International Journal of
Environmental Research and Public Health

MDPI

Article

Utilizing Crowdsourced Data for Studies of Cycling and Air Pollution Exposure: A Case Study Using Strava Data

Yeran Sun [1,*] and Amin Mobasheri [2]

[1] Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, Glasgow G12 8RZ, UK
[2] GIScience Research Group, Institute of Geography, Heidelberg University, D-69120 Heidelberg, Germany; a.mobasheri@uni-heidelberg.de
* Correspondence: yeran.sun@glasgow.ac.uk; Tel.: +44(0)141-330-3283

Academic Editor: Kim Natasha Dirks
Received: 17 November 2016; Accepted: 27 February 2017; Published: 8 March 2017

Abstract: With the development of information and communications technology, user-generated content and crowdsourced data are playing a large role in studies of transport and public health. Recently, Strava, a popular website and mobile app dedicated to tracking athletic activity (cycling and running), began offering a data service called Strava Metro, designed to help transportation researchers and urban planners to improve infrastructure for cyclists and pedestrians. Strava Metro data has the potential to promote studies of cycling and health by indicating where commuting and non-commuting cycling activities are at a large spatial scale (street level and intersection level). The assessment of spatially varying effects of air pollution during active travel (cycling or walking) might benefit from Strava Metro data, as a variation in air pollution levels within a city would be expected. In this paper, to explore the potential of Strava Metro data in research of active travel and health, we investigate spatial patterns of non-commuting cycling activities and associations between cycling purpose (commuting and non-commuting) and air pollution exposure at a large scale. Additionally, we attempt to estimate the number of non-commuting cycling trips according to environmental characteristics that may help identify cycling behavior. Researchers who are undertaking studies relating to cycling purpose could benefit from this approach in their use of cycling trip data sets that lack trip purpose. We use the Strava Metro Nodes data from Glasgow, United Kingdom in an empirical study. Empirical results reveal some findings that (1) when compared with commuting cycling activities, non-commuting cycling activities are more likely to be located in outskirts of the city; (2) spatially speaking, cyclists riding for recreation and other purposes are more likely to be exposed to relatively low levels of air pollution than cyclists riding for commuting; and (3) the method for estimating of the number of non-commuting cycling activities works well in this study. The results highlight: (1) a need for policymakers to consider how to improve cycling infrastructure and road safety in outskirts of cities; and (2) a possible way of estimating the number of non-commuting cycling activities when the trip purpose of cycling data is unknown.

Keywords: crowdsourced data; Strava Metro; cycling purpose; air pollution exposure; particulate matter

1. Introduction

Over the last two decades, researchers have provided much evidence of the benefits of cycling as a health-enhancing physical activity [1–7]. Recently, volunteered geographic information (VGI), user-generated content (UGC) and crowdsourced data are becoming promising data sources for transport and health research [8,9]. Traditional methods of collecting cycling data, including manual

counts, stated preference surveys [10,11] and annual average daily bicycle (AADB) volumes [12], are expensive and time-consuming. Each of these methods has its advantages, but each is almost impossible to accomplish over a broad area simultaneously, which is why crowdsourced methods are gaining interest in planning [9]. Through the expansion of Global Positioning Systems (GPS) new methods for collecting detailed cycling route information have emerged [13]. GPS-enabled mobile devices, such as smartphones, allow individuals to track and map their cycling routes [13–16]. More recently, crowdsourced cycling data are used to analyze cycling behavior [13,17] and make associations between cycling and health [9]. Strava is a popular website used to track users' cycling and running activity via GPS-enabled devices, such as smart phones and smart watches. Millions of people upload their rides and runs to Strava every week via their smartphones or other GPS devices [18]. Strava launched a data service called Strava Metro that offers aggregated data sets after anonymizing and aggregating individual's GPS traces. In earlier studies that use traditional data collection methods, research on the role of cycling for health through physical activity has been limited by the lack of information on where bicyclists ride [9]. With a high spatial resolution, Strava Metro data is able to provide new opportunities for research into active travel, sustainable travel and public health. This could benefit studies of cycling behavior and public health, and further help policymakers in urban planning, especially designing urban infrastructures aiming to make urban residents healthier and cities more sustainable. For instance, knowing where people like to cycle could help policymakers to improve cycling infrastructure more effectively (e.g., availability of cycle parking in areas of high demand) and promote road safety by giving priority to roads where there are more cycling trips. In several recent studies, Strava data has been used to map ridership over a city [13], evaluate the impact of bicycle infrastructure on cycling behavior [17], and investigate impacts of residential and employment density, land use diversity, cycling facilities and terrain on cycling behavior [9].

The impact of outdoor and traffic-related air pollution on health is an important issue in transport and health [19–28]. Typically, impacts of active travel (cycling and walking) and inactive travel (traveling by car, bus or train) on health are compared [22–27]. Although earlier studies offer much evidence on the health benefits of cycling or walking due to increased physical activity [1–7], some other studies reveal cycling also carries some potential health risks, including air pollution, accidents and noise [29–31]. One of the most important risks is from poor air quality [29,30]. Exposure to air pollution is harmful to human health [32–44] and more than 80% of people living in urban areas that monitor air pollution are exposed to air quality levels that exceed World Health Organization (WHO) safe limits [32]. Cyclists riding in urban areas are therefore likely to be exposed to high levels of air pollution. Recent studies use health impact modelling (HIM) to estimate the health benefits and risks of active travel (cycling, walking), and reveal that the total benefits of active travel outweighed the risks [28,45,46]. Particularly, a very recent study reveals that benefits of active travel outweighed the harm caused by air pollution in all but the most extreme air pollution concentrations [28]. It is also becoming widely accepted that increasing cycling time tends to increase health improvements. However many such studies assess cyclists' exposure to air pollution based on city-level air pollution values when in fact, air pollution levels vary spatially over a city. Relating cycling activities to air pollution at a larger scale (e.g., street-level) could promote assessment of air pollution exposure when it is known where and when cyclists ride in a city. Ideally, urban planners and policymakers could use this knowledge of where cyclists ride to devise cycling and walking routes that minimize the risks faced by active commuters, and to decrease volume of cyclists riding in the environments that are associated with the highest exposures [29,47].

Moreover, Strava Metro data also indicate cycling purpose (commuting or non-commuting) of cycling activities. Researchers might make use of this in studies of cycling purpose and health. In this paper, we explore the potential of Strava Metro in research of active travel and health by using the data to investigate spatial patterns of non-commuting cycling activities and associations between cycling purpose (commuting and non-commuting) and air pollution exposure at a large scale. Additionally, as some cycling trip data sets (e.g., crowdsourced GPS trajectories or bike-sharing

origin-destination trips) lack trip purpose (commuting or non-commuting), we can't directly relate cycling purpose to air pollution exposure when utilizing those data sets. However, we might estimate the number of non-commuting cycling trips based on the number of all-purpose trips and environmental characteristics that affect cycling behavior. In this paper, we try to estimate the number of non-commuting cycling trips based on the number of trips for all purposes and environmental characteristics, as the estimation model could be validated by the Strava Metro data. If this method of estimating the number of non-commuting cycling trips is shown to be good, we may then use it to estimate the number of other non-commuting cycling trip datasets where trip purpose is unknown.

In this paper, we use the Strava Metro data in Glasgow, UK to carry out an empirical analysis. Firstly, in order to explore spatial patterns of non-commuting cycling activities, we investigate where non-commuting cycling activities are more likely to be than commuting cycling activities by identifying clusters where there are high rates of non-commuting cycling activities. Afterward, to associate cycling purpose with air pollution exposure at a large scale (i.e., the street intersection level), we investigate whether cyclists riding for recreation and other purposes (excluding commuting) are more likely to be exposed to relatively low levels of air pollution than cyclists riding for commuting. Note that levels of air pollution also might also vary over time in the study area. We focus on spatial variations of air pollution levels, not spatio-temporal variations of air pollution levels as temporal resolution of the air pollution data is one year. In this study, we focus on the difference in air pollution exposure during cycling, not the difference in health effects of cycling. Additionally, we strive to improve the estimation of the number of non-commuting cycling activities by using different regression methods (linear and non-linear methods) as the estimation models.

2. Materials and Methods

In this section, the methods used for spatial analysis of non-commuting cycling activities and air pollution are presented. Section 2.1 introduces the data and study area, Section 2.2 explores spatial patterns of non-commuting cycling activities and associations between cycling purpose and air pollution exposure, and Section 2.3 presents how we estimate the number of non-commuting activities according to total cycling activities and locational characteristics.

2.1. Data and Study Area

2.1.1. Strava Metro Data

Strava (San Francisco, CA, USA) is a popular online social network for cyclists and runners with a user base larger than other similar sites like MapMyRide (Under Armoour, Baltimore, MD, USA), MapMyRun (Under Armour, Baltimore, MD, USA) or RideWithGPS (Ride with GPS, Portland, OR, USA). Strava consists of a mobile app and a website, allowing users can track their rides, runs, walks and hikes on a smartphone or another GPS device. The Strava app records distance, time, average speed and route (GPS trajectory) of each activity. Users can also add textual information like titles and tags to describe their trips. The 'commute' flag indicates walking or riding journeys to or from work. Users can upload their GPS-tracked activities recorded by their Strava apps to the Strava's website (https://www.strava.com). Strava's database comprises nearly a trillion GPS points globally and is growing by over 8 million activities every week [48].

