# Exploring potential of crowdsourced geographic information in studies of active travel and health: Strava data and cycling behaviour

Yeran Sun*

Urban Big Data Centre, University of Glasgow, 7 Lilybank Gardens, Glasgow G12 8RZ, United Kingdom - yeran.sun@glasgow.ac.uk

**KEY WORDS:** Crowdsourced Geographic Information, Strava Metro, Cycling, Mobility, Spatial analysis

**ABSTRACT:**

In development of sustainable transportation and green city, policymakers encourage people to commute by cycling and walking instead of motor vehicles in cities. One the one hand, cycling and walking enables decrease in air pollution emissions. On the other hand, cycling and walking offer health benefits by increasing people's physical activity. Earlier studies on investigating spatial patterns of active travel (cycling and walking) are limited by lacks of spatially fine-grained data. In recent years, with the development of information and communications technology, GPS-enabled devices are popular and portable. With smart phones or smart watches, people are able to record their cycling or walking GPS traces when they are moving. A large number of cyclists and pedestrians upload their GPS traces to sport social media to share their historical traces with other people. Those sport social media thus become a potential source for spatially fine-grained cycling and walking data. Very recently, Strava Metro offer aggregated cycling and walking data with high spatial granularity. Strava Metro aggregated a large amount of cycling and walking GPS traces of Strava users to streets or intersections across a city. Accordingly, as a kind of crowdsourced geographic information, the aggregated data is useful for investigating spatial patterns of cycling and walking activities, and thus is of high potential in understanding cycling or walking behavior at a large spatial scale. This study is a start of demonstrating usefulness of Strava Metro data for exploring cycling or walking patterns at a large scale.

## 1. INTRODUCTION

By means of enhancing physical activity, active travel (cycling or walking) produces health benefit (Forsyth et al., 2015; Oja et al., 1998; Oja et al., 2001; Pucher et al., 2010; Wen and Rissel, 2008). 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 (Griffin and Jiao, 2015). Specifically, travel survey data tends to have a low spatial granularity as geography level of travel survey data is usually census tract; whilst traffic counts data have a high spatial granularity but a low spatial coverage as traffic counts points are usually located in major roads other than minor roads. In recent years, GPS-enabled mobile devices, such as smartphones and smartwatches, allow individuals to track their cycling GPS traces with fine spatial granularity (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 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 it launched a data service called Strava Metro. 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 of each city. Strava Metro data is of high potential in a wide range of applications, including mapping cycling activities over cities (Jesticoa et al., 2016), assessing effects of environmental factors on cycling behavior (Griffin and Jiao, 2015; Heesch et al., 2016) and assessing air pollution during cycling (Sun and Mobasheri, 2017). Moreover, by comparing cyclist counts between Strava data and manual count data in count stations, some studies have revealed that Strava Metro data is a good representation of cycling population (Jesticoa et al., 2016; Herrero, 2016). As a result, due to a high level of spatial granularity and a large spatial coverage Strava Metro provides an opportunity to depicting cycling behaviour. This study aims to demonstrate usefulness of Strava Metro data in depicting cycling behaviour over a city by taking account of cycling activities and daytime population. Moreover, this study could offer implications for policies to help policymakers to consider investment priority in bicycle infrastructure of the areas where cyclists are likely to go.

## 2. MATERIALS AND METHODS

In this section, research data and methods are presented. Specifically, sub section 2.1 introduces the research data, and sub section 2.2 introduces the approach to investigating spatial patterns of cycling behaviour.

### 2.1 Research Data

The Strava Metro dataset (Urban Big Data Centre, 2016) has 287, 833 cycling activities within the Glasgow Clyde Valley Planning area (including Glasgow City and seven contiguous council areas) in 2015. This dataset contains three sub sets with three different formats: Streets, Origin-Destination, Nodes (Strava Metro, 2015). This study uses the Nodes sets. The Nodes set was created based on a street network which is extracted from OpenStreetMap. Specifically, the Node set contains all nodes of the street network, and each node represents an intersection of streets (see Figure 1). Table 1 lists attributes of nodes, including count of cycling activities (regardless of unique riders) at the node (street intersection) at a specific time. Note that the temporal granularity is the minute level (Strava Metro, 2015).

