value of 0.092, which shows minimal linear relationship. Meantime, the p-value gives 0.4174, thus we cannot accept the hypothesis that $\log(1+U)$ is linear to $\log(1+PP)$. However, if we consider the spatial autocorrelation independent variable $\log(1+U)$, where the autocorrelation exists as illustrated above, we see a big difference. By adding the lagged dependent variable, the model becomes $\log(1+U) \sim \rho W \log(1+U) + \log(1+PP)$, which is actually a SAR model. Using the *lagsarlm* function built in the *spdep* package we obtain 0.81551 for the spatial autoregressive parameter (ρ), with p-value less than 2.22e-16 on an asymptotic T-test, which shows a high significance.

For comparison of the linear regression and the spatial lag models, we construct spatially lagged dependent variables using the *lag.listw* function. Including the spatial lag as a predictor in the linear regression model we observe that lagged Y OLS is much better than OLS model, no matter the significance of predictors or fit of data. However, the AIC value of this lagged OLS model is 152.2112, which is only a little better than the OLS model. As the fit of the lagged model is quite good, we can say that this lagged OLS model is overfitting, where the lagged dependent variable occupies considerable weight, which is meaningless for practical models.

From observations and the linear regression model, we have seen no obvious relationship between $\log(1+U)$ and $\log(1+UEY)$ and $\log(1+UEM)$ respectively, but from residual tests we can see a significance between them. Thus when considering these two predictors, we need to consider the residual effects. The SEM model is much better than the OLS model. In terms of p-value, the OLS model is a failure, and actually, these two variables cannot be employed as predictors at all. Compared to an AIC value of 244.64 in the OLS model, the AIC value in the SEM model is 189.89. However, both of these models are not that good compared to the lagged model based on the significance of $\Pr(|t|)$ or $\Pr(|z|)$ value. Thus, we cannot employ these two predictors into our final model.

The SLX model is similar to the SAR model, with the difference that the lagged variables are predictors, not dependent variables. We observe that the language groups, especially European and Asian ones, and three age groups are particularly spatially correlated. As a whole, the lagged X OLS model is better than the OLS model. Both multiple and adjusted R-squared values prove that. For AGY and lagged AGY in the lagged model, the Intercept also shows significance. In addition, the AIC value of the lagged model is 185.9607, compared to the OLS model with a value of 194.3642.

5.4 Final Regression and Spatial Formulas for Population Estimation

As a result of the previous considerations, we construct spatial models for the different cities, integrating all potential variables or lagged ones, together with error spatial autocorrelation. The final results are listed in Table 13.

Table 13. Final Spatial Regression Formula

City	**Formula**
Melbourne	*U~PP + GC+AGY+AGM + lagged U*
Sydney	*U~PP + GC + AGY + AGM + UEM + lagged U*
Brisbane	*U~PP + ASG + AGM*
Perth	*U~PP + MDI +AGM + error*

After constructing the linear model or the spatial model in a given city, we are subsequently able to make informed predictions of the population for the suburbs. The visualization between the real population and the prediction after applying these formulas is shown in Fig. 11 and Table 14. The first six suburbs are in Melbourne, followed by four suburbs in Sydney, three suburbs in Perth, and finally three suburbs in Brisbane. As can be observed, the predictions are closely aligned with actual (official) population. We get an approximate error range of +/-15%, and an absolute average error of 7%. This is by far more accurate than other models for micro-population estimation.

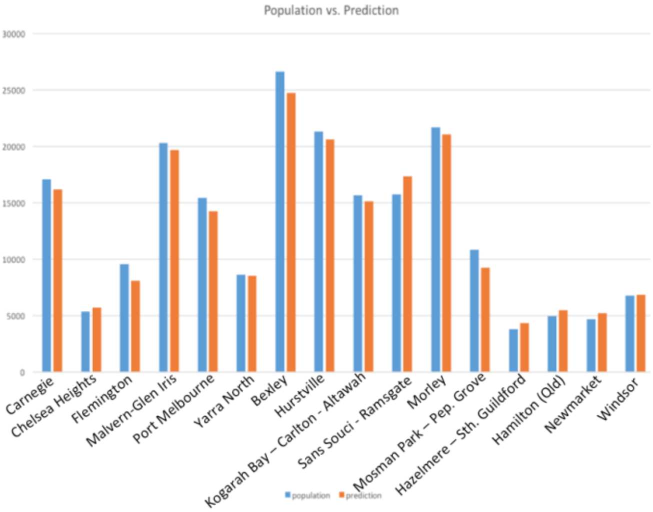


Fig. 11 Actual vs Predicted Population

Table 14. Final Actual vs Predicted Population for Suburbs

City	Suburb	Pop.	Prediction
Melbourne	Carnegie	17047	16184
	Chelsea Heights	5329	5676
	Flemington	9547	8090
	Malvern-Glen Iris	20279	19654
	Port Melbourne	15413	14213
	Yarra – North	8621	8544
	Bexley	26629	24735
Sydney	Hurstville	21329	20615
	Kogarah Bay - Carlton Allawah	15637	15104
	Sans Souci - Ramsgate	15730	17318
	Morley	21665	21037
Perth	Mosman Park-PeppermintGrove	10828	9234
	Hazelmere - South Guildford	3788	4298
Brisbane	Hamilton (Qld)	4930	5491
	Newmarket	4670	5183
	Windsor	6762	6836

