Measuring Activity-Based Social Segregation using Public Transport Smart Card Data

L. Kolkowski*[1], M. Dixit[2], O. Cats[2, 4], T. Verma[3], and E. Jenelius[4]

[1]Graduate student M.Sc. Transport, Infrastructure and Logistics, Delft University of Technology, The Netherlands

[2]Delft University of Technology, Faculty of Civil Engineering and Geosciences, The Netherlands

[3]Delft University of Technology, Faculty of Technology, Policy and Management, The Netherlands

[4]KTH Royal Institute of Technology, Division of Transport Planning, Sweden

SHORT SUMMARY

While social segregation is often assessed using static data concerning residential areas, the extent to which people with diverse background travel to the same destinations may offer an additional perspective on the extent of urban segregation. This study further contributes to the measurement of activity-based social segregation between multiple groups using public transport smart card data. In particular, social segregation is measured using the ordinal information theory index to measure the income group mix at public transport journey destination zones. The method is applied to the public transport smart card data of Stockholm County, Sweden. Applying the index on 2017-2020 smart card data sets for a selected week, shows significant differences between income groups' segregation along the radial public transport corridors. The overall slight decrease in income segregation over the years can be linked to declining segregation in the city center as a travel destination and its public transport hubs. Increasing zonal segregation is observed in suburban and rural zones with commuter train stations. This method helps to quantify social segregation, enriching the analysis of urban segregation and can aid in evaluating policies based on the dynamics of social life.

Keywords: Social segregation, Public transport, Ex-post transport appraisal, Smart card data analysis

1. INTRODUCTION

Social segregation often leads to disparities in essential living conditions (Leonard, 1987; Acevedo-Garcia & Lochner, 2003; Marques, 2012). Spatial segregation of social groups is conventionally measured using segregation indices applied mostly on residential socioeconomic data (Bischoff & Reardon, 2014), i.e. static data of one social space. While static data such as income, education and housing as well as spatial distance between groups are key drivers of segregation (Tan, Chai, & Chen, 2019; United Nations, 2020), considering only these offers a limited view. Given the societal relevance of social segregation, it is necessary to go beyond static measures to better reflect the extent to which different people from different backgrounds are likely to encounter each other.

Recent studies utilize mobility data to measure activity-based segregation (Farber, O'Kelly, Miller, & Neutens, 2015). Often, self-reported travel diary data is used to measure activity-based segregation which can involve accuracy, privacy, and availability issues, as well as incomplete data sets (Bagchi & White, 2005; Pelletier, Trépanier, & Morency, 2011). In addition, it can require immense efforts and high costs to obtain sufficient data sets. Conversely, public transport travel

data offer valuable mobility traces to measure activity-based segregation. In particular, passively collecting smart card data offer unprecedented large data sets of real transactions, thus observed mobility traces (Utsunomiya, Attanucci, & Wilson, 2006). Insofar, only Abbasi et al. (2021) measure multiple group's social segregation using public transport smart card data (Abbasi, Ko, & Min, 2021). Based on the fare reduction for children, seniors and passengers with disabilities stored on the smart cards, they were able to extract social characteristics to form social groups. For many transport authorities and countries, this kind of personal information is not available or extracting it would raise data privacy concerns. As a result, social information often cannot be retrieved directly from smart cards. Even richly equipped smart cards usually do not contain the desired socio-economic information.

A method connecting social and mobility data would allow for quantifying the activity-based social segregation and also to empirically measure the impacts of different interventions and policies. Limited access to transportation results in lesser access to essential amenities and opportunities to participate both socially and economically (Lucas, 2011). Transport disadvantage is strongly correlated to social exclusion as found by studies such as (Church, Frost, & Sullivan, 2000). Public transport can potentiality reduce activity-based segregation by offering an affordable mean of transport.

This study aims to answer how multi-group activity-based social segregation could be measured using large-scale disaggregated mobility data such as public transport smart card data. Connecting social information to public transport user's mobility patterns would enable measuring activity-based segregation and therewith facilitate ex-post transport appraisal from a social segregation perspective.

