Supplementary Information:

Insights from self-organizing maps for predicting accessibility demand for healthcare infrastructure

1. Methodology

1.1. Assigning census data to hexagonal grid cells

In brief, the centroid of each cell was intersected with the DA polygons to determine spatial collocation, and the DA census data were assigned to their appropriate hexagonal cells. DA polygons were typically larger than the hexagonal cells in our case, so census data had to be shared or split (depending on data type) across multiple hexagonal cells. For averaged or proportional data (e.g. income, age), we assigned the same value to all cells from the same DA. For discrete data (e.g. population counts), we divided the value equally among constituent cells. This downscaling method is simple to implement and does not make assumptions about how populations are distributed according to land use. We acknowledge that heterogeneous population downscaling – which allows for more realistic spatial allocation of facilities and populations – could be achieved by integrating different scenarios with models of urban land use (e.g. Meiyappan et al., 2014), but this is beyond the scope of our study.

1.2. Hospitals and walk-in clinics

We define hospitals as medical institutions providing diagnostic and treatment services for people whose illnesses/injuries require bed occupation for at least one night, and walk-in clinics as providing treatment services for people with minor illnesses/injuries that do not require a visit to a hospital emergency department or urgent care facility. Apart from the two Surrey hospitals examined in this study, two other hospitals (in New Westminster to the northwest and Langley to the southeast) are potentially accessible to some of Surrey's residents. However, we ignored these in the present study because they fall outside the city's jurisdiction and investment strategy. We ran our catchment area analysis to calculate how many Surrey residents could reach each hospital and clinic in New Westminster and Langley within 30 minutes using public transport. In all cases, these external healthcare facilities could be reached by < 4% of Surrey's population, so the marginal increase in theoretical accessibility to healthcare for residents near Surrey's boundaries should not significantly impact our findings.

1.3. Defining an upper bound for income distributions

Income data are reported categorically by Statistics Canada, into 16 income groups defined by upper and lower income bounds. The highest income level ('$250,000 and above') is unbounded on one side. For the purposes of converting categorical census data to continuous data, a single value characterising the open-ended income category was estimated using Pareto's Law of Income Distribution (Parker & Fenwick, 1983). The Pareto Curve can be linearly represented as:

$$log(Z) = log(A) - vlog(X)$$

Equation 1

where Z is the number of units with incomes over a certain amount, X is the amount of income, and A and v are equivalent, respectively, to the intercept and the unstandardized regression coefficient, and are parameters to be solved for. The Pareto Curve has been shown to be linear only at the upper tail of an income distribution, so we calculated v as (Henson, 1967):

$$v = \frac{log(H_L + H_{L-1}) - log(H_L)}{log(W_L) - log(W_{L-1})}$$

Equation 2

where H_L is the number of households in the open-ended category, H_{L-1} is the number of households in the category immediately preceding the open-ended one, W_L is the lower limit of the open-ended category and W_{L-1} is the lower limit of the category immediately preceding the open-ended one. Using this estimate of v, the mean income for the open-ended category (y_L) was obtained by (Henson, 1967):

$$y_L = W_L\left(\frac{v}{v-1}\right)$$

Equation 3

Using income data combined from all DAs in 2016, the value of y_L (i.e. the midpoint of the last, open-ended income category) was calculated as $266,950.

1.4. Accessibility analysis

We calculated travel-time estimates (for an optimal combination of public transportation and walking) between every pair of grid cells – an 'origin' and a 'destination' (O-D) – using OpenTripPlanner (OTP, 2017). To account for fluctuations in service availability throughout the day, we averaged travel-time estimates across 36 O-D matrices for a typical working day (Tuesday, 19th September 2017), with departures every 20 minutes between 7am and 7pm. The OTP routing engine used in this study does not account for traffic congestion levels, which can slow travel times; instead, we consider accessibility via public transportation based on current service timetabling in the GTFS dataset. Higher resolution estimates of travel-time can be acquired from vehicle GPS data (e.g. Wessel et al., 2017), although this is beyond the scope of our paper.

