Journal of Transport & Health 6 (2017) 93–104

Contents lists available at ScienceDirect

Journal of Transport & Health

journal homepage: www.elsevier.com/locate/jth

Exploiting crowdsourced geographic information and GIS for assessment of air pollution exposure during active travel

Yeran Sun[a,*], Yashar Moshfeghi[a], Zhang Liu[b]

[a] Urban Big Data Centre, University of Glasgow, Glasgow G12 8RZ, United Kingdom
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Keywords:
Crowdsourced geographic information
Strava Metro
Active travel
Air pollution exposure
Particulate matter

ABSTRACT

Improvement on assessment of air pollution exposure will enhance assessment of health risk-benefit when active travel (cycling and walking). Earlier studies assessed air pollution exposure according to travel time and city-level air pollution. The lack of spatially fine-grained travel data is a barrier to an accurate assessment of air pollution exposure. Due to a high-level spatial granularity, Strava Metro provides an opportunity to assessing air pollution exposure in combination with spatially varying air pollution concentrations. Strava Metro anonymized and aggregated a large volume of users' traces to streets for each city. In this study, to explore the potential of crowdsourced geographic information in research of active travel and health, we used Strava Metro data and GIS technologies to assess air pollution exposure in Glasgow, UK. Particularly, we incorporated time of the trip to assess average inhaled dose of pollutant during a single cycling or pedestrian trip. Empirical results demonstrate that Strava Metro data provides an opportunity to an assessment of average air pollution exposure during active travel. Additionally, to demonstrate the potential of Strava Metro data in policy-making, we explored the spatial association of air pollution concentration and active travel. As a result, we identified areas that require investment priority, and finally offered implications for policies.

1. Introduction

Through enhancing physical activity, active travel (cycling or walking) produces health benefit (Forsyth et al., 2012; Oja et al., 1998, 2011; Pucher et al., 2010; Wen and Rissel, 2008). At the same time, as outdoor physical activities cycling and walking are also of risks, including traffic accidents and air pollution exposure (Weichenthal et al., 2011; de Nazelle et al., 2013; Hollingworth et al., 2014). Generally speaking, recent studies support both empirically and theoretically that the total benefits of active travel tend to outweigh the risks (Tainio et al., 2016; Doorley et al., 2015; Mueller et al., 2015). As the impact of air pollution exposure on health is not as noticeable as that of traffic accidents, people are likely to ignore the harmful effect of air pollution exposure. In fact, air pollution exposure tends to have both a short-term and long-term effects on human health (WHO 2016; Anderson et al., 2012; Lipsett et al., 2011; Pope et al., 2015; Chen et al., 2012; Li et al., 2016; Shah et al., 2015).

Earlier studies assessed air pollution exposure according to travel time and city-level air pollution. However, air pollution concentrations tend to spatially vary over a city due to heterogeneously spatial distributions of motor vehicles, industrial emissions, and reconstructions. Improvement on assessment of air pollution exposure was limited due to lacks of spatially fine-grained travel data. Traditionally, stated preference surveys data (Forsyth and Oakes, 2015; Sener et al., 2009) have a good spatial coverage but a low

* Correspondence to: 7 Lilybank Gardens, Glasgow G12 8RZ, United Kingdom.
E-mail address: yeran.sun@glasgow.ac.uk (Y. Sun).

http://dx.doi.org/10.1016/j.jth.2017.06.004
Received 20 March 2017; Received in revised form 3 June 2017; Accepted 11 June 2017
Available online 16 June 2017

spatial granularity, such as census tract level; while manual counts and annual average daily bicycle (AADB) data (El Esawey, 2014) has a high spatial granularity, such as street level, but a poor spatial coverage as major roads (e.g., highways and motorways) other than minor roads are covered. Volunteer geographic information (VGI) or crowdsourced geographic information (CGI) paves a new way to tracking mobility and travel activities of people with a large spatial granularity (Sun and Li, 2015; Steiger et al., 2015; Li et al., 2016a; Griffin and Jiao, 2015; Siła-Nowicka et al., 2015; Thakuriah et al., 2016). GPS-enabled mobile devices, such as smartphones and smartwatches, allow individuals to track and share their cycling and walking tracks (Jesticoa et al., 2016; Broach et al., 2012; Casello and Usyukov, 2014; Hood et al., 2011). In the era of Big Data, a large volume of cycling and walking traces generated by individuals are becoming potential data for studies of travel and health (Prins et al., 2014; Duncan et al., 2009; Dill, 2009; Griffin and Jiao, 2015; Sun and Mobasheri, 2017).

Recently, as a popular platform dedicated to tracking users' cycling, walking, running and hiking activities, Strava is gaining attention from both researchers and planners after a data service called Strava Metro was launched. There are millions of users uploading their rides, walks, runs and hikes to Strava each week (Strava Metro, 2016). To protect user privacy, Strava Metro anonymized and aggregated users' traces to streets for each city. Although being aggregated, Strava Metro data is of high potential in a wide range of applications, including mapping cycling activities (Jesticoa et al., 2016), investigating effects of environmental factors on cycling behavior (Griffin and Jiao, 2015; Heesch et al., 2016) and assessing air pollution when cycling (Sun and Mobasheri, 2017). Due to a high-level spatial granularity, Strava Metro provides an opportunity to assessing air pollution exposure in combination with spatially varying air pollution concentrations (Sun and Mobasheri, 2017). Moreover, by comparing cyclist counts between Strava data and realistic traffic data in count stations, some studies have demonstrated that Strava Metro data tends to be a good representation of cycling population (Jesticoa et al., 2016; Herrero, 2016). Therefore, we could utilise Strava Metro data for assessing air pollution exposure based on local air pollution concentrations instead of city-level air pollution concentrations. Additionally, we could make use of Strava Metro data to support policy-making. As Strava Metro data can indicate spatial distribution of cycles or walks across a city, we could combine the map of cycles and walks with the map of air pollution to identify: 1) areas with low-level air pollution and low-volume cycles and walks; and 2) areas with high-level air pollution and high-volume cycles and walks. Policymakers would consider investment priority in bicycle infrastructure of those areas to 1) attract more people to cycle or walk in areas of low-level air pollution; or 2) separate cyclists and pedestrians from motor vehicles in areas of high-level air pollution (e.g., major roads).

Sun and Mobasheri (2007) used Strava Metro data to estimate instantaneous air pollution exposure when cycling considering the only spatial distribution of cycling activities but no time spent cycling. In this study, we aimed to estimate average air pollution exposure during a single cycling or pedestrian trip. Particularly, we incorporated time of trip into an assessment of air pollution exposure when cycling and walking. Specifically, we assessed inhaled dose of pollutant during a cycling or pedestrian trip by considering air pollutant concentration, the duration of exposure and the ventilation rate. To demonstrate that Strava Metro data could provide an opportunity to an assessment of average air pollution exposure during a single cycling or pedestrian trip, we conducted empirical experiments in Glasgow, UK. Additionally, to demonstrate the potential of Strava Metro data in policy-making, we explored spatial association of air pollution concentration, cycling count and pedestrian count. Consequently, we identified areas with low-level air pollution and low-volume cycles and walks, and areas with high-level air pollution and high-volume cycles and walks. Finally, we discussed the empirical results and offered implications for policies.

