BIG BICYCLE DATA PROCESSING:
FROM PERSONAL DATA TO URBAN APPLICATIONS

C. J. Pettit, Lieske S. N., Leao S. Z.

City Futures Research Centre, The University of New South Wales:
c.pettit@unsw.edu.au, s.lieske@unsw.edu.au, s.zarpelonleao@unsw.edu.au.

Commission II: Theme Session 17 Smart Cities

KEY WORDS: Big data, little data, data processing, data visualisation

ABSTRACT:

Understanding the flows of people moving through the built environment is a vital source of information for the planners and policy makers who shape our cities. Smart phone applications enable people to trace themselves through the city and these data can potentially be then aggregated and visualised to show hot spots and trajectories of macro urban movement. In this paper our aim is to develop procedures for cleaning, aggregating and visualising human movement data and translating this into policy relevant information. In conducting this research we explore using bicycle data collected from a smart phone application known as RiderLog. We focus on the RiderLog application initially in the context of Sydney, Australia and discuss the procedures and challenges in processing and cleaning this data before any analysis can be made. We then present some preliminary map results using the CartoDB online mapping platform where data are aggregated and visualised to show hot spots and trajectories of macro urban movement. We conclude the paper by highlighting some of the key challenges in working with such data and outline some next steps in processing the data and conducting higher volume and more extensive analysis.

1. INTRODUCTION

1.1 Big Data and Smart Cities

In an age of big data, little data and smart cities (Batty, 2015) there is an imperative to provide more accessible evidence to planners and policy makers to better shape our cities. Our research aim is to develop procedures for cleaning, aggregating and visualising human movement data and translating this into policy relevant information. In this paper we focus on movement data of bicyclists who are using the RiderLog application to capture data about their individual bicycle journeys across the City of Sydney, Australia. This little data (individual record bicycle journeys) can then be aggregated across a city, which then becomes big data. With big data there are challenges in processing and cleaning which we outline in this paper. Next, we present some preliminary findings visualised using the Carto DB online mapping platform. Finally, we conclude the paper by highlighting some of the key challenges in working with such data and outline next steps in the research.

1.2 From personal to urban applications: the need for Big Data pre- and post-processing

The growing volumes of data available from sensors, social media, and other digital interconnected systems are seen as ‘remarkable opportunities for researchers and policy analysts’ (Shneiderman and Plaisant 2015, p. 1). Indeed, the idea of ‘smart cities’ heavily relies on the possibility of integrating and understanding big (geospatial) data and turning it into knowledge and intelligence which is used to shape better and more effective urban environments (Batty, 2013; Li et al., 2015).

In this study we are focused on data produced by individuals motivated by personal goals. This type of data is becoming more prominent with the ubiquity of smart phones with multiple sensors, and the increasing use of mobile phone applications for daily routine activities in society (Lane et al. 2010). When combined in a crowd scale these data may have the capacity to reveal macro behavioural patterns which are of interest for city planners and policy makers alike. The transposition of these big datasets into urban research or urban planning, however, is not a simple exercise. The different purposes between data production and application, together with issues associated to privacy, human inconsistencies, and device inaccuracies, pose challenges to its practical use.

Contemporary datasets, according to Laney (2001), are characterised by their volume (data size is large), their velocity (data is created rapidly and continuously), and variety (data comprises multiple types and is captured from different sources), also known as the 3Vs of big data. IBM estimates that 2.5 quintillion bytes of data are generated every day, and that 90 percent of today’s data has been created in the last two years alone (Zhang et al. 2012). However, more data does not necessarily mean more useful information, since big data is also highly heterogeneous, complex, unstructured, incomplete, and noisy (Ma et al. 2014), and most current information systems or methods are unable to handle and process big data (Tsai et al. 2015).

Knowledge discovery in databases has always required a number of operations and processes to turn data from a raw state into a more appropriate format for analysis and visualisation (Fayyad et al. 1996), even when they had smaller size and complexity. Big data brings some additional challenges. Pettit et al. (2012) introduced a number of visualisation techniques for representing urban space and place; however, these are not specifically focused on the application of big data.

Tsai et al. (2015) presented a comprehensive review on efforts attempting to produce new methods that are able to handle big

doi:10.5194/isprsannals-III-2-173-2016

data during the input, analysis and output stages of knowledge discovery. They identified that most of the recent literature is focused on innovative methods for data mining and analysis, with much less attention to the pre- and post-analysis processing methods.

Data detection, selection, cleaning, filtering, correction, completion, and transformation are some of the pre-processing methods applied to prepare databases with the objective to obtain more accurate, complete and compatible information sets (Fayyad et al. 1996). In the context of big data, new approaches are combining those traditional methods with strategies to reduce its size and complexity. These can include the extraction of relevant records, event types, or key events; folding data to make cyclic patterns such as days or weeks clear; and pattern simplification strategies to simplify complex sequence of events (Shneiderman and Plaisant 2015). The question that remains is to what extent reduction can be made before losing important meanings. Pre-processing of big data can be so demanding that Kandel et al. (2011) refers to it as ‘data wrangling’.