We applied a modified version of the isochronic or cumulative-opportunity measure (Wachs & Kumagai, 1973; Pereira, 2018) to estimate the number of residents who could theoretically access healthcare facilities. Accessibility was evaluated only from the perspective of the origin (i.e. the 'active' accessibility of a population towards a facility) rather than the destination (i.e. the 'passive' accessibility of the facilities with respect to the population) (Papa & Coppola, 2012). Active accessibility for each origin grid cell (for a total n grid cells) was calculated as:

$$F_{o,T} = \sum_{d=1}^{n} F_d f(t_{odr})$$

$$f(t_{odr}) = \begin{cases} 1 if\, t_{odr} \leq T \\ 0 if\, t_{odr} > T \end{cases}$$

Equation 4

where $F_{o,T}$ is the number of facilities F that can be reached from origin o within time threshold T, F_d is the number of facilities in destination cell d, and $f(t_{odr})$ is a time threshold function whose value (either zero or one) depends on whether travel-time t_{odr} is greater or smaller than time threshold T.

There is much discussion as to the appropriate time thresholds that should be used for cumulative opportunity measures (Neutens, 2015). Acceptable travel-times are known to vary according to travel mode, as well as demographic, socioeconomic and lifestyle factors (Guagliardo, 2004; Milakis et al., 2015). We set our travel-time threshold T to 30 minutes, based on sensitivity testing reported in Mayaud et al. (2018). Whilst results were somewhat sensitive to the choice of threshold, and the appropriateness of different thresholds has been much discussed in this field, the majority of metropolitan transport plans use time thresholds of 30–40 minutes when considering accessibility to hospitals via public transit (Boisjoly & El-Geneidy, 2017).

Limitations of the cumulative opportunity measure include the fact that it does not account for the size (or 'attractiveness') of the destination, nor the impedance (or 'friction') of travel time, cost and effort beyond the threshold variable. Numerous other accessibility measures exist in the literature, including gravity-based (Hansen, 1959) and place rank measures (El-Geneidy & Levinson, 2011), some of which address these limitations. Nevertheless, in comparison with these other methods, the cumulative-opportunity measure makes few assumptions about user behaviour and preference, and is most easily interpreted (Geurs & van Wee, 2004; Neutens et al., 2010).

Incorporating the explicit cost of travel is also not an objective of our study. The public transportation system in Metro Vancouver is run by a single operator, Translink. Surrey lies within a single 'travel zone' as defined by Translink, the regional transit operator meaning that all transit journeys starting and finishing in Surrey cost the same price, including any transfers made within a 90-minute window from the start of the first journey (Translink, 2018). Since our catchment analyses are based on travel-time cut-offs of 30 minutes, the price of using public transport is essentially fixed for an individual user. However, affordability of public transport has been shown to influence individuals' usage (El-Geneidy et al., 2016), so this should be more carefully considered in future studies.