2. Materials and methods

In this section, the approach to assessment of air pollution exposure during a single cycling or pedestrian trip is presented. Section 2.1 introduces the research data, and Section 2.2 introduces the approach to assessing air pollution exposure by using Strava Metro data and GIS technologies. Section 2.3 introduces the exploration of spatial associations of air pollution concentration, cycles and walks.

2.1. Research data

2.1.1. Strava Metro data

Strava (Strava, San Francisco, CA, USA) consists of a website and a mobile app. Strava apps installed in GPS devices such as smartphones and smartwatches can record distance, time, average speed and GPS route of each ride, run, walk or hike. Users also can upload their GPS-tracked activities from their Strava apps to the Strava's website, a social media platform where Strava users share their rides, runs, walks and hikes. Users can further add textual information to portrait their trips. Strava has collected approximately a trillion GPS points globally and keeps collecting millions of activities every week (Riordan, 2016). To exploit such a huge amount of GPS-tracked activities, Strava launched Strava Metro, a suite of data services that enables cutting-edge views into cycling and pedestrian patterns (Strava Metro, 2015). To prevent privacy issues and keep geographic information, Strava Metro anonymized and aggregated users' GPS-tracked activities to streets for each city (Strava Metro, 2015).

Recently, the Urban Big Data Centre, UK publicly released a Strava Metro dataset (Urban Big Data Centre, 2016). This dataset has 287, 833 cycling activities and 156,002 pedestrian activities (including walks, runs and hikes) within the Glasgow Clyde Valley Planning area (including Glasgow City and seven contiguous council areas) in 2015. This dataset contains three subsets with three different formats: Streets, Origin-Destination, Nodes (Strava Metro, 2015). In this study, we used the Streets and Nodes sets. Both the Streets and Nodes sets were created based on a street network which was extracted from OpenStreetMap. The Streets set contains all edges of the street network; while the Node set contains all nodes of the street network. An edge represents a street, and a node

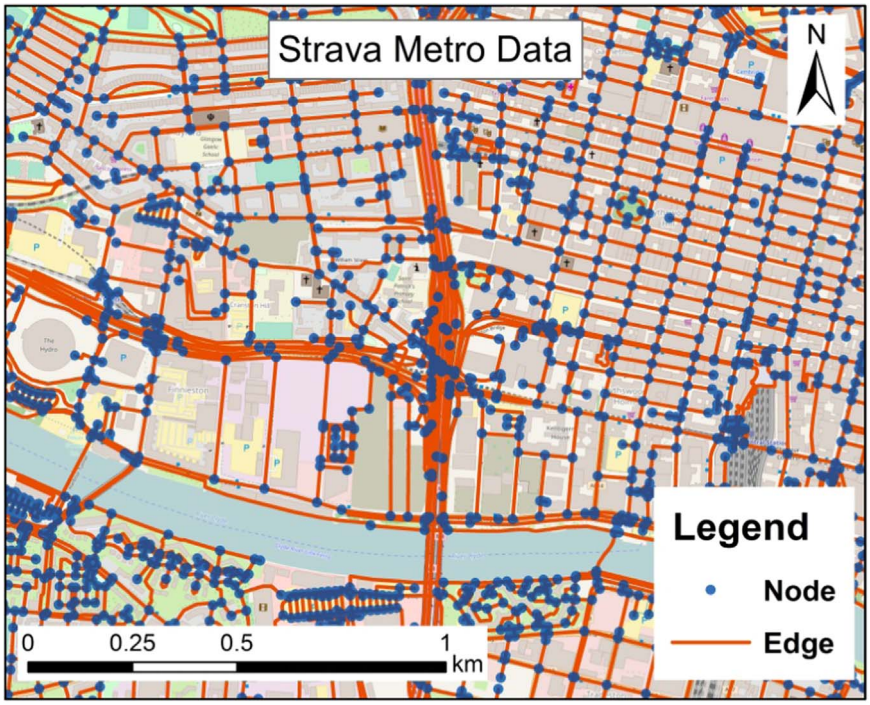

Fig. 1. Nodes and edges of Strava Metro data (Basemap: OpenStreetMap, licensed under the Open Database License).

represents an intersection of streets (see Fig. 1). Table 1 lists attributes of nodes and edges, including the count of cycling activities (regardless of unique riders) in the node (street intersection) and count of activities (regardless of unique riders) in the edge (street) at a specific time. Note that the temporal granularity is the minute level. Additionally, Strava Metro explains that median time instead of mean time was used because there were instances where cyclists or pedestrians will stop on a piece of road to go into a store or talk with a friend while their GPS is still running (Strava Metro, 2015).

Additionally, the dataset contains a file that offers demographics of the cycling and pedestrian trips (see Table 2), including average trip distance, average trip time, and user base structure by sex and age. There are over 280 thousand cycling trips and over 150 thousand pedestrian trips contributed by over 10 thousand of cyclists and over 10 thousand of pedestrians respectively. It is noted that, although this data set has a large user sample set, average annual cycling frequency and average annual pedestrian frequency of Strava users seems to be much smaller than the real frequencies. Specifically, on average, each cyclist has 21 cycling trips while each pedestrian had 14 pedestrian trips in 2015. Moreover, average annual cycling frequency is 1.5 times of average annual pedestrian frequency. Interestingly, male cyclists outnumber male pedestrians; while female pedestrians outnumber female cyclists. The largest male group for both cyclists and pedestrians is aged 35–44 while the largest female group for both cyclists and pedestrians is aged 25–34. Almost half of cycling and pedestrian trips were contributed by users aged 25–44 (25–34 and 35–44). Additionally, a large portion of trips is recreational trips (Strava Metro, 2015). Therefore, the majority of the Strava users are likely to be young and sporty cyclists and pedestrians.

2.1.2. Air pollution data

In this study, fine particulate matter (PM2.5) was used to measure levels of air pollution concentrations as it is a common air pollutant used in related studies (Anderson et al., 2012; Pope et al., 2015; Chen et al., 2012; Li et al., 2016b). Although annual mean PM2.5 estimates in 2015 are available in the Air Quality in Scotland (Ricardo Energy & Environment, 2016), the spatial resolution of the PM2.5 concentration data is 1 km × 1 km. This seems to be low considering high spatial granularity of the Strava Metro data. To better exploit spatial granularity of the Strava Metro data, we used a 100 m × 100 m PM2.5 concentration map (SAHSU, 2016) to represent annual average estimates of background PM2.5 concentrations over the study area (see Fig. 2). This 100 m annual PM2.5 map is based on European-wide models for PM2.5, developed for 2010 which are based on routine air pollution monitoring data (AIRBASE database, 2013) incorporating satellite-derived and chemical transport model estimates plus road and land use data (de Hoogh, et al. 2016). The latest version of the 100 m PM2.5 map is for 2010 but not for 2015. In this study, we used 2010 PM2.5 map to roughly represent PM2.5 concentrations across Glasgow in 2015. We mapped the annual average estimates of background PM2.5 concentrations in 2010 over the study area (see Fig. 2). Areas with low-level background PM2.5 concentrations are mainly situated in green spaces (parks, gardens or hills) or along rivers; while areas with the highest level of background PM2.5 concentrations are situated along motorways.

Table 1
Fields in Nodes and Edges files (Strava Metro, 2015).