Similarly, big data also raises new challenges for the post-processing analysis stage of knowledge discovery in databases, which in most cases are associated to the visualisation of patterns and processes. ProfitBricks (2015) presented a brief review of 39 data visualisation tools for big data currently available, including open-source, free, and commercial platforms. Some examples with geographic mapping capabilities include GoogleMaps[1], CartoDB[2], ProcessingGIS[3], and Leaflet[4]. With varied levels of sophistication, these developments demonstrate that this is a field in expansion. Interestingly, Kendal et al. (2011) argues that ‘analysts might more effectively wrangle data through new interactive systems that integrate data verification, transformation, and visualization’ (p. 1); therefore, bringing pre- and post-processing close together. This is the underlying approach for the research undertaken in our focus on bicycling data.

1.3 Big Cycling Data from participatory sensing via mobile phones

There is a growing trend to use mobile devices and applications to collect data relating to fitness activities (Clarke and Steele, 2014). Some mobile phone applications currently available for bicycling include MapMyRide[5], iBike[6], Cycle Meter[7], Strava[8], and RiderLog[9]. They vary in their format, purpose and functionalities; some save routes and monitor progress of ordinary riders, some are designed for professional riders, while others are more focused on bicycling for transport. What they all have in common is that they produce large amounts of complex data that document riding journeys. Individually, each application comprises data records with locations, time and intervals, and other attributes, which are organised into the specific application’s format and purpose. Daily, new records are captured from registered users, new users join the system, and some previous users may disconnect from the system.

At the same time, there is an increasing interest from city planners and policy makers in evidenced based research in active transportation. This is due to contemporary issues associated to health (chronical disease associated with physical inactivity) (Pratt et al. 2014), and the environment (transition to less carbon intensive cities) (Haines and Wilkinson 2014).

Recent research has been focused on the development and evaluation of mobile applications associated to bicycling, such as BikeNet (Eisenman et al. 2009), Biketastic (Reddy et al. 2010), and SocialCycle (Navarro et al. 2013). However, most of this research describes the mobile applications within the context of individual use, making only brief mention to its potential implications for aggregation both spatially and temporally to assist with city planning and policy making across urban geographies. In fact we could not find any reported study directly concerned with the transposition of data produced by individuals with a personal goal, into a database useful for the wider purpose of urban planning analysis.

Many factors can cause noise, inconsistencies, errors, inaccuracies, and incompleteness in the personal tracking data collected by people using their mobile phones via a specific application for bicycling. Inaccuracies can come from weak signal (i.e. in urban canyons in the CBD or in underground areas); incomplete data can occur if the signal is completely lost, if the batteries of the phone go flat, or if an incoming call interrupts the app (depending on the app). Incompleteness is also related to the fact that some people do not record all of their rides, or riders use different mobile applications, or do not change options when undertaking different types of journeys. The rider may have multiple purposes in a trip, simplifying its answer to the system with a single purpose. Some riders may also do their cycling as part of a multi-modal journey, placing the bike, for example, temporarily inside a train. All these sources or noise in the data are not a concern for individual users of the mobile application for the purpose of monitoring their fitness progress. However, these varied sources of inconsistencies, accumulated to millions or billions of records, can have a great adverse impact when this data is aimed to be used for city planning or policy making.

2. RESEARCH DESIGN AND METHODS

Bicycle Network have developed the RiderLog application[10] which is a free smart phone application and captures the location of a bicyclist every two seconds. For this research project we have bicycling route data covering all of Australia from 2010 to 2014. This includes 148,769 bicycle journeys undertaken by 9,727 cyclists. In this study we focus on 26,242 routes completed by 1,923 unique cyclists from the year 2010 to year 2014 in New South Wales (NSW). In this paper we focus specifically on its application for Sydney, the capital city for the State of New South Wales.

Data processing steps and flow are summarized in Figure 1. Original data are a 421 MB text file. In order to structure and clean the data as well as begin the process of fixing errors we brought the data into Microsoft Excel (text Import Wizard, limited with character, “|”). We then separated the data into smaller files based on Australian state or territory: New South Wales, Queensland, Northern Territory, South Australia, Tasmania, Victoria and Western Australia. These files were

[1] *https://developers.google.com/maps/*
[2] *https://cartodb.com/*
[3] *http://processingjs.org/*
[4] *http://leafletjs.com/*
[5] *http://www.mapmyride.com/*
[6] *https://itunes.apple.com/au/app/ibike/id369550718?mt=8*
[7] *http://itunes.apple.com/us/app/cyclemeter-gps-bikecomputer/id330595774?mt=8*
[8] *www.strava.com*
[9] *https://www.bicyclenetwork.com.au/general/programs/1006/*
[10] *https://www.bicyclenetwork.com.au/general/programs/1006/*

saved individually as, “rider_STATE.xlsx”. The New South Wales (NSW) data rider_NSW.xlsx file at this stage was 58.1 MB.