2. Results

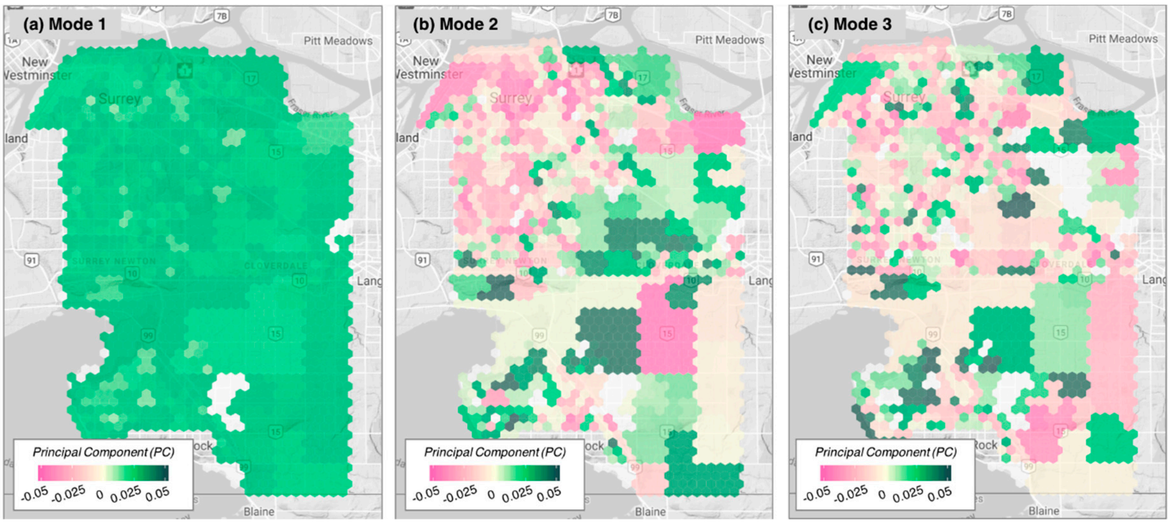

Figure S1 Spatial distribution of principal components (PCs, i.e. weightings) associated with each of the three main modes of the 2016 dataset.

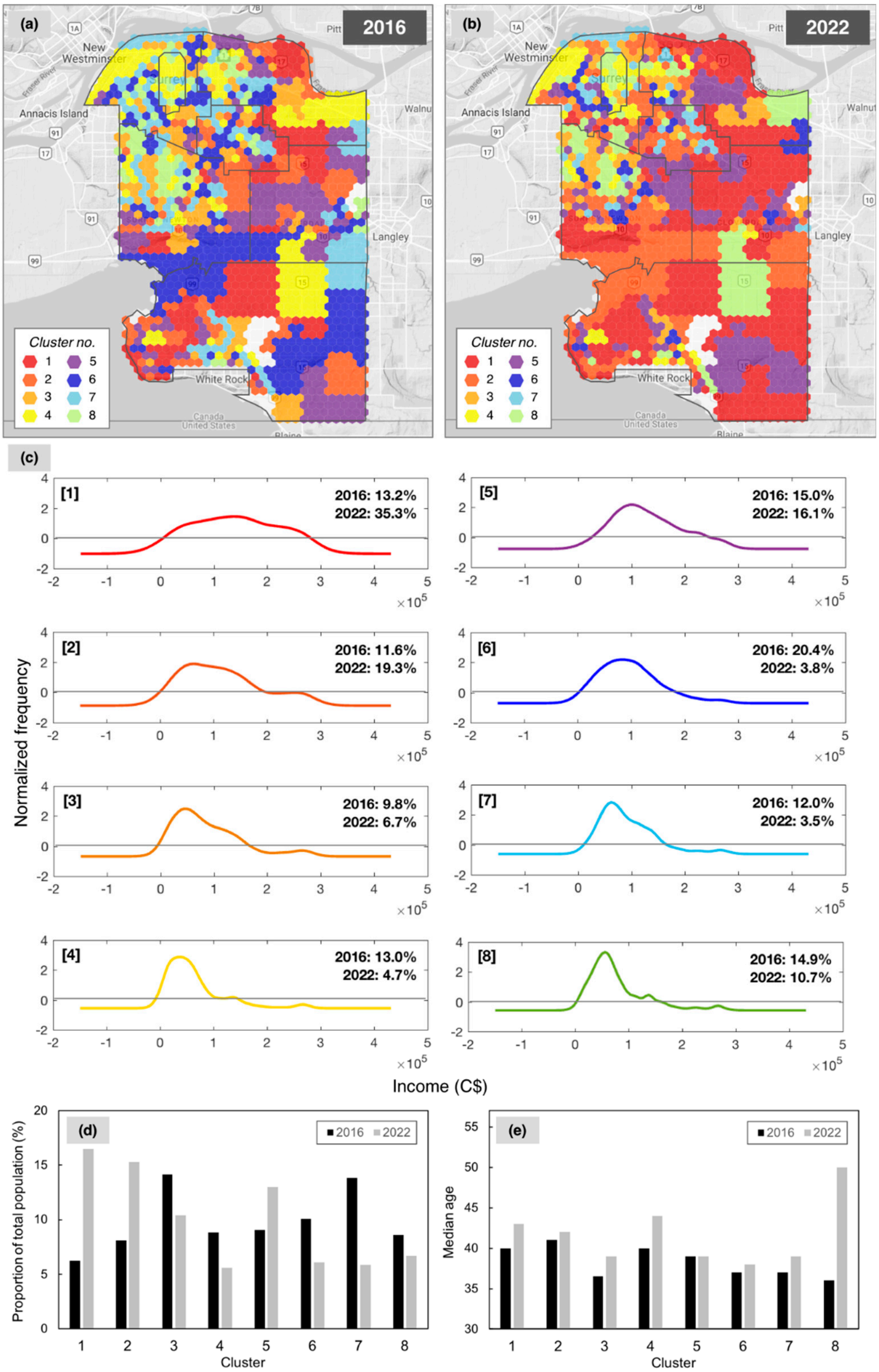

Figure S2 Spatially mapped SOM topology, coloured according to clustering, for (a) 2016, and (b) 2022. The 2016 data were used to train the SOM algorithm, and this was used to classify both the 2016 and 2022 data. White cells show no data; (c) Frequency distributions for each SOM cluster, showing in bold the frequency that they occur in the 2016 and 2022 maps. The y-axis on these plots varies about zero because the inputs are demeaned and normalised using their standard deviations; (d) Proportion of total city population belonging to each cluster, for 2016 and 2022; (e) Median age of population belonging to each cluster, for 2016 and 2022.

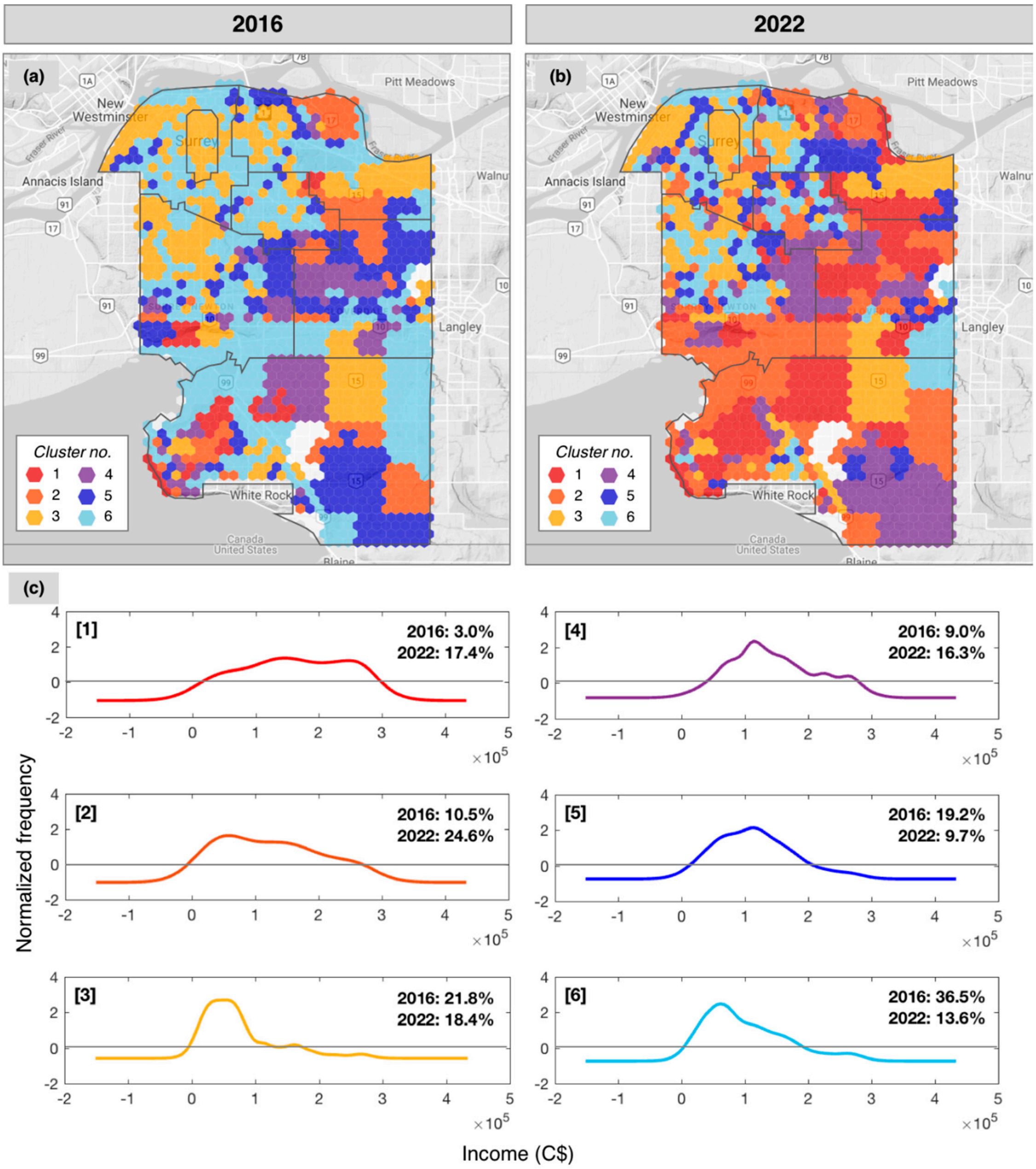

Figure S3 Spatially mapped SOM topology, coloured according to clustering, for (a) 2016, and (b) 2022. The 2022 data were used to train the SOM algorithm, and this was used to classify both the 2016 and 2022 data. White cells show no data; (c) Frequency distributions for each SOM cluster, showing in bold the frequency that they occur in the 2016 and 2022 maps. The y-axis on these plots varies about zero because the inputs are demeaned and normalised using their standard deviations.

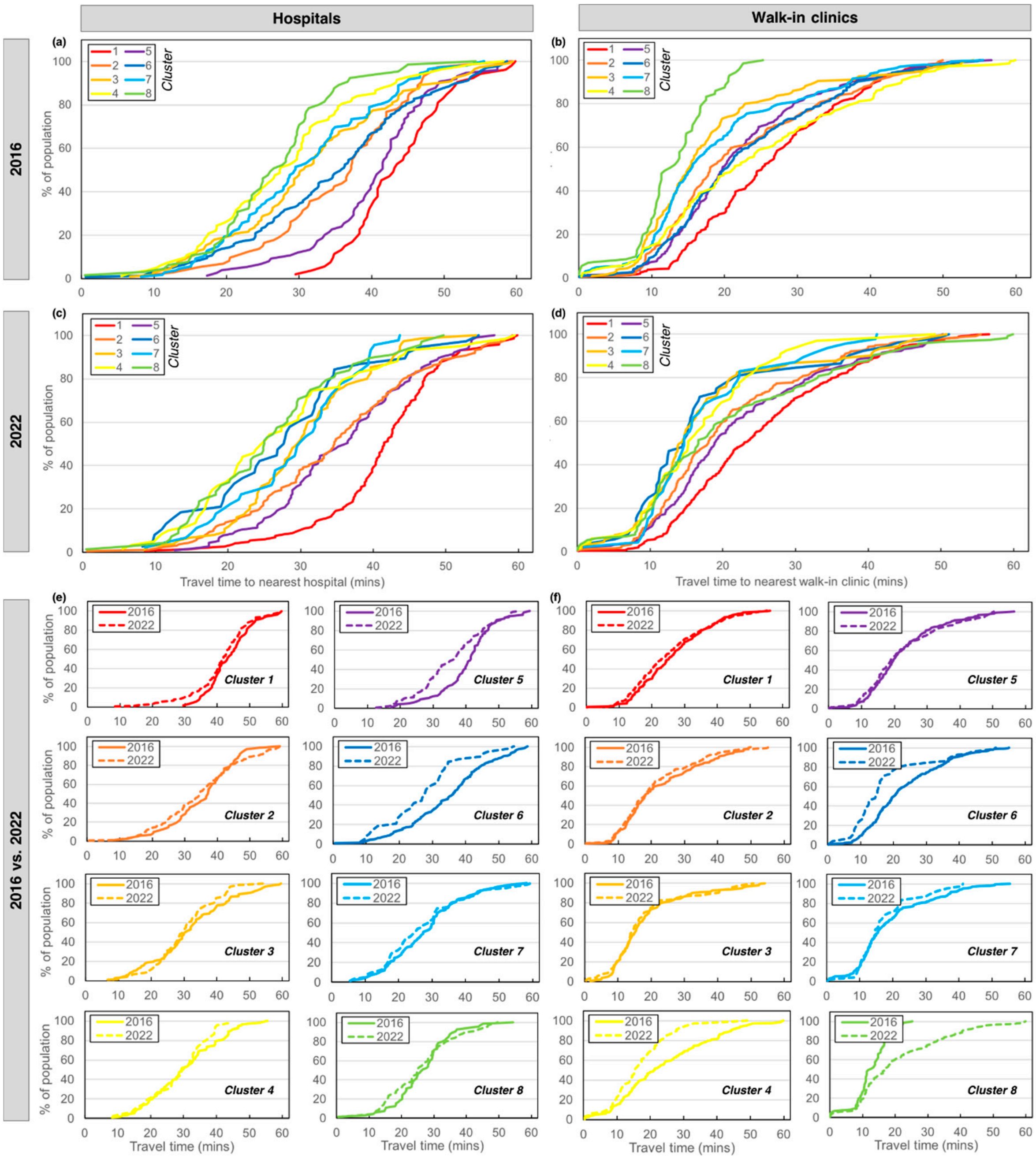

Figure S4 Cumulative frequency distributions of travel-time to closest facility for each cell, grouped by cluster for both 2016 and 2022. Plots show distributions for: all clusters to hospitals in (a) 2016 and (c) 2022; all clusters to walk-in clinics in (b) 2016 and (d) 2022; individual clusters comparing 2016 and 2022 for (e) hospitals and (f) walk-in clinics.