Field	Description
node_id	Unique and permanent Node ID number for delivery.
year	Numerical year format (yyyy).
day	Numerical day format (1–365).
hour	Numerical hour format (0–24).
minute	Numerical minute format (0–59).
$Num\_act_{Bike}^{Node}$	Count of cycling activities (cycling trips) at the intersection for the day, hour and minute. This number represents the number of cycling activities (cycling trips) that meet at the intersection.
$Med\_time\_wait_{Bike}^{Node}$	Median wait time of cycling activities at the intersection for that minute.
$Num\_act_{Walk}^{Node}$	Count of pedestrian activities (pedestrian trips) at the intersection for the day, hour and minute. This number represents the number of pedestrian activities (pedestrian trips) that meet at the intersection.
$Med\_time\_wait_{Walk}^{Node}$	Median wait time of pedestrian activities at the intersection for that minute.

Field	Description
edge_id	Unique and permanent Street ID number for delivery.
year	Numerical year format (yyyy).
day	Numerical day format (1–365).
hour	Numerical hour format (0–24).
minute	Numerical minute format (0–59).
$D\_num\_act_{Bike}^{Edge}$	Count of cycling trips (regardless of unique riders) on the piece of street for the day, hour and minute. This number represents the number of cycling trips going the direction the street was digitized.
$D\_med\_time\_move_{Bike}^{Edge}$	Median time in seconds that cycling trips on the piece of street for the day, hour and minute. This number represents the time of cyclists going the direction the street was digitized.
$R\_num\_act_{Bike}^{Edge}$	Count of cycling trips (regardless of unique riders) on the piece of street for the day, hour and minute. This number represents the number of cycling trips going against direction the street was digitized.
$R\_med\_time\_move_{Bike}^{Edge}$	Median time in seconds that cycling trips on the piece of street for the day, hour and minute. This number represents the time of cyclists going against direction the street was digitized.
$D\_num\_act_{Walk}^{Edge}$	Count of pedestrian trips (regardless of unique pedestrians) on the piece of street for the day, hour and minute. This number represents the number of pedestrian trips going the direction the street was digitized.
$D\_med\_time\_move_{Walk}^{Edge}$	Median time in seconds that pedestrian trips on the piece of street for the day, hour and minute. This number represents the time of pedestrians going the direction the street was digitized.
$R\_num\_act_{Walk}^{Edge}$	Count of pedestrian trips (regardless of unique pedestrians) on the piece of street for the day, hour and minute. This number represents the number of pedestrian trips going against direction the street was digitized.
$R\_med\_time\_move_{Walk}^{Edge}$	Median time in seconds that pedestrian trips on the piece of street for the day, hour and minute. This number represents the time of pedestrians going against direction the street was digitized.

2.1.3. Comparison of Strava cycling volumes and regular cycling volumes

To examine whether Strava Metro data could be a good proxy for estimating real cycling volumes across space, we compare Strava ridership with realistic regular ridership at the street-level. UK Department for Transport offers annual average daily flow (AADF) data covering some major roads in Glasgow (Department for Transport, 2016). An AADF is an average over a full year of the number of vehicles passing a point in the road network each day. The data provides the volume of cycles and the volume of motor vehicles. We could measure the correlation between cycling volume from AADF data and cycling volume from Strava Metro data by calculating Pearson's R coefficient.

2.2. Assessment of air pollution exposure

In this study, we mainly aimed to calculate inhaled dose of pollutant during a cycling or pedestrian trip. We are not able to know realistic annual cycling frequency and pedestrian frequency of Strava users because the majority of the users tend to upload part of their cycling and pedestrian trips. We thus did not assess annual inhaled dose of Strava users when they were riding or walking. The inhaled dose of pollutant is a product of the air pollutant concentration, the duration of exposure and the ventilation rate (Schepers et al., 2015; Tainio et al., 2016). We first could calculate total inhaled dose of pollutant (PM2.5) of all Strava cyclists and that of all Strava pedestrians respectively. Afterwards, as the total numbers of Strava cycling and pedestrian trips are known, we could figure out average inhaled dose of pollutant (PM2.5) during a single cycling trip and that during a single pedestrian trip respectively.

Suppose a cycling or pedestrian trip could be divided into different segments based on edges (streets) and nodes (intersections). There are two statuses for cyclists or pedestrians: waiting in nodes or moving in edges. Time of trip equals time waiting in nodes plus time moving in edges. In Fig. 3, a cycling trip starts from n_1 and ends at n_4. Simply, time of the trip equals to the sum of time waiting

Table 2
Demographics of cycles and walks of Strava users in 2015.

	Cycling	Pedestrian
Athlete ID count (User count)	13,684	11,249
Activity count (Trip count)	287,833	156,002
Average distance of trips	24 km	7 km
Average time of trips	81 min	46 min

Gender	Male	Female	Blank Gender
Cycling	11,212	1698	770
Pedestrian	7546	2973	727

		Under 25	25–34	35–44	45–54	55–64	Over 64	No Birth date
Cycling	Male	718	2176	2957	2028	448	73	2812
	Female	141	417	346	217	44	2	531
Pedestrian	Male	513	1750	2118	1034	180	30	1921
	Female	243	779	635	260	44	4	1008

in the four nodes (n_1, n_2, n_3, n_4) and time moving in the three edges (e_1, e_2, e_3). As cycling and pedestrian trips were aggregated to streets and nodes, times of trips were aggregated to streets and nodes. Specifically, moving times of trips were aggregated to streets while waiting times of trips were aggregated to nodes. We could cumulate inhaled dose of pollutant during all cycling trips or all pedestrian trips to figure out total inhaled dose of Strava users when they were cycling or walking (including running, hiking). After that, we were able to calculate average inhaled dose of Strava users during a single cycling or pedestrian trip after dividing the total inhaled dose for Strava cyclists or pedestrians by a total number of cycling trips or the total number of pedestrian trips. The rest of this section will elaborate on how to calculate average inhaled dose of cyclists during a single cycling trip and average inhaled dose of pedestrians during a single pedestrian trip.

