Supplementary Information:

A Job, Indeed!

Accessibility Equity to Advertised Employment in Cascadia

1. METHODS

1.1. Data Sources

1.1.1. Extracting Job Information From Indeed.com

While Indeed.com operates an open-source API for access to job information, this service is restricted to publishers with typically large viewership who receive approval through the Indeed Publisher Program (Indeed Engineering, 2019). We therefore extracted the latitude, longitude, and salary for each relevant job posting using a combination of manual extraction and open-source geocoding services.

We began by manually obtaining the Indeed URL that filters for full-time positions within a specific salary range in a specific city and state/province. For example:

https://indeed.com/jobs?q=${salary_min}+-
+${salary_max}&l={city},+{state}&jt=fulltime&radius=50&start={page}

This returns an HTML page containing 10–15 job postings and an index of the total number of jobs within the selected salary window. For each posting on the page, the job title, company name, and salary (if available, as not all postings report salaries) were extracted. As salaries are typically reported as “US$15 an hour” or “US$40K–50K a year,” they were extrapolated into a per annum salary. In calculating the per annum salary, we assumed that there are 261 working days in a year and 8 work hours in a day. Where salary ranges were present, the average per annum salary was calculated. The next page was visited until the number of jobs extracted equaled the index of the total jobs available. Once this limit was reached, we incremented the salary range by $500 (in the relevant currency), created a new URL with an updated salary window, and repeated the process. Careful attention was paid to ensure duplicate jobs were not counted.

As our routing method requires that jobs be spatially located, latitude, and longitude coordinates must be assigned to each posting. This information is not present on Indeed.com,

so we used nested techniques to infer coordinates from job postings. OpenStreetMap's (2019) Nominatim API service was used to extract a street address from each company listing. The API query for a specific address on Nominatim resembles:

https://nominatim.openstreetmap.org/search.php?q={address}&format=json&polygon=1&addressdetails=1

To determine the exact latitude and longitude coordinates for the given address, two different methods were used depending on the city. For Vancouver-based companies, latitude and longitude information was extracted from the website Canada411.com using the OpenLookup library from OpenStreetMap (2019). The HTML query for a specific address on Canada411.com resembles:

https://canada411.yellowpages.ca/search/si/1/{company_name}/vancouver%20bc

For companies based in the United States (i.e., Portland and Seattle), the Find Place service from the Google Places API (Google, 2019) was used to query place information from a text input (this can be a name, address, or phone number).

There is no limit to the number of jobs that can be geocoded using OpenLookup, although there is a computational constraint since OpenStreetMap's requirements limit usage to a maximum of one request per second. In contrast, the Google API only permits a maximum of 11,000 free queries per month. The queries were therefore distributed evenly between Portland and Seattle, such that a maximum of 5,500 job postings per city were extracted. For all three cities, the extracted data were processed into tables containing the company, job title, latitude/longitude coordinates, parsed salary, and relevant salary range.

1.1.2. Census Data

To reflect broader moves to make urban analytics research more transparent and transferable, this study almost exclusively makes use of open data sources and open-source code. For all three cities, we obtained current-year estimates of socioeconomic and demographic data for 2018 at the smallest standard geographic unit for which all census data were available: the Census Block Group (CBG) level in the United States (typically composed of 600–3,000 residents), and the Dissemination Area (DA) level in Canada (typically composed of 400–700 residents). The census data in each city's boundaries were spatially reorganized into a grid of

equally sized hexagons (500 m diameter, or 0.16 km^2). Hexagons were used to reduce sampling bias from edge effects and to permit spatial connectivity between data points (Birch et al., 2007; Pereira, 2018). A well-established methodology described by Mayaud et al. (2018) was used to assign the appropriate census data to each hexagonal cell.

1.1.3. Transportation Network

The spatial layout of road networks and pedestrian infrastructure was determined using OpenStreetMap (2019), and this formed the basis for performing the door-to-door routing algorithm in the catchment area analysis (Figure S1). Geolocated timetable data of public transit stops and routes were acquired in General Transit Feed Specification (GTFS) format for September 2017 from the three regional transit operators (TriMet in Portland, King County Metro in Seattle, and Translink in Vancouver). The static GTFS data used here do not account for traffic congestion.

1.2. Data Analysis

1.2.1. Assigning Census Data to Hexagonal Grid Cells

In brief, the centroid of each cell was intersected with the Census Block Group (CBG) or Dissemination Area (DA) polygons to determine spatial collocation, and the census data were assigned to their appropriate hexagonal cells. CBG/DA polygons were typically larger than the hexagonal cells in our case (i.e., one or more hexagons corresponded to each census polygon). In these cases, 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, transport expenditure), we assigned the same value to all cells from the same CBG/DA. For discrete data (e.g., population counts), we divided the value equally among constituent cells. In cases where a hexagon was larger than the underlying CBGs/DAs (i.e., a hexagon was comprised of two or more census polygons), discrete data were simply summed, while averaged or proportional data were combined as population-weighted means. In the case of household income, we calculated the median household income of combined polygons as the midpoint of the income category in which the 50th percentile of households was located.

The downscaling method we describe is simple easy 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.

1.2.2. Defining an Upper Bound for Income Distributions

Income data are reported categorically into ten income groups by the U.S. Census Bureau and into sixteen income groups by Statistics Canada. These income groups are defined by upper- and lower-income bounds, with the highest group (US$200,000 and C$250,000) being unbounded on one side. For the purposes of distribution fitting in our SOM analysis, a single value characterizing the open-ended income category was estimated for each city using Pareto's Law of Income Distribution (see Parker & Fenwick, 1983). The Pareto Curve can be linearly represented as:

$$\log(Z) = \log(A) - v\log(X)$$

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 census units for each city, the value of y_L was calculated as US$204,930 for Portland, US$205,110 for Seattle, and C$256,120 for Vancouver.

To enable intercity comparisons of household spending, we normalized household data using median household incomes. Median household income for each city was calculated by multiplying the number of households in each income category by the appropriate category midpoint, and then dividing by the total number of households, giving estimates of US$82,922 for Portland, US$95,844 for Seattle, and C$81,390 for Vancouver.

1.2.3. Self-Organizing Maps

To provide more nuance to our income analysis than traditional measures of central tendency, we applied an unsupervised machine learning technique, the self-organizing map (SOM) (Kohonen, 2001), to summarize key features in the income datasets. Mayaud et al. (2019, in review) also used SOMs to relate income distributions to accessibility metrics and outlined their advantages compared with principal component analysis (PCA). The SOM algorithm used in this study is adapted from the code developed by Radic et al. (2015), drawing on the open-source MATLAB-based SOM toolbox by Vesanto et al. (1999, 2000). We refer the reader to Mayaud et al. (2019) for a full overview of the algorithm.

To perform SOM analysis on categorical census data, we converted the data to a continuous data distribution assuming Pareto's Law of Income Distribution (see Mayaud et al., 2019). We assigned a basic midpoint estimator for each income bin and duplicated each estimator into a vector, according to the frequency of occurrence in the given bin. We used MATLAB's *ksdensity* kernel smoothing method to perform a kernel density estimation (KDE) (Rosenblatt, 1956; Parzen, 1962) for each income vector, which was then demeaned and normalized by its standard deviation.

We trained a separate SOM for each city to produce city-specific characteristic income distributions that were assigned to nodes on a 2-D map. The SOM for each city was composed of six clusters (three rows x two columns) (see Figure S2). We opted to group clusters one and four into a “high income” category and clusters three and six into a “low income” category. For ease of analysis, we chose not to consider clusters two and five, which are considered “middle income.”

The population proportions residing in each income grouping are similar across the cities: 50.7%, 44.4%, and 48.0% of residents live in low-income cells in Portland, Seattle, and Vancouver, respectively, while 22.7%, 22.3%, and 11.9% (respectively) live in high-income cells.

1.2.4. Catchment Area Analysis

We performed catchment area analysis to reveal the sociodemographic makeup of the populations that can or cannot reach transportation hubs from their homes within certain time thresholds. Travel-time estimates were computed between every pair of grid cells (an “origin” and a “destination”) to create an O–D matrix. This involved optimally combining public transit and walking in OpenTripPlanner (OTP, 2017), an open-source routing engine called within the travel-time matrix algorithm of Pereira (2017). The OTP routing engine does not account for traffic congestion levels, but these could be incorporated into the methodology in future using GPS data (e.g., Wessel et al., 2017).

We then applied a modified version of the isochronic or cumulative-opportunity measure (Wachs & Kumagai, 1973) to estimate the number of jobs accessible from each cell. Advantages of the cumulative opportunity measure include its low computational costs, no requirements for prior information about residents’ travel behavior, and results that are easily communicable to stakeholders and policymakers. Limitations of the approach 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 behavior and preference, and is most easily interpreted (Neutens et al., 2010).

A notable drawback of the cumulative opportunity measure is that it relies on defining arbitrary cutoff times for access, which can vary depending on travel mode, socioeconomic status, and lifestyle factors (Neutens, 2010; Boisjoly & El-Geneidy, 2017). For the sake of simplicity and to remain within the scope of our aim, we chose to use a threshold of 30 minutes in this study, on the basis that residents should reasonably expect to reach a connector hub within this time from their home. However, our replicable framework will allow multiple time thresholds to be considered in future research.

The time of day for which the O–D matrices are calculated is also important to consider, because service levels and transit departure times vary throughout the data. While some studies (e.g., Mayaud et al., 2018; Pereira, 2018, 2019) have calculated average travel time matrices based on travel at regular intervals throughout the day, Boisjoly & El-Geneidy (2017) showed that calculating accessibility for travel at 8 a.m. only was a reliable indicator of relative accessibility at other times in Toronto. Given that our study examines accessibility to jobs, we consider Boisjoly & El-Geneidy's (2017) single-time approach to reliably reflect commuting dynamics in our study cities. We therefore calculated door-to-door travel times as beginning a journey at 8 a.m. on a typical working day (Tuesday, 19 September, 2017).

1.2.5. Equity Analysis

First, we evaluated equality in our job datasets by drawing a Lorenz curve to describe the distribution of total accessibility over a population. This involved plotting the cumulative proportion of a city's population on the x-axis and the cumulative distribution of origin-oriented accessibility (i.e., how many jobs each person can reach) along the y-axis. It should be noted that in this particular case, two or more individuals are able to attend the same job, meaning that multiple counting frequently occurs. As a result, summing accessibility over all individuals does not technically define a true "total" level of accessibility.

A theoretically equal society experiences one percentage point increase in accessibility for every percentage point increase in population; this is depicted by a "line of equality" at 45^{o}. The further the Lorenz curve deviates from the diagonal line of equality, the higher the level of inequality. The Gini coefficient is calculated as the area between the Lorenz curve and the line of equality, divided by the triangle bounded by the line of equality, the x-axis, and the y-axis. While this can be a useful way of comparing accessibility across our three cities, the

index measures inequality only, not equity. Capturing equity requires the introduction of a threshold value below which people are deemed to be socially excluded from a service.

Defining a minimum number of jobs to be “sufficient” for social inclusion (akin to a minimum wage, in income terms) is a complex task. There is little applicable literature on different countries’ definitions of minimum access to jobs. Moreover, the concept of social exclusion is in itself relational (i.e., dependent on the societal norms at hand), so will inevitably be person- and context-specific (Lucas, 2012).