| Field | Description |
|---|---|
| node_id | Unique 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_ride | Number of cycling activities |

Table 1. Fields in the Nodes file (Strava Metro, 2015)

Additionally, the dataset contains a file that offers demographics of the cycling 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 contributed by over 10 thousand of cyclists. It is noted that, although this data set has a large user sample set, average annual cycling frequency of Strava users seems to be much smaller than the real frequencies. Specifically, on average, each cyclist has 21 cycling trips in 2015. Unsurprisingly, male cyclists outnumber female cyclists. Specifically, number of male cyclists is 5 times of number of female cyclists. The largest age group of male cyclists is 35-44 whilst the largest age group of female cyclists is 25-34. Generally, almost half of cycling trips were contributed by users aged 25-44 (25-34 and 35-44). Additionally, a large portion of trips are recreational trips (Strava Metro, 2015). Therefore, the majority of the Strava users are likely to be young and sporty cyclists.

| Cycling | |
|---|---|
| Athlete ID count (User count) | 13,684 |
| Activity count (Trip count) | 287,833 |
| Average distance of trips | 24 km |
| Average time of trips | 81 minute |

| Age | Male | Female |
|---|---|---|
| Under 25 | 718 | 141 |
| 25-34 | 2,176 | 417 |
| 35-44 | 2,957 | 346 |
| 45-54 | 2,028 | 217 |
| 55-64 | 448 | 44 |
| Over 64 | 73 | 2 |
| No Birth date | 2,812 | 531 |
| Total | 11,212 | 1,698 |

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

In this study, daytime population is used as background population. The daytime population data is downloaded from Scotland's Census (2016). The geography level of daytime population data is census output area. Daytime population is estimated based on the 2011 census data. Specifically, the daytime population is an estimate of the population of an area during the working day. It includes everybody who works or studies in the area, wherever they usually live, and all respondents who live in the area but do not work or study. People who work or study mainly at or from home, or who do not have a fixed place of work or study, are included in the area containing their home address. The daytime population will include shift and night workers such as hospital staff and security guards. Figure 3 maps density of daytime population at the census output area level. Areas with high-density daytime population are not particularly situated around the city centre.

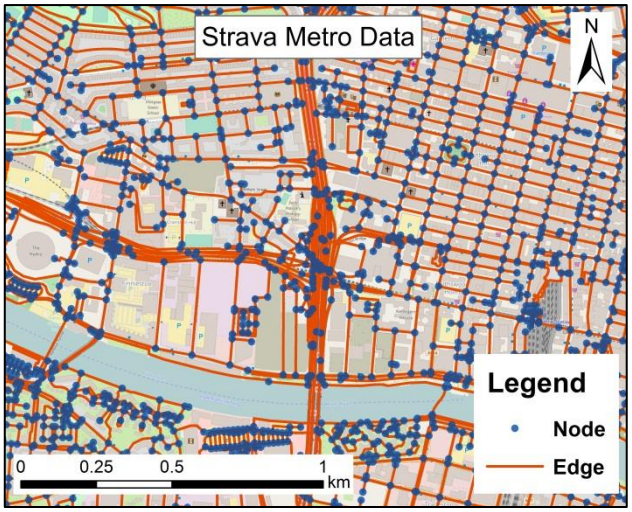


Figure 1. Nodes and edges of Strava Metro data (Basemap: OpenStreetMap, licensed under the Open Database License).

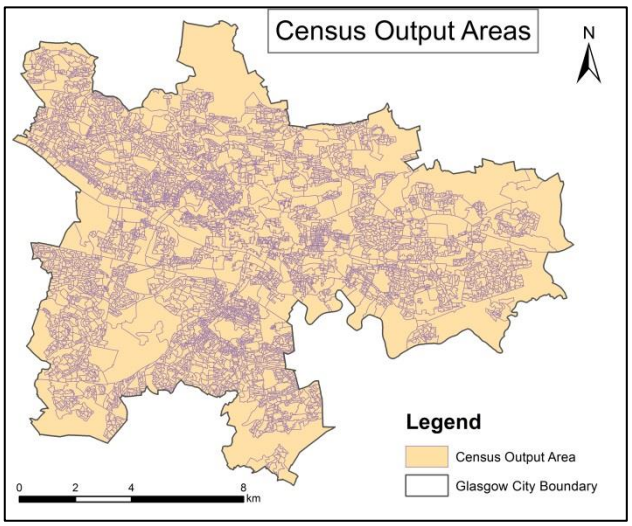


Figure 2. Census output areas in Glasgow.

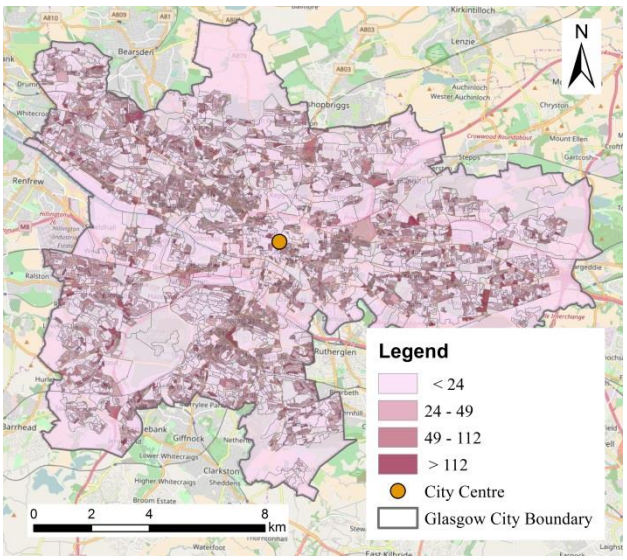


Figure 3. Density of daytime population in Glasgow (Basemap: OpenStreetMap, licensed under the Open Database License).

### 2.2 Investigation of spatial patterns

This study explores spatial patterns of cycling behaviour over a city by identifying spatial clusters of cycling activities. By considering background population this study uses the ratio of cycling activities to daytime population to identify clusters of high density cycling activities. Specifically, an improved AMOEBA (A Multidirectional Optimum Ecotope-Based) algorithm developed by Duque et al. (2011) is used to identify clusters of high ratio of cycling activities to daytime population. Then this study associates clusters with locally environmental characteristics such as land use type. As the population data is available at the census area level, this study calculates the ratio of cycling activities to daytime population at the census area level.

Firstly, this study defines the ratio of cycling activities to daytime population (RCADTP) within an area (census output area). Suppose *i* is an area, RCADTP of *i* is computed as

$$ratio\_ride\_pop(i) = \frac{num\_ride^{Area}(i)}{DT\_pop^{Area}(i)} \quad (1)$$

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

where $num\_ride^{Area}(i)$ is the number of cycling activities in the area *i*, and $DT\_pop^{Area}(i)$ is the daytime population in the area *i*. $N_i$ is the set of nodes that are located within the area *i*, and $num\_ride^{Node}(j)$ is the number of cycling activities in the node *j*.

In this paper, the improved AMOEBA (A Multidirectional Optimum Ecotope-Based) algorithm developed by Duque et al. (2011) is used to identify clusters of high RCADTP. This algorithms suits for the task in this study as it is applicable to classification of a large number of areas and identification of irregularly shaped clusters. This study briefly introduces the improved AMOEBA algorithm based on Duque et al. (2011). 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 $\bar{x}$ 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. (2011) rewrite the formulation of $G_i^*$ as follows:

$$G_R^* = \frac{\sum_{i \in R} x_i - n\bar{x}}{S\sqrt{\frac{Nn-n^2}{N-1}}} \quad (3)$$

Basically, $G_R^*$ depends on the areas that are in the region *R* and the parameters *N*, $\bar{x}$ 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.

## 3. RESULTS AND DISCUSSION

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

### 3.1 Spatial patterns of cycling behaviour

First of all, annual total cycling activities at each node is calculated after aggregating number of cycling activities at different times throughout the year 2015. Second, Nodes and census output area boundaries are overlapped. Then total number of cycling activities and RCADTP of each census output area are calculated according to Equations (1)-(2). Figure 4 maps number of cycling activities in census output areas. Areas with high-density cycling activities are situated around the city centre.