2. METHODOLOGY

We develop a method to measure activity-based social segregation by enriching mobility data in such a way that it connects to travelers' social characteristics. General requirements regarding mobility data as well as socioeconomic data sets for segregation studies with disaggregate mobility data are formulated. Next, socioeconomic residential data is linked to each disaggregate element of the large-scale mobility data. Lastly, multi-group segregation measures are applied to the enriched disaggregated mobility data. This sequence of steps (shown in Figure 1) constitutes a method for measuring social segregation using large-scale disaggregate mobility data and socioeconomic data. Residential socioeconomic data and its abstracted groups are connected to observed disaggregated mobility data by using the same spatial units, e.g. statistical census zones. Once the socioeconomic data is assigned to the mobility data via the inferred travelers' home zones (see (Kholodov et al., 2021)), different segregation measures can be applied.

As many relevant segregation variables pertain to ordinal social groups, the "ordinal information theory index" is used (Reardon, 2009). The ordinal information theory index measures segregation as the ratio of between-category variation to total variation. As a result, travelers' experienced segregation at the journey destination zone, assessed by the segregation measure, depends on their home zone's social status - the category travellers are assigned to.

The following notations are introduced to calculate the ordinal information theory index given in Equation 1. The index is based on the ordinal variation function ν shown in Equation 2 which relies on the distribution function f presented in Equation 3.

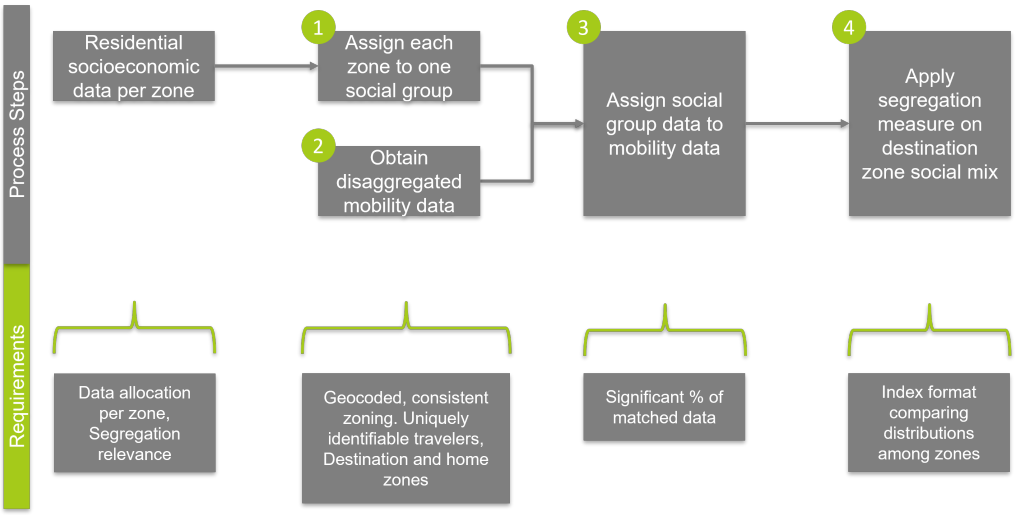

Figure 1: Framework for measuring social segregation by connecting mobility data to socioeconomic data: The first row illustrates the steps in the methodology and the second row marks the requirements for each step.

- k: Ordered categories (social groups)
- m: Unordered categories (neighborhoods, zones)
- t_m: Total population in m
- T: Total Population
- c_m: $[K-1]$ Tuple of cumulative population distribution in m
- v: ordinal variation

$$\Lambda = \sum_{m=1}^{M} \frac{t_m}{Tv}(v - v_m) \tag{1}$$

$$v = \frac{1}{K-1} \sum_{j=1}^{K-1} f(c_j) \tag{2}$$

$$f(c) = -[clog_2(c) + (1-c)log_2(1-c)] \tag{3}$$

The ratio of $[K-1]$ ordinal classes is multiplied with the sum of $K-1$ values of f, the distribution function defined in Equation 3. The closer v_m is to 1 the less homogeneity there is in the unordered group m. Therefore, 1 represents the maximum social segregation. Contrarily, v_m=0 indicates the maximum amount of homogeneity in m, thus no segregation.

By tracking the mobility of users over time and comparing similar time spans, it is possible to measure the evolution of segregation. Since the index allows to calculate contributions to the segregation index at the zonal level, the evolution of segregation can be measured even for a single zone. This allows observing a zone's social mix of travelers over time.

