urban science

MDPI

Article

Building a National-Longitudinal Geospatial Bicycling Data Collection from Crowdsourcing

Simone Z. Leao [1,*], Scott N. Lieske [2], Lindsey Conrow [3], Jonathan Doig [1], Vandana Mann [1] and Chris J. Pettit [1]

[1] City Futures Research Centre, University of New South Wales, Sydney NSW 2052, Australia; j.doig@unsw.edu.au (J.D.); v.mann@unsw.edu.au (V.M.); c.pettit@unsw.edu.au (C.J.P.)

[2] School of Earth and Environmental Sciences, University of Queensland, Brisbane QLD 4072, Australia; scott.lieske@uq.edu.au

[3] School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85281, USA; lconrow@asu.edu

* Correspondence: s.zarpelonleao@unsw.edu.au; Tel.: +61-29-385-4828

Received: 17 May 2017; Accepted: 26 June 2017; Published: 28 June 2017

Abstract: To realize the full potential of crowdsourced data collected by smartphone applications in urban research and planning, there is a need for parsimonious, reliable, computationally and temporally efficient data processing routines. The literature indicates that the opportunities brought by crowdsourced data in generating low-cost, bottom-up, and fine spatial and temporal scale data, are also accompanied by issues related to data quality, bias, privacy concerns and low accessibility. Using an exemplar case of RiderLog, a crowdsourced GPS tracked bicycling data, this paper describes and critiques the processes developed to transform this urban big data. Furthermore, the paper outlines the important tasks of formatting, cleaning, validating, anonymizing and publishing this data for the capital cities of each state and territory in Australia. More broadly, this research contributes to the foundational underpinnings of how to process and make available crowdsourced data for research and real world urban planning purposes.

Keywords: crowdsourced data; smartphone; bicycle; RiderLog; big data

1. Introduction

Understanding the flows of people moving through the built environment is a vital source of information in assisting transportation planners and managers to mitigate congestion, optimize modal flows, coordinate multiple transport modes and improve urban quality of life. While data collection has been both expensive and difficult in the past [1], today technology yields opportunities for relatively easy and low-cost transportation data collection for both motorized and active travel. Example technologies include public transport smart card records (e.g., [2,3]), passive positioning counts for car navigation systems (e.g., [4]), bike share program records (e.g., [5,6]) and individual-level transportation data collected through smartphone applications (commonly known as "apps").

Mapping has been a task developed by state organizations for centuries, but recently citizens have started to become involved in mapping their immediate environment [4]. These new distributed, bottom-up mapping methods fall under the umbrella of crowdsourcing. After evaluating a large number of understandings of crowdsourcing reported in the literature, mostly based on domains of specific applications, [7] proposed a general definition of crowdsourcing as:

> a type of participative online activity in which an individual, and institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task.

> The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.
>
> (p. 197)

Apps can capture locational information and enable people to view, record and share travel routes. These data can be aggregated, analysed and visualised to provide insights into where and when people are navigating through the built environment, offering an opportunity to improve our understanding of urban movements. Data generated by individuals who use the apps are becoming more prominent with the ubiquity of smartphones, the variety of embedded sensors in these devices, and the increased use of mobile phone applications for daily routine activities in society [8]. There are many apps that specifically target cycling activity. Of these many use GPS technology for recording route information, including MapMyRide, Cycle Meter, Strava, RiderLog, Runtastic, MotiFit, Ride Report, Ride Star, iBiker, Bike Computer, and Ryde. Each application has a distinctive look and feel, focus, function, format and intended audience. Some, for example, target exercise enthusiasts, others target people using cycling as transport. Regardless of differences, each application produces substantial amounts of data that document cycling journeys. Data are typically comprised of locations, dates and times, and other attributes that fit a particular application's structure and purpose [9].

Data from such apps have the potential to reveal localized urban-scale transportation and behavioural patterns. This fine-scale data can be aggregated across a city as well as across time, which then becomes big data. Transmuting such large and complex datasets into forms that are usable for urban research and planning, however, entails numerous challenges including device inaccuracies, human inconsistency, sampling bias, and privacy issues [9].

To realize the full potential of crowdsourced data for improved urban research and planning, there is a need for parsimonious, reliable, computationally and temporally efficient data processing routines. Using the exemplar case of RiderLog, a crowdsourced GPS tracked bicycling data covering the whole Australia for four years, this paper describes and critiques the full processes developed to transform the raw database generated from individual volunteers to a robust dataset of collective cycling activity. The data curation and processing included several operations for data formatting, cleaning, validation, anonymization, and publishing.

Most literature on crowdsourcing addresses specific data operations individually, or exemplar applications of the crowdsourced data. To the best of our knowledge, there is a lack of studies addressing the full process of converting a raw crowdsourced data into a polished and published data collection. This full process involves different operations, each one with their own challenges. This research contributes to the discussion on the opportunities and challenges involved in this full process, with the objective to support similar initiatives to increase and facilitate the use of crowdsourced big data for research and real world planning purposes.

2. Background

As an extension to traditional methods of transport surveying, there is an emerging body of literature documenting how new developments in information and communication technologies can be used to map and monitor human movement. Opportunities and challenges of crowdsourcing approaches to movement traces, including walking, cycling, public transport, and private vehicles, have been reported in the literature. Due to the focus of this study and paper size restrictions, this background review is limited to previous research on crowdsourcing bicycling movement.

The first studies on bicycling transport used GPS receivers provided to a small number of cyclists [10–13]. A study in Portland, Oregon, from March through November 2007, collected data on 166 cyclists [10]. Participants carried GPS devices for seven days, producing 1955 bicycling trips. New trips were initiated every time a cyclist was stationary for five minutes. After completing and visualizing their tracks, participants were asked about the accuracy of the recorded route and reasons for choosing that route. Data was captured on 100 cyclists in Waterloo, Ontario, using GPS receivers and a web survey [11]. They collected data for 1232 trips over the course of five weeks. Reference [12] studied 48 cyclists for four days in 2010. These data were combined with a 2005 study where participants hand drew their most frequently used bicycling route on a map. By combining data sets they developed records of 278 bicycle trips. A GPS-based survey in [13] consisted of 79 participants surveyed over one week, resulting in 941 recorded trips.

The development of GPS enabled smartphones made it possible to extend sample sizes of cycling-related studies to a very large number, although giving less control to researchers on the sampling process as the recruitment of cyclists became self-selecting. In San Francisco, data from 952 cyclists using the smartphone app CycleTracks were collected between November 2009 and April 2010 generating a total of 5178 bicycle trips [14]. The GPS recorded tracks were associated with characteristics of the road network, compared to shortest paths options, and a path size logit route model was proposed. Estimated results indicated that cyclists in the San Francisco area strongly prefer bike lanes, disfavour climbing hills, turning, and deviating excessively from the minimum distance path. CycleTracks smartphone app was also used to capture data of 3600 cycling routes in Austin, Texas, from May through October 2011 [15]. The findings of this study gleaned detailed information about routes that cyclists in the study area used by matching the routes to the underlying street network. While the data required challenging cleaning and processing before use as input to match routes, the finalized data set captured more detailed cycling information than possible with conventional data collection methods [15].