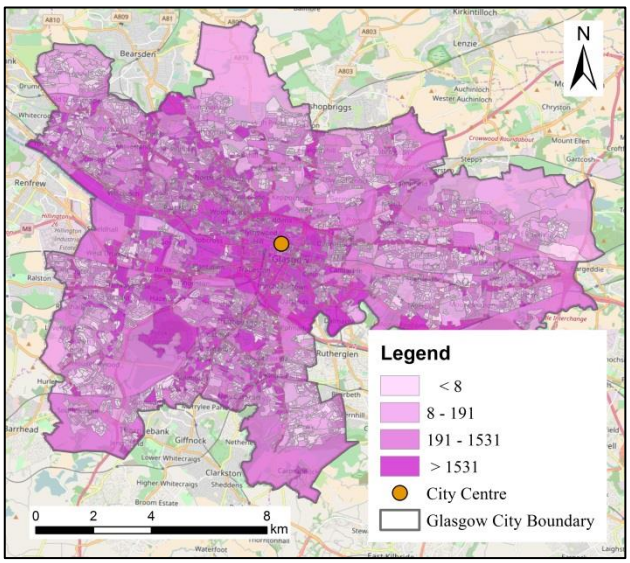


Figure 4. Number of cycling activities in census output areas.

In the AMOEBA algorithm, an observation is the RCADTP of an area (census output areas). In this paper, running AMOEBA is conducted using ClusterPy (RiSE group). The AMOEBA algorithm identifies statistically significant clusters of high value and clusters of low value. Figure 5 maps the cluster of high RCADTP. In the top map, clusters of high value and low value represent cluster of high RCADTP and low RCADTP respectively. This study then associates clusters of high RCADTP with locally environmental characteristics such as main land use types by overlapping the clusters and basemap such as GoogleMap and OpenStreetMap. As a consequence, clusters of high RCADTP mainly surround green spaces such as parks and gardens, as well as the river crossing the city (see the bottom map in Figure 5). Strava cyclists are likely to go to green spaces and the riverside. This implies that large portion of Strava cycling trips tend to be recreational cycling trips.

### 3.2 Discussion

Moreover, this study could offer an implication for policies that improvement on bicycle infrastructure in clusters of high RCADTP to increase road safety for cyclists and attract more recreational cyclists. Nevertheless, there are still some limitations in this paper. First, there is representativeness bias in cycling 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; whilst 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. Second, 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, this study would select GPS traces of cycles created by a number of Strava users who compose a cohort. This would enable a cohort study of cyclists in a city.