Strava Metro is a suite of data services that enables cutting-edge views into cycling and pedestrian (running, walking, hiking) patterns [49]. Aiming to produce state-of-the-art spatial data products and services to make cycling, running, and walking in cities better, Metro anonymizes and aggregates activity data from Strava's millions of users and then collaborates with departments of transportation and city planning groups to improve infrastructure for bicyclists and pedestrians [49]. To acquire the number of commuting activities, Metro first count 'commute' flags. As some users tend to not select the 'commute' flag, Metro further uses two ways to detect commuting activities from activities where 'commute' is not flagged: (1) activities with keywords in the titles such as "To Work" and "Commute

To"; and (2) starting and ending points more than 1 km apart within a distance and time threshold (This can be user defined but we find 30 miles and 90 min to be a good top threshold) [48].

More than half of the Strava Metro's dataset for dense metropolitan areas corresponds to commuting trips [49]. Some studies suggest crowdsourced data may be a good proxy for estimating daily, categorical cycling volumes by comparing cyclist counts between Strava data and manual count data in count stations [13,50]. For example, the Oregon Department of Transportation (ODOT) found the month-to-month correlation on the Hawthorne Bridge between total number of bicycles counted with a bike counter and total number of Strava bicycles trips over a one-year period was an adjusted R-Squared 0.91 [50]. Although crowdsourced cyclists represent a small portion of all cyclists, comparison with manual counts revealed a linear relationship between crowdsourced cyclists and total ridership in Victoria, Canada [13]. Due to an ability of tracking cycling activities at a high level of spatial granularity and a fairly high representation of the population of cyclists, Strava Metro data seem to have a high potential for studies of active travel and health in urban areas.

The Urban Big Data Centre, UK offers a Strava Metro dataset to researchers [51]. This dataset records 287,833 cycling activities (174,758 commuting activities and 113,075 non-commuting activities) contributed by 13,684 users (11,216 males, 1698 females and 770 blank-gender users) within the Glasgow Clyde Valley Planning area (including Glasgow City and seven council areas) in 2015. This dataset contains three sub sets in three formats: Streets, Origin-Destination and Nodes (see [49]). In this study, we select the sub set Nodes. A node represents an intersection of street and an edge represents a street (see Figure 1). The street network is extracted from OpenStreetMap (OpenStreetMap Foundation, West Midlands, UK). Attributes of a node includes count of all-purpose cycling activities at the intersection in 2015 and counts of commuting activities at the intersection (node) in 2015. Note that Strava Metro uses the number of all-purpose cycling trips (regardless of unique riders) that meet at the intersection to represent count of all-purpose cycling activities at the intersection, and uses the number of commuting cycling trips (regardless of unique riders) that meet at the intersection to represent count of commuting cycling activities at the intersection. We can infer that count of the non-commuting activities at the intersection (node) equals the difference between count of all-purpose cycling activities and count of commuting activities at the intersection (node). Additionally, the dataset contains a file that offers demographics of the bike trips, including average distance (24 km), median distance (15 km), average time (1.34 h), and median time (0.77 h).

Additionally, we look at the distributions of numbers of all-purpose cycling activities, inferred non-commuting cycling activities and commuting cycling activities in a node. Figure 2 shows cumulative distributions of numbers of all-purpose cycling activities, non-commuting cycling activities and commuting cycling activities in a node by using the complementary cumulative distribution function (CCDF). The CCDF of a variable X at x represents probability that X takes a value more than x, i.e., $P(X > x)$. For instance, CCDF of the number of all-purpose cycling activities at 2500, i.e., $P(X > 2500)$, equals to 0.1. This means that 10% of nodes have a relatively large number of all-purpose cycling activities (e.g., more than 2500); whilst the other 90% of nodes have a relatively small number of all-purpose cycling activities (e.g., less than or equal to 2500). Intuitively, the cumulative distributions of numbers of all-purpose cycling activities, non-commuting cycling activities and commuting cycling activities all seem to approximately follow an exponential law as they look like straight lines in the log-linear plot (see Figure 2). Note the log-linear plot is a semi-log plot with a logarithmic scale on the y-axis and a linear scale on the x-axis.

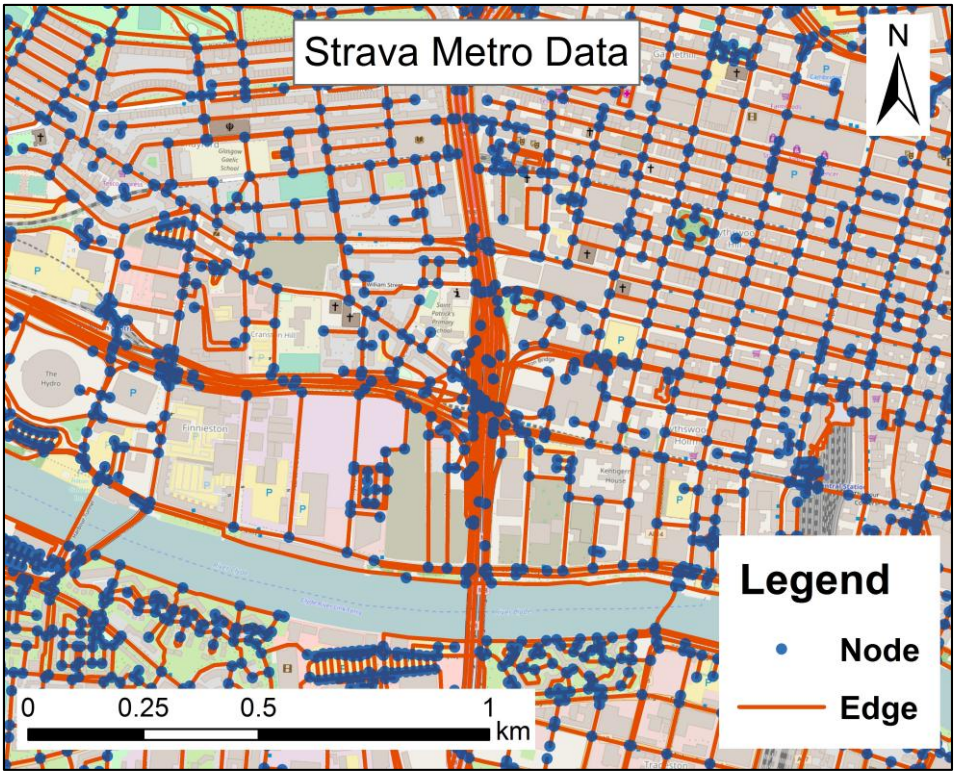

Figure 1. Nodes and edges of Strava Metro data (Basemap: OpenStreetMap, licensed under the Open Database License (ODbL)).

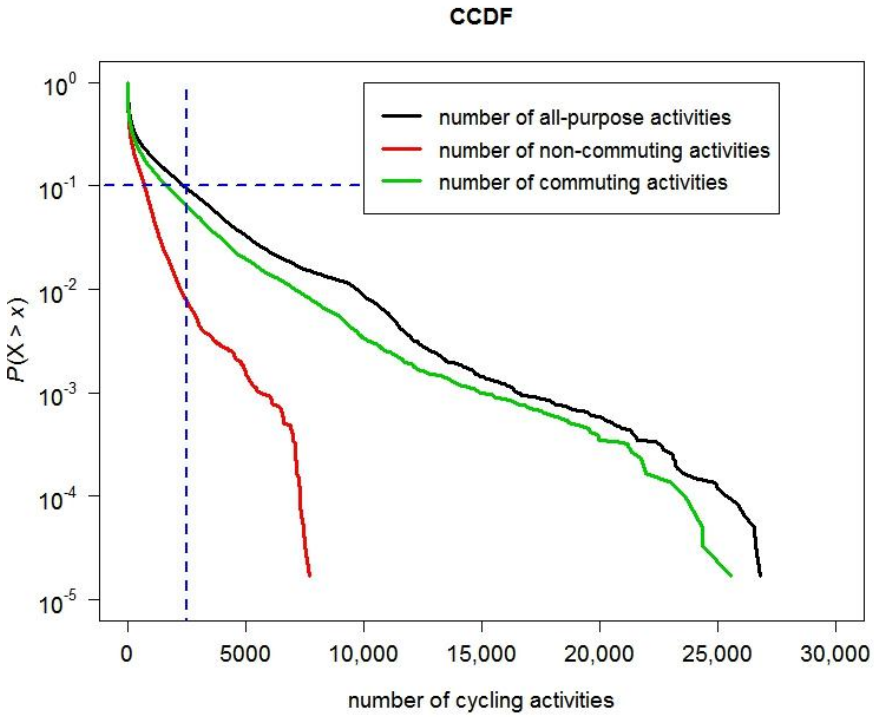

Figure 2. Cumulative distributions of numbers of all-purpose cycling activities, non-commuting cycling activities and commuting cycling activities in the log-linear plot.

Within the administrative boundaries of Glasgow, there are 59,718 nodes. Among those nodes, a few of nodes seem to have null or incorrect records. For instance, in some nodes count of commuting

cycling activities is larger than count of all-purpose cycling activities; while in some other nodes counts of all-purpose cycling activities is zero. After removing these noise nodes, 50,057 nodes are kept as the data set. In the spatial analysis, we first need to aggregate nodes to areas as we want to identify clusters consisting of contiguous areas. In this study, the area unit is census output area. There are 5486 census output areas in Glasgow (see Figure 3). The census output areas data is downloaded from DATA.GOV.UK [52].