The ubiquity of smartphones in society together with the growing adoption of fitness and leisure apps have immensely increased the amount of detailed data on human mobility, including cycling. Examples include the commercial apps MapMyFitness, with more than 20 million members in 2013, and Strava, with an average of 2.5 million weekly downloads in 2014 [16]. Big data has begun to create significant impacts in transport planning, illuminating processes that were previously under-sampled and poorly understood. Reference [16] also stress that although GPS tracked data by smartphones is one of the "most fruitful in terms of the application models which can link directly to transport policy on a city or country level" (p. 116), these types of data pose challenges with respect to accuracy and volume. Additional challenges include potential bias in the sample from self-selection recruitment [17,18], and privacy concerns when personal data is transferred to third party uses [19]. These challenges are explored as follows.

GPS is known to produce error and location imprecision for varied reasons. An example is error resulting from a cold start where trips commence before the GPS or smartphone app begins to accurately record location, and from GPS drift when the device is stationary [20,21]. Another quality issue is influenced by overcast skies, proximity to large buildings, and whether the smartphone is open and accessible or stowed away in a bag or other container. The study developed by [11] reported GPS accuracy varying in some cases by 10 to 15 m. Signal interference when recording cycling activity with GPS devices in "urban canyons" was also found in [10,21]. Reference [13] acknowledges trade-offs associated with the frequency of GPS point capture noting, on one hand, greater accuracy in capturing a route but balanced against increased battery drain and increasing file sizes.

Previous studies that collect bicycling route data with GPS receivers and smartphone apps present valuable lessons and recommendations for processing bicycling route data, including data cleaning and quality checks. Reference [21] argue that because the collection of data by GPS devices generates very large quantities of records, it is necessary to develop automated procedures to analyze the data, and convert it to readily visualized information for respondents, analysts, and decision makers. This leads to the need for rule-based algorithms. For example, a study developed by [10] omitted GPS data points with high error values, as determined by their software, and records new trips if a rider is stopped for more than five min. Trips that occurred outside the spatial and temporal extent of the study, and trips where a measure of horizontal or vertical accuracy exceeded a threshold value, were deleted in [15]; and they also split trips if there was more than three minutes or more than 1000 ft. between points, and deleted trips with fewer than five points. Trips less than 100 m, and trips where participants did not provide trip details as prompted by the user interface were removed in [20]. The GPS locations captured by [11] contained elevation values which allowed filtering out points where elevation was less than zero, greater than 600 m or where the change in elevation between consecutive points was greater than 35 m (Ontario, Canada). Points less than 200 m and greater than 4200 m above sea level (Switzerland) were removed in [22]. A speed threshold where any points that indicated travel faster than 75 km/h were removed in [11]. And points with unreasonable values for speed and acceleration were eliminated in [12]. Similarly, [22] calculated speed and acceleration from GPS location and associated time stamp information, and then removed points that would have required a speed greater than 50 m per second using an error buffer of 30 m.

Automating the process of reducing or removing errors from GPS tracked data is not an ordinary task, particularly with large volumes of data generated by smartphone apps [9,23]. File management issues may include simply opening and viewing large files, as big data files might extend beyond the capabilities of conventional systems, requiring specialized software and skills [19]. Despite several analytic tools that have been developed, handling large volumes of data varying in time remains problematic [24]. Methods for processing 26,242 routes collected using the RiderLog app by 1923 cyclists in New South Wales, Australia, between May 2010 and May 2014, were described by [9]. The bulk of their data processing used Microsoft Excel, which they concluded was insufficient for handling the volume of data involved. Knowledge discovery in databases has always required several operations and processes to turn data from a raw state into a more appropriate format for analysis and visualisation [25], even when data had smaller volume and complexity. Contemporary datasets, including crowdsourced data, are characterised by their volume (large size of the data), their velocity (data are created rapidly and continuously), and variety (data are of multiple types and acquired from various sources) [26]. These are known as the 3Vs of big data, and have been updated in the literature to 5Vs, through the addition of veracity and value [27]. While the first 3Vs (volume, velocity, variety) emphasise the issues associated to the origin and characteristics of big data, the additional two (veracity, value) highlight issues related to the use and application of the data to relevant purposes.

Evidence of bias has been identified in sample formation in volunteered geographic information (VGI) [17]. Using crowdsourced geotagged records, their analysis demonstrated an unbalanced representation of urban and rural populations, with bias towards urban perspectives. They argue that the lack of information from some groups in crowdsourced data can have implications in studies based on those type of data, which are increasingly being used by commercial and scientific research and urban planning. People who do not add data effectively to user-generated content (UGC) databases are excluded from studies, as their points of view and activity patterns are not recorded. In this context, veracity of big data is not only related to the quality of individual records, but also the representativeness of the overall collection.

Sampling bias specifically addressing bicycling behaviour from smartphone application surveys was the topic of a study developed by [18]. Bias occurs when the demographic characteristics of the smartphone application users do not match the characteristics of the greater bicycling population, and it is mostly caused by the volunteered nature of the sample recruitment. Comparing smartphone application samples (self-selected sample) to traditional travel surveys (statistically defined sample) for seven cities in the USA, [18] identified that smartphones tended to under-sample females, older adults, and lower-income populations, and to oversample some minority ethnicity populations. Despite this, the study concludes a promising future for smartphone crowdsourced bicycling data, based on the increasing adoption of smartphones, and the much larger number of records produced when compared to bicycling representation in traditional travel survey. This implies, however, that the bias of any crowdsourced data should be assessed and informed to the data users, so analysis and decision-making based on the data takes the bias into consideration.

Privacy is another relevant concern associated with crowdsourced data. Technological developments have had a profound impact on the level of privacy experienced by individuals, due to contemporary computer systems providing intrusive capabilities for companies, governments and individuals. This is particularly relevant for data generated and released as a continuous stream of locations. With specific focus on privacy issues in geographic information data, [19] recommends reduction of spatial and temporal resolution to avoid re-identification attacks. Limiting the amount of information presented may also be required to avoid re-identification; for example, ethnicity/age/gender together with a place of regular visit (home or work, depending of the time of the day) may be enough to identify an individual with a unique set of attributes. This is also reinforced by [28] where mobility data are linked to a user. They argue that even if the data are made anonymous in that the true identities of individuals are replaced by random identifiers; the individuals are still highly identifiable when partial knowledge of their whereabouts are publicly observable through their trajectories.

A study developed by [28] addressing the anonymization of bicycling trajectories data recorded by smartphone application replaced long routes by splitting them into several smaller routes, each one with its own random identifier, as if they were recorded by different people, and having different origin, destinations, distances and duration. Although this makes the re-identification of a rider impossible, it also changes the data significantly, producing distributions of trips by distance or duration, or origin-destination locations different than reality. Analysis based on these data can accurately look at locations most regularly crossed by cyclists, but inferences on average distances performed by cyclists would lead to wrong and misleading results. Therefore, the success of the anonymization method heavily depends on the success in preserving the data utility for specific purposes.