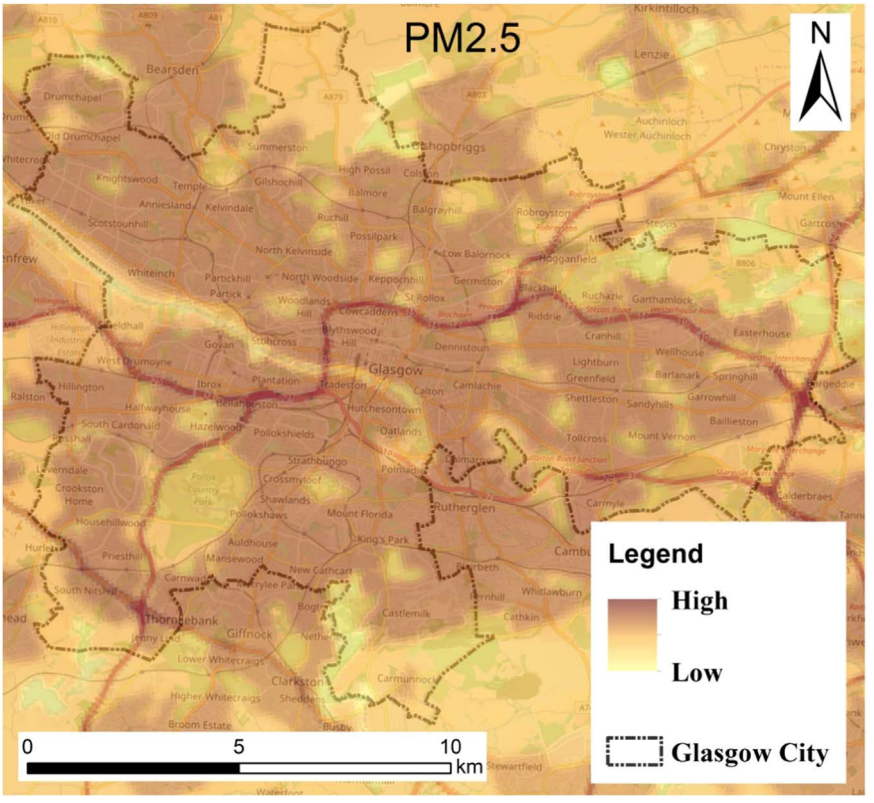

Fig. 2. 100 m × 100 m map representing annual average background PM2.5 concentrations in 2010 (Basemap: OpenStreetMap, licensed under the Open Database License).

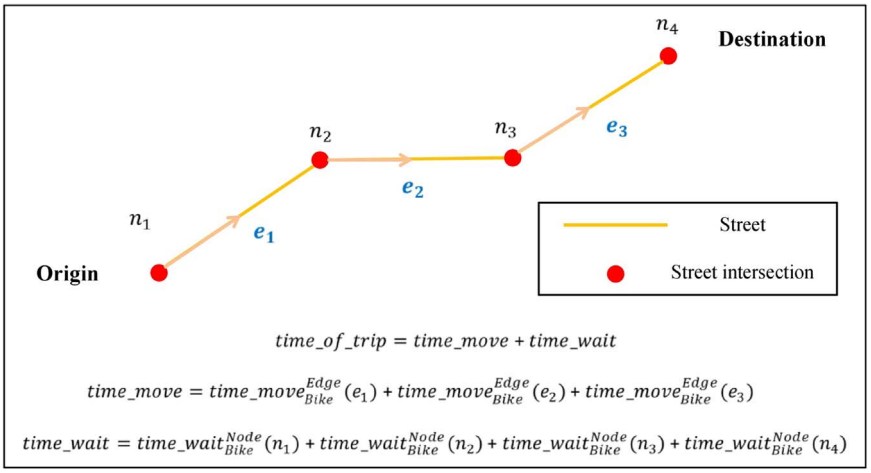

Fig. 3. Composition of a cycling or pedestrian trip.

2.2.1. Inhaled dose when waiting in nodes

First of all, we could calculate concentrations of cyclists and pedestrians when they were waiting in nodes. Suppose there was one cycling activity at a node, meaning that there was one cyclist at this node at a particular time. We could convert background PM2.5 concentration at a node (street intersection) to travel mode specific exposure concentrations at the moment when the cyclist or pedestrian was at that node. Earlier studies suggest that the mode specific exposure concentrations could be estimated by multiplying background PM2.5 concentration by 2.0 for cycling or 1.1 for walking (Tainio et al., 2016; Kahlmeier et al., 2014). For cycling and walking, we used ventilation rates of 2.55 and 1.37 (unit: m^3/h) respectively, according to earlier studies (de Nazelle et al., 2009; Kahlmeier et al., 2014; Tainio et al., 2016).

We could start with calculating total waiting time for cyclists and pedestrians when they were at the node *i* by

$$Time\_wait_{Bike}^{Node}(i) = \sum_{t \in T_i} Num\_act_{Bike}^{Node}(i, t) * Med\_time\_wait_{Bike}^{Node}(i, t) \quad (1)$$

$$Time\_wait_{Walk}^{Node}(i) = \sum_{t \in T_i} Num\_act_{Walk}^{Node}(i, t) * Med\_time\_wait_{Walk}^{Node}(i, t) \quad (2)$$

where $Num_act_{Bike}^{Node}(i, t)$and $Num\_act_{Walk}^{Node}(i, t)$ are the number of cyclists and number of pedestrians at the node i for the time t (minute-level); and $Med\_time\_wait_{Bike}^{Node}(i, t)$ and $Med\_time\_wait_{Walk}^{Node}(i, t)$ are median waiting time of cyclists and median waiting time of pedestrians at the node i for the time t. Besides, T_i is the set of t at the node i.

Then we could calculate total inhaled dose of cyclists and pedestrians when they were waiting at the node *i* by

$$ID\_wait_{Bike}^{Node}(i) = PM_{BG}^{Node}(i) * Convert_{Bike}^{PM} * Time\_wait_{Bike}^{Node}(i) * VR_{Bike}^{PM} \quad (3)$$

$$ID\_wait_{Walk}^{Node}(i) = PM_{BG}^{Node}(i) * Convert_{Walk}^{PM} * Time\_wait_{Walk}^{Node}(i) * VR_{Walk}^{PM} \quad (4)$$

where $PM_{BG}^{Node}(i)$ is the background PM2.5 concentration at the node *i*, and it equals the background PM2.5 concentration at the PM2.5 grid where the node *i* is situated; $Convert_{Bike}^{PM}$ and $Convert_{Walk}^{PM}$ are the converting factors for cycling and walking; VR_{Bike}^{PM} and

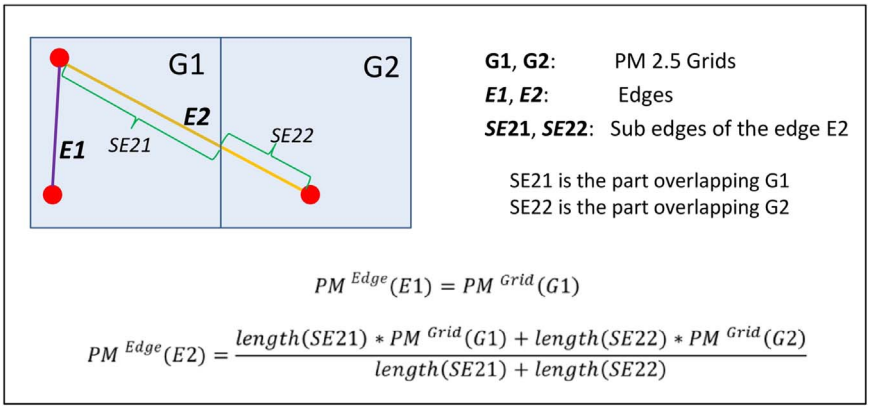

Fig. 4. Calculation of the background PM2.5 concentration in an edge intersecting two different PM2.5 grids.

VR_{Walk}^{PM} are ventilation rates for cycling and walking. In this study, $Convert_{Bike}^{PM} = 2$, $Convert_{Walk}^{PM} = 1.1$, $VR_{Bike}^{PM} = 2.55$ and $VR_{Walk}^{PM} = 1.37$.

Accordingly, we could calculate total inhaled dose for cyclists and pedestrians when they were waiting in all nodes by

$$Total\_ID\_wait\_Bike = \sum_{i \in S_N} ID\_wait_{Bike}^{Node}(i) \tag{5}$$

$$Total\_ID\_wait\_walk = \sum_{i \in S_N} ID\_wait_{Walk}^{Node}(i) \tag{6}$$

where S_N is the set of all nodes.