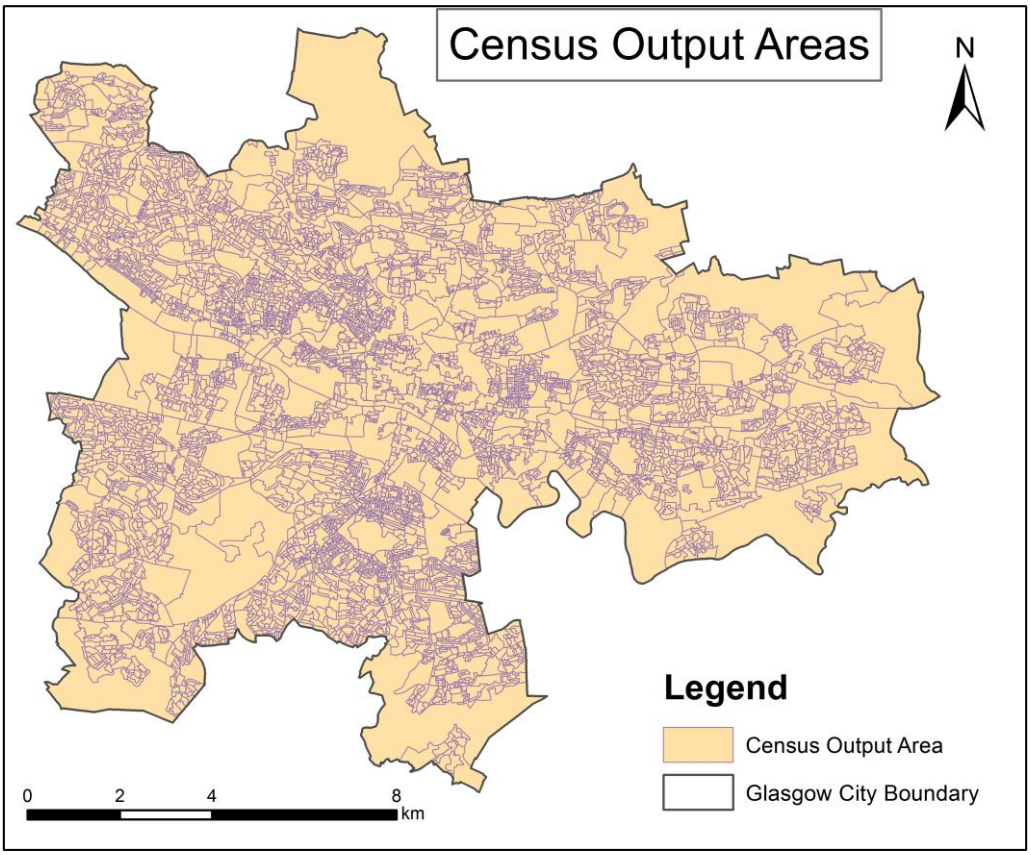

Figure 3. Census output areas in Glasgow (source: DATA.GOV.UK [52]).

2.1.2. Air Pollution Data

In this paper, particulate matter (PM), including PM_{10} (coarse PM) and $PM_{2.5}$ (fine PM) are used as the air pollutants to measure levels of air pollution [34–39]. PM_{10} is PM with a diameter of 10 micrometers or less; while $PM_{2.5}$ is PM with a diameter of 2.5 micrometers or less. Exposure to air pollution tends to increase risk of disease and mortality [33,44]. For instance, earlier studies provide much evidence that long-term exposure to PM is associated with an increase in cardiovascular and respiratory diseases [33–40] and recently, some studies also reveal short-term exposure to PM is associated with increased mortality risk [41–44]. According to a report released by WHO [32], Glasgow is one of the worst UK cities for both PM_{10} and $PM_{2.5}$. In this study, background maps for $PM_{2.5}$ and PM_{10} are downloaded from Air Quality in Scotland [53]. The background pollutant concentration maps are presented in 1 km $\times$ 1 km grid squares across Scotland. The background maps (reference year 2013) contain estimates of pollutant concentrations (PM_{10} and $PM_{2.5}$) based on an average over a year (annual average) for 2015, 2020, 2025 and 2030 from a base year of 2013 [54]. The 2013 maps are based on ambient monitoring and meteorological data for 2013 [54]. Air pollution background concentration maps for 2015 are used in this study. Specifically, each grid has values to represent annual average estimates of $PM_{2.5}$ and PM_{10} concentrations in 2015 (see Figure 4). The unit of $PM_{2.5}$ and PM_{10} is $\mu g/m^3$. The modelling methodology of PM_{10} background maps is based on Scottish

monitoring data and Scottish meteorological data, used to model the annual mean background and roadside concentrations for Scotland; whilst the modelling methodology of $PM_{2.5}$ background maps is based on the UK Pollution Climate Mapping (PCM) approach, used to model the annual mean background and roadside concentrations of $PM_{2.5}$ for the UK as a whole [54].

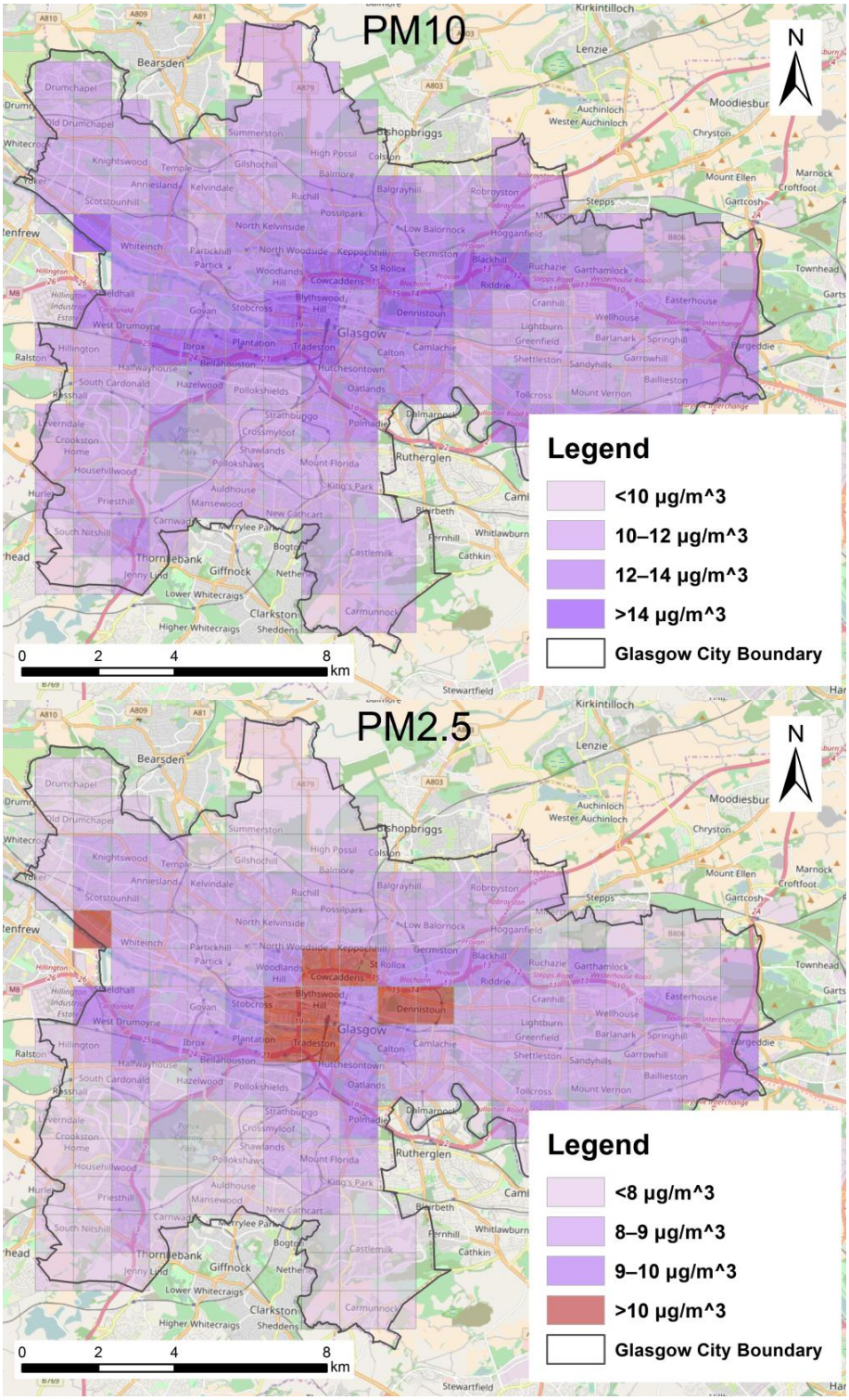

Figure 4. PM grids in Glasgow to represent levels of PM_{10} and $PM_{2.5}$ (source: Scottish Air Quality Database, SAQD).