3. RESULTS AND DISCUSSION

Case study description

The steps outlined in section 2 are applied on the multi-modal public transport system of Stockholm County, Sweden which is home to 2.4 million inhabitants. Segregation in Stockholm is mostly connected to findings on residents' ethnics and income (Andersson & Kährik, 2015). To determine social groups, each so-called DeSo (Demographic statistics areas) zone is assigned to one of four income quantiles. The median of the distribution of income per 20+ years old inhabitant per zone is used from the 2017 Swedish income and tax register (SCB, 2017).

The segregation index is measured at the destination of journeys. For week 5 (27-01 till 03-02) in 2020, 8.45 million journeys including a destination stop area and an inferred home zone are obtained. Similar journey data sets are derived for the corresponding weeks in 2017, 2018 and 2019. To connect the residential-based social data to public transport smart card data, the income groups are assigned to the journeys' home zones. Once the social information is connected to every smart card transaction via card IDs, the smart card data set is enriched to apply a destination-based measurement of the social mixture.

Results

The social segregation index score for each day of the week in 2017 averages 0.1923. Compared to 2017, the average segregation drops by 2.4% to 0.1877 in 2018 and by 3.3% to 0.1856 in 2019. In 2020, the segregation level averages 0.1888, up slightly from 2019 and 2018 but still 1.8% lower than in 2017. This implies that segregation levels were the lowest in 2019. Looking at 2020 the index displays lower levels than 2017 but higher segregation than in 2019 and for Monday to Thursday in 2018.

Figure 2 illustrates that people mix to a similar extent on Monday to Thursday. Conversely, the lowest segregation is observed on Fridays and Saturdays. These can be related to the combination of work, leisure, and shopping activities. On Sundays, travelers mix less and experience more segregation as the index is higher than on other days. Sundays are considered as rest days with the least working activity which leads inhabitants to stay more within their home zone. These findings match other activity-based conclusions, such as that work-related activities reduce segregation (Ellis, Wright, & Parks, 2004).

Figure 3 and Figure 4 display the weighted and "absolute" segregation contribution of each zone in 2017, respectively. The absolute contribution is the result of the differences between the total ordinal variation and the zone-specific variation $v - v_m$. The weighted value corresponds to each zone's contribution to the segregation index calculated by the absolute contribution in relation to the population affected, i.e. $\frac{t_m}{Tv}(v - v_m)$. In other words, the absolute value expresses the contribution before being set into relation with the number of passengers affected, while the weighted value accounts for the number of passengers affected compared to the overall amount of passengers.

As can be seen, the weighted segregation contribution is highest in central zones and suburban centers. What stands out, is the general pattern of zones with both high absolute and weighted contributions to the segregation index, which indicates that many passengers experience segregation at these destinations. For 2017, outskirt neighborhoods have high absolute and weighted contributions to the segregation index. Contribution to the segregation index mostly comes from the load of passengers in the city and suburban centers.

By taking differences of weighted segregation contributions between years, it can be seen whether a specific zone contributed to a decline or rise of the segregation index. For each zone and day of

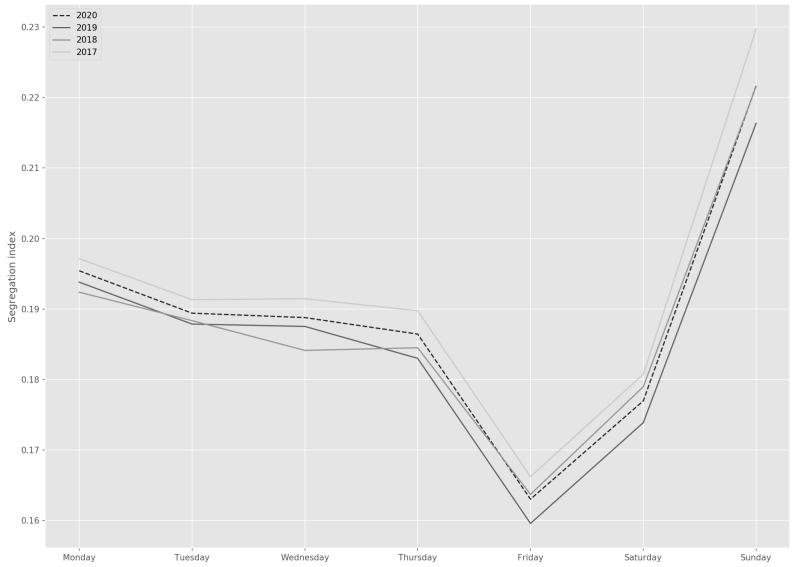

Figure 2: Segregation throughout week 5: Segregation index ranges from 0 to 1 with 1 indicating no income group mixing in a destination zone and therefore maximum segregation and 0 indicating equally distributed income groups over all destination zones, thus no segregation