2.2.2. Inhaled dose when moving in edges

We further could calculate inhaled dose of cyclists and pedestrians when moving in edges. Unlike nodes, edges might spatially intersect more than one PM2.5 grids (see Fig. 4). Assuming cyclists and pedestrians are moving at a constant speed on an edge, the time cyclists or pedestrians are staying within a grid is proportional to the length of the edge's overlapping part with that grid. In this case, a weighted mean of background PM2.5 concentrations in grids intersecting an edge was used to represent the background PM2.5 concentration at the edge. Fig. 4 shows a simple case that an edge intersects two different PM2.5 grids.

We could start with calculating total moving time for cyclists and pedestrians when they were at the edge j by

$$\begin{aligned} Time\_move_{Bike}^{Edge}(j) &= D\_time\_move_{Bike}^{Edge}(j) + R\_time\_move_{Bike}^{Edge}(j) D\_time\_move_{Bike}^{Edge}(j) \\ &= \sum_{t \in T_j} D\_num\_act_{Bike}^{Edge}(j, t) * D\_med\_time\_move_{Bike}^{Edge}(j, t) R\_time\_move_{Bike}^{Edge}(j) \\ &= \sum_{t \in T_j} R\_num\_act_{Bike}^{Edge}(j, t) * R\_med\_time\_move_{Bike}^{Edge}(j, t) \end{aligned} \tag{7}$$

$$\begin{aligned} Time\_move_{Walk}^{Edge}(j) &= D\_time\_move_{Walk}^{Edge}(j) + R\_time\_move_{Walk}^{Edge}(j) D\_time\_move_{Walk}^{Edge}(j) \\ &= \sum_{t \in T_j} D\_num\_act_{Walk}^{Edge}(j, t) * D\_med\_time\_move_{Walk}^{Edge}(j, t) R\_time\_move_{Walk}^{Edge}(j) \\ &= \sum_{t \in T_j} R\_num\_act_{Walk}^{Edge}(j, t) * R\_med\_time\_move_{Walk}^{Edge}(j, t) \end{aligned} \tag{8}$$

where $D_num_act_{Bike}^{Edge}(i, t)$and $D\_num\_act_{Walk}^{Edge}(i, t)$ are the number of cyclists and number of pedestrians at the edge j for the time t when cyclists and pedestrians going the direction the street was digitized; $R_num_act_{Bike}^{Edge}(i, t)$and $R\_num\_act_{Walk}^{Edge}(i, t)$ are the number of cyclists and number of pedestrians at the edge j for the time t when cyclists and pedestrians going against the direction the street was digitized; similarly, $D\_med\_time\_wait_{Bike}^{Edge}(i, t)$, $D\_med\_time\_wait_{Walk}^{Edge}(i, t)$, $R\_med\_time\_wait_{Bike}^{Edge}(i, t)$ and $R\_med\_time\_wait_{Walk}^{Edge}(i, t)$ are median moving time of cyclists and median moving time of pedestrians at the edge j for the time t with a direction along and against the direction the street was digitized respectively. Besides, T_j is the set of t at the edge j.

Then we could calculate total inhaled dose of cyclists and pedestrians when they were moving at the edge j by

$$ID\_move_{Bike}^{Edge}(j) = PM_{BG}^{Edge}(i) * Convert_{Bike}^{PM} * Time\_move_{Bike}^{Edge}(j) * VR_{Bike}^{PM} \tag{9}$$

$$ID\_move_{Walk}^{Edge}(j) = PM_{BG}^{Edge}(j) * Convert_{Walk}^{PM} * Time\_move_{Walk}^{Edge}(j) * VR_{Walk}^{PM} \tag{10}$$

$$PM_{BG}^{Edge}(j) = \frac{\sum_{l=1}^{K} length(s_l) * PM_{BG}^{Edge}(s_l)}{\sum_{l=1}^{K} length(s_l)} \tag{11}$$

where $PM_{BG}^{Edge}(i)$ is the background PM2.5 concentration at the edge j; $Convert_{Bike}^{PM} = 2$, and $Convert_{Walk}^{PM} = 1.1$; s_l represents a segment of the edge j and it intersects a unique PM2.5 grid, K means that the edge j spatially intersects K different PM2.5 grids, in other words, the edge j is split into K segments by PM2.5 grids; $length(s_l)$ is the length of the segment s_l and $PM_{BG}^{Edge}(s_l)$ is the background PM2.5 concentration at the grid that intersects s_l.

Accordingly, we could calculate total inhaled dose of cyclists and pedestrians when they were moving in all edges by

$$Total\_ID\_move\_bike = \sum_{j \in S_E} ID\_move_{Bike}^{Edge}(j) \tag{12}$$

$$Total\_ID\_move\_walk = \sum_{j \in S_E} ID\_move_{Walk}^{Edge}(j) \tag{13}$$

where S_E is the set of all edges.

2.2.3. Total inhaled dose and average inhaled dose

Finally, we could figure out total inhaled dose for cyclists and pedestrians by combining total inhaled dose of cyclists and pedestrians when they were waiting in nodes and total inhaled dose of cyclists and pedestrians when they were moving in edges.

$$Total\_ID\_bike = Total\_ID\_wait\_bike + Total\_ID\_move\_bike \quad (14)$$

$$Total\_ID\_walk = Total\_ID\_wait\_walk + Total\_ID\_move\_walk \quad (15)$$

Accordingly, average inhaled dose of cyclists during a cycling trip and average inhaled dose of pedestrians during a pedestrian trip could be computed as

$$Ave\_ID\_bike = \frac{Total\_ID\_bike}{Num\_bikes} \quad (16)$$

$$Ave\_ID\_walk = \frac{Total\_ID\_walk}{Num\_walks} \quad (17)$$

where *Num_bikes* and *Num_walks* are the total number of cycling trips and the total number of pedestrian trips respectively.

2.3. Spatial associations of cycles, walks and air pollution concentration

In this study, spatial associations of cycles, walks and air pollution concentration were explored by using the bivariate local Moran's *I* statistic method. The bivariate local Moran's *I* statistic method (Anselin, 1995) is widely used to measure the spatial association of two distinct attributes. Specifically, we explored the spatial association of background PM2.5 concentration and cycling count, and spatial association of background PM2.5 concentration and pedestrian count respectively. For simplicity, we used the number of cycling activities, the number of pedestrian activities and background PM2.5 concentration at each node to represent the spatial distributions of cycling activities, pedestrian activities and background PM2.5 concentration. We could identify: 1) areas with low-level air pollution and low-volume cycles; 2) areas with low-level air pollution and low-volume walks; 3) areas with high-level air pollution and high-volume cycles, and 4) areas with high-level air pollution and high-volume walks.

3. Results and discussions

This section demonstrates the empirical results in the study area and makes discussions about the results.

3.1. Comparison of Strava cycling volumes and regular cycling volumes

Unlike Strava Metro data which has a large spatial coverage, AADF data covers only 119 links (major roads) across Glasgow. Therefore, we matched those links with streets from the Strava Metro data based on spatial proximity (a 5-m threshold), road name, start junction name and end junction name. Accordingly, the correlation between AADF's annual average daily cycling volume and Strava's annual cycling volume equals 0.83, indicating spatial distribution of Strava cycling volume is fairly proportional to that of realistic cycling volume on major roads across Glasgow.