We estimated the severity of social exclusion from the area bounded by the Lorenz curve, line of equality, and social exclusion threshold. We calculated this area using a classic trapezoid integration function, and then divided it by the total area bounded by the line of equality, the x-axis, and the y-axis to give a “social exclusion ratio.” Higher values of the social exclusion ratio represent lower equity overall.

2. FIGURES

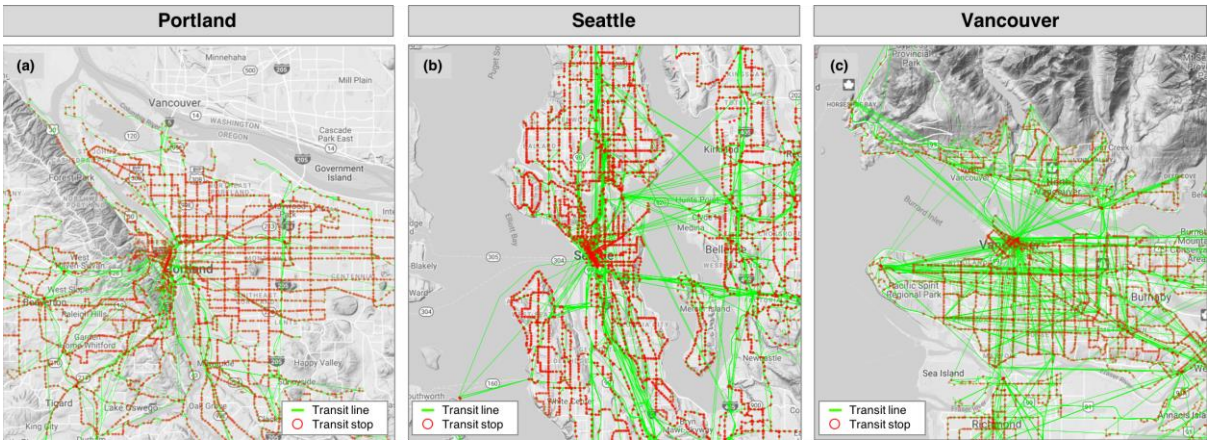

Figure S1: Transit Stops and Lines in Each City. Figure prepared by authors using Google Maps as a basemap.