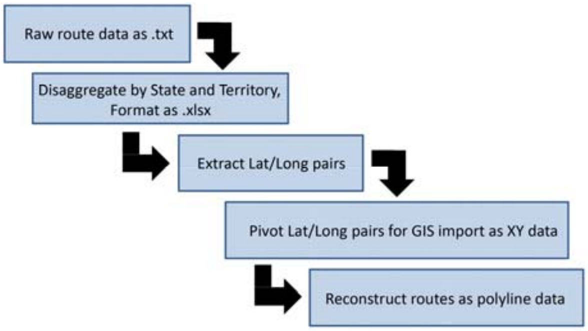
Raw route data as .txt
Disaggregate by State and Territory, Format as .xlsx
Extract Lat/Long pairs
Pivot Lat/Long pairs for GIS import as XY data
Reconstruct routes as polyline data

Figure 1. Data Processing Steps and Flow

Extracting geographic data from these tables was a challenge due to data size and formatting. In the Excel files geographic data are contained in a text formatted column, ‘Route’. For NSW, individual route records contain between 62 to 32,753 characters with latitude longitude pairs separated by commas and irregularly interspersed with miscellaneous words and other characters. Cleaning the data required removing words and characters using Excel’s find and replace capabilities. For NSW 68,153 unneeded strings of varying length were removed from the Route column.

Separating the often long strings of latitude and longitude information to Lat/Long pairs and saving them within individual data columns were accomplished with a series of two Excel functions, LEFT and MID. These functions enable extraction of a subset of a string from the left-most character or from the middle of a string using an indicated position, respectively. LEFT and MID were used in two separate formulas where the LEFT equation extracts a Lat/Long pair from the long string and the MID equation copies the original string less the Lat/Long pair in the first formula. Together, these formulas allow all Lat/Long pairs to be extracted to single columns from the original string in a recursive fashion. The 32,753 characters in the longest NSW cycling routes result in 915 Lat/Long pairs. If developed as a single data table these 915 pairs multiplied by the 26,243 data records along with the 27 columns already in the spreadsheet would result in a spreadsheet containing over 1,856 columns and 48,707,000 cells. This large data volume required what would be an easy series of iterations within a data table for a small dataset to proceed in a stepwise fashion. The first step in processing this large volume of data was to determine the length of the geographic data (route) strings in each data record then sort from smallest to largest. For the first 125 Lat/Long pairs we were only able to process 25 pairs at a time. For each grouping of 25 Lat/Long pairs we used the LEFT and MID formulas above to extract the geographic data then used a VBA script to replace formulas with data values in the worksheet. We then deleted ‘no data’ values (where the long string had been full parsed). Next, the intermediate data generated by the MID equations were deleted which substantially reduced the file size. In the case of the second iteration (the first 50 Lat/Long pairs) this stepwise process of replacing formulas with values and deleting intermediate data reduced the data set from 186 MB to 85.1 MB. After having completed five of these iterations, for the first 125 columns, the vast majority of the data (22,000 of 26,243 routes) including nearly all route records of 4,000 characters or less had been processed (Figure 2). The remaining routes, although the strings were longer, were able to be processed considerably more quickly and were therefore processed in groups of 100 Lat/Long pairs.

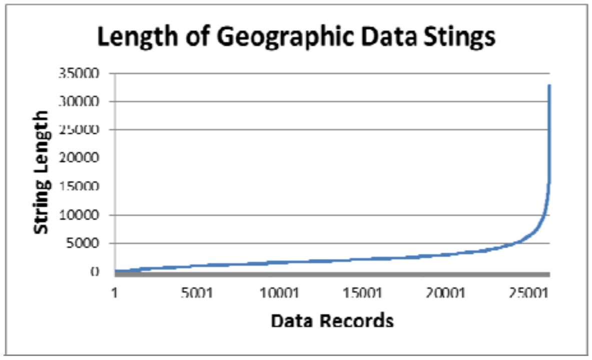

Figure 2. Length of Geographic Data Strings

Another part of data structuring and cleaning involved working with time stamp data. The original data for ride start and finish time were amalgamated with date and year in the format: “10/1/2011 6:40:16 AM”. From these data we generated several time variables through a process of extracting a subset of a string from a string in a process similar to that described above. We generated a Year variable, a Month variable as an integer indicating one through 12, a Date variable as a combination date and month where, for example, 10/31 indicates 31 October, and day of the week. We also separated start and finish time from the original data resulting in HH:MM:SS format, e.g. “06:40:16” that excluded the day, month and year included in the original data. The duration variable included with the original data was incorrect so we re-calculated trip duration by subtracting start time from finish time.

The next step in cleaning the data was to delete several columns that became redundant or are otherwise not necessary: DateStarted, DateFinished, TimePaused, Route, and year of birth. TimePaused contained no data. Year of birth data are captured by rider age and date of ride which are both retained in the data set.

After data structuring, cleaning and the preliminary processing indicated above data we transposed data from records based on routes to geographic data where records are based on Lat/Long coordinates suitable for CartoDB and GIS import. We restructured the data from routes to geographic coordinates with a pivot table type data summarization using visual basic for applications (VBA) within Microsoft Excel. Data were first saved in a new worksheet as values rather than formulas. The VBA script then transposed the data from records based on route to a more standard spatial data format where data indicate point locations of a bicycling route. Along with the pivot of the Lat/Long data the VBA script associated attributes originally linked with individual cycling routes to each Lat/Long pair. These attributes include rider ID, top speed, ride purpose, the aforementioned time variables, start local government area (LGA), end LGA, age and gender.

The challenges, solutions and processing volumes in working with the NSW RiderLog data are summarized in Table 1. The solutions used to address the problems of extra text in the Lat/Long strings, correcting the start and finish time, fixing the incorrect rider duration and splitting the Lat/Long data from one to two columns, although sometimes time consuming to execute, were easily accomplished with standard MS Excel commands that could be applied simultaneously to data columns or entire worksheets.

Other errors required more sophisticated techniques including occasions where rides took place across two days resulting in duration errors, the lack of a time stamp on Lat/Long pairs and the overall volume of data. Fixing duration errors associated with multi-day rides required replacing the simple ride duration formula (finish time minus start time) with a separate formula that is able to calculate duration while accounting for multiple days. There were 68 of these errors in the NSW data.

The solution to the lack of time stamps associated with Lat/Long observations required estimating the time of each location observation. Calculation of a new data field, "TimeModel" was also embedded within the pivot script. Timemodel was calculated by dividing trip duration by number of segments, summing this value for the segment number in question and adding this value to start time in order to get an approximate time stamp for each point in a rider's route.

The overall challenge of dealing with the large volume of data, especially reducing the complexity of route data formatted as long text strings, required numerous iterations performed in such a way as to minimize file sizes at each given stage. The volume of data also required the use of multiple files (4 .xlsx files) to accommodate the entire New South Wales RiderLog 2010 – 2014 dataset. We estimate the RiderLog 2010 – 2014 dataset the covers the entirety of Australia would require 21 .xlsx files. At both the New South Wales and national scales this volume of data exceeds the definition of big data presented in Batty (2013); any dataset which cannot fit into an Excel spreadsheet.

Challenges	Solution	Volume
Extra Text	Find/Replace	68,153 strings
Start/Finish Time Incorrect	Switch labels	2 columns
Ride Across two Days	Change formula	68
Ride Duration Incorrect	Re-calculate	26,243 records
Amalgamated Lat/Long	Separate	1 column
No Time Stamp on Lat/Long data	TimeModel	All records
Data Volume	Iteration; multiple data files, 4 .xlsx files for NSW	>48,700,000 possible cells

Table 1. Big Data Challenges, Solutions and Volume for the NSW Rider data

CartoDB offers multiple ways to view cycling route data, as points which may be displayed in time sequence and as routes (lines) which display well in still images. The final step in preparing our data tables for upload into CartoDB was to split the amalgamated Lat/Long pairs into two columns. Bringing point data into CartoDB required connecting, or uploading, the data and georeferencing, a simple matter of specifying the latitude and longitude columns in the data. Once correctly georeferenced data could be displayed and manipulated within CartoDB. Uploading route data as lines into CartoDB was facilitated by bringing the point data into GIS and converting points to lines based on RouteID.

In CartoDB, both temporally dynamic torque maps based on point data and polyline based cycling route maps may be displayed by category. The CartoDB Map Layer Wizard may be used to select among several category options, select the column that contains the data one wishes to view, and adjust the legend to symbolize categories by colour.

3. RESULT AND ANALYSIS

In this section of the paper we illustrate how the processed RiderLog data can be visualised and then undertake some preliminary analysis of the results with a focus on the City of Sydney. User profile data which includes gender and year of birth can be used to create specific map visualisations of the flow of bicyclists across the city. Also, the user can specify the purpose of each individual trip and again this can provide an interesting breakdown of who is bicycling where for what purpose. Other views of the data can be made on ride duration and origin of the trip. Using CartoDB we have created a frame where these parameters may be toggled on or off to visually analyse bicycle behaviour across the city.