There is a vast literature on data privacy protection, including several techniques developed for data anonymization. Although based on different methods and achieving varied results, they all seem to have in common some trade-off between increased level of privacy protection with decreased levels of data integrity [29]. The latter being due to processes of data aggregation, generalisation, addition of random values, partial removal of attributes or sparse data. This seems to undermine the full potential promised by big data with fine spatial and temporal resolution in supporting research and planning; however, [19] argues that location-aware applications often require much less spatial and temporal resolution and coverage than the ones provided by underlying location system. This implies that some levels of spatial generalisation and temporal aggregation may still be very effective for most research and planning applications.

This review demonstrates that there are many opportunities in the use of crowdsourced data collected via GPS-enabled smartphone applications for both research and real world city planning related to bicycling and active transportation. This is supported by the progressive technological developments of these devices, the growing social adoption of smartphones worldwide, and the increased interest of people in fitness, health and leisure applications. These opportunities, however, come accompanied by challenges associated with large volumes of data, the low quality of some records, sampling bias, and privacy concerns. Contributing to a growing body of research addressing these challenges, the following sections of this paper describe, assess and discuss a sequence of processes developed with the objective of turning a large crowdsourced bicycling database into a high-quality, validated, anonymised and easily accessible collection of datasets. The overall goal is to develop data that are useful for a variety of research and planning purposes and that can assist Australia to move forward towards improved levels of active transportation adoption.

3. Methods and Application—Crowdsourced Data Processing

3.1. RiderLog: Crowdsourced Bicycling Data

Bicycle Network [30] is an Australian nonprofit organisation dedicated to "getting more people cycling more often". As part of their Ride2Work campaign, Bicycle Network developed the RiderLog application, a free smartphone application that captures GPS locations of cyclists as they complete cycling trips. The app also collects information associated with the rider, such as age and gender, and the cycling ride, such as trip purpose. Additional functionality of the application calculates some attributes of the cycling route when they are completed, such as overall distance, duration, average speed, and top speed. The RiderLog app is used by people interested in monitoring their personal mobility and physical activity performance. This research uses RiderLog bicycling route data covering all of Australia from May 2010 to May 2014, including 148,769 bicycle journeys undertaken by 9727 cyclists. If analysed as a collection, these data have the potential to unravel macro patterns across Australian cities that can assist better understanding of human mobility and better planning of cycling infrastructure. However, in its raw format, the data are not suitable for any analysis or visualisation. Several processing operations are required to transform the data from the raw format to one that has good quality, validity, privacy and compatibility with digital analytical systems, such as geographical information systems (GIS) and statistical packages.

3.2. Research Design

Figure 1 presents the framework of data processing sequence applied to the RiderLog database. Two products are generated: RiderLog-Public and RiderLog-Licensed. RiderLog-Public results are generated from a preliminary cleaning process based on automated scripts, followed by a broad anonymization process in which routes are converted into spatially generalised points within a 500 m grid, and published openly to the public for non-commercial uses through the CityData store platform, linked to other data and metadata portals. RiderLog-Licensed extends the automated cleaning process using additional manual editing; this is necessary as this dataset represents cycling tracks as line features. The fine anonymization uses random trimming to blur origins and destinations, but still maintains details of the rest of the tracks. Its metadata is published and the data can be made available for research and city planning purposes through a license agreement. A validation process is developed with the final cleaned data to assess the representativeness or bias of the sample against the population. Each processing stage indicated in the framework in Figure 1 is described and assessed as follow.

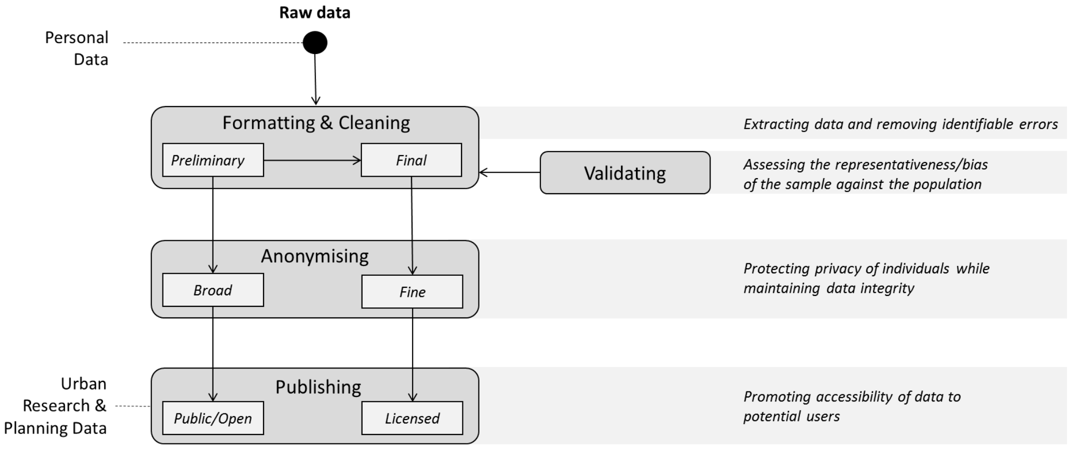

Raw data
Personal Data
Formatting & Cleaning
Preliminary
Final
Validating
Extracting data and removing identifiable errors
Assessing the representativeness/bias of the sample against the population
Anonymising
Broad
Fine
Protecting privacy of individuals while maintaining data integrity
Urban Research & Planning Data
Publishing
Public/Open
Licensed
Promoting accessibility of data to potential users

Figure 1. Crowdsourced data processing sequence to transform data from personal to urban research and planning purposes.

3.3. Processing 1—Formatting and Cleaning the Data

The unprocessed RiderLog data were made available by the Bicycle Network as a 421 MB plain text flat file. The data consisted of a continuous string with latitude/longitude pairs and other attributes associated with the rider and the ride, such as rider identification code (ID), age, gender, route ID, state, date, time, purpose, distance, duration, average speed, and top speed. The strings containing latitude/longitude pairs were irregularly interspersed with words and other characters. In this format, the data are unsuitable for any statistical analysis, spatial analysis or visualisation in a geographical information system.

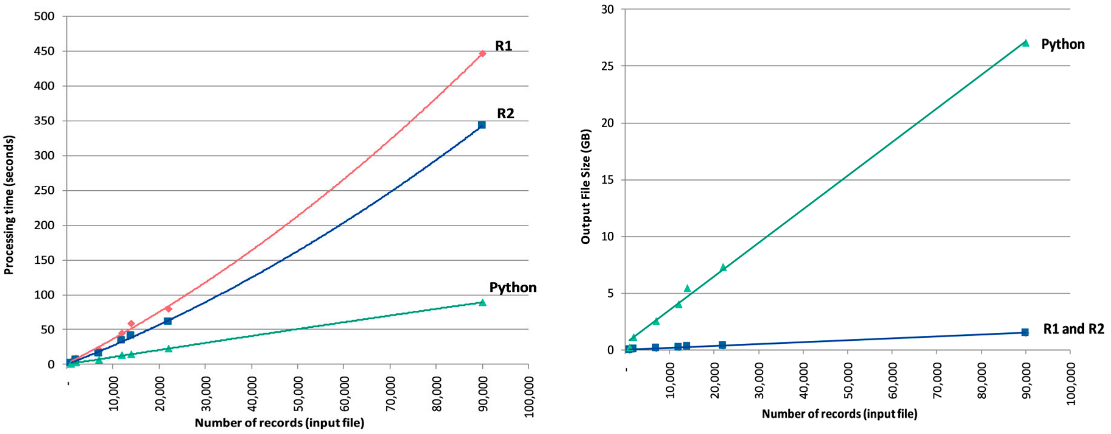
In order to turn this file into a collection of datasets that can be visualised and analysed, a series of formatting operations are required. They include removing miscellaneous characters; extracting XY coordinates; extracting and formatting time; estimating timestamps for all coordinate points; extracting other attributes associated with records; and structuring the data so they are compatible inputs for geographic information systems and other analytical packages. Developing these formatting processes was challenging due to original data format which was unstructured and very large. Previous work detailed in [9], developed initial efforts in RiderLog data formatting based on Microsoft Excel and Visual Basic for Applications (VBA) scripting. This method allowed the partial analysis and visualisation of RiderLog data, but proved insufficient to handle to full database.