Only grids that have more than half of their area included in the administrative boundaries of Glasgow are included in the study, resulting in a total of 175 grids. In Glasgow, air pollution levels are observed to have been decreasing in recent years. For instance, the PM_{10} annual mean has decreased from 2010 to 2015 [55]. The main source of air pollution produced is road traffic while the other sources are of less significance [55]. Accordingly, areas with relatively high levels of PM_{10} and $PM_{2.5}$ are in or closed to the city center; while levels of PM_{10} and $PM_{2.5}$ tend to decrease from the city center to outskirts of the city (see Figure 4) where the volume of motor vehicles would be expected to be lower. WHO set 20 $\mu g/m^3$ as a safety guideline for annual mean of PM_{10}, and 10 $\mu g/m^3$ as a safety guideline for annual mean of $PM_{2.5}$ [32] and Glasgow is below the safety guideline for PM_{10} for all the areas (20 $\mu g/m^3$). Annual average $PM_{2.5}$ for several areas (grids) that are located in the city center is above the safety guideline (10 $\mu g/m^3$), while annual average $PM_{2.5}$ for the other areas (grids) is below the safety guideline (10 $\mu g/m^3$). We mark areas (grids) with an annual average $PM_{2.5}$ exceeding the safety guideline in red (see Figure 4).

2.2. Non-Commuting Cycling Activities and Air Pollution Exposure

2.2.1. Spatial Patterns of Non-Commuting Cycling Activities

Firstly, we define the *non-commuting rate* to measure dominance of non-commuting activities within an area (census output area). Suppose i is an area, *non-commuting rate* of i is computed as

$$rate\_non\_act(i) = \frac{num\_non\_act(i)}{num\_non\_act(i) + num\_com\_act(i)} \quad (1)$$

$$num\_non\_act(i) = \sum_{j \in N_i} num\_non\_act^{Node}(j) \quad (2)$$

$$num\_com\_act(i) = \sum_{j \in N_i} num\_com\_act^{Node}(j) \quad (3)$$

where $num\_non\_act(i)$ and $num\_com\_act(i)$ are the number of non-commuting and commuting cycling activities in the area i. $num\_non\_act^{Node}(j)$ and $num\_com\_act^{Node}(j)$ are the number of non-commuting and commuting activities in the node j. N_i is the set of nodes that are located within the area i.

In this paper, the improved AMOEBA (A Multidirectional Optimum Ecotope-Based) algorithm developed by [56] is used to identify clusters of high *non-commuting rate*. As a spatially constrained clustering method, this algorithm is applicable to classification of a large number of areas and identification of irregularly shaped clusters. Here we briefly introduce the improved AMOEBA algorithm based on [56]. Essentially, a region or ecotope is a spatially linked group of areas. A region can thus be defined as a spatially contiguous set of areas. The value of the G_i^* statistic is used to measure the level of clustering of an attribute x around an area. Suppose we run AMOEBA on a study region with N areas and an attribute x with elements x_i, indicating the value of x at area i. Let us denote this set of areas as M, and $\overline{x}$ and S as the mean and the standard deviation of the attribute x and let R be a sub region of M with n areas. Duque et al. [56] rewrite the formulation of G_i^* as follows:

$$G_R^* = \frac{\sum_{i \in R} x_i - n\overline{x}}{S\sqrt{\frac{Nn - n^2}{N-1}}} \quad (4)$$

Basically, G_R^* depends on the areas that are in the region R and the parameters N, $\overline{x}$ and S that are obtained from the areas in M.

Accordingly, a positive (negative) and statistically significant value of G_i^* statistic indicates the presence of a cluster of high (low) values of attribute x around area i. Thus, AMOEBA identifies high-valued, or low-valued, ecotopes (regions) by looking for subsets of spatially connected areas with a high absolute value of the G_i^* statistic. There is only one parameter, i.e., the significance level

threshold, that is required to run the AMOEBA algorithm. The significance level threshold was set to 0.01, meaning only clusters with a p-value less than 0.01 are statistically significant.

2.2.2. Comparison of Air Pollution Exposure by Cycling Purpose

We quantitatively investigate whether cyclists riding for non-commuting (recreation and other purposes) are more likely to be exposed to lower level of PMs (PM_{10} and $PM_{2.5}$) than cyclists riding for commuting purpose at the node level (the intersection level). First, we compare means of instantaneous exposure to PMs for non-commuting and commuting cycling activities. Second, we compare percentages of 'high exposure' activities for non-commuting and commuting cycling activities. As some areas exceed the WHO guideline value for annual $PM_{2.5}$ level (see Figure 4), we call activities that are located within areas (PM grids) of 'high' $PM_{2.5}$ levels (annual $PM_{2.5}$ > the WHO guideline: 10 $\mu g/m^3$) 'high exposure' activities in this paper.

A more reasonable approach to assess the exposure of cyclists to PM concentrations should take account of not only where cyclists are riding but also the time spent cycling to calculate cyclists' inhaled dose of PM_{10} or $PM_{2.5}$ air pollution [28]. Ideally, the comparison of air pollution exposure by cycling purpose should be based on cyclists' inhaled dose of air pollution when riding for commuting and non-commuting. As the Strava Metro do not contain data on time spent cycling, trip distance, or trip speed at the individual level (level of the cyclist), they can't support assessment of cyclists' intake of PM_{10} or $PM_{2.5}$ air pollution. Therefore, in this paper, we use instantaneous exposure to PM_{10} or $PM_{2.5}$ air pollution based on only locations of cycling activities.

Suppose there is one cycling activity at a node, meaning that there is one cyclist at this node at a particular time. We could use levels of PM_{10} or $PM_{2.5}$ at a node (street intersection) to represent exposure of the cyclist to PM_{10} or $PM_{2.5}$ air pollution at the moment when he or she is at that node. In other words, levels of PM_{10} or $PM_{2.5}$ at a node where a cycling activity is located could be used to represent instantaneous exposure of the cyclist to PM_{10} or $PM_{2.5}$ air pollution. We could calculate instantaneous exposure of the cyclist to PM_{10} or $PM_{2.5}$ air pollution for each non-commuting or commuting cycling activity. Accordingly, we could calculate means of instantaneous exposure of the cyclist to PM_{10} or $PM_{2.5}$ air pollution for all non-commuting cycling activities and for all commuting cycling activities by

$$\overline{PM}_{10}^{NON} = \frac{\sum_{i \in S} Num^{NON}(i) * PM_{10}(i)}{\sum_{i \in S} Num^{NON}(i)} \quad (5)$$

$$\overline{PM}_{10}^{COM} = \frac{\sum_{i \in S} Num^{COM}(i) * PM_{10}(i)}{\sum_{i \in S} Num^{COM}(i)} \quad (6)$$

$$\overline{PM}_{2.5}^{NON} = \frac{\sum_{i \in S} Num^{NON}(i) * PM_{2.5}(i)}{\sum_{i \in S} Num^{NON}(i)} \quad (7)$$

$$\overline{PM}_{2.5}^{COM} = \frac{\sum_{i \in S} Num^{COM}(i) * PM_{2.5}(i)}{\sum_{i \in S} Num^{COM}(i)} \quad (8)$$

where i is a node and S is set of nodes. $Num^{NON}(i)$ and $Num^{COM}(i)$ are numbers of non-commuting activities and commuting activities in node i. $PM_{2.5}(i)$ and $PM_{10}(i)$ are the PM_{10} and $PM_{2.5}$ values in node i.

Moreover, percentages of 'high exposure' activities for non-commuting and commuting cycling activities are calculated by

$$Per_{2.5}^{NON} = \frac{\sum_{j \in H_{2.5}} Num^{NON}(j)}{\sum_{i \in S} Num^{NON}(i)} \quad (9)$$

$$Per_{2.5}^{COM} = \frac{\sum_{j \in H_{2.5}} Num^{COM}(j)}{\sum_{i \in S} Num^{COM}(i)} \quad (10)$$

where i is a node and S is set of nodes. $Num^{NON}(i)$ and $Num^{COM}(i)$ are numbers of non-commuting activities and commuting activities in node i. $H_{2.5}$ is a sub set of S, and it consists of nodes that are located in areas (PM grids) with 'high' $PM_{2.5}$ levels (annual $PM_{2.5}$ > the WHO guideline: 10 $\mu g/m^3$). And j is a node in sub set $H_{2.5}$.

2.3. Estimation of Non-Commuting Cycling Activities

In this paper, we try to estimate the number of non-commuting cycling activities at the node level. This means that (1) the dependent variable is the number of non-commuting cycling activities in a node; and (2) the independent variables are the number of all-purpose cycling activities in the node and locational characteristics of the node. Earlier studies investigate impacts of environmental characteristics, including land use mix, cycling facilities, volume or mix of motor vehicle, green space and water, on cycling behavior [57–61]. Table 1 lists the independent variables used in the estimation of non-commuting cycling activities. We select the locational characteristics as the independent variables according to the environmental characteristics from earlier studies and data availability. Specifically, *Dis_to_Greenspace* and *Dis_to_Waterbody* representing distance from node to its nearest rail station and to its nearest water body are used to measure presence or proximity of green space and water. *Dis_to_Citycentre* representing distance from node to the city center is used to reflect levels of land use mix, assuming that the closer to the city center, the higher the level of land use mix is likely to be. *Num_nearest_busstops* representing number of bus stops within a distance of 100 m to node is used to reflect volume or mix of motor vehicles, assuming that the more bus stops there are around, the higher the volume or mix of motor vehicle is likely to be.

Table 1. Independent variables in the estimation of non-commuting cycling activities.

Variable	Meaning
Num_cycling	Number of cycling activities at the node
Dis_to_Greenspace	Distance from node to its nearest rail station
Dis_to_Waterbody	Distance from node to its nearest water body
Dis_to_Citycentre	Distance from node to the city center
Num_nearest_busstops	Number of near bus stops (number of bus stops within a distance of 100 m to node)

In addition to ordinary least squares (OLS), the most widely used model, three other regression models: multilayer perceptron neutral network (MLP), support vector machine (SVM) and random forest (RF) are used to estimate non-commuting cycling activities according to the independent variables when independent variables are not linearly correlated with the dependent variable. We used a cross validation method to evaluate the performances of the four models.

3. Results and Discussion

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

3.1. Non-Commuting Cycling Activities and Air Pollution Exposure

3.1.1. Spatial Patterns of Non-Commuting Cycling Activities

In this section, an investigation of spatial patterns of non-commuting cycling activities in Glasgow is demonstrated. In the input of the AMOEBA algorithm, an observation is the *non-commuting rate* of an area (census output area) (see Equations (1)–(3)). As census output area is the area unit, the entire study region (Glasgow) consists of 5486 areas (census output areas) (see Figure 3). In this paper, running AMOEBA is conducted using ClusterPy. ClusterPy is a Python library of spatially constrained clustering algorithms [62].

In the output of AMOEBA, there are 'solution values' representing clusters of high values or low values. Specifically, areas with positive 'solution values' belong to high value clusters; areas with negative 'solution values' belong to low value clusters; and areas with 'solution values' of zero are those outside the clusters. Table 2 shows how we group areas with 'solution values' to three cluster types: *cluster of high value*, *cluster of low value* and *outside of cluster*. In this empirical study, the *value* here is the *non-commuting rate* of an area.

Table 2. Cluster types and associated solution values.

Solution Value	Cluster Type
≥ 1	Cluster of High value
0	Outside of Cluster
≤ -1	Cluster of Low value

As a result, we map clusters of high and low value *non-commuting rate* in Glasgow (see Figure 5) which reveals that clusters of high value *non-commuting rate* tend to be located in outskirts of the city, away from the city center. This indicates that non-commuting cycling activities are more likely to be located in outskirts of the city.

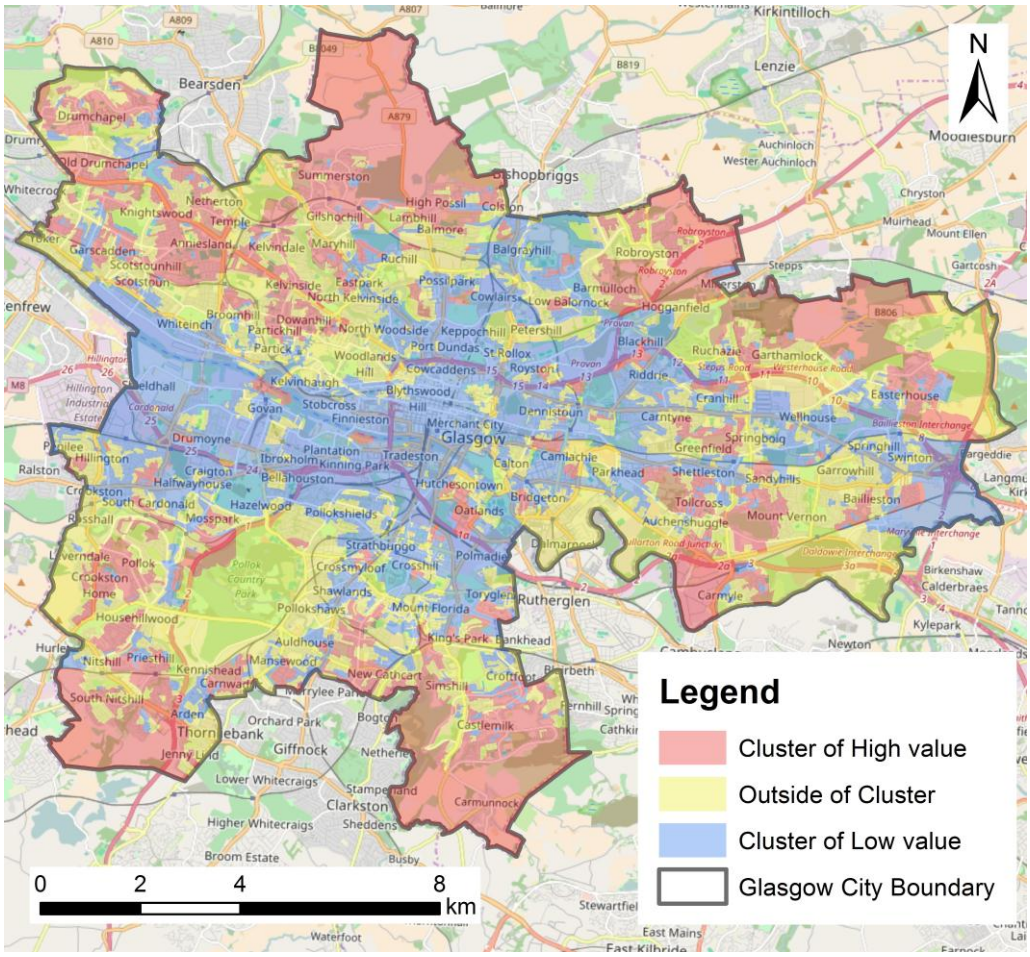

Figure 5. Clusters of high and low *non-commuting rate.*

3.1.2. Comparison of Air Pollution Exposure by Cycling Purpose

Through viewing Figures 4 and 5 in combination we can infer that non-commuting cycling activities are less likely to be in areas of relatively high levels of PM than commuting cycling activities at the census output area level. Table 3 lists percentages of areas of clusters of high and low *non-commuting rate* with different levels of PM_{10} and $PM_{2.5}$. 80% of clusters of high *non-commuting rate* and only 30% of clusters of low *non-commuting rate* are located within grids with a relatively low PM_{10} level (e.g., below 12 $\mu g/m^3$); whilst 70% of clusters of low *non-commuting rate* and only 20% of clusters of high *non-commuting rate* are located within grids with a relatively high PM_{10} level (e.g., 12 $\mu g/m^3$ and above). Similarly, 96% of clusters of high *non-commuting rate* and only 58% of clusters of low *non-commuting rate* are located within grids with a relatively low $PM_{2.5}$ level (e.g., <9 $\mu g/m^3$); whilst 42% of clusters of low *non-commuting rate* and only 4% of clusters of high *non-commuting rate* are located within grids with a relatively high $PM_{2.5}$ level (e.g., 9 $\mu g/m^3$ and above).

Table 3. Percentages of areas of clusters of high and low *non-commuting rate* with different levels of PM_{10} and $PM_{2.5}$.

Percent of Areas		Clusters of High and Low *Non-Commuting Rate*		
		Cluster of High Value	Outside of Cluster	Cluster of Low Value
PM_{10} (Unit: $\mu g/m^3$)	<0	0%	0%	0%
	10–12	80%	61%	30%
	12–14	19%	36%	46%
	>14	1%	3%	24%
$PM_{2.5}$ (Unit: $\mu g/m^3$)	<8	60%	33%	8%
	8–9	36%	58%	50%
	9–10	4%	8%	28%
	>10	0%	1%	14%

Here we quantitatively investigate whether cyclists riding for recreation and other purposes are more likely to be exposed to lower levels of PMs (PM_{10} and $PM_{2.5}$) than cyclists riding for commuting purpose at the node level (the intersection level). Firstly, the means of instantaneous exposure to PM_{10} and $PM_{2.5}$ for non-commuting and commuting cycling activities are calculated by Equations (5)–(8). Table 4 lists the means of instantaneous exposure to PM_{10} and $PM_{2.5}$ for non-commuting and commuting cycling activities. $\overline{PM}_{10}^{NON} < \overline{PM}_{10}^{COM}$ and $\overline{PM}_{2.5}^{NON} < \overline{PM}_{2.5}^{COM}$ indicate that the means of instantaneous exposure to PM_{10} and $PM_{2.5}$ for non-commuting cycling activities are smaller than those for commuting cycling activities. We use the Wilcoxon test to statistically test whether the mean of one group is substantially larger or smaller than the mean of the other group. The Wilcoxon test is used as an alternative to the *T*-test when the data cannot be assumed to be normally distributed. Table 4 also lists results of the Wilcoxon test. In the results of the Wilcoxon test, the *p*-values corresponding to PM_{10} and $PM_{2.5}$ are all less than 0.001. This indicates that the means of instantaneous exposure to PM_{10} and $PM_{2.5}$ for non-commuting cycling activities are both statistically significantly smaller than those for commuting cycling activities at the 0.01 level. This indicates that spatially speaking, cyclists riding for non-commuting purposes tend to be exposed to lower levels of instantaneous PM_{10} and $PM_{2.5}$ air pollution than cyclists riding for commuting purposes.

Table 4. Means of instantaneous exposure to PM_{10} and $PM_{2.5}$ for non-commuting and commuting cycling activities.

Air Pollution Exposure	$\overline{PM}_{10}^{NON}$	$\overline{PM}_{10}^{COM}$	$\overline{PM}_{2.5}^{NON}$	$\overline{PM}_{2.5}^{COM}$
Unit: $\mu g/m^3$	11.823	12.631	8.233	8.753
Wilcoxon Test *p*-value	<0.001		<0.001	

Moreover, we calculate and compare percentages of 'high exposure' activities for non-commuting and commuting cycling activities by Equations (9) and (10). Table 5 lists percentages of 'high exposure' activities for non-commuting and commuting cycling activities. $Per_{2.5}^{COM}$ is two times larger than $Per_{2.5}^{NON}$, indicating that cyclists riding for commuting purposes are more likely to pass through areas of 'high' $PM_{2.5}$ levels than cyclists riding for non-commuting purposes.

Table 5. Percentages of 'high exposure' activities for non-commuting and commuting cycling activities.

$Per_{2.5}^{NON}$	$Per_{2.5}^{COM}$
6.4%	15.0%

In summary, the empirical results reveal that: spatially speaking, cyclists riding for recreation and other purposes are more likely to be exposed to lower level of PMs than cyclists riding for commuting purposes. The means of instantaneous exposure to PM_{10} and $PM_{2.5}$ for non-commuting cycling activities are smaller than those for commuting cycling activities. In addition to encouraging commuters to ride bikes on working days, encouraging people to ride bikes for recreation on non-working days and holidays may contribute to development of urban sustainability. As recreational cyclists are more likely to be in outskirts of cities, policymakers might consider how to improve cycling infrastructure and road safety in those areas when designing or changing urban infrastructure.

3.2. Estimation of Non-Commuting Cycling Activities

The 50,057 nodes used in the spatial analysis constitute the data set used in the estimation of non-commuting cycling activities as well. We further calculate the locational characteristics for each node (see Table 1 in Section 2). An explorative analysis is made to know whether each independent variable is linearly correlated with the dependent variable. Figure 6 shows the scatterplots generated for each independent variable with the dependent variable. Apart from *Num_cycling*, other independent variables do not have a significant correlation with the dependent variable as absolute values of the corresponding Pearson correlation coefficients are less than 0.1. Therefore, in addition to OLS, three other models are also used to estimate non-commuting cycling activities.

We run a 10-folder cross validation to measure the performances of the four estimation models. Correlation of predicted and actual number of non-commuting activities is used to measure estimation performance. Random forest (RF) outperforms other algorithms with a correlation coefficient of 0.981 (see Table 6). Figure 7 plots the predicted and actual number of non-commuting activities. The estimation results indicate that the estimation of the number of non-commuting cycling activities is fairly good in this study. This suggests that we may be able to estimate the number of non-commuting cycling trips when the trip purpose of cycling data is unknown.

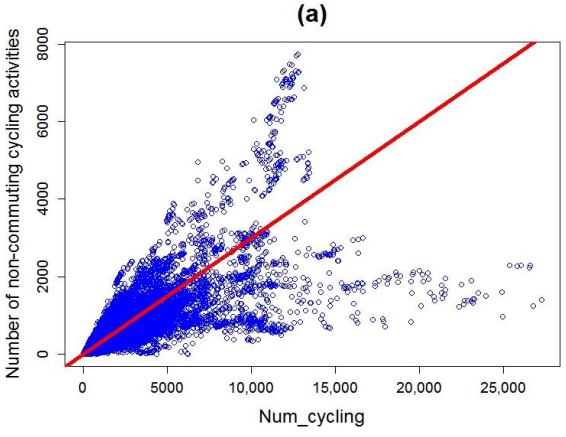

Figure 6. *Cont.*

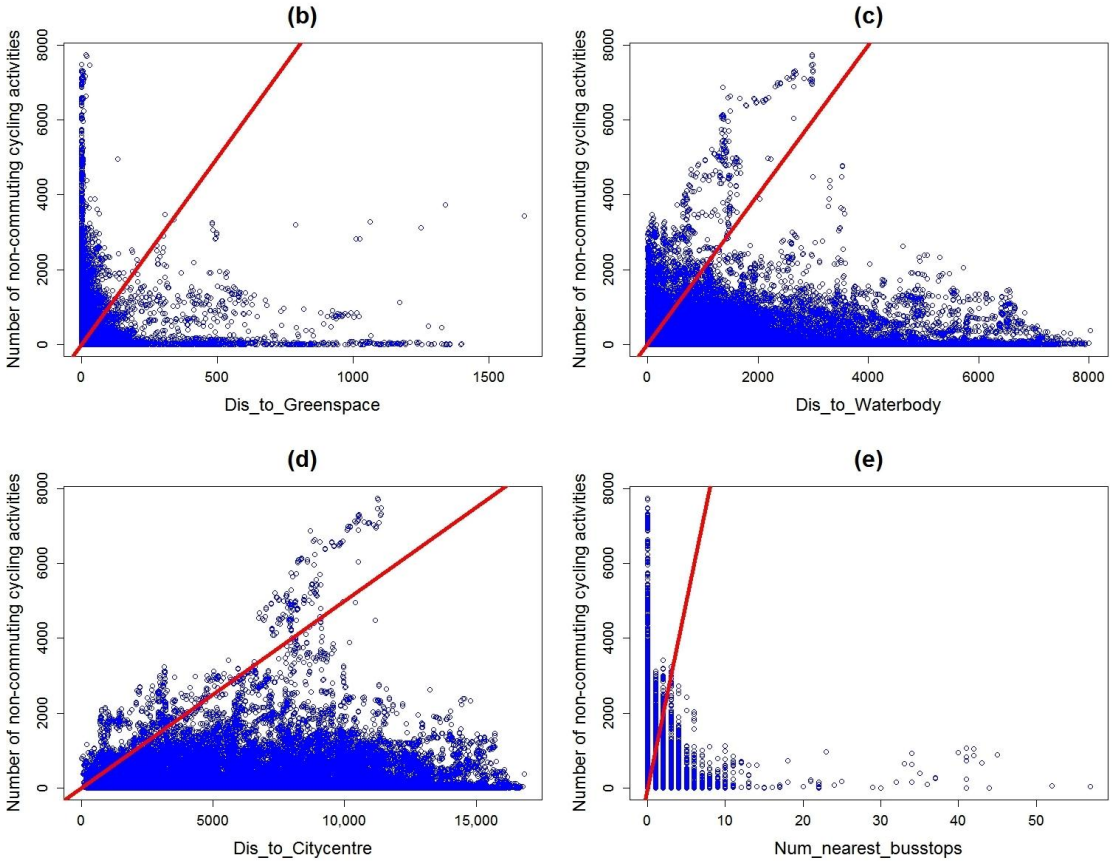

Figure 6. Scatterplots generated for each independent variable with the dependent variable: (**a**) Num_cycling vs. the dependent variable; (**b**) Dis_to_Greenspace vs. the dependent variable; (**c**) Dis_to_Waterbody vs. the dependent variable; (**d**) Dis_to_Citycentre vs. the dependent variable; (**e**) Num_nearest_busstops vs. the dependent variable.

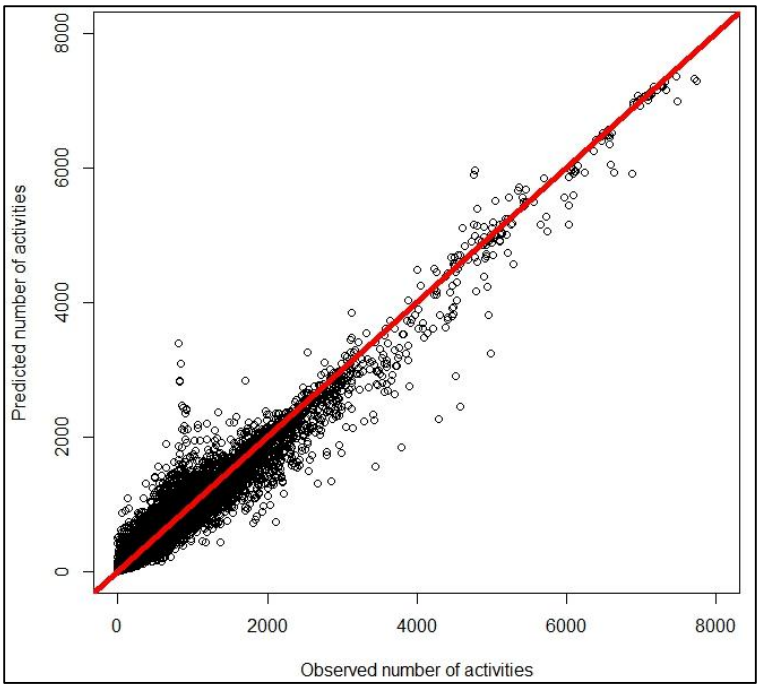

Figure 7. Predicted and observed number of non-commuting activities.

Table 6. Estimation results of non-commuting cycling activities by different algorithms.

Accuracy	OLS	MLP	SVM	RF
Correlation coefficient (Pearson's *R*)	0.814	0.904	0.807	0.981

OLS: ordinary least squares; MLP: multilayer perceptron neutral newtork; SVM: support vector machine; RF: random forest.

4. Conclusions

In this study, we investigate spatial patterns of cycling activities and associations between cycling purpose and air pollution exposure in Glasgow, UK by using Strava Metro data. Empirical results reveal some findings that (1) compared with commuting cycling activities, non-commuting cycling activities are more likely to be located in outskirts of the city; (2) spatially speaking, cyclists riding for recreation and other purposes are more likely to be exposed to relatively low levels of air pollution than cyclists riding for commuting; and (3) the method for estimating of the number of non-commuting cycling activities works well in this study. The results suggest that (1) policymakers might consider how to improve cycling infrastructure and road safety in outskirts of cities; and (2) we may be able to estimate the number of non-commuting cycling activities when trip purpose of cycling data is unknown. We conclude that this study is a good start in utility of crowdsourced cycling data for studies of cycling and air pollution exposure.

4.1. Limitations

This paper does present a few limitations. First, a census output area is used as the area unit in in the identifying clusters. The modifiable areal unit problem (MAUP) might influence the cluster identification in this study. Second, although the estimation of non-commuting cycling activities is good at the node level, the estimation at the edge (street) level is unknown. Third, although the models work well in estimating non-commuting cycling activities, there might be some potential to improve the estimation. Ideally, we could improve the estimation by incorporating more attributes such as land use mix, residential density, traffic count, road type, road width, etc., into the models. We are not able to include those attributes due to present data availability. Fourth, instantaneous assessment of air pollution is used in this study. In fact, cumulative assessment of air pollution makes more sense to studies of health effects of active travel. Ideally, the inhaled dose of air pollution during a cycling trip should be assessed according to not only where the trip takes place but also the time spent travelling. Furthermore, long-term air pollution exposure of a cyclist should also take account of the number of his or her commuting trips and non-commuting trips within a longer period (e.g., one year or more). As Strava Metro doesn't offer individual-level trips, we know neither how long a commuting or non-commuting cycling trip takes nor how many commuting or non-commuting cycling trips each biker takes in one year. Thus, we are not able to assess cumulative exposure of a cyclist to air pollution. Finally, since VGI, UGC and crowdsourced data are collected and shared by individuals, there are arguments about the quality and fitness for use of such data in projects [63]. While we are aware of this issue, we do not tackle this in this study, as it requires separate study on this topic.

4.2. Future Works

In the future, we will take account of some aspects to enhance this study. First, as the Strava database offered by Urban Big Data Centre offers the number of cycling activities in distinct daily time slots, we could model spatio-temporal variations in non-commuting cycling activities according to spatio-temporal characteristics; Second, we will incorporate other air pollutants such as sulphur dioxide, nitrogen dioxide and ozone into the analysis. In addition, although Strava Metro doesn't offer individual-level trips, we may be able to assess cumulative air pollution exposure by using the sub data set 'Streets' in Strava Metro. Based on the number of trips on each street segment, we could use

length of the street segment to represent the length of sub cycling trip, and then estimate the time of sub cycling trip based on average speed of cycling for commuting or recreation and other purposes in Glasgow, a figure we could possibly obtain from Strava Metro or other data sources (e.g., travel surveys). Accordingly, we could estimate the total air pollution exposure to all of Glasgow's Strava cyclists when riding for commuting or recreation and other purposes during one year, and further estimate annual average air pollution exposure of one cyclist when riding for commuting or recreation and other purposes. This would represent a large amount of future work and potential to gain a much greater understanding of the impact of city air pollution.

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. The authors would like to express gratitude to the anonymous referees for their comments and thank Keith Maynard at Urban Big Data Centre University of Glasgow for proof reading this article.

Author Contributions: Yeran Sun conceived and designed the experiments; Yeran Sun and Amin Mobasheri preprocessed and analyzed the data; Yeran Sun performed the experiments in this article; Yeran Sun and Amin Mobasheri wrote the manuscript.

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

References

1. Cavill, N.; Davis, A. *Cycling and Health: What's the Evidence?* Cycling England: London, UK, 2007.
2. Forsyth, A.; Krizek, K.J.; Agrawal, A.W.; Stonebraker, E. Reliability testing of the Pedestrian and Bicycling Survey (PABS) method. *J. Phys. Act. Health* **2012**, *9*, 677–688. [CrossRef] [PubMed]
3. Oja, P.; Vuori, I.; Paronen, O. Daily walking and cycling to work: Their utility as health-enhancing physical activity. *Patient Educ. Couns.* **1998**, *33* (Suppl. 1), S87–S94. [CrossRef]
4. Oja, P.; Titze, S.; Bauman, A.; de Geus, B.; Krenn, P.; Reger-Nash, B.; Kohlberger, T. Health benefits of cycling: A systematic review. *Scand. J. Med. Sci. Sports* **2011**, *21*, 496–509. [CrossRef] [PubMed]
5. Pucher, J.; Buehler, R.; Bassett, D.R.; Dannenberg, A.L. Walking and cycling to health: A comparative analysis of city, state, and international data. *Am. J. Public Health* **2010**, *100*, 1986–1992. [CrossRef] [PubMed]
6. Taddei, C.; Gnesotto, R.; Forni, S.; Bonaccorsi, G.; Vannucci, A.; Garofalo, G. Cycling Promotion and Non-Communicable Disease Prevention: Health Impact Assessment and Economic Evaluation of Cycling to Work or School in Florence. *PLoS ONE* **2015**, *10*, e0125491. [CrossRef] [PubMed]
7. Wen, L.M.; Rissel, C. Inverse associations between cycling to work, public transport, and overweight and obesity: Findings from a population based study in Australia. *Prev. Med.* **2008**, *46*, 29–32. [PubMed]
8. Elwood, S.; Goodchild, M.F.; Sui, D.Z. Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Assoc. Am. Geogr.* **2012**, *102*, 571–590. [CrossRef]
9. Griffin, G.P.; Jiao, J. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *J. Transp. Health* **2015**, *2*, 238–247. [CrossRef]
10. Forsyth, A.; Oakes, J.M. Cycling, the Built Environment, and Health: Results of a Midwestern Study. *Int. J. Sustain. Transp.* **2015**, *9*, 49–58. [CrossRef]
11. Sener, I.N.; Eluru, N.; Bhat, C.R. An analysis of bicycle route choice preferences in Texas, US. *Transportation* **2009**, *36*, 511–539. [CrossRef]
12. El Esawey, M. Estimation of annual average daily bicycle traffic with adjustment factors. *Transp. Res. Rec.* **2014**, *2443*, 106–114. [CrossRef]
13. Jesticoa, B.; Nelsona, T.; Wintersb, M. Mapping ridership using crowdsourced cycling data. *J. Transp. Geogr.* **2016**, *52*, 90–97. [CrossRef]
14. Broach, J.; Dill, J.; Gliebe, J. Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transp. Res. Part A Policy Pract.* **2012**, *46*, 1730–1740.
15. Casello, J.M.; Usyukov, V. Modeling cyclists' route choice based on GPS data. *Transp. Res. Rec.* **2014**, *2430*, 155–161. [CrossRef]
16. Hood, J.; Sall, E.; Charlton, B. A GPS-based bicycle route choice model for San Francisco, California. *Transp. Lett.* **2011**, *3*, 63–75. [CrossRef]
17. Heesch, K.C.; James, B.; Washington, T.L.; Zunig, K.; Burke, M. Evaluation of the Veloway 1: A natural experiment of new bicycle infrastructure in Brisbane, Australia. *J. Transp. Health* **2016**, *3*, 366–376. [CrossRef]

18. Strava Metro. Data-Driven Bicycle and Pedestrian Planning. 2016. Available online: http://metro.strava.com/ (accessed on 15 January 2016).
19. Bascom, R.; Bromberg, P.A.; Costa, D.L.; Devlin, R.; Dockery, D.W.; Frampton, M.W.; Lambert, W.; Samet, J.M.; Speizer, F.E.; Utell, M.; et al. Health effects of outdoor air pollution. *Am. J. Respir. Crit. Care Med.* **1996**, *153*, 477–498.
20. Cao, J.; Yang, C.; Li, J.; Chen, R.; Chen, B.; Gu, D.; Kan, H. Association between long-term exposure to outdoor air pollution and mortality in China: A cohort study. *J. Hazard. Mater.* **2011**, *186*, 1594–1600. [CrossRef] [PubMed]
21. Künzli, N.; Kaiser, R.; Medina, S.; Studnicka, M.; Chanel, O.; Filliger, P.; Herry, M.; Horak, F., Jr.; Puybonnieux-Texier, V.; Quénel, P.; et al. Public-health impact of outdoor and traffic-related air pollution: A European assessment. *Lancet* **2000**, *356*, 795–801. [CrossRef]
22. Briggs, D.J.; de Hoogh, K.; Morris, C.; Gulliver, J. Effects of travel mode on exposures to particulate air pollution. *Environ. Int.* **2008**, *34*, 12–22. [CrossRef] [PubMed]
23. Chertok, M.; Voukelatos, A.; Sheppeard, V.; Rissel, C. Comparison of air pollution exposure for five commuting modes in Sydney—Car, train, bus, bicycle and walking. *Health Promot. J. Aust.* **2004**, *15*, 63–67.
24. De Nazelle, A.; Fruin, S.; Westerdahl, D.; Martinez, D.; Ripoll, A.; Kubesch, N.; Nieuwenhuijsen, M. A travel mode comparison of commuters' exposures to air pollutants in Barcelona. *Atmos. Environ.* **2012**, *59*, 151–159. [CrossRef]
25. Gulliver, J.; Briggs, D. Time-space modeling of journey-time exposure to traffic-related air pollution using GIS. *Environ. Res.* **2005**, *97*, 10–25. [CrossRef] [PubMed]
26. Rojas-Rueda, D.; de Nazelle, A.; Tainio, M.; Nieuwenhuijsen, M.J. The health risks and benefits of cycling in urban environments compared with car use: Health impact assessment study. *BMJ* **2011**, *343*, d4521. [CrossRef] [PubMed]
27. Zuurbier, M.; Hoek, G.; Oldenwening, M.; Lenters, V.; Meliefste, K.; van den Hazel, P.; Brunekreef, B. Commuters' exposure to particulate matter air pollution is affected by mode of transport, fuel type, and route. *Environ. Health Perspect.* **2010**, *118*, 783–789. [CrossRef] [PubMed]
28. Tainio, M.; de Nazelle, A.J.; Götschi, T.; Kahlmeier, S.; Rojas-Reuda, D.; Nieuwenhuijsen, M.J.; De sa Herick, T.; Kelly, P.; Woodcock, J. Can air pollution negate the health benefits of cycling and walking? *Prev. Med.* **2016**, *87*, 233–236. [CrossRef] [PubMed]
29. De Nazelle, A.; Seto, E.; Donaire-Gonzalez, D.; Mendez, M.; Matamala, J.; Nieuwenhuijsen, M.J.; Jerrett, M. Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environ. Pollut.* **2013**, *176*, 92–99. [CrossRef] [PubMed]
30. Weichenthal, S.; Kulka, R.; Dubeau, A.; Martin, C.; Wang, D.; Dales, R. Traffic-related air pollution and acute changes in heart rate variability and respiratory function in urban cyclists. *Environ. Health Perspect.* **2011**, *119*, 1373–1378. [CrossRef] [PubMed]
31. Hollingworth, M.; Harper, A.; Hamer, M. An Observational Study of Erectile Dysfunction, Infertility, and Prostate Cancer in Regular Cyclists: Cycling for Health UK Study. *J. Men's Health* **2014**, *11*, 75–79. [CrossRef]
32. WHO. Air Pollution Levels Rising in Many of the World's Poorest Cities. 2016. Available online: http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/ (accessed on 15 May 2016).
33. Anderson, J.O.; Thundiyil, J.G.; Stolbach, A. Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health. *J. Med. Toxicol.* **2012**, *8*, 166–175. [CrossRef] [PubMed]
34. Dockery, D.W.; Pope, C.A.; Xu, X.; Spengler, J.D.; Ware, J.H.; Fay, M.E.; Ferris, B.G., Jr.; Speizer, F.E. An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.* **1993**, *329*, 1753–1759. [CrossRef] [PubMed]
35. Dominici, F.; Peng, R.D.; Bell, M.L.; Pham, L.; McDermott, A.; Zeger, S.L.; Samet, J.M. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* **2006**, *295*, 1127–1134. [CrossRef] [PubMed]
36. Hoek, G.; Brunekreef, B.; Goldbohm, S.; Fischer, P.; van den Brandt, P.A. Association between mortality and indicators of traffic-related air pollution in the Netherlands: A cohort study. *Lancet* **2002**, *360*, 1203–1209. [CrossRef]

