Computers, Environment and Urban Systems 54 (2015) 204–218

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Synthesizing spatial interaction data for social science research: Validation and an investigation of spatial mismatch in Wichita, Kansas

Michael A. Niedzielski [a,*], Morton E. O'Kelly [b], E. Eric Boschmann [c]

[a] Department of Geography and Geographic Information Science, University of North Dakota, 221 Centennial Drive Stop 9020, Grand Forks, ND 58202, United States
[b] Department of Geography, The Ohio State University, 1036 Derby Hall, 154 N Oval Mall, Columbus, OH 43210, United States
[c] Department of Geography and the Environment, University of Denver, 2050 E. Iliff Ave., Denver, CO 80208-0710, United States

ARTICLE INFO

Article history:
Received 9 November 2014
Received in revised form 13 August 2015
Accepted 5 September 2015
Available online 21 September 2015

Keywords:
Commuting
Spatial interaction
Accessibility
Disaggregated
Race
Income

ABSTRACT

Rising economic segregation suggests a need to examine constraints to job access by race/ethnicity and economic inequality simultaneously. This often requires detailed socio-spatial interaction data to make progress on theoretical and modeling development, empirical studies and policy insights. Commuting data are commonly used because of its wide availability. Despite excellent work trip datasets from the U.S. Census such as the Census Transportation Planning Package and the Longitudinal Employer-Household Dynamics (LEHD) data, there are often gaps between the data that are available and ideal detailed commuting data suited to models and data analysis. This is because commuting data are available for a limited set of socio-economic dimensions and this coarseness limits researchers in their ability to uncover nuances of place-based generalizations about commuting, either socially or spatially. In one promising approach, an information minimizing technique was proposed as a workable practical method to synthesize disaggregated work trip flows. Because the strength of fit between predicted and observed trips is unknown, this paper validates this method using real commutes disaggregated by income and then synthesizes race–income work trips using LEHD data for the Wichita, Kansas metropolitan statistical area. We find that low-income Whites travel longer distances and have more dispersed travel patterns than all African-American and Asian income groups and that both low- and middle-income groups of all race groups have spatially constrained flows.

1. Introduction

Urban research has long contended that access to jobs is not evenly distributed within U.S. metropolitan areas due to the effects of racial/ethnic residential segregation patterns and urban-economic restructuring (Kain, 1968; Kasarda, 1989; Preston & McLafferty, 1999; Wilson, 1987). This body of research has further expanded to examine job access constraints that occur due to gender and domestic responsibilities (Hanson & Pratt, 1995; Kwan, 1999), access to automobiles and public transit dependence (Taylor & Ong, 1995), or differences based upon job skills or wage levels (Immergluck, 1998).

In more recent years residential segregation in the U.S. has increased along lines of class/income (Massey, Rothwell, & Domina, 2009). For example, the changing spatial distributions of metropolitan area poverty include more people living in concentrated poverty neighborhoods, a growth in high-poverty neighborhoods, and a decentralization and suburbanization of poverty (Hafley & Boschmann, 2015; Jargowsky, 2003; Kneebone & Garr, 2010). Overall, income inequality has reached its highest levels since the Great Depression (Fry & Taylor, 2012) leading to greater recognition of economic segregation in U.S. metropolitan areas (Florida & Mellander, 2015). Thus, in addition to on-going effects of racial/ethnic residential segregation on employment access, the rising economic segregation suggests a need to examine constraints to job access by race/ethnicity and economic inequality simultaneously.

One particular challenge to this agenda is the lack of adequate datasets that provide detailed socio-economic commuting information at spatially refined scales. Empirical studies of spatial mismatch, segregation, social interaction potential, excess commuting, commute reduction potential, and jobs–housing balance (Boschmann, 2011; Farber, Neutens, Miller, & Li, 2013; Horner & Marion, 2009; Horner & Mefford, 2007; Niedzielski, 2006; Niedzielski & Boschmann, 2014; O'Kelly & Niedzielski, 2008; O'Kelly, Niedzielski, & Gleeson, 2012; Widener, Farber, Neutens, & Horner, 2013) demonstrate the *requirement* of very detailed spatially and socially disaggregated data to make progress on theoretical and modeling development, empirical studies and policy insights. Among the data needed are measurements of residential/employment locations and commuting trip lengths. There are typically two choices for using fine-grained spatial interaction data. One choice is to develop individual-level models based on activity-travel data. While these highly disaggregated data provide important insights about the context and contingency of person-based spatial deprivation, it does not generally allow for comparison of intra-metropolitan spatial

* Corresponding author.
E-mail addresses: michael.niedzