Using crowdsourced data to map bicycling behaviour

Vanessa Brum-Bastos*[1], Colin J. Ferster[2] and Trisalyn Nelson[1]

[1] School of Geographical Sciences and Urban Planning, Arizona State University - USA
[2] Department of Geography, University of Victoria - Canada
*Email: vanessa.bastos@asu.edu

Abstract

Growing interest in data for bicycling planning has many jurisdictions aiming to develop a sampling strategy for measuring bicycle counts. However, it can be difficult to determine where bicycling counters should be placed to generate a representative sample of bicycling, and bicycle counts tend to only occur on heavily used bicycle routes. Our goal, is to utilize crowdsourced data on bike ridership, from Strava Metro, and continuous signal processing data mining techniques to map regions of bicycling behaviour in San Francisco – CA, US. Bicycling behaviour regions can be used to stratify bicycle count sampling. Our results indicate it is possible to differentiate bicycling behaviour from Strava and we mapped routes categorized into six unique ridership behaviours including commute to work, commute to school, leisure, 8 am peak , 11 am peak and least utilised. We recommend sampling ridership from each of the categories when developing count programs.

Keywords: Strava, mobility, transportation, urban planning, smart cities

1. Introduction

Planning bicycling transport options is challenging because of the lack of information on bicycling behaviour (Gosse and Clarens, 2014). Data on bicycling ridership, used for transportation planning, is typically obtained via manual counts and automatic counters (Butler, Skinner and Perl, 2013; Nordback *et al.*, 2013). Automatic counters are usually installed at a single location and measure ridership with high temporal resolution, whilst manual counts typically include more locations but are measured for a short period of time. When planning bicycle counts it is problematic to determine where and when counts should occur. A systematic approach to stratifying the city to enable representative sampling of bicycle ridership would help planners ensure data is representative and collection efficient.

The advent of crowdsourced data on bike ridership represents an opportunity to develop methodologies to capture the spatial-temporal patterns in bicycling behaviour (Jestico, Nelson and Winters, 2016). Crowdsource bicycling data has the benefit of being high resolution in both space and time. Most research using crowdsourced bike data have focused on counts in daily or seasonal time steps (Romanillos *et al.*, 2016), yet there is potential to analyse continuous time series signals to make more detailed characterizations of bicycling behaviour. Space-time behaviour patterns can be used to partition bicycling routes and streets into categories. By obtaining bicycle counts in each category a more representative sample of bicycling can be obtained. In this work, we used crowdsourced data

on bike ridership and data mining to map the spatial-temporal behaviour of cyclists in 2017 in San Francisco city - California, USA.

2. Methods

2.1 Crowdsourced data from Strava

Strava Metro provided bicyclist count data from the 1st of January 2017 to the 31st of December 2017 at one-minute temporal resolution along each street segment in San Francisco city - California, USA. Data includes 102,552 bicyclists of which 18.03 % were female, 77.30% were male and 4.67% did not declare their gender.

2.2 Extracting cycling temporal profiles

We aggregated the data by computing total hourly counts for each day and then calculating the mean for each hour of the day for all street segments for weekdays and weekends separately. The aggregation produced 18,027 curves representing the temporal profile of bicycling behaviour for each street segment for a given hour of the day.

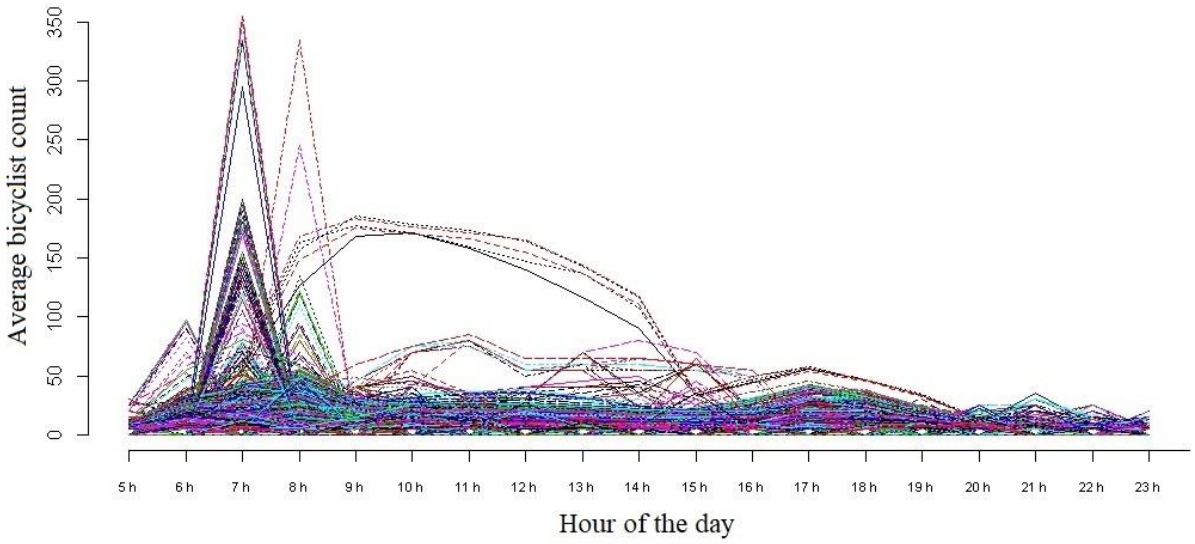

Figure 1: Mean hourly bicyclist count for weekdays in 2017 for each street segment in San Francisco city.

2.3 Calculating Dynamic Time Warping (DTW) distances

We used dynamic time warping (DTW) distance to assess the pairwise similarity between the temporal profiles of all street segments. DTW is one of the most used metrics of similarity between two time series (Zhang *et al.*, 2017) ; it finds the optimal global alignment between two time series by exploiting temporal distortions between them (Sakoe and Chiba, 1978).

The best alignment is found by first applying Equation 1 to compute a pairwise distance matrix (M) for all points in the two time series (Figure 2).

$$D = |A_i - B_j| + min\begin{Bmatrix} D\,[i-1, j-1] \\ D\,[i-1, j] \\ D\,[i, j-1] \end{Bmatrix}$$

Equation 1

Here A_i is the mean cyclist count for street segment A at hour i , B_j is the mean cyclists count for street segment B at hour j, $D\,[i-1, j-1]$ is the previously computed distance between mean cyclist counts in the previous hour for both time series, $D\,[i-1, j]$ is the previously computed distance between mean cyclist counts at the current hour for series B and the previous hour for series A, and $D\,[i, j-1]$ is the previously computed distance between mean cyclist counts at the current hour for series A and the previous hour for series B.

The DTW distance between the two street segments is calculated by finding the least cost path in matrix M (Figure 2). In order to avoid unrealistic least cost paths (Zhang *et al.*, 2017) we used a Sakoe-Chiba band (Sakoe and Chiba, 1978) (See yellow polygon in Figure 2) to constraint the warping to a one-hour interval before or after the hour being warped.

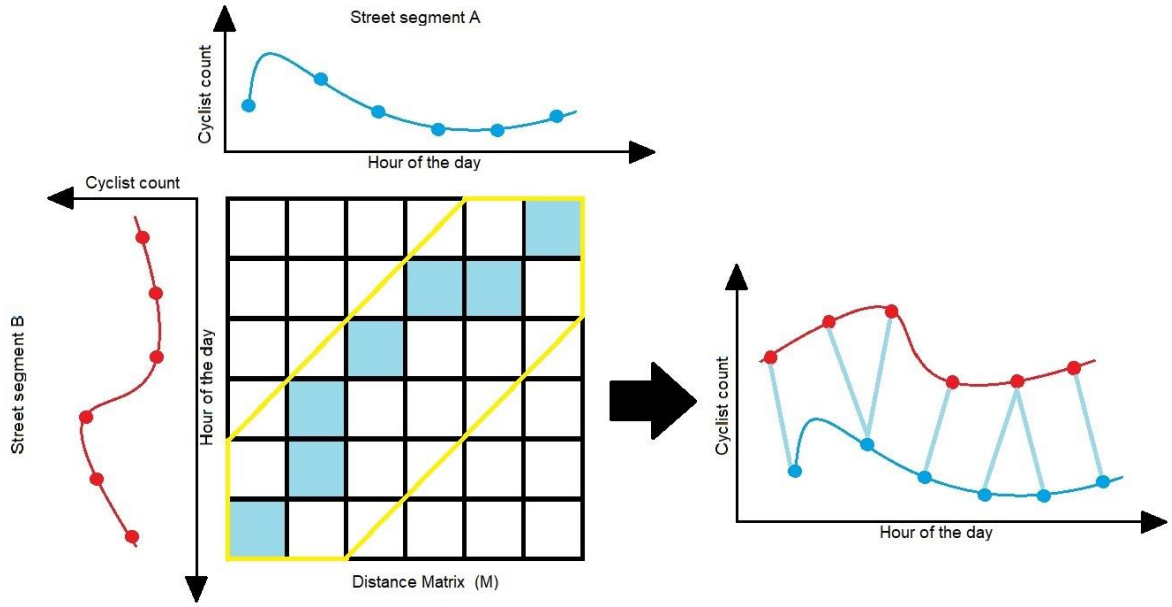

Figure 2 – Distance matrix (M) for the hourly mean cyclist count of street segments A and B. The optimal warping is highlighted in bright blue in the matrix (M) and in the graphical representation on the right. The yellow polygon represents the warping constraint delineated by the one-hour Sakoe-Chiba band.

We calculated the pairwise DTW distance for 18.027 street segments, which generated an 18.027 X 18.027 DTW similarity matrix (W) accounting for the differences amongst all curves.

2.4 Identifying types of bicycling temporal profiles

The dissimilarity matrix (W) can be used to find whether street segments were showing similar cyclist use pattern during the weekdays. For this, we applied Ward's clustering, a hierarchical bottom-up

algorithm that computes dissimilarities between two groups as the increase in the error sum of squares after merging those groups. The algorithm starts with each street segments as their own group and successively merges them into clusters based on the minimum increase in the error sum of squares, until it becomes a single cluster (Murtagh and Legendre, 2011). For selecting the optimal number of clusters, we used the Calinski-Harabaz Index (CHI) (Calinski and Harabasz, 1974) that considers the within and between groups dispersion as shown in Equation 2.

$$CHI = \frac{trace\ (B)}{trace\ (H)}$$

Equation 2

here H and B are the within and between group dispersion matrices, the trace of H is the sum of the within cluster variance and the trace of B is the sum of the between cluster variances; a higher CHI indicates a better data partition (Ahmed, 2012) because it shows that the within group distances are lower and the between groups distances are higher.

We varied the number of clusters from the number of street segments (i.e. the maximum possible number of clusters, if every street segment is allocated to its own cluster) to one and used the configuration with highest CHI. We plotted the profiles within each cluster and analysed their main characteristics so that we could meaningfully label them before mapping. Finally, we interpreted the bicycling behaviour in each cluster and assigned labels that are interpretable by city planners.

3. Results and discussion

3.1 Identifying types of bicycling temporal profiles

The optimal partitioning of the street segments based on their temporal profile was achieved with six clusters, which were labelled according to the main characteristics we expected for each bicycling behaviour (Table 1). In Figure 3, we show the temporal profiles for all street segments in the "Commute" and "School" clusters, which are the most relevant for transportation planners.

Cycling behaviour	Number of peaks	Approximate peak time	Slope
Commute	2	8 am | 6 pm	Dramatic rises
School	2	8 am | 4 pm	Dramatic rises
Leisure	1	1 pm	Gradual rise

Table 1: Main characteristics used link cycling behaviours to clusters.

The commute cluster (n = 1109) was labelled based on its peaks around eight in the morning and six in the afternoon, which indicates that these street segments are being used by bicyclists on their way to work and back. Most of these street segments are located near picturesque parts of the city, such as green areas and shorelines (Figure 3). This finding supports the argument that scenery plays a role in route choices for cyclists, for not only leisure bicycling but also commute.

The school cluster (n=2532) was named based on its peaks around eight in the morning and three in the afternoon, which indicates that these street segments are being used by bicyclists on their way to

school and back. Most of these street segments are located near educational institutions and some of them incorporate green areas.

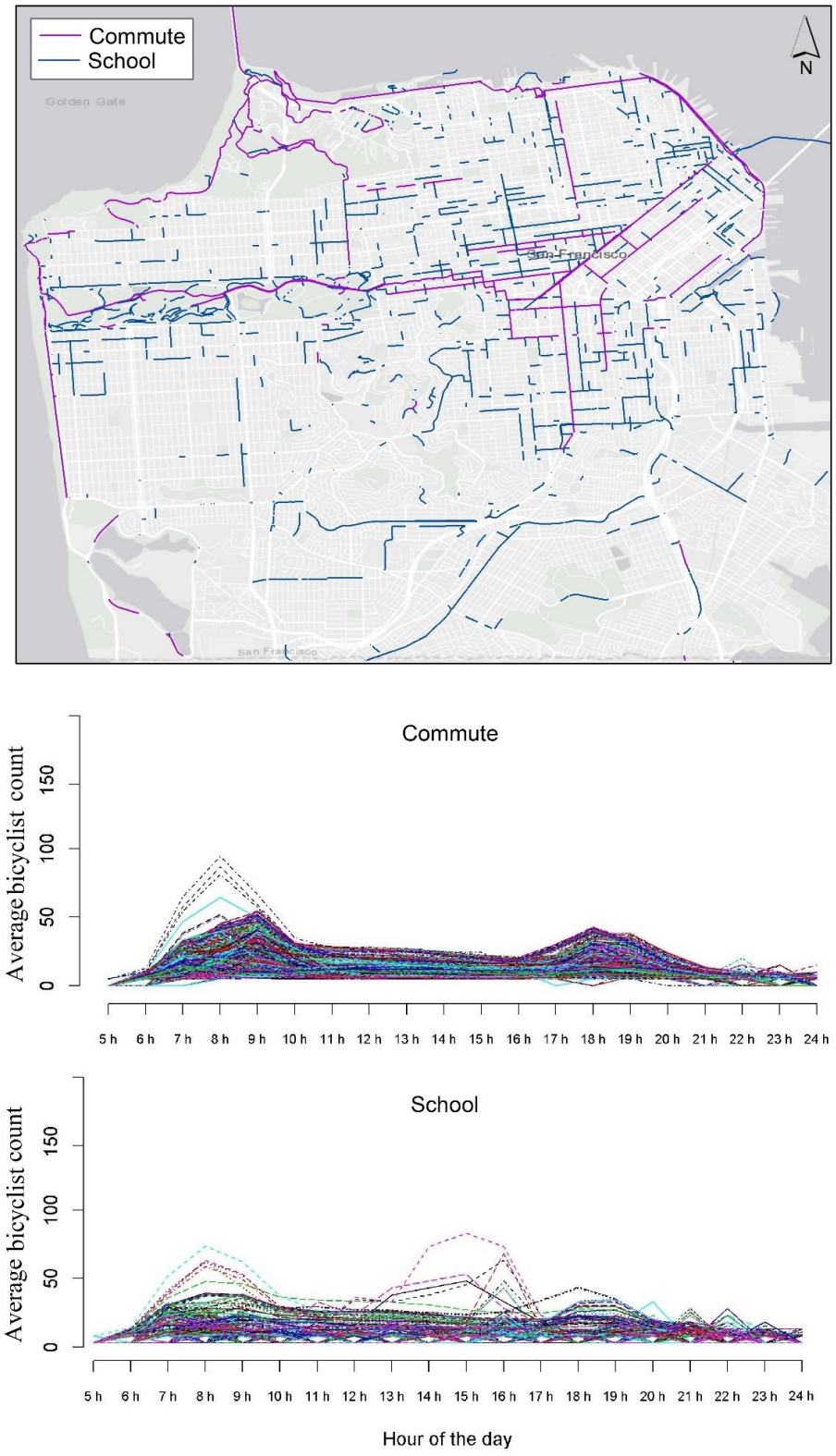

Figure 3 – San Francisco base map and street segments classified as “Commute” and “School” with respective temporal profiles displayed by graphs below the map.