37. Miller, K.A.; Siscovick, D.S.; Sheppard, L.; Shepherd, K.; Sullivan, J.H.; Anderson, G.L.; Kaufman, J.D. Long-term exposure to air pollution and incidence of cardiovascular events in women. *N. Engl. J. Med.* **2007**, *356*, 447–458. [CrossRef] [PubMed]
38. Lipsett, M.J.; Ostro, B.D.; Reynolds, P.; Goldberg, D.; Hertz, A.; Jerrett, M.; Smith, D.F.; Garcia, C.; Chang, E.T.; Bernstein, L. Long-Term Exposure to Air Pollution and Cardiorespiratory Disease in the California Teachers Study Cohort. *Am. J. Respir. Crit. Care Med.* **2011**, *184*, 828–835. [CrossRef] [PubMed]
39. Pope, C.A.; Burnett, R.T.; Thurston, G.D.; Thun, M.J.; Calle, E.E.; Krewski, D.; Godleski, J.J. Cardiovascular mortality and long-term exposure to particulate air pollution: Epidemiological evidence of general pathophysiological pathways of disease. *Circulation* **2004**, *109*, 71–77. [CrossRef] [PubMed]
40. Pope, C.A.; Turner, M.C.; Burnett, R.T.; Jerrett, M.; Gapstur, S.M.; Diver, W.R.; Krewski, D.; Brook, R.D. Relationships between fine particulate air pollution, cardiometabolic disorders and cardiovascular mortality. *Circ. Res.* **2015**, *116*, 108–115. [CrossRef] [PubMed]
41. Bell, M.L.; Ebisu, K.; Peng, R.D.; Walker, J.; Samet, J.M.; Zeger, S.L.; Dominici, F. Seasonal and regional short-term effects of fine particles on hospital admissions in 202 US counties, 1999–2005. *Am. J. Epidemiol.* **2008**, *168*, 1301–1310. [CrossRef] [PubMed]
42. Chen, R.; Kan, H.; Chen, B.; Huang, W.; Bai, Z.; Song, G.; Pan, G. Association of particulate air pollution with daily mortality: The China Air Pollution and Health Effects Study. *Am. J. Epidemiol.* **2012**, *175*, 1173–1181. [CrossRef] [PubMed]
43. Li, M.H.; Fan, L.C.; Mao, B.; Yang, J.W.; Choi, A.M.; Cao, W.J.; Xu, J.F. Short-Term Exposure to Ambient Fine Particulate Matter Increases Hospitalizations and Mortality in COPD: A Systematic Review and Meta-Analysis. *Chest* **2016**, *149*, 447–458. [CrossRef] [PubMed]
44. Shah, A.S.; Lee, K.K.; McAllister, D.A.; Hunter, A.; Nair, H.; Whiteley, W.; Langrish, J.P.; Newby, D.E.; Mills, N.L. Short term exposure to air pollution and stroke: Systematic review and meta-analysis. *BMJ* **2015**, *350*, h1295. [CrossRef] [PubMed]
45. Doorley, R.; Pakrashi, V.; Ghosh, B. Quantifying the health impacts of active travel: Assessment of methodologies. *Transp. Rev.* **2015**, *35*, 559–582. [CrossRef]
46. Mueller, N.; Rojas-Rueda, D.; Cole-Hunter, T.; de Nazelle, A.; Dons, E.; Gerike, R.; Götschi, T.; Int Panis, L.; Kahlmeier, S.; Nieuwenhuijsen, M. Health impact assessment of active transportation: A systematic review. *Prev. Med.* **2015**, *76*, 103–114. [CrossRef] [PubMed]
47. Lippmann, M. Exposure science in the 21st century: A vision and a strategy. *J. Expo. Sci. Environ. Epidemiol.* **2013**, *23*, 1. [CrossRef] [PubMed]
48. Riordan, B. Strava Metro: Better Data for Better Cities. Metro.Strava.com. 2016. Available online: http://ubdc.ac.uk/media/1416/uofg-training.pdf (accessed on 15 October 2016).
49. Strava Metro. Strava Metro Comprehensive User Guide Version 2.0. 2015. Available online: http://ubdc.ac.uk/media/1323/stravametro_200_user_guide_withoutpics.pdf (accessed on 15 January 2016).
50. Herrero, J. Using Big Data to Understand Trail Use: Three Strava Tools. TRAFx Research. 2016. Available online: https://www.trafx.net/insights.htm (accessed on 15 May 2016).
51. Urban Big Data Centre, UK. Data Services: Strava Metro Data. Available online: http://ubdc.ac.uk/data-services/data-catalogue/transport-data (accessed on 15 January 2016).
52. DATA.GOV.UK. Datasets: Output Areas. Available online: https://data.gov.uk/dataset (accessed on 15 October 2015).
53. Ricardo Energy & Environment. 2015 Air Quality in Scotland. Available online: http://www.scottishairquality.co.uk (accessed on 15 March 2016).
54. UK Department for Environment Food & Rural Affairs. Background Concentration Maps User Guide. GOV.UK, 2016. Available online: http://laqm.defra.gov.uk (accessed on 15 June 2016).
55. Ricardo Energy and Environment. Technical Reports—Air Quality Scotland Brochure 2015. The Scottish Government, 2016. Available online: http://www.scottishairquality.co.uk (accessed on 15 August 2016).
56. Duque, J.C.; Aldstadt, J.; Velasquez, E.; Franco, J.L.; Betancourt, A. A computationally efficient method for delineating irregularly shaped spatial clusters. *J. Geogr. Syst.* **2011**, *13*, 355–372. [CrossRef]
57. Carver, A.; Salmon, J.; Campbell, K.; Baur, L.; Garnett, S.; Crawford, D. How do perceptions of local neighborhood relate to adolescents' walking and cycling? *Am. J. Health Promot.* **2005**, *20*, 139–147. [CrossRef] [PubMed]
58. Hunt, J.D.; Abraham, J.E. Influences on bicycle use. *Transportation* **2007**, *34*, 453. [CrossRef]

59. De Vries, S.I.; Hopman-Rock, M.; Bakker, I.; Hirasing, R.A.; van Mechelen, W. Built environmental correlates of walking and cycling in Dutch urban children: Results from the SPACE study. *Int. J. Environ. Res. Public Health* **2010**, *7*, 2309–2324. [CrossRef] [PubMed]
60. Fraser, S.; Lock, K. Cycling for transport and public health: A systematic review of the effect of the environment on cycling. *Eur. J. Public Health* **2010**, *21*, 738–743. [CrossRef] [PubMed]
61. Mäki-Opas, T.E.; Borodulin, K.; Valkeinen, H.; Stenholm, S.; Kunst, A.E.; Abel, T.; Härkänen, T.; Kopperoinen, L.; Itkonen, P.; Prättälä, R.; et al. The contribution of travel-related urban zones, cycling and pedestrian networks and green space to commuting physical activity among adults—A cross-sectional population-based study using geographical information systems. *BMC Public Health* **2016**, *16*, 760. [CrossRef] [PubMed]
62. RiSE Group. ClusterPy: Library of Spatially Constrained Clustering Algorithms. Available online: www.rise-group.org/risem/clusterpy (accessed on 15 January 2011).
63. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 139–167. [CrossRef]