Therefore, new scripts using R and Python have been developed to perform the tasks described above and applied to all RiderLog datasets for 2010–2014. Although all three scripts performed significantly better than the previous VBA method, their tests demonstrated the extent that different languages, platforms, and coding structures can significantly impact performance, in terms of processing time and output file size.

Two scripts were developed in R using three packages: Dplyr, Lubridate, and Splitstackshape. Dplyr provided basic tools for data manipulation (e.g., rearranging columns), while Splitstackshape [31,32] was used to split and stack the route column. The latter is particularly suited for this task as it is designed to handle unbalanced data that do not have the same lengths for all values, as is the case with the route variable [32]. Lubridate is designed to assist parsing dates and times [33]. The first script (R1) operated by first creating all variables required for a single trip, namely the parsed dates and times, as well as calculated duration. Each row was then split and stacked on the route variable, so that each coordinate pair represented one row in the data frame. This step had the effect of multiplying the length of the data frame by the total number of coordinate pairs across all trips. The stacked route variable was split again to create separate columns for the latitude and longitude values. The final processing steps used a function to create the sequential leg variable for each trip, another function to create the modelled time stamp, and a loop to create the origin and destination label column for each trip. The second R script, R2, was similar to the first, though rather than split and stack the route within the original data frame, the route ID and route variables were subset into a separate data frame. The split and stack function was then performed on the separate data frame and then joined to the original data frame once completed. It is more efficient to split and stack only the relevant columns since it avoids replicating each column during the stack processing.

The Python script required two packages, Pandas and Datetime. The Pandas package is designed to simplify data handling by providing data structures to manipulate and analyse data frame objects. Datetime handles extraction and manipulation of date and time variables. The script was written using a similar approach as the R2 script in terms of creating a separate route object to handle splitting and stacking the latitude and longitude variables, and joining them back to the original data frame. One exception was that the Python script required additional processing to format dates so they could be used by the Datetime package. Python version 2.7 (Python Software Foundation, Wilmington, DE, USA) was run using Jupyter notebook launched via the Anaconda Navigator. Running the Python script resulted in some errors (out of bounds date time, value error, type error, and memory errors) and required considerable troubleshooting. This indicates that additional preparation of the data may be necessary.

The VBA script took 80 min to process just one of the states of Australia—New South Wales (NSW) dataset, which had to be split into four files due to size of the output dataset produced. The format processing time for the same dataset was reduced to 79 s for script R1, 61 s for script R2, and 22 s for the Python script. All scripts were run in the same desktop computer with an Intel(R) Xeon(R) CPU E5-1660 v3 @ 3.00 GHz, Core i5-3320, 64 GB RAM, and 64-bit Windows 7 Enterprise Service Pack 1 OS. Comparison of the total processing time for the whole RiderLog data by size of the original file in number of records, is illustrated in Figure 2a. Another criterion for comparison is the size of the output file generated by the script. While the VBA and R scripts produced a similar result (an output file with approximately 400 MB for the NSW dataset), the Python script generated a file much larger (17 times larger for NSW). This is illustrated in Figure 2b.



(**a**) Processing time by number of records (**b**) Output file size by number of records

Figure 2. Comparative performance of R and Python scripts.

Each method has its own strengths and weaknesses. VBA allowed the preliminary work, but proved slow and lacked full automation which led to additional time cost. The Python script is the fastest; however, it produces the largest output files, and requires significant troubleshooting due to data complexity and inconsistencies. R scripts are slightly slower than the Python script, but still reasonably fast, and produce files with a size that is easier to handle than Python's, particularly in an online and cloud environment. Another aspect to consider is that the R scripts reduce the potential for human error as more steps are automated. Moreover, R is isolated from other software on the computer and runs in the background, allowing other concurrent uses of computer resources without significant loss of run time performance. At the current stage of development, R2 script has been selected as the most appropriate for automating the formatting future batches of RiderLog crowdsourced data.

The formatting phase produces a file for each state of Australia as a CSV table with Lat-Long pairs of all points for all routes with associated date-time stamp and characteristics of the riders and routes. Based on this file, another dataset is produced, consisting of a feature file representing all routes as lines across the city, with associated characteristics, and start and end date-time. While the first dataset is very useful for visualisation of bicycling dynamics along time, the latter is a rich database for analysis of cycling route choice, gender inequalities, spatial distributions and variances, and so forth.

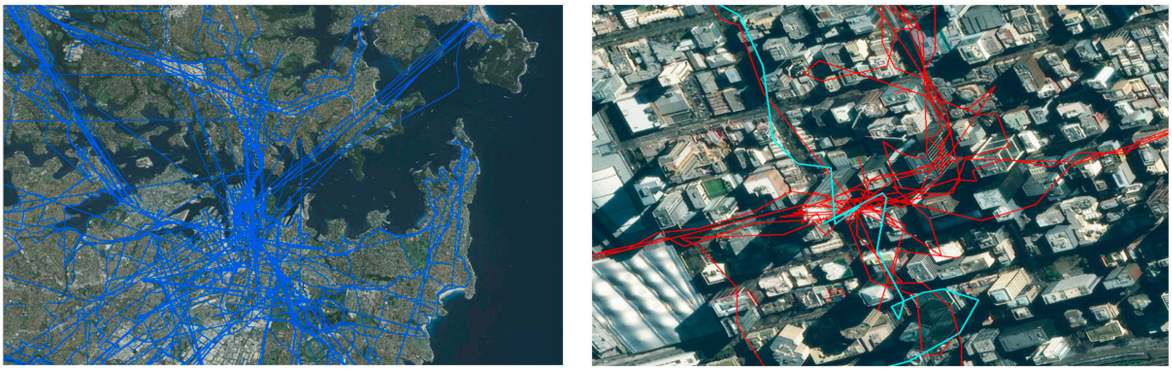
The line feature file is produced within a GIS by linking GPS coordinates of a cycling track with the same ID into a route event for all cycling tracks recorded. However, when this file is visualised geographically, overlaying a map with the road network or a satellite image, several issues are noticeable. These consist of imprecisions from the smartphones used to collect data, variances associated to GPS drift when the device is stationary (for example in a road intersection with traffic lights), weak or loss of signal due to urban canyons or excessive clouds, interruption of a track due to flat battery, accidentally turning the application on or off, etc. Therefore, before the data are analysed, they need to be cleaned.