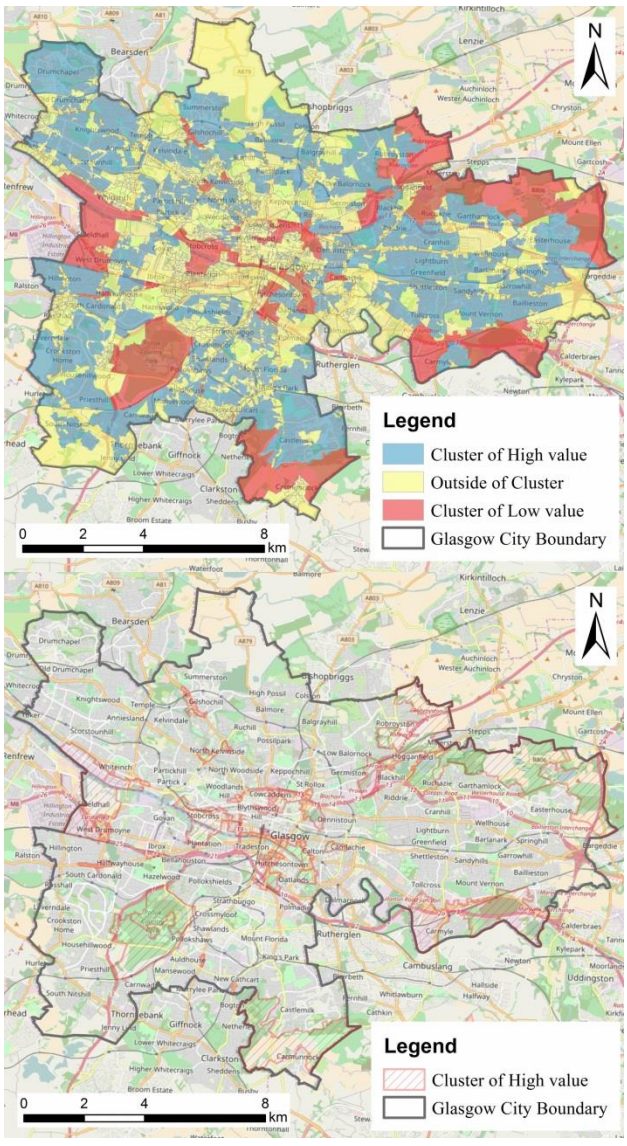


Figure 5. Clusters of high RCADTP (Basemap: OpenStreetMap, licensed under the Open Database License).

## 4. CONCLUSIONS

This study demonstrates usefulness of Strava Metro data in depicting cycling behaviour over a city. The representativeness of Strava cyclists are potentially biased in age and probably income or education. In the future, we will take account of some aspects to enhance this study. First, the effect of potential biased issues on the fitness of use for Strava Metro data needs to be investigated. Second, 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). 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.

## ACKNOWLEDGEMENTS

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

Broach, J., Dill, J., Gliebe, J., 2012. Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, 46 (10), pp. 1730–1740.

Casello, J.M., Usyukov, V., 2014. Modeling cyclists' route choice based on GPS data. *Transportation Research Record*, 2430, pp. 155–161.

Dill, J., 2009. Bicycling for transportation and health: the role of infrastructure. *Journal of Public Health Policy*, 30(Suppl 1), pp. S95-110.

Duncan, MJ., Badland, H.M., Mummery, W.K., 2009. Applying GPS to enhance understanding of transport-related physical activity. *Journal of Science and Medicine in Sport*, 12(5), pp. 549-56.

Duque, J.C., Aldstadt, J., Velasquez, E., Franco, J.L., Betancourt, A., 2011. A computationally efficient method for delineating irregularly shaped spatial clusters. *Journal of Geographical Systems*, 13, pp. 355–372.

Forsyth, A., Oakes, J.M., 2015. Cycling, the Built Environment, and Health: Results of a Midwestern Study. *International Journal of Sustainable Transportation*, 9(1), pp. 49-58.

Griffin, G.P., Jiao, J., 2015. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport & Health*, 2 (2), pp. 238–247.

Herrero, J., 2016. Using big data to understand trail use: three Strava tools. TRAFx Research. https://www.trafx.net/insights.htm

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, Australia. *Journal of Transport & Health*, 3(3), pp. 366–376.

Hood, J., Sall, E., Charlton, B., 2011. A GPS-based bicycle route choice model for San Francisco, California. *Transportation Letters*, 3, pp. 63–75.

Jesticoa, B., Nelsona, T., Wintersb, M., 2016. Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, 52, pp. 90–97.

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. *Journal of Science and Medicine in Sport*, 21(4), pp. 496–509.

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, pp. 127-33.

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. *American Journal of Public Health*, 100(10), pp. 1986-92.

RiSE group. ClusterPy: Library of spatially constrained clustering algorithms. http://www.rise-group.org/risem/clusterpy.

Scotland's Census, 2016. Daytime population statistics. http://www.scotlandscensus.gov.uk/ods-web/data-warehouse.html#additionaltab

Strava Metro., 2015. Strava Metro Comprehensive User Guide Version 2.0. 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 http://metro.strava.com/

Sun, Y., Mobasheri, A., 2007. Utilizing Crowdsourced Data for Studies of Cycling and Air Pollution Exposure: A Case Study Using Strava Data. *International Journal of Environmental Research and Public Health*, 14(3), pp. 274.

Urban Big Data Centre, 2016. Data services: Strava Metro data. Urban Big Data Centre, Glasgow, UK.

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. *Preventive Medicine*, 46(1), pp. 29–32.