3.2. Assessment of air pollution exposure

By the Eqs. (1)–(17) we calculated average inhaled dose of Strava users during a cycling trip and a pedestrian trip step by step (see Table 3). For both cyclists and pedestrians, the total inhaled dose when moving in edges is more than two times of the total inhaled dose when waiting in nodes. Average inhaled dose of Strava users during a single cycling trip is four times of that of Strava users during a pedestrian trip. According to annual variation of PM2.5 concentration at the monitoring site Glasgow Kerbside (urban traffic) in Glasgow (Ricardo Energy & Environment, 2016), it is likely that annual PM2.5 concentration across Glasgow decreases from 2011 to 2015 as well. Therefore, it is likely that realistic intake of PM2.5 in 2015 is smaller than the one assessed in this study as the 100 m PM2.5 concentration map used is for 2010.

Moreover, one the one hand, almost half of cycling and pedestrian trips were contributed by users aged 25–44, and a large portion of Strava cycling and pedestrian trips are recreational trips (Strava Metro, 2015). On the other hand, the spatial distribution of Strava cycles is fairly proportional to that of realistic cycles on major roads across Glasgow. Therefore, we might use average inhaled dose of Strava users during a cycling trip to roughly represent average inhaled dose of young and sporty cyclists during a cycling trip. Once

Table 3
Total and average inhaled dose of Strava cyclists and pedestrians.

Inhaled dose	*Total_ID_bike*	*Total_ID_walk*	*Ave_ID_bike*	*Ave_ID_walk*
Unit: μg	9,436,394	1,246,793	33	8

Inhaled dose	*Total_ID_move_bike*		*Total_ID_wait_bike*	*Total_ID_move_walk* *Total_ID_wait_walk*
Unit: μg	6,722,580	2,713,814	867,121	379,672

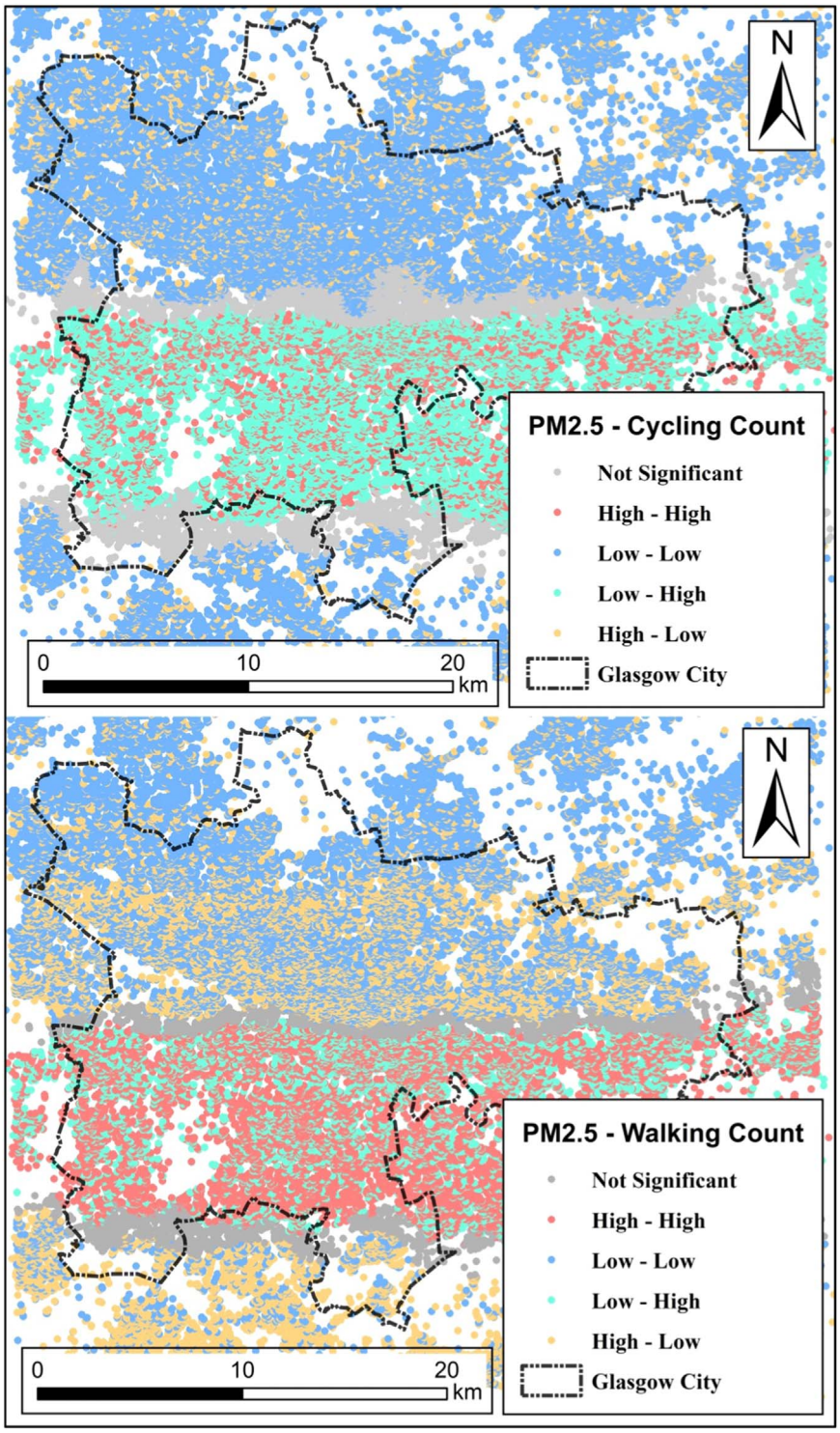

Fig. 5. Clusters and outliers representing spatial associations of air pollution, cycling count and walking count in Glasgow.

the annual frequency of those cyclists was known from other data sources such as travel survey, we could further try to assess average annual inhaled dose of young and sporty cyclists in Glasgow.

3.3. Spatial associations of cycles, walks and air pollution concentration

The bivariate local Moran's I method was implemented in the open software GeoDa (Center for Spatial Data Science, 2016). First, 59,722 nodes with cycling count, pedestrian count and background PM2.5 concentration were input into GeoDa. Second, 600 nearest nodes of a node are used to represent the neighbours of this node. In other words, each node has 600 neighbours. As there are approximately 60,000 nodes, 600 neighbours (nodes) take 1% of the total number of nodes. Finally, after 999 Monte Carlo simulations, statistically significant clusters and outliers were detected. Fig. 5 displays the clusters and outliers detected in Glasgow. In Fig. 5, a High-High cluster indicates spatial clustering of high-value PM2.5 concentration and high-value cycling count (or pedestrian count), whereas a Low-Low cluster indicates spatial clustering of low-value PM2.5 concentration and low-value cycling count (or pedestrian count). A Low-High outlier indicates that low-value PM2.5 concentration is associated with high-value cycling count (or pedestrian count); whereas a High-Low outlier indicates that high-value PM2.5 concentration is associated with low-value cycling count (or pedestrian count). Besides, the category Not Significant means that there is no statistically significant spatial association between PM2.5 concentration and cycling count (or pedestrian count). Fig. 5 shows that 1) areas with low-level air pollution and low-volume cycles are mainly situated in the north of the city and southern outskirts of the city; 2) areas with low-level air pollution and low-volume walks are mainly situated in the north of the city and southern outskirts of the city; 3) areas with high-level air pollution and high-volume cycles are mainly situated in the south of the city excluding southern outskirts of the city; and 4) areas with high-level air pollution and high-volume walks are mainly situated in the south of the city excluding southern outskirts of the city.