4. Conclusions

The use of Strava Metro data on bike ridership and data mining of continuous temporal signals allowed us to differentiate bicycling behaviour profiles in San Francisco – CA, US. The use of maps to analyse the spatial distribution and label these clusters can help transportation planners to find the areas of similar bicycling patterns, which can help target interventions and stratify count programs. While crowdsourced fitness data over represents competitive cyclists, whose behaviour is not representative of all bicyclists (Barratt, 2017), by processing the continuous temporal patter in data we identified locations where commute patterns dominated. For example, we identify signals that represented bicycling to school and bicycling to work (identified based on the timing of peaks of activity). The analysis of two core categories, "Commute" and "School", highlighted the functional importance of the Golden Gate Park for connecting downtown and residential areas for both types of bicycling behaviour. Measuring and understanding the behaviour of bicyclists can improve the chances of success of strategies implemented in a transportation plan.

2. Acknowledgements

The authors would like to acknowledge the Strava team led by Rodrigo Davis for providing the Strava ridership data for San Francisco city.

3. References

Ahmed, K. I. (2012) *Acoustic data optimisation for seabed mapping with visual and computational data mining*. National University of Ireland, Maynooth.

Barratt, P. (2017) 'Healthy competition: A qualitative study investigating persuasive technologies and the gamification of cycling', *Health and Place*. Pergamon, 46, pp. 328–336. doi: 10.1016/j.healthplace.2016.09.009.

Berkeley, U. C. *et al.* (2014) *Guidebook on Pedestrian and Bicycle Volume Data Collection. National Cooperative Highway Research Program (NCHRP) Report 797*.

Butler, D. H., Skinner, R. E. and Perl, A. D. (2013) *Monitoring bicyclist and pedestrian travel and behavior, current research and practice. Transportation Research Circular Number E-C183.* Available at: www.TRB.org (Accessed: 23 May 2019).

Calinski, T. and Harabasz, J. (1974) 'A dendrite method for cluster analysis', *Communications in Statistics - Theory and Methods*, 3(1), pp. 1–27. doi: 10.1080/03610927408827101.

EcoCounter (2017) *Case Study: Using Permanent and Temporary Counters in Quebec City - Eco-Counter*, *EcoCounter : Le Blog*. Available at: https://www.eco-compteur.com/en/blog/permanent-temporary-counters-quebec-city/ (Accessed: 23 May 2019).

Gosse, C. A. and Clarens, A. (2014) 'Estimating Spatially and Temporally Continuous Bicycle Volumes by Using Sparse Data', *Transportation Research Record: Journal of the Transportation Research Board*, 2443(1), pp. 115–122. doi: 10.3141/2443-13.

Jestico, B., Nelson, T. and Winters, M. (2016) 'Mapping ridership using crowdsourced cycling data', *Journal of Transport Geography*, 52, pp. 90–97. doi: 10.1016/j.jtrangeo.2016.03.006.

Murtagh, F. and Legendre, P. (2011) 'Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm'. doi: 10.1007/s00357-014-9161-z.

Nordback, K. *et al.* (2013) 'Estimating Annual Average Daily Bicyclists', *Transportation Research Record: Journal of the Transportation Research Board*, 2339(1), pp. 90–97. doi: 10.3141/2339-10.

Romanillos, G. *et al.* (2016) 'Big Data and Cycling', *Transport Reviews*. Routledge, 36(1), pp. 114–133. doi: 10.1080/01441647.2015.1084067.

Sakoe, H. and Chiba, S. (1978) 'Dynamic Programming Algorithm Optimization for Spoken Word Recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), pp. 43–49. doi: 10.1109/TASSP.1978.1163055.

Zhang, Z. *et al.* (2017) 'Dynamic Time Warping under limited warping path length', *Information Sciences*. Elsevier, 393, pp. 91–107. doi: 10.1016/J.INS.2017.02.018.