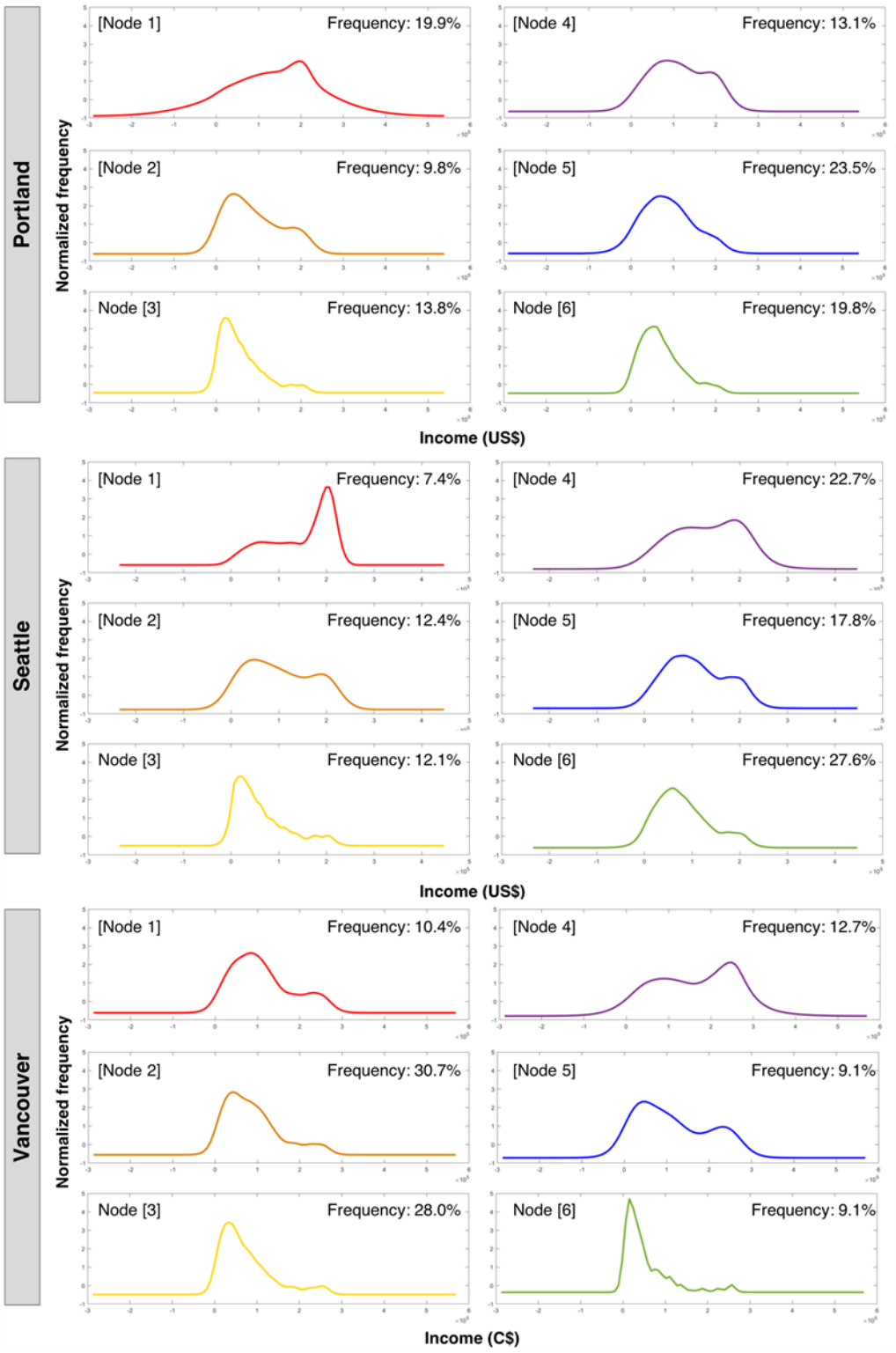

Figure S2: Frequency Distributions for Each Self-Organizing Map (SOM) Cluster (Node), Showing in Bold the Frequency (in Terms of Number of Cells) that they Occur in Each City; the Y-Axis Varies About Zero Because the Inputs are Demeaned and Normalized Using Their Standard Deviations

3. REFERENCES

Boisjoly, G., & El-Geneidy, A. M. (2017). How to get there? A critical assessment of accessibility objectives and indicators in metropolitan transportation plans. *Transport Policy, 55*, 38–50.

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

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

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

Henson, M. F. (1967). Trends in the income of families and persons in the United States: 1947 to 1964. *U.S. Bureau of the Census, Technical Paper No. 17, Washington, USA.*

Mayaud, J. R., Tran, M., Pereira, R. H. M., & Nuttall, R. (2018). Future access to essential services in a growing smart city: The case of Surrey, British Columbia. *Computers, Environment and Urban Systems.* doi:10.1016/j.compenvurbsys.2018.07.005

Milakis, D., Cervero, R., van Wee, B., & 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, B. C., & Jain, A. K. (2014). Spatial modeling of agricultural land use change at global scale. *Ecological Modelling, 291*, 152–174.

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

OpenTripPlanner (2017) Available at: https://github.com/opentripplanner/OpenTripPlanner

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

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

Pereira, R. H. M. (2018). Transport legacy of mega-events and the redistribution of accessibility to urban destinations. *Cities, 81,* 45–60. 10.1016/j.cities.2018.03.013

Wachs, M., & Kumagai, T. G. (1973) Physical accessibility as a social indicator. *Socioeconomic Planning Science, 7*, 327–456.