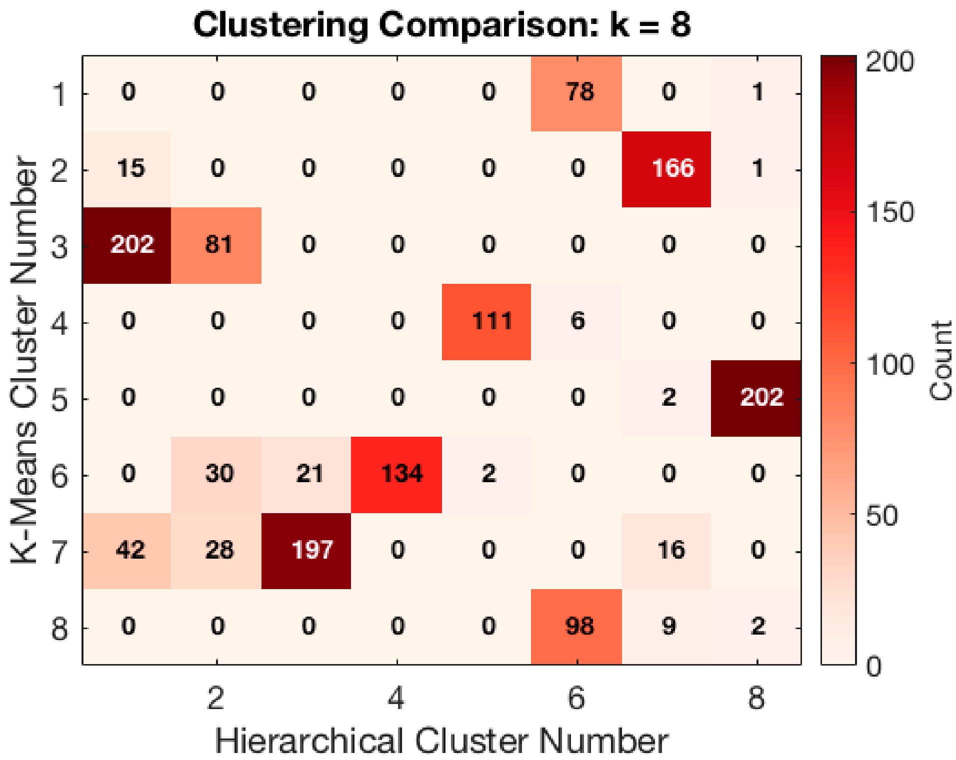

Figure S5 Comparison of cluster results when using k-means clustering and hierarchical clustering.

3. Supplementary References

El-Geneidy A and Levinson D (2011) Place rank: Valuing spatial interactions. *Networks and Spatial Economics* 11(4): 643–659.

El-Geneidy A, Levinson D, Diab E, Boisjoly G, Verbich D and Loong C (2016) The cost of equity: Assessing transit accessibility and social disparity using total travel cost. *Transportation Research Part A: Policy and Practice* 91: 302–316.

Guagliardo MF (2004) Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics* 3(3): 1–13.

Hansen WG (1959) How accessibility shapes land use. *Journal of the American Institute of Planners* 25, 73–76.

Henson MF (1967) Trends in Income of Families and Persons in the United States: 1947 to 1960. *U.S. Bureau of Census, Technical Paper No. 17, Washington, USA.*

Milakis D, Cervero R, Van Wee B and Maat K (2015) Do people consider an acceptable travel time? Evidence from Berkeley, CA. *Journal of Transport Geography,* 44: 76–86.

Meiyappan P, Dalton M, O'Neill BC and Jain AK (2014) Spatial modeling of agricultural land use change at global scale. *Ecological Modelling,* 291: 152–174.

Milakis D, Cervero R, Van Wee B and Maat K (2015) Do people consider an acceptable travel time? Evidence from Berkeley, CA. *Journal of Transport Geography,* 44: 76–86.

Neutens T (2015) Accessibility, equity and health care: review and research directions for transport geographers. *Journal of Transport Geography*, 43: 14–27.

Neutens T, Schwanen T, Witlox F and De Maeyer P (2010) Equity of urban service delivery: A comparison of different accessibility measures. *Environment and Planning A*, 42(7): 1613–1635.

Papa E and Coppola P (2012) Gravity-based accessibility measures for integrated land-use planning (GraBAM). In: Hull, A., Silva, C. & Bertolini, L. (eds). *Accessibility Instruments for Planning Practice.* COST Office, 117–124.

Parker RN and Fenwick R (1983) The Pareto Curve and its utility for open-ended income distributions in survey research. *Social Forces* 61(3): 872–885.

Translink (2018) Avaialble at: https://m.translink.ca/Fares-and-Passes/ *[accessed January 2018]*

Wessel N, Allen J and Farber S (2017) Constructing a routable retrospective transit timetable from a real-time vehicle location feed and GTFS. *Journal of Transport Geography* 62: 92–97.