6 CONCLUSIONS AND FUTURE WORK

In this paper, we provide novel solutions for analysing the micro-populations of skyscrapers and suburbs based on geotagged Twitter data. Given that raw Twitter data can be noisy, we support preprocessing of those data, both for suburbs and skyscrapers. Based on this we construct linear models for suburbs of four cities and tackled spatial autocorrelation issues between Twitter data and the official Census data. Using these models, we give predictions of the population of the suburbs of the selected four cities. Our results show that Twitter can indeed be used for micro-population estimation with quantifiable degrees of accuracy. We applied these models to estimate the population of skyscrapers. The next step would of course be to validate these results with the actual micro-population inhabitants in the skyscrapers. However these actual (live) statistics are not available. Refinements to this work would include modeling the relationship between the estimated predicted micro-populations with the water/energy usage of the skyscrapers, but here again the live access to such data is generally not available. Instead yearly aggregated statistics are made available. This system can be adapted to any other situational scenario simply by reconfiguring the harvesters with other bounding boxes for the areas of interest. The polygons/shapefiles are freely available online.

ACKNOWLEDGMENTS

The authors wish to thank the NeCTAR project for the use of the Cloud systems underpinning this paper, and the AURIN project for the Census and suburb Shapefiles. The corresponding author is Prof. Richard O. Sinnott.

REFERENCES

[1] B. Zhan, et al. "Crowd analysis: a survey." Machine Vision and Applications 19.5-6 (2008): 345-357.

[2] D.L. Swets and B. Punch. "Genetic algorithms for object localization in a complex scene." IEEE International Conference on Image Processing. 1995.

[3] S.Z. Li, et al. "Statistical learning of multi-view face detection." Computer Vision—ECCV 2002. Springer Berlin Heidelberg, 2002. 67-81.

[4] M. Jones and P. Viola. "Fast multi-view face detection." Mitsubishi Electric Research Lab TR-20003-96 3 (2003): 14.

[5] X. Huang, L. Li, and T. Sim. "Stereo-based human head detection from crowd scenes." Image Processing, 2004. ICIP'04. 2004 International Conference on. Vol. 2. IEEE, 2004.

[6] P.S.F. Yip, et al. "Estimation of the number of people in a demonstration."Australian & New Zealand Journal of Statistics 52.1 (2010): 17-26.

[7] H. Jacobs, Herbert. "To count a crowd." Columbia Journalism Review 6.1 (1967): 37.

[8] J. Seidler, K. Meyer and L.M. Gillivray. "Collecting data on crowds and rallies: A new method of stationary sampling." Social Forces 55.2 (1976): 507-519.

[9] E. Swank and J. D. Clapp. "Some methodological concerns when estimating the size of organizing activities." Journal of Community Practice6.3 (1999): 49-69.

[10] D. Ryan, et al. "An evaluation of crowd counting methods, features and regression models." Computer Vision and Image Understanding 130 (2015): 1-17.

[11] A.C. Davies, J.H. Yin and S.A. Velastin. "Crowd monitoring using image processing." Electronics & Communication Engineering Journal7.1 (1995): 37-47.

[12] C.S. Regazzoni, A. Tesei and V. Murino. "A real-time vision system for crowding monitoring." Industrial Electronics, Control, and Instrumentation, 1993. Proceedings of the IECON'93., International Conference on. IEEE, 1993.

[13] A.N. Marana et al. "Estimation of crowd density using image processing."Image Processing for Security Applications (Digest No.: 1997/074), IEE Colloquium on. IET, 1997.

[14] D. Kong, D. Gray and H. Tao. "A viewpoint invariant approach for crowd counting." Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 3. IEEE, 2006.

[15] C. Ratti et al. "Mobile landscapes: using location data from cell phones for urban analysis." Environment and Planning B: Planning and Design 33.5 (2006): 727-748.

[16] M.C. Gonzalez, C.A. Hidalgo and A.L. Barabasi. "Understanding individual human mobility patterns." Nature 453.7196 (2008): 779-782.

[17] C. Song et al. "Limits of predictability in human mobility." Science327.5968 (2010): 1018-1021.

[18] F. Calabrese et al. "The geography of taste: analyzing cell-phone mobility and social events." Pervasive computing. Springer Berlin Heidelberg, 2010. 22-37.

[19] Y. Liang et al. "How big is the crowd?: Event and location based population modeling in social media." Proceedings of the 24th ACM Conference on Hypertext and Social Media. ACM, 2013.