It is important to understand the purpose of a ride when city planners and policy makers are considering new bicycle infrastructure. Different Smart Phone applications are used for different purposes. The RiderLog application is predominantly used by those commuting and travelling for transport, rather than recreational purposes. Hence, the results in Figure 3 show a predominance in bicycle trips made for transport. We can see the Central Business District (CBD) of Sydney is a hot spot of activity as one would expect for bicycle movements for travel. It also highlights the CBD provides limited recreational opportunities for cyclists. However, examining Figure 3 we can see that the roadway within Centennial Park, which is located 1 kilometre out of the CBD, is a major attractor for recreational bicyclists. Using such data we can begin to understand which part of the city's bicycle network is being used for what purpose and this can in turn provide valuable information on future infrastructure planning provisions across the city.

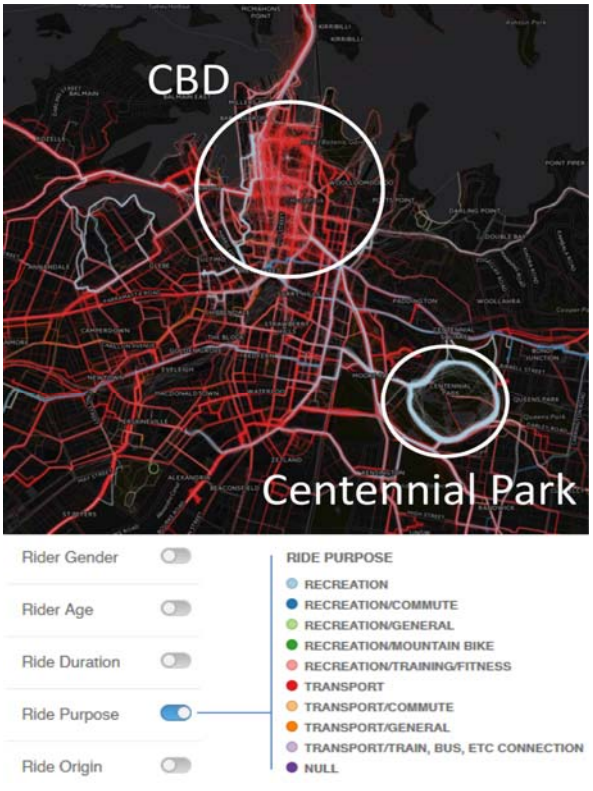

Figure 3. Sydney Area Rider Routes by Ride Purpose

One of the user profile questions in RiderLog is related to the age of the bicyclist. We can see in Figure 4 that a significant number of bicyclists using the RiderLog application are around 40 years of age. It is difficult for such an application to record journeys made by children travelling to school independently as most will not have their own smart phone application to record their journey. There is also no mechanism in the Application for a bicyclist to log if there might have been a child on board a bike being dropped to school. So the use of such applications is difficult in trying to find out information about bicycling behaviour of children. However, preliminary analysis of this aged based data would suggest policy makers might like to target bicycle promotion programs to those aged below 40 if the goal was to increase bicyclist numbers across the City.