There is a limit of how much the data can be improved due to limitations of the device capturing the raw data or the environmental context in which the individuals are when collecting data. The cleaning process should eliminate identifiable errors in the database, though intrinsic imprecisions of the data due to the current technology would remain. Eliminating identifiable errors will avoid the use of wrong data in analysing cycling patterns.

Table 1 summarises the most commons errors identified in the data, and describe scripts developed in Python to totally or partially automate the cleaning processes. Figure 3 illustrates some of these errors.

Table 1. Common errors in RiderLog data and respective cleaning process.

Error Type	Error Description	Cleaning Process and Scripts
Out of place tracks	Cycling tracks which are outside the boundaries of its respective state.	The script identifies GPS points which location is outside the geographic boundaries of the respective state, then removes these records and saves them in another file so they can be incorporated into their respective state's data later.
Too short tracks	Cycling tracks with two GPS points, due to the accidental turning on and off of the app.	The script filters routes with only two points and removes the corresponding GPS points.
Too long segments	Long-distance tracks with (partial or total) straight lines crossing street blocks due to GPS signal.	The script calculates the distance between two consecutive points along a route and selects tracks with legs 500 m or longer; it then assesses the intersections between the selected tracks with long legs and urban blocks to identify if the long legs are errors or genuine data. Errors are removed (the long segment), and the remaining genuine data is edited manually.
Inconsistent origin and destination (O-D) locations	Origins and destinations of tracks are input by users, having inconsistent names	Through editing in GIS, origin and destination locations of all tracks are replaced by standardized area codes, using the Statistical Areas Level 1 (SA1) and Level 2 (SA2) by the Australian Bureau of Statistics.



(**a**) Cycling tracks with long legs crossing blocks and water bodies

(**b**) Cycling tracks with location imprecisions due to urban canyon and GPS drift in intersections

Figure 3. Types of errors and noise present in the RiderLog data before cleaning.

3.4. Processing 2—Validating the Data

Currently, there is no conventional information on cyclists' specific routes for commuting trips that are complete (including the entire population) and geographically detailed. The Australian Census's Journey to Work (JTW) dataset produced by the Australian Bureau of Statistics (ABS) provides commuting flows across regions through an origin-destination matrix for the entire population aggregated into statistical areas, but without route details. The data from the RiderLog app provide a potential alternative or supplement to the JTW as it provides a large sample of tracked cycling journeys. These crowdsourced data, however, must be validated to investigate to what extent the crowdsourced sample is representative of the overall cycling population. The aim of this processing phase is to undertake a validation to ensure the data are used in appropriate analytical and planning contexts.

If the sample is representative, it means that the Riderlog can be used for planning and management of bicycling infrastructure. If the sample is not representative, it still can be used for research and planning, but the direction of the bias must be made explicit to users, and this information should be taken into consideration in any analysis of the data.

In order to compare JTW data and RiderLog, some modelling is required so their formats are comparable. An agent-based model (ABM) has been developed to transform the Origin-Destination matrix from JTW data into geographical tracks using shortest paths from origin to destination along the road network [34]. Then, these tracks can be compared to the RiderLog tracks. Only JTW Origin-Destinations pairs with bicycle as the main mode of transport are used in the modelling, and only RiderLog routes with transport purpose are used in the validation. There are no current data available to validate recreational cycling trips.

A validation experiment using the agent-based model described above indicated that RiderLog crowdsourced data is a representative sample of bicycling distributions in Sydney, when compared to the Census 2011 JTW data. There is a strong correlation (R^2 0.94) between the best fit functions of percentage of cyclists per distance for the RiderLog database and the ABM simulated JTW, as shown in Figure 4. Moreover, both datasets resulted in very similar parameters: average journey distance (8.5 km for RiderLog and 8.4 km for ABM simulation); average journey duration (35 min for both databases); and average speed (15.4 km/h for RiderLog and 14.9 km/h for ABM simulation). The results of the validation process for the RiderLog datasets are included as a comment in the respective published metadata.

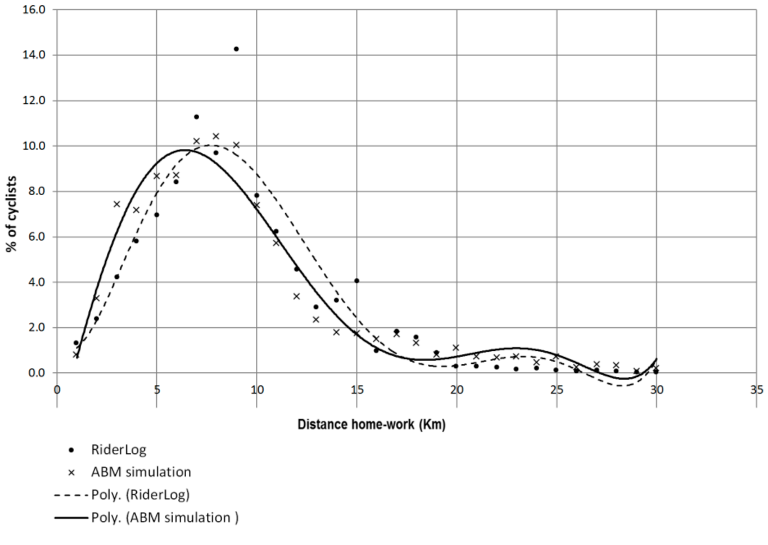


Figure 4. Comparison of best fit functions of % of cyclists per distance for RiderLog data and agent-based model (ABM)-simulated Journey to Work survey data (ABM/JTW).

This result is promising, since RiderLog is updated daily with geographically detailed data. In contrast, JTW data is aggregated to large regions and updated in Australia only every five years. RiderLog can provide data with spatial and temporal scales that are much more useful for urban planning and place making activities. It can assist in identifying those routes which are most traversed by riders which provides evidence for new or upgrade bicycling infrastructure.

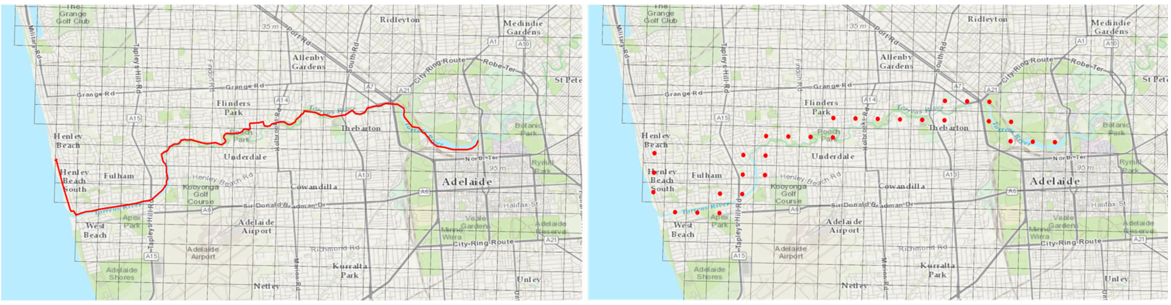
This process needs to be extended for all state datasets, and it can be repeated every five years following the census data, or used with other data sources that may become available. Further improvements of the validation process would involve a comparison between RiderLog and JTW by selected demographics (gender and age group, for example).