3.4. Discussions

Altitude of northern Glasgow is higher than that of southern Glasgow, and northern Glasgow is hillier than southern Glasgow. Cycle commuters and pedestrian commuters seem to rely more on the road network in southern Glasgow than recreational cyclists and pedestrians. Therefore, we could offer implications for policies that 1) improvement of bicycle infrastructure in northern Glasgow to attract more recreational cyclists and recreational pedestrians; 2) improvement of bicycle infrastructure in southern Glasgow to separate cycle commuters and pedestrian commuters from motor vehicles; and 3) reducing traffic-related air pollution in southern Glasgow to decrease journey-time exposures of cyclists and pedestrians. Furthermore, although there is a difference of population structure between Strava cyclists and regular cyclists, Strava Metro data has a good spatial granularity and spatial coverage across a city. As it is expensive and time-consuming to conduct a travel survey every year, Strava Metro data offers a good opportunity to explore the annual variations of cycles and walks, which could be used to roughly evaluate the realistic effects of policies or interventions on modal shift from inactive travel (motorized vehicles) to active travel (cycles or walks), and decrease in journey-time exposures of cyclists and pedestrians with reduction of air pollution emissions.

There are still some limitations in this paper. First, the median time was used to represent the average time of moving in edges and waiting in nodes. The gap between median time and mean time (realistic average time) is likely to influence the accuracy of air pollution exposure assessment. Second, the assessment of air pollution in this study is also sensitive to some parameters such as converting factor and ventilation rate. We used the values for converting factor and ventilation rate based on earlier studies; however, how to better determine more appropriate values for them needs more efforts. Third, there is representativeness bias in both cycling and pedestrian trips. The population structure (gender, age and other socio-economically personal characteristics) between Strava cyclists and regular cyclists is likely to be different. As young people are more active in social media, old cyclists and pedestrians are likely to be under-represented by Strava users. Some users like to upload a large proportion of their cycling or pedestrian trips; while other users might upload a small proportion of their trips. As they upload a small proportion of their trips, their realistic trips are under-represented by trips of Strava. Fourth, although Strava has the original GPS traces of cycles and walks, it only offers aggregated data to researchers due to a risk of privacy issues. The original GPS trace data has a larger potential than the aggregated data. Ideally, we would select GPS traces of cycles and walks created by the number of Strava users who compose a cohort. This would enable a cohort study of cyclists or pedestrians in a city.

4. Conclusions

In this study, to explore the potential of Strava Metro data in research of active travel and health, we used Strava Metro data to estimate air pollution exposure in Glasgow, UK. Empirical results demonstrate that Strava Metro data provides an opportunity to the assessment of air pollution exposure during active travel. Additionally, to demonstrate the potential of Strava Metro data in policy-making, we explored the spatial association of air pollution concentration and active travel. As a consequence, we identified areas that require investment priority, and finally offered implications for policies.

In the future, we will take account of some aspects to enhance this study. First, with health impact modelling we will investigate the risk–benefit balance between air pollution exposure and physical activity when people are riding or walking. Particularly, knowing cycling and walking time enables not only assessment of air pollution exposure but also assessment of physical activity (Tainio et al., 2016; Ainsworth et al., 2011). Second, as some people think that harm from air pollution might exceed benefit from physical activity of active travel in severely polluted regions or bad weather conditions, we will examine reverse impacts of air pollution on active travel, particularly in extremely polluted regions. Typically, we are interested to explore whether high-level air

pollution would reduce outdoor activities. Third, although Strava only offer aggregated data to researchers due to privacy issue, it is still possible to publicize original GPS traces of some Strava users. As some Strava users probably are glad to make their traces publicly and be used for research, Strava might send requests to users and ask whether they are glad to publicize their original GPS traces. Once some original GPS traces were available, Strava data would have a larger potential in studies of active travel and health.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

This work is supported by the UK Economic and Social Research Council (Grant No. ES/L011921/1). The authors are thankful to the Urban Big Data Centre University of Glasgow for offering data services