[20] P. Georgiev, A. Noulas, and C. Mascolo. "The call of the crowd: Event participation in location-based social services." arXiv preprint arXiv:1403.7657 (2014).

[21] Z. Cheng et al. "Exploring Millions of Footprints in Location Sharing Services." ICWSM 2011 (2011): 81-88.

[22] S. Scellato et al. "Socio-Spatial Properties of Online Location-Based Social Networks." ICWSM 11 (2011): 329-336.

[23] A.M. MacEachren et al. "Senseplace2: Geotwitter analytics support for situational awareness." Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on. IEEE, 2011.

[24] T. Sakaki, M. Okazaki and Y. Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World Wide Web. ACM, 2010.

[25] T. Terpstra et al. "Towards a realtime Twitter analysis during crises for operational crisis management". Simon Fraser University, 2012.

[26] R. Lee and K. Sumiya. "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection." Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks. ACM, 2010.

[27] J. Gomide et al. "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter." Proceedings of the 3rd international web science conference. ACM, 2011.

[28] F. Botta, Federico, H.S. Moat and T. Preis. "Quantifying crowd size with mobile phone and Twitter data." Royal Society open science 2.5 (2015): 150162.

[29] D. Brockmann, L. Hufnagel and T. Geisel. "The scaling laws of human travel." Nature 439.7075 (2006): 462-465.

[30] R. Tanton et al. "Small area estimation using a reweighting algorithm."Journal of the Royal Statistical Society: Series A (Statistics in Society) 174.4 (2011): 931-951.

[31] L.J. Hubert, R.G. Golledge and C.M. Costanzo. "Generalized procedures for evaluating spatial autocorrelation." Geographical Analysis 13.3 (1981): 224-233.

[32] M. Sawada, "Global Spatial Autocorrelation indices–Moran's I, Geary's C and the General Cross-Product Statistic." Laboratory of Paleoclimatology and Climatology, Dept. Geography, University of Ottawa, (Mimeo) (2001).

[33] P.A.P Moran, "Notes on continuous stochastic phenomena." Biometrika37.1/2 (1950): 17-23.

[34] [R.C. Geary, "The contiguity ratio and statistical mapping." The incorporated statistician 5.3 (1954): 115-146.

[35] A.D. Cliff and J. K. Ord. "The choice of a test for spatial autocorrelation."Display and analysis of spatial data (1975): 54-77.

[36] V. Halleck Vega, Solmaria, and J. Paul Elhorst. "The SLX model." Journal of Regional Science 55.3 (2015): 339-363.

[37] J.C. Anderson, J. Lehnardt and N. Slater. CouchDB: the definitive guide. " O'Reilly Media, Inc.", 2010.

[38] R.O. Sinnott and W. Chen. "Estimating crowd sizes through social media." 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops). IEEE, 2016.

[39] Y. Gong, F. Deng and R.O. Sinnott, "Identification of (near) Real-time Traffic Congestion in the Cities of Australia through Twitter", Understanding the City with Urban Informatics, CIKM 2015, Melbourne, Australia, October 2015.

[40] J.P. Zaldumbide and R.O. Sinnott, "Identification and Verification of Real-Time Health Events through Social Media", IEEE International Conference on Data Science and Data Intensive Systems, Sydney, Australia, December 2015.

[41] R.O. Sinnott and S. Yin, "Accident Black Spot Identification, Verification and Prediction through Social Media", IEEE International Conference on Data Science and Data Intensive Systems, Sydney, Australia, December 2015.

[42] R.O. Sinnott, et al, "The Australian Urban Research Gateway", Journal of Concurrency and Computation: Practice and Experience, April 2014, doi: 10.1002/cpe.3282.

[43] H. Akaike, 2011. Akaike's Information Criterion. In International Encyclopedia of Statistical Science (pp. 25-25). Springer Berlin Heidelberg.

Professor Richard O. Sinnott is Director of eResearch at the University of Melbourne and holds a Professorial role in Applied Computer Systems. He was formerly technical director of the National e-Science Centre, UK and director of e-Science at the University of Glasgow. He has a PhD in Distributed Systems; a Master of Science in Software Engineering and a Bachelor of Science in Theoretical Physics (Hons). He has published over 300 peer-reviewed papers in conferences/journals across a wide range of computing science areas with specific focus over the last fifteen years in supporting communities demanding finer-grained access control (security).

Wei Wang is a recent graduate of the University of Melbourne. He has a Master of Science in Computer Science and a Bachelor of Electrical Engineering from University of Science and Technology Beijing. His interests lie in analysis of Big Data; parallel, distributed and cloud computing and the operation and maintenance of servers.

Author/s:
Sinnott, RO;Wang, W

Title:
Estimating micro-populations through social media analytics

Date:
2017-04-19

Citation:
Sinnott, R. O. & Wang, W. (2017). Estimating micro-populations through social media analytics. SOCIAL NETWORK ANALYSIS AND MINING, 7 (1), https://doi.org/10.1007/s13278-017-0433-6.