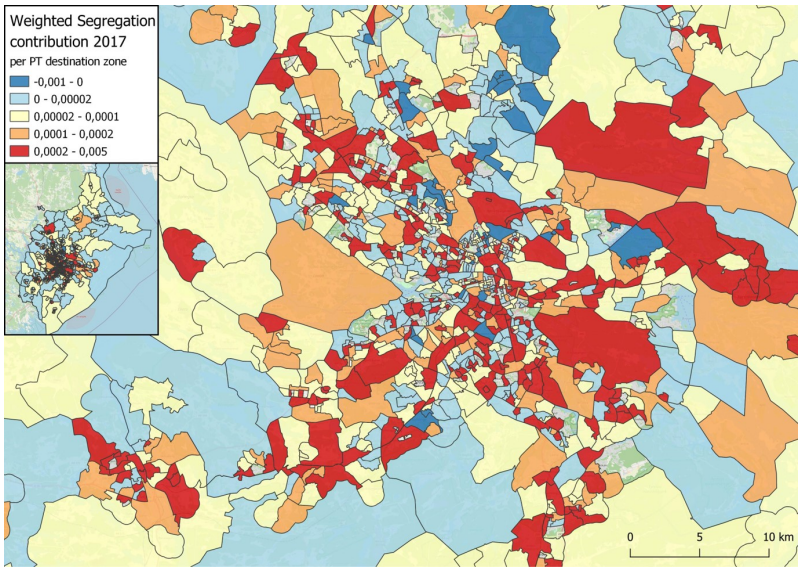

Figure 3: Weighted segregation contribution 2017: each DeSo zone's arriving PT passenger mix contribution to the weighted segregation index level

the week, the difference is calculated by taking the more recent year's contribution and subtracting the 2017 contribution. Then, the average of differences is determined per zone over all days of the week. Thereby, the evolution of segregation can be assessed at the zonal level. A negative difference indicates thus a decline in segregation. Contrarily, a positive difference shows an increase in contribution to segregation.

By the setup of this case study, changes in segregation levels are related to a more diverse use of the PT system and to the population affected. Lower experienced segregation levels of passengers traveling to the city center outweigh the higher suburban and rural segregation experiences, due to the higher number of passengers affected in central areas.

The dispersed structure and inner-city zone size make it difficult to immediately spot patterns and

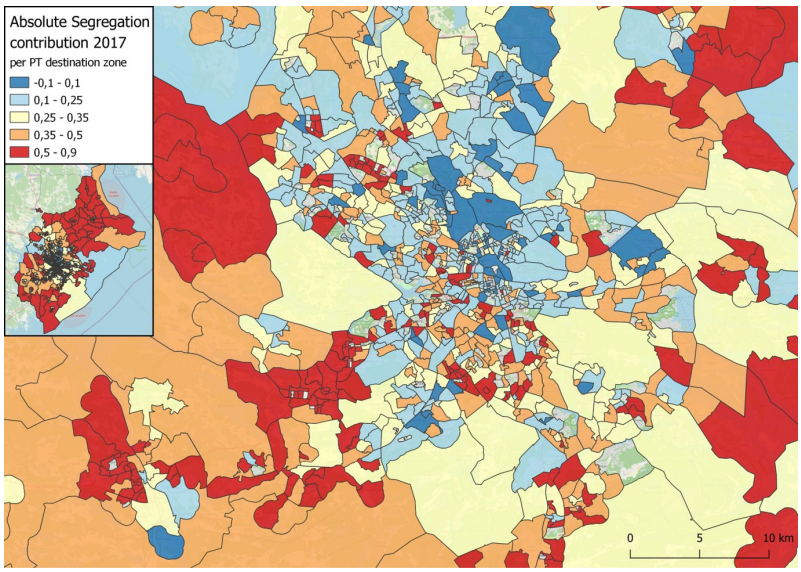

Figure 4: Absolute segregation contribution 2017: each DeSo zone's arriving PT passenger mix contribution to the absolute segregation index level

inspect effects in Figure 5. In addition, the zone sizes have an impact on the impression of segregation levels, even though it does not indicate the number of passengers affected. In the city center, few sharply increased segregation zones can be detected, accompanied by strong decreasing and slightly to not decreasing segregation levels. Urban zones with segregation reductions outnumber the ones with rises for 2018.