3.5. Processing 3—Anonymising the Data

A good anonymization procedure needs to balance the protection of people's privacy while maintaining as much of the integrity of the data as possible. Different anonymization methods have been applied to the two RiderLog data products: spatial generalisation for RiderLog-Public, and randomised trimming for Riderlog-Licensed.

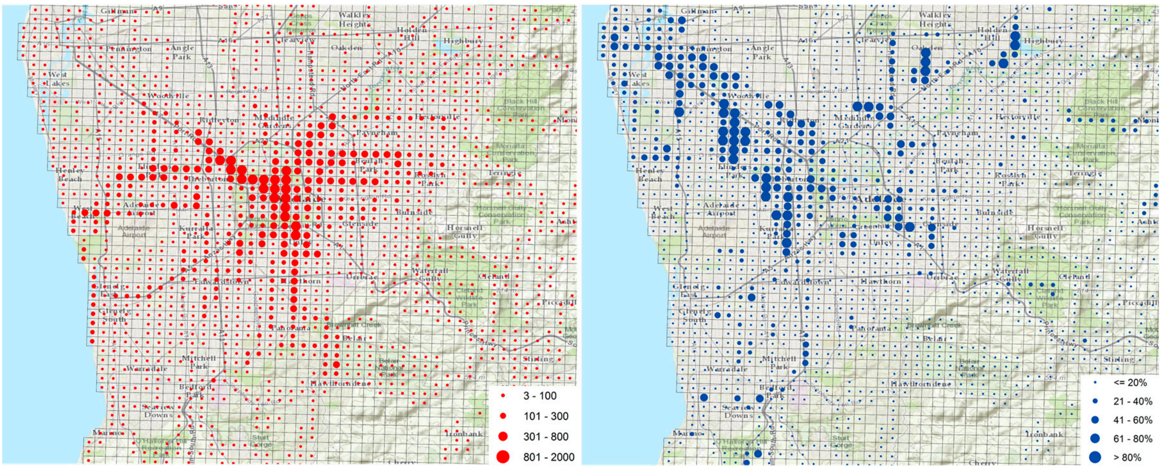
Riderlog-Public data has been anonymised through a spatial generalisation of routes using a 500 m × 500 m grid. In this method, each cycling route has been decomposed into a sequence of points, in which each point is not the exact location of the route (Figure 5a) but the centroid of the grid cell crossed by the route (Figure 5b). A criterion used here was the intention to preserve the cycling routes, as this is an information relevant for planning, but rarely available. Strava Data, for example, does not allow the reconstruction of routes, as it divides cycling itineraries in road segments without a linking ride identifier. In the anonymised RiderLog-Public data, routes can be reconstructed as a sequence of cells with a size large enough to protect privacy, but also with a size appropriate for planning purposes. Therefore, although without a precise location, route selection can be assessed in terms of the characteristics of the cells, such as existing cycling infrastructure, topographic profile, density of points of interest, and so forth.

Data is removed for cells with less than three cyclists. All information of the rider and route are linked to each point. Therefore, summations and averages can be calculated for all the routes by cell. Figure 6a, for example, illustrates the distribution of total cycling routes by cell for the period 2010–2014 in Greater Adelaide, indicating a high concentration around the city centre. Figure 6b presents the proportion of cycling routes done by female riders by cell for the same area and period. Although the overall rate of female riders in Greater Adelaide is low (18.2%), some areas have significant presence of female riders, with a proportion above 60%. Moreover, this anonymised format allows that characteristics of the rider and the ride be combined into summations or averages by cell, such as "distribution of recreational cycling trips by males with age equal or above 56 years".



(**a**) Original route ID 1498 (**b**) Spatially generalized route ID 1498

Figure 5. Spatial generalisation of a route.

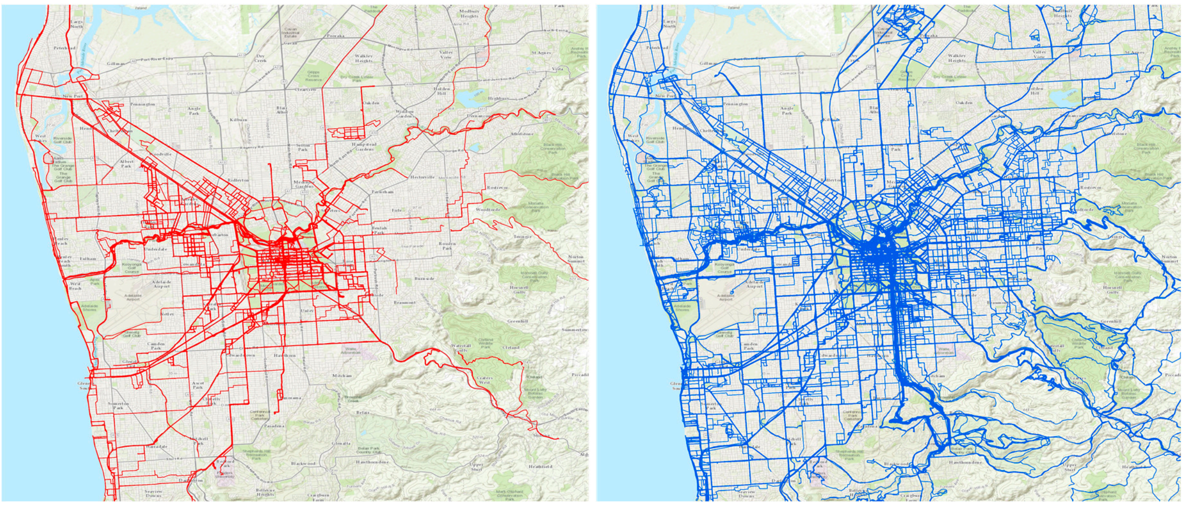


(**a**) Total routes (**b**) % of female riders

Figure 6. Anonymised RiderLog-Public data for Greater Adelaide, 2010–2014.

Riderlog-Licensed data consist of line features describing the location and characteristics of individual cycling routes. This data has been anonymised through a randomised trimming process of origins and destinations within individual cycling routes. In this process, a new location (latitude-longitude pair) for the origin and/or destination is defined as the intersection point of the route and a buffer around the actual origin or destination. The size of the buffer is defined randomly considering a given value range. The points comprising the route between the new origin and new destination remain unchanged. The overall distance and duration of the routes, as well as average and top speed, are preserved as the original values, despite the fact that the geometry of the route after anonymization may have become shorter. Therefore, places of residence, employment, or usual visit for any cyclist cannot be identified. However, since the data still presents details of itineraries, it is restricted to be accessed only by users with non-commercial purposes, no sharing with third parties, and under license agreement that allows only the publication of the aggregated analysis.

Figure 7 illustrates the RiderLog Licensed data, including all the route features for Greater Adelaide for the period 2010–2014 filtered by gender.



(**a**) Female rides (**b**) Male rides

Figure 7. Anonymised RiderLog-Licensed data for Greater Adelaide, 2010–2014.

3.6. Processing 4—Publishing the Data

In recent years, several open data initiatives have been implemented worldwide, with transparency and data reuse as major aims [35]. The Australian Urban Infrastructure Network (AURIN) is a significant example of a national initiative to promote an online workbench of data and tools to assist data sharing and research [36,37].