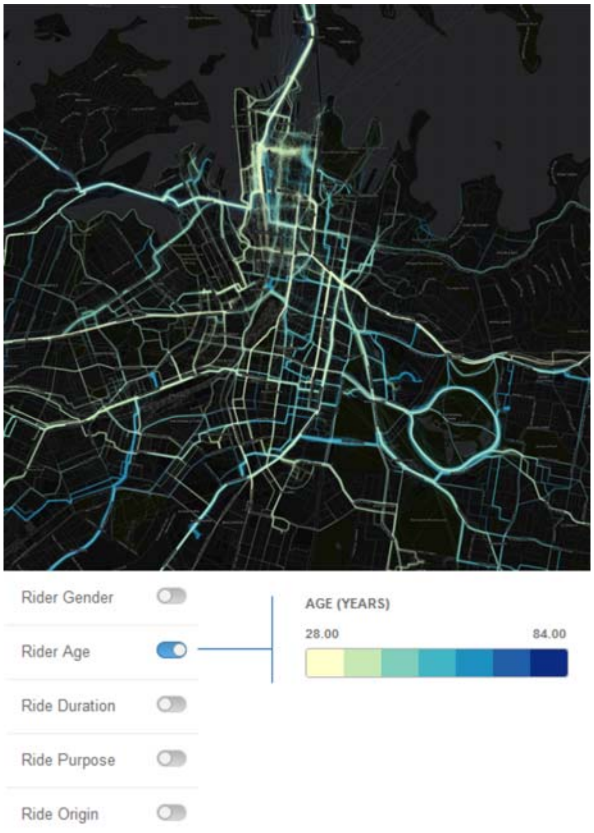

Figure 4. Sydney Area Routes by Rider Age

Understanding gender patterns in bicycling is very important in order to develop and maintain bicycle infrastructure which is used by both male and female riders. In Figure 5 we can see the females (in red) follow distinctive and a more limited set of routes when bicycling as compared to males. If we visually ground truth this information we see that female bicyclist activity is more restricted to those routes where there is dedicated and indicated bicycling lanes. Whereby male bicyclist seems to be more comfortable bicycling in areas where bicycle lanes may not be available. This preliminary analysis suggests that by providing more dedicated bicycling infrastructure this could likely result in an increase of female bicyclists across the city.

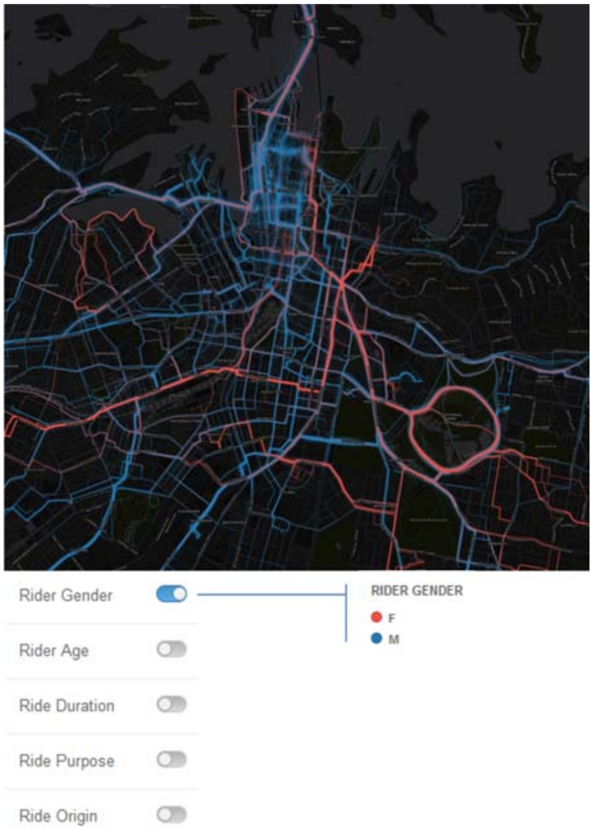

Figure 5. Sydney Area Rider Routes by Gender

Understanding how far people are willing to bicycle and from which origin to destination is also an important piece of evidence when planning for travel behaviour. Figure 6 illustrates the travel time taken by bicyclists to arrive at their destination. It indicates that most trips range from 0-15km. Such information is important in understanding bicycle catchment areas which is useful in the formulation of city wide transport and planning strategies.

To further understand bicycle movements across the city origin data can be most useful (Figure 7). This data can highlight suburbs and precincts where bicyclists reside. This can further assist planners in developing municipal specific bicycle infrastructure strategies. For example those areas where bicyclists live and are in close proximity to a train station might then be a candidate for a bicycle parkiteer (a secure parking station) which supports multi-modal travel behaviour as illustrated in Figure 8.