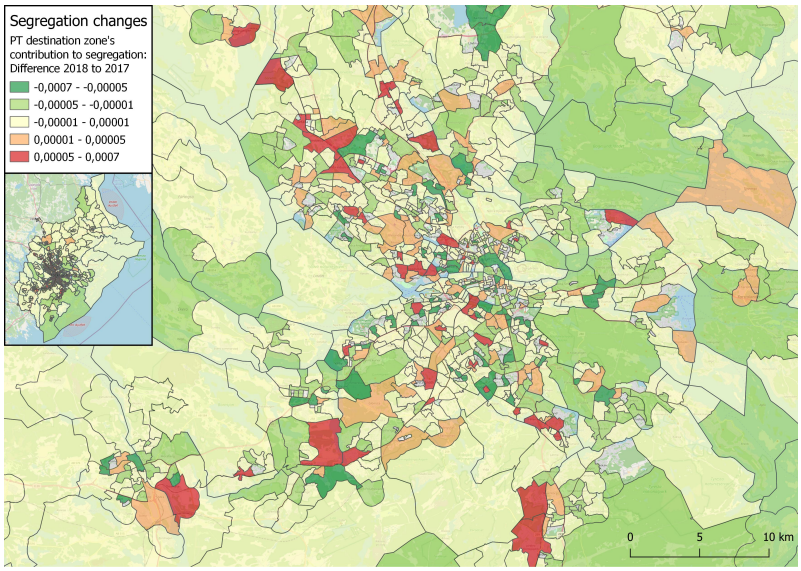

Figure 5: Segregation changes 2017-2018 - Change of each DeSo zone's arriving PT passenger mix contribution to the segregation index level. Decreasing levels indicate less segregation contribution

For both 2018, as well as in the 2020 comparison to 2017 (not shown here), there are more zones where less segregation is experienced than zones that experienced an increase in segregation. Even though for both years there are some zones which experienced an increase in segregation (shown in dark red). The overall change in 2020 is about 25% less compared to 2018 changes which matches the overall fallback trend mentioned earlier.

Weighted differences between the years' weekly average index allow drawing conclusions on the segregation development. For the Stockholm case, the segregation index suggests less activity-

based segregation in the years after 2017, especially in the city center. These effects potentially relate to the enhanced public transport system due the opening of a new commuter train infrastructure and frequency increase after the "Citybanan" tunnel opening in July 2017.

City center inbound PT passengers are found to be more income-diverse in 2018 and 2020 than in 2017, while outbound passengers towards the suburbs have more and more uniform income backgrounds, especially when traveling to commuter train stations. Both stronger increasing and decreasing effects are indicated for the northwest and southwest corridors. Increasing segregation levels in these suburban and peri-urban zones could be linked to general trends of urbanization and gentrification, as well as PT dependency and the transport disadvantage of low-income groups.

Main limitation of this study are the unrevealed direct causal effects as well as only assessing PT travel patterns. In addition, the assumptions made about the homogeneity of groups in an area could lead to inadequacies in capturing the precise social composition.

4. CONCLUSIONS

We demonstrate how connecting mobility data to social data could potentially lead to a more detailed understanding of social segregation and examine segregation developments in relation to transport or policy changes. Daily or weekly segregation levels help evaluate overall levels and trends. Weighted segregation levels are suitable for analyses in which the relation of zone segregation plays a role. The absolute segregation contribution should be used for detailed, zonal assessment. Particularly for urban planners and policymakers, it could be of interest to measure social segregation effects and assess the impacts of various interventions. In addition, the index format facilitates comparisons of segregation levels with other cities and regions.