The aim of publishing data and metadata on the Web is ultimately to enable potential consumers to discover, access and use existing data in various applications. This means that resources saved from avoiding the reproduction of existing data can be allocated to data exploration and analysis. Several factors can influence the success of a publishing system employed [35]. First, data should be discoverable, and this depends in part on the portals in which they are available or advertised. Second, good quality metadata should also be published, so potential users can discover data contents and characteristics, and assess if it is a suitable fit for their intended uses. Moreover, data should be made accessible through formats easily compatible with computer data analytical systems, preferably non-proprietary.

Implicit in what could be termed the social contract for crowdsourcing data is the return of the data to the crowd. The data should be made available to participants, and also to researchers and other interested parties and the public, in an open format and free of charge. RiderLog Public datasets and metadata are made available as open data.

The Geonode open source geospatial content management system was chosen as the data publishing platform for the CityData store, the primary location where RiderLog-Public is published (https://citydata.be.unsw.edu.au/). This platform provides an intuitive user interface for uploading and documenting data, for data discovery and download, and it automatically provides standard web service Application Programming Interfaces (APIs) on published data. The Geonode platform has an active and responsive developer community. It has been developed with funding from World Bank and other sponsors on the back of a suite of component open source products: Django Python web framework; PostGIS spatial database; GeoServer OGC-compliant spatial web services; pyCSW CSW-compliant metadata catalogue; Geospatial Python libraries; OpenLayers and GeoExt web mapping libraries.

Geonode's support for the Open Geospatial Consortium (OGC) standard Catalogue Services for the Web (CSW) API allows metadata to be syndicated to other catalogues, such as Research Data Australia, which is maintained by the Australian Government's Australian National Data Service (ANDS). The data itself may similarly be accessed and used by other systems such as AURIN via the OGC's open standards for web access to spatial data: Web Map Service (WMS) for maps of data in common image formats; Web Feature Service (WFS) for vector data, and; Web Coverage Service (WCS) for raster data. Geonode is increasingly used with the QGIS free and open source Geographic Information System (GIS) which also supports these standards, as do all the major commercially available GIS packages.

The CityData store provides a home page for each RiderLog-Public dataset, including title, abstract and other metadata, a graphic thumbnail and direct access to download, web service connect and other actions on the dataset. This home page is used to mint a digital object identifier (DOI) for the data, and is linked from the corresponding research DOI (e.g., for this paper). Importantly, the data home pages are accessible to web search engines, facilitating far more effective discovery than is possible through closed catalogues.

An open licence from the Creative Commons suite is applied to the RiderLog-Public data, specifically the Creative Commons Attribution-Non-commercial 4.0 International (CC BY-NC 4.0) licence. This allows anyone to "share—copy and redistribute the material in any medium or format", and "adapt—remix, transform, and build upon the material" under the following terms: (a) Attribution—a user must give appropriate credit, provide a link to the license, and indicate if changes were made. A user may do so in any reasonable manner, but not in any way that suggests the licensor endorses the user or the data use; and (b) Non-commercial—a user may not use the material for commercial purposes.

RiderLog-Licensed will have metadata published in Research Data Australia with a DOI. Access to this data for download will be granted by the UNSW City Futures Research Centre for non-commercial purposes through a restrictive license agreement (approved project, credit attribution, not sharing the data with third parties).

Finally, in order to disseminate RiderLog-Public to a wider audience, a visualisation of the data through dashboards with synchronised maps and graphs providing some relevant analytics is made available to any user interested in exploring the data without downloading the data or advanced skills in data analytics. The visualisation platform is built combining Tableau Public and ESRI StoryMaps platforms, and made accessible through the urban big data visualisation website "CityViz" (https://cityfutures.be.unsw.edu.au/cityviz/).

4. Discussion and Conclusions

The value of RiderLog has been identified by the authors as a unique source of data on cycling for the Australian context due to its extensive coverage combined with fine scale, both in space and time, and its bottom-up nature. If made available, these data can complement traditional aggregated data sources on transportation, and potentially assist research and planning related to topics such as mobility behaviour, active transportation, urban infrastructure, and urban well-being.

This paper described and assessed a series of data processing steps put in place to convert a raw crowdsourced data by smartphones into a robust national and longitudinal database of urban mobility with bicycles. These processes included data formatting, cleaning, validation, anonymization and publishing. With these complete set of processes the data from individuals are now available to academia, government, non-governmental organisations, industry, and the broader community so that it may be used to support better city planning and policymaking.

The processed RiderLog datasets are a significant contribution to transport research and urban planning. Data processing produced data in a format that is compatible with non-proprietary analytical software, eliminated easily identifiable errors, assessed bias in the data against census data, protected privacy of individuals with as much preservation of data integrity as possible, created keys to spatially link the data to other traditional data sources such as the census, released the data in open and free manner, published the data and metadata in multiple portals to increase discoverability, and produced an online data visualisation for non-expert audience. Since most of these processes for the RiderLog-Public are automated, they can be applied to new batches of data with ease and low cost, contributing to the maintenance of a national and longitudinal open database on cycling activity.

Despite these significant achievements, the work still has limitations that should be addressed in future research. Specifically, the cleaning and anonymization routines written to date do not fully automate the workflow from data collection to data publication. This presents a challenge as the Riderlog app is continually in use. An automated workflow which includes the establishment of a direct data service to the underlying database is required to ensure the resultant published data and map visualisation remain current, and that rapid analytics can be undertaken by the city planner or policymaker. Future advanced scripting is also required to clean RiderLog-Licensed complex route formation due to GPS imprecisions, such as scripts to automate snapping routes to the road network.

In this paper, we have explained the underlying workflow and challenges in processing urban big data derived from a smartphone app known as Riderlog. As we live in an increasingly urbanized world, it is crucial that we better plan our cities and realise the potential of big data. In this research we have successfully acquired, formatted, cleaned, anonymised, visualised and published crowdsourced fine-scale cycling data for cities across Australia. This paper demonstrates that research has a key role in building the bridge between data generation and data quality, ethics, and availability as we move to an era of open data and city analytics.

Acknowledgments: The authors would like to acknowledge the support of Bicycle Network for providing the crowdsourced RiderLog data; the PLUS Alliance initiative for promoting and funding the collaboration between UNSW and the Arizona State University; and the Australian National Data Service (ANDS) for funding the High Value Collection Project.