References

Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., et al., 2011. 2011 compendium of physical activities: a second update of codes and MET values. Med. Sci. Sport. Exerc. 43 (8), 1575–1581.
AIRBASE database, 2013. The European air quality database. European Environment Agency, European Union.
Anderson, J.O., Thundiyil, J.G., Stolbach, A., 2012. Clearing the air: a review of the effects of particulate matter air pollution on human health. J. Med. Toxicol. 8 (2), 166–175.
Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27, 93–115.
Broach, J., Dill, J., Gliebe, J., 2012. Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transp. Res. A-Pol. 46 (10), 1730–1740.
Casello, J.M., Usyukov, V., 2014. Modeling cyclists' route choice based on GPS data. Transp. Res. Rec. 2430, 155–161.
Center for Spatial Data Science, 2016. GeoDa – An Introduction to Spatial Data Analysis. Center for Spatial Data Science, the University of Chicago, Chicago, USA.
Chen, R., Kan, H., Chen, B., Huang, W., Bai, Z., Song, G., Pan, G., 2012. Association of particulate air pollution with daily mortality: the China air pollution and health effects study. Am. J Epidemiol. 175 (11), 1173–1181.
De Nazelle, A., Rodríguez, D.A., Crawford-Brown, D., 2009. The built environment and health: impacts of pedestrian-friendly designs on air pollution exposure. Sci. Total Environ. 407 (8), 2525–2535.
De Nazelle, A., Seto, E., Donaire-Gonzalez, D., Mendez, M., Matamala, J., Nieuwenhuijsen, M.J., Jerrett, M., 2013. Improving estimates of air pollution exposure through ubiquitous sensing technologies. Environ. Pollut. 176, 92–99.
Department for Transport, 2016. Traffic counts. Department for Transport, London, UK.
Dill, J., 2009. Bicycling for transportation and health: the role of infrastructure. J Public Health Policy 30 (Suppl 1), S95–S110.
Doorley, R., Pakrashi, V., Ghosh, B., 2015. Quantifying the health impacts of active travel: assessment of methodologies. Transp. Rev. 35 (5), 559–582.
Duncan, M.J., Badland, H.M., Mummery, W.K., 2009. Applying GPS to enhance understanding of transport-related physical activity. J. Sci. Med. Sport. 12 (5), 549–556.
El Esawey, M., 2014. Estimation of annual average daily bicycle traffic with adjustment factors. Transp. Res. Rec. 2443, 106–114.
Forsyth, A., Oakes, J.M., 2015. Cycling, the built environment, and health: results of a midwestern study. Int. J. Sustain. Transp. 9 (1), 49–58.
Forsyth, A., Krizek, K.J., Agrawal, A.W., Stonebraker, E., 2012. Reliability testing of the Pedestrian and Bicycling Survey (PABS) method. J. Phys. Act. Health 9 (5), 677–688.
Griffin, G.P., Jiao, J., 2015. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. J. Transp. Health 2 (2), 238–247.
Heesch, K.C., James, B., Washington, T.L., Zunig, K., Burke, M., 2016. Evaluation of the Veloway 1: a natural experiment of new bicycle infrastructure in Brisbane. Aust. J. Transp. Health 3 (3), 366–376.
Herrero, J., 2016. Using big data to understand trail use: three Strava tools. TRAFx Res (Available from). ⟨https://www.trafx.net/insights.htm⟩.
Hollingworth, M., Harper, A., Hamer, M., 2014. An observational study of erectile dysfunction, infertility, and prostate cancer in regular cyclists: cycling for health UK study. JOMH 11 (2), 75–79.
Hood, J., Sall, E., Charlton, B., 2011. A GPS-based bicycle route choice model for San Francisco, California. Transp. Lett. 3, 63–75.
de Hoogh, K., Gulliver, J., van Donkelaar, A., et al., 2016. Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ. Res 151, 1–10.
Jesticoa, B., Nelsona, T., Wintersb, M., 2016. Mapping ridership using crowdsourced cycling data. J. Transp. Geogr. 52, 90–97.
Kahlmeier, S., Schweizer, C., Rojas-Rueda, D., Nieuwenhuijsen, M., Nazelle, A., de, Bode, O., 2014. Development of the Health Economic Assessment Tools (HEAT) for walking and cycling — meeting background document. Consensus workshop: meeting report, Bonn, Germany, 2014.
Li, M., Sagl, G., Mburu, L., Fan, H., 2016a. A contextualized and personalized model to predict user interest using location-based social networks. Comput. Environ. Urban Syst. 58, 97–106.
Li, M.H., Fan, L.C., Mao, B., Yang, J.W., Choi, A.M., Cao, W.J., Xu, J.F., 2016b. Short-term exposure to ambient fine particulate matter increases hospitalizations and mortality in COPD: a systematic review and meta-analysis. Chest 149 (2), 447–458.
Lipsett, M.J., Ostro, B.D., Reynolds, P., Goldberg, D., Hertz, A., Jerrett, M., Smith, D.F., Garcia, C., Chang, E.T., Bernstein, L., 2011. Long-term exposure to air pollution and cardiorespiratory disease in the california teachers study cohort. Am. J. Respir. Crit. Care Med. 184 (7), 828–835.
Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., de Nazelle, A., Dons, E., Gerike, R., Götschi, T., Int Panis, L., Kahlmeier, S., Nieuwenhuijsen, M., 2015. Health impact assessment of active transportation: a systematic review. Prev. Med. 76, 103–114.
Oja, P., Vuori, I., Paronen, O., 1998. Daily walking and cycling to work: their utility as health-enhancing physical activity. Patient Educ. Couns. 33 (Suppl 1), S87–S94.
Oja, P., Titze, S., Bauman, A., de Geus, B., Krenn, P., Reger-Nash, B., Kohlberger, T., 2011. Health benefits of cycling: a systematic review. Scand. J. Med. Sci. Sports 21 (4), 496–509.
Pope, C.A., Burnett, R.T., Thurston, G.D., Thun, M.J., Calle, E.E., Krewski, D., Godleski, J.J., 2015. Relationships between fine particulate air pollution, cardiometabolic disorders and cardiovascular mortality. Circ. Res. 116, 108–115.
Prins, R.G., Pierik, F., Etman, A., Sterkenburg, R.P., Kamphuis, C.B., van Lenthe, F.J., 2014. How many walking and cycling trips made by elderly are beyond commonly used buffer sizes: results from a GPS study. Health Place 27, 127–133.
Pucher, J., Buehler, R., Bassett, D.R., Dannenberg, A.L., 2010. Walking and cycling to health: a comparative analysis of city, state, and international data. Am. J. Public Health 100 (10), 1986–1992.
Ricardo Energy & Environment, 2016, 2015. Air Quality in Scotland. Ricardo Energy & Environment, Harwell, UK.
Riordan, B., 2016. Strava Metro: Better Data for Better Cities. Strava Metro, San Francisco, USA. Available from: ⟨http://ubdc.ac.uk/media/1416/uofg-training.pdf⟩.
SAHSU, 2016. Environmental data: NO2 and PM2.5 air pollution grids for Europe, 100m resolution (annual means, ug/m3), 2010. MRC-PHE Centre for Environment

and Health, London, UK.
Schepers, P., Fishman, E., Beelen, R., Heinen, E., Wijnen, W., Parkin, J., 2015. The mortality impact of bicycle paths and lanes related to physical activity, air pollution exposure and road safety. J. Transp. Health 2, 460–473.
Sener, I.N., Eluru, N., Bhat, C.R., 2009. An analysis of bicycle route choice preferences in Texas, US. Transportation 36, 511–539.
Shah, A.S., Lee, K.K., McAllister, D.A., Hunter, A., Nair, H., Whiteley, W., Langrish, J.P., Newby, D.E., Mills, N.L., 2015. Short term exposure to air pollution and stroke: systematic review and meta-analysis. BMJ 350, h1295.
Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J.A., Demšar, U., Fotheringham, A.S., 2015. Analysis of human mobility patterns from GPS trajectories and contextual information. Int. J. Geogr. Inf. Sci. 30, 881–906.
Steiger, E., Westerholt, R., Resch, B., Zipf, A., 2015. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. Comput. Environ. Urban Syst. 54, 255–265.
Strava Metro, 2015. Strava Metro Comprehensive User Guide Version 2.0. Available from: ⟨http://ubdc.ac.uk/media/1323/stravametro_200_user_guide_withoutpics.pdf⟩.
Strava Metro, 2016. Data-Driven Bicycle and Pedestrian Planning. Strava Metro, San Francisco, USA. Available from: ⟨http://metro.strava.com/⟩.
Sun, Y., Li, M., 2015. Investigation of travel and activity patterns using location-based social network data: a case study of active mobile social media users. ISPRS Int. J. Geo-Inf. 4 (3), 1512–1529.
Sun, Y., Mobasheri, A., 2007. Utilizing crowdsourced data for studies of cycling and air pollution exposure: a case study using strava data. Int. J. Environ. Res. Public Health 14 (3), 274.
Tainio, M., De Nazelle, A.J., Götschi, T., Kahlmeier, S., Rojas-Reuda, D., Nieuwenhuijsen, M.J., De sa Herick, T., Kelly, P., Woodcock, J., 2016. Can air pollution negate the health benefits of cycling and walking? Prev. Med. 87, 233–236.
Thakuriah, P., Sila-Nowicka, K., Gonzalez, Paule, J., 2016. Sensing spatiotemporal patterns in urban areas: analytics and visualizations using the integrated multimedia city data platform. Built Environ. 42 (3), 415–429.
Urban Big Data Centre, 2016. Data services: Strava Metro data. Urban Big Data Centre, Glasgow, UK.
Weichenthal, S., Kulka, R., Dubeau, A., Martin, C., Wang, D., Dales, R., 2011. Traffic-related air pollution and acute changes in heart rate variability and respiratory function in urban cyclists. Environ. Health Perspect. 119 (10), 1373–1378.
Wen, L.M., Rissel, C., 2008. Inverse associations between cycling to work, public transport, and overweight and obesity: findings from a population based study in Australia. Prev. Med. 46 (1), 29–32.
WHO, 2016. Air pollution levels rising in many of the world's poorest cities. Available from: ⟨http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/⟩.