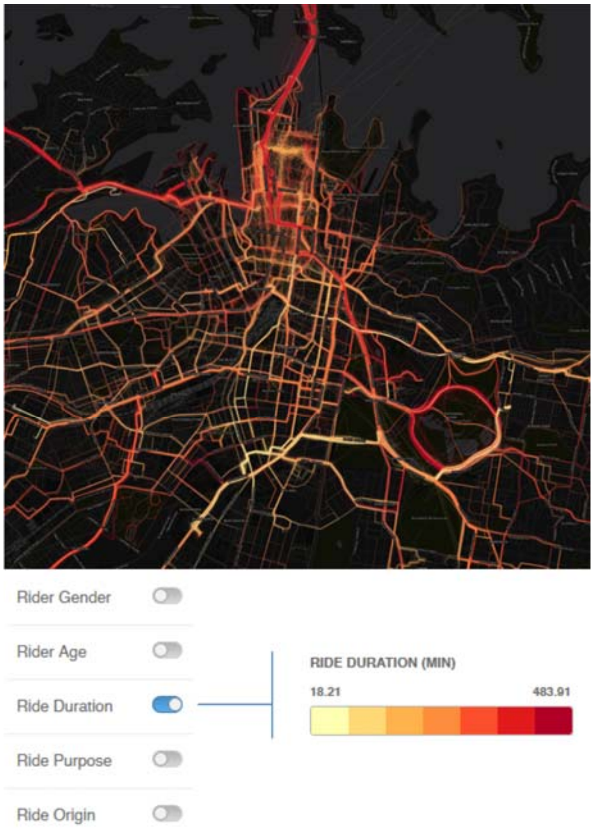

Figure 6. Sydney Area Rider Routes by Rider Duration

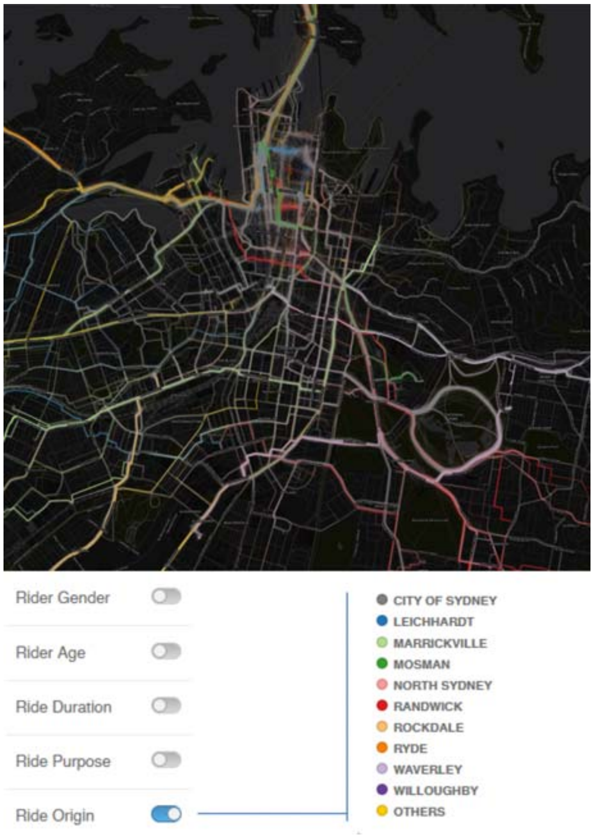

Figure 7. Sydney Area Rider Routes by Ride Origin

Figure 8. Bicycle parkiteer situated at a train station. Photo attributed to Bicycle Network

4. CONCLUSIONS

In this paper we have discussed the processing, visualisation and potential application of bicycling data acquired from individuals using the RiderLog Smart Phone Application. We have discussed methods for processing and cleaning the data, which have initially been undertaken using Excel. However, given the need to use multiple files and iterative processing to accommodate one data set we conclude Excel is not sufficient for handling such data efficiently. When we move to the next steps of the research we will be looking at upscaling this approach to include Riderlog data from across all of Australia. Next steps will look at migrating the database into a platform which handles big datasets and better support processing and cleaning, possibly R Project through the pbdR initiative (Ostrouchov et al. 2012).

Through our cleaning processes we have identified a number of challenges in aggregating this data from individual rider journeys to a city wide analysis. In addition to efficiently handling the large volume of data, the most notable of these challenges was separating long strings of text containing geographic data on bicycling routes. While solutions to some data problems were easily implemented and others were more challenging, it is important to be cognisant of how data manipulations at any given stage would impact the data at future stages of development as well as in the final analysis. When processing big data it was also important to be cognisant of one's end goal and/or desired product outputs as a means of evaluating the potential impact of data manipulations at any given step.

Once we have the resultant data processed and cleaned, the CartoDB online mapping platform was used to visualise the results across the City of Sydney. In future work with real-time data feeds, potentially at the scale of analysis of the entire country of Australia, we will exploit CartoDB's real-time Big Data Connectivity and CartoDB's Deep Insights Technologies. Using Deep Insights one may manipulate and visualise hundreds of millions of spatial data points. The use of such online mapping platforms provide a powerful vehicle for city planners and policy makers to interact with the data and make more evidence-based decisions about the shaping of our cities (Pettit et al. 2015). In this paper we have focused on data and a visualisation platform which can be used to support city planning in relation to active transport and providing recreational opportunities. However, it is important to note the data collected from smart phone applications such as Riderlog are not considered statistically rigorous. Future research will examine data fusion techniques that can be deployed to combine Riderlog data with more systematic bicycle count data, household travel survey data and other sources to provide a richer picture and more robust data source to support evidenced based decision making.

In an increasingly urbanised world we continue to plan for the sustainable growth of our cities. This includes promoting active transport and looking for solutions to alleviate congestion. There is a critical need for evidenced based city planning and policy making which uses data from a rich variety of sources to address such concerns. The potential of smart phone collected data such as the Riderlog data presented in this research provides an important source of truth which can be further interrogated to understand the flow of people and how they interact with each other and the built environment.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge Bicycle Network for supplying the RiderLog data which has made this analysis possible. Peter Patterson, from the City Futures Research Centre is also acknowledged for his preliminary exploratory work with the Riderlog data.

6. REFERENCES

Batty M, 2013. Big data, smart cities and city planning, *Dialogues in Human Geography*, 3 (3): 274–279.