Author Contributions: Simone Z. Leao, Scott N. Lieske and Chris J. Pettit have conceived, designed and coordinated this research project. Lindsey Conrow has developed the main script for data extraction and formatting. Vandana Tomar has developed the cleaning scripts and performed manual editing of the datasets. Simone Z Leao has developed the validation and anonymization processes. Jonathan Doig has implemented the City Data store where the data is published. All of the authors contributed to the writing of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Banister, D. *Transport Planning*, 2nd ed.; Spon: London, UK, 2002.
2. Alsger, A.; Assemi, B.; Mesbah, M.; Ferreira, L. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transp. Res. Part C* **2016**, *68*, 490–506. [CrossRef]
3. Gschwender, A.; Munizaga, M.; Simonetti, C. Using smart card and GPS data for policy and planning: The case of Transantiago. *Res. Transp. Econ.* **2016**, *59*, 242–249. [CrossRef]
4. Heipke, C. Crowdsourcing geospatial data. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 550–557. [CrossRef]
5. O'Brien, O.; Cheshire, J.; Batty, M. Mining bicycle sharing data for generating insights into sustainable transport systems. *J. Transp. Geogr.* **2014**, *34*, 262–273. [CrossRef]
6. Fishman, E. Bikeshare: A Review of Recent Literature. *Transp. Rev.* **2016**, *36*, 92–113. [CrossRef]
7. Estelles-Arolas, E.; Gonzales-Ladron-de-Guevara, F. Towards an integrated crowdsourcing definition. *J. Inf. Sci.* **2012**, *38*, 189–200. [CrossRef]
8. Lane, N.D.; Miluzzo, E.; Lu, H.; Peebles, D.; Choudhury, T.; Campbell, A.T. A survey of mobile phone sensing. *IEEE Commun. Mag.* **2010**, *48*, 140–150. [CrossRef]
9. Pettit, C.J.; Lieske, S.N.; Leao, S.Z. Big bicycle data processing: From personal data to urban applications. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 173–179. [CrossRef]
10. Dill, J. Bicycling for transportation and health: The role of infrastructure. *J. Public Health Policy* **2009**, *30*, S95–S110. [CrossRef] [PubMed]
11. Casello, J.; Akram, O.N.; Rewa, K.C.; Hill, J. An analysis of stated preference and GPS data for bicycle travel forecasting. In Proceedings of the Transportation Research Board 90th Annual Meeting, Washington, DC, USA, 23–27 January 2011.
12. Krenn, P.J.; Oja, P.; Titze, S. Route choices of transport bicyclists: A comparison of actually used and shortest routes. *Int. J. Behav. Nutr. Phys. Act.* **2014**, *11*. [CrossRef] [PubMed]
13. Yeboah, G. *Understanding Urban Cycling Behaviours in Space and Time*; Northumbria University: Newcastle upon Tyne, UK, 2014.
14. Hood, J.; Sall, E.; Charlton, B. A GPS-based bicycle route choice model for San Francisco, California. *Transp. Lett.* **2011**, *3*, 63–75. [CrossRef]
15. Hudson, J.G.; Duthie, J.C.; Rathod, Y.K.; Larsen, K.A.; Meyer, J.L. *Using Smartphones to Collect Bicycle Travel Data in Texas (No. UTCM 11-35-69)*; University Transportation Center for Mobility, Texas Transport Institute: College Station, TX, USA, 2012.
16. Romanillos, G.; Austwick, M.Z.; Ettema, D.; De Kruijf, J. Big data and cycling. *Transp. Rev.* **2016**, *36*, 114–133. [CrossRef]
17. Hecht, B.; Stephens, M. A tale of two cities: Urban biases in volunteered geographic information. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014.
18. Blanc, B.; Figliozzi, M.; Clifton, K. How representative of bicycling population are smartphone application surveys of travel behaviour? *Transp. Res. Rec.* **2016**, *2587*, 78–89. [CrossRef]
19. Beresford, A. Privacy issues in geographic information technologies. In *Frontiers of Geographic Information Technology*; Rana, S., Sharma, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Chapter 13.
20. Ellison, R.; Greaves, S. Travel time competitiveness of cycling in Sydney, Australia. *Transp. Res. Rec.* **2011**, *2247*, 99–108. [CrossRef]

21. Stopher, P.R.; Jiang, Q.; FitzGerald, C. Processing GPS data from travel surveys. In Proceedings of the 28th Australasian Transport Research Forum, Sydney, Australia, 28–30 August 2005.
22. Schuessler, N.; Axhausen, K.W. Processing raw data from Global Positioning Systems without additional information. *Transp. Res. Rec.* **2009**, *2105*, 28–36. [CrossRef]
23. Kandel, S.; Heer, J.; Plaisant, C.; Kennedy, J.; van Ham, F.; Riche, N.H.; Weaver, C.; Lee, B.; Brodbeck, D.; Buono, P. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Inf. Vis.* **2011**, *10*, 271–288. [CrossRef]
24. Shneiderman, B.; Plaisant, C. Sharpening analytic focus to cope with big data volume and variety. *IEEE Comput. Graph. Appl.* **2015**, *35*, 10–14. [CrossRef] [PubMed]
25. Fayyad, U.; Piatetsky-Shapir, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **1996**, *17*, 37–54.
26. Laney, D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Available online: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed on 10 January 2017).
27. Demchenko, Y.; Grosso, P.; Laat, C.; Membrey, P. Addressing big data issues in scientific data infrastructure. In Proceedings of the 2013 International IEEE Conference on Collaboration Technologies and Systems, San Diego, CA, USA, 20–24 May 2013.
28. Song, Y.; Dahlmeier, D.; Bressan, S. Not so unique in the crowd: A simple and effective algorithm for anonymizing location data. In Proceedings of the PIR@SIGIR, Gold Coast, Australia, 6–11 July 2014.
29. Li, T.; Li, N. On the tradeoff between privacy and utility data in data publishing. In Proceedings of the 15th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining, KDD'09, Paris, France, 28 June–1 July 2009; pp. 517–528.
30. Bicycle Network. Riderlog—Make Every Ride Count. Available online: https://www.bicyclenetwork.com.au/general/programs/1006/ (accessed on 10 August 2016).
31. Wickham, H.; Francois, R. Dplyr: A Grammar of Data Manipulation. Available online: https://github.com/tidyverse/dplyr (accessed on 10 December 2016).
32. Mahto, A. Splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values. Available online: https://CRAN.R-project.org/package=splitstackshape (accessed on 10 December 2016).
33. Grolemund, G.; Wickham, H. Dates and times made easy with lubridate. *J. Stat. Softw.* **2011**, *40*, 1–25. [CrossRef]
34. Leao, S.Z.; Pettit, C.J. Mapping bicycle patterns with an agent-based model, Census and crowdsourced data. In *Agent-Based Modelling of Urban Systems, Lecture Notes in Artificial Intelligence*; Namazi-Rad, M.R., Padgham, L., Perez, P., Nagel, K., Bazzan, A., Eds.; Springer: Cham, Switzerland, 2017; pp. 112–128.
35. Attard, J.; Orlandi, F.; Scerri, S.; Auer, S. A systematic review of open government data innitiatives. *Gov. Inf. Q.* **2015**, *32*, 399–418. [CrossRef]
36. Sinnott, R.O.; Bayliss, C.; Bromage, A.; Galandg, G.; Grazioli, G.; Greenwood, P.; Macaulay, A.; Morandini, L.; Nogoorani, G.; Nino-Ruiz, M.; et al. The Australian urban research gateway. *J. Concurr. Comput.* **2014**, *27*, 358–375. [CrossRef]
37. Pettit, C.J.; Barton, J.; Goldie, X.; Sinnot, R.; Stimson, R.; Kvan, T. The Australian urban intelligence network support smart cities. In *Planning Support Systems and Smart Cities, Lecture Notes in Geoinformation and Cartography*; Geertman, S., Ferreira, J., Jr., Goodspeed, R., Stillwell, J., Eds.; Springer: Cham, Switzerland, 2015.