The results help evaluating the segregation situation in Stockholm and at the same time raise the question of why segregation - as measured in terms of the diversity of income mix-up at travel destinations in this study - is appearing more or less in certain areas. By disentangling what led to change in segregation levels, it could be assessed whether direct effects can be linked to specific changes. Further, segregation effects might be incorporated into a multi-criteria policy analysis and investment assessment.

ACKNOWLEDGMENT

The authors would like to thank Region Stockholm for sharing the data used within this project. The work was supported by Stockholm County Council Trafik och Region call 2019. Specifically, the authors would like to thank Isak Jarlebring Rubensson. Further, the authors thank Matej Cebecauer from KTH Royal Institute of Technology for his support on the data setup.

REFERENCES

Abbasi, S., Ko, J., & Min, J. (2021, 4). Measuring destination-based segregation through mobility patterns: Application of transport card data. *Journal of Transport Geography*, *92*. doi: 10.1016/j.jtrangeo.2021.103025

Acevedo-Garcia, D., & Lochner, K. A. (2003). Residential segregation and health. *Neighborhoods and health*, 265–87.

Andersson, R., & Kährik, A. (2015). Widening gaps : Segregation dynamics during two decades of economic and institutional change in stockholm. In *Socio-economic*

segregation in european capital cities (pp. 134–155). New York: Routledge.

Bagchi, M., & White, P. R. (2005, 9). The potential of public transport smart card data. *Transport Policy*, *12*, 464-474. doi: 10.1016/j.tranpol.2005.06.008

Bischoff, K., & Reardon, S. F. (2014). Residential segregation by income, 1970-2009. *Diversity and disparities: America enters a new century*, *43*.

Church, A., Frost, M., & Sullivan, K. (2000). Transport and social exclusion in London. *Transport policy*, *7*(3), 195–205. Retrieved from www.elsevier.com/locate/tranpol

Ellis, M., Wright, R., & Parks, V. (2004, 9). Work together, live apart? geographies of racial and ethnic segregation at home and at work. *Annals of the Association of American Geographers*, *94*, 620-637. doi: 10.1111/j.1467-8306.2004.00417.x

Farber, S., O'Kelly, M., Miller, H. J., & Neutens, T. (2015, 12). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography*, *49*, 26-38. doi: 10.1016/j.jtrangeo.2015.10.009

Kholodov, Y., Jenelius, E., Cats, O., van Oort, N., Mouter, N., Cebecauer, M., & Vermeulen, A. (2021, 5). Public transport fare elasticities from smartcard data: Evidence from a natural experiment. *Transport Policy*, *105*, 35-43. doi: 10.1016/j.tranpol.2021.03.001

Leonard, J. S. (1987). The interaction of residential segregation and employment discrimination. *Journal of Urban Economics*, *21*(3), 323-346.

Lucas, K. (2011, 11). Making the connections between transport disadvantage and the social exclusion of low income populations in the tshwane region of south africa. *Journal of Transport Geography*, *19*, 1320-1334. doi: 10.1016/j.jtrangeo.2011.02.007

Marques, E. (2012, 9). Social networks, segregation and poverty in são paulo. *International Journal of Urban and Regional Research*, *36*, 958-979. doi: 10.1111/j.1468-2427.2012.01143.x

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*, 557-568. doi: 10.1016/j.trc.2010.12.003

Reardon, S. F. (2009). Measures of ordinal segregation. In *Occupational and residential segregation.* Emerald Group Publishing Limited.

SCB. (2017). *SCB:s Open data for DeSO – Demographic Statistical Areas.* (Retrieved from https://www.geodata.se/geodataportalen/srv/swe/catalog.search;jsessionid=42B5AAC3339638A205A27724ECF960BF#/search?resultType=swe-details&_schema=iso19139*&type=dataset%20or%20series&from=1&to=20 using the explanation from https://www.scb.se/en/services/open-data-api/open-geodata/deso--demographic-statistical-areas/)

Tan, Y., Chai, Y., & Chen, Z. (2019, 10). Social-contextual exposure of ethnic groups in urban china: From residential place to activity space. *Population, Space and Place*, *25*. doi: 10.1002/psp.2248

United Nations. (2020). *World social report 2020: inequality in a rapidly changing world.* United Nations Department of Economic and Social Affairs. New York: United Nations publication. Sales No. E.20.IV.1.

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation research record*, *1971*(1), 118–126.