Batty, 2015. Data About Cities: Redefining Big, Recasting Small, Data and the City Workshop, The Programmable City Project at National University of Ireland, Maynooth, Agu 31- Sept 1st, 2015. http://www.spatialcomplexity.info/files/2015/08/Data-Cities-Maynooth-Paper-BATTY.pdf

Clarke A and Steele R, 2014. Health Participatory Sensing Networks, *Mobile Information Systems*, Volume 10 (3): 229-242

Eisenman S B, Miluzzo E, Lane N D, Peterson R A, and Campbell A T, 2009. BikeNet: A Mobile Sensing System for Cyclist Experience Mapping, *ACM Transactions on Sensor Networks*, 6 (1): Article 6, December 2009.

Fayyad U M, Piatetsky-Shapiro G, Smyth P, 1996. From data mining to knowledge discovery in databases. *AI Magazine*, 17 (3): 37-54.

Haines A and Wilkinson P, 2014. Health in the 'Low-Carbon' Economy, Chapter 75 In: Freedman B (ed.), Global Environmental Change, Springer Science+Business Media, Dordrecht.

Kandel S , Heer J , Plaisant C , Kennedy J , van Ham F , Riche N , Weaver C , Lee B , Brodbeck D and Buono P, 2011.Research directions in data wrangling: visualisations and transformations for usable and credible data, *Information Visualization*, 10 (4): 271-288.

Lane D, Miluzzo E, Lu H, Peebles D, Choudhury T, and Campbell A T, 2010. A Survey of Mobile Phone Sensing, *IEEE Communications Magazine*, September 2010, p. 140-150.

Laney D, 2001. 3D data management: controlling data volume, velocity, and variety. META Group, Technical Report, 2001.

Li, S., Dragicevic, S., Castro, F.A., Sester, M. Winter, S. Coltekin, A, Pettit, C.J. Jiang, B., Haworth, J., Stein, A. (2015). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*. rXiv:1511.03010 [physics.soc-ph].

Ma C, Zhang H H, Wang X, 2014. Machine learning for big data analytics in plants. *Trends in Plant Science*, 19 (12): 798-808.

Navarro K F, Gay V, Golliard L, Johnston B, Leijdekkers P, Vaughan E, Wang X, and Williams M-A, 2013. SocialCycle: What Can a Mobile App Do To Encourage Cycling? *In Proceedings of the Second IEEE International Workshop on Global Trends in Smart Cities*, pages: 24-30.

Ostrouchov, G., Chen, W.-C., Schmidt, D., Patel, P., 2012. Programming with Big Data in R [WWW Document]. URL http://r-pbd.org/

Pettit, C.J. Barton, J, Goldie, X, Sinnott, R. Stimson, R, Kvan, T. (2015) The Australian Urban Intelligence Network supporting Smart Cities, in Geertman S, Stillwell J, Ferreira J and Goodspeed J (eds) Planning Support Systems and Smart Cities, Lecture Notes in Geoinformation and Cartography, pp 243 – Springer, pp 243-259.

Pettit, C. Widjaja, I, Russo, P, Sinnott, R, Stimson, R, Tomko, M. (2012) Visualisation support for exploring urban space and place, *XXII ISPRS Congress*, Technical Commission IV 25 August – 01 September 2012, Melbourne, Australia Editor(s): M. Shortis, J . Shi, E. Guilbert, ISPRS Annals Vol 1-2, pp 153-158.

Pratt M, Norris J, Lobelo F, Roux L, Wang G, 2014. The cost of physical inactivity: moving into the 21st century, *British Journal of Sports Medicine*, 48: 171–173.

ProfitBricks Blog, 2015. 39 data visualization tools for big data. http://blog.profitbricks.com/39-data-visualization-tools-for-big-data/ Accessed on 02 Dec 2015.

Reddy S, Shilton K, Denisov G, Cenizal C, Estrin D, and Srivastava M, 2010. Biketastic: Sensing and Mapping for Better Biking, In *Proceedings of CHI 2010: Bikes and Buses*, April 10-15, 2010, Atlanta, Georgia, USA.

Shneiderman B and Plaisant C, 2015. Sharpening analytic focus to cope with Big Data volume and variety. Visualization Viewpoints, *IEEE Computers Society*, May/June 2015, pages 10-14.

Tsai C-W, Lai C-F, Chao H-C, and Vasilakos A, 2015. Big Data analytics: a survey. *Journal of Big Data*, 2 (21): 1-33.

Zhang L, Stoffel A, Behrisch M, 2012. Visual analytics for the Big Data Era – a comparative review of state-of-the art commercial systems. *IEEE Symposium on Visual Analytics Science and Technology*, Seatle, WA, USA, October 14-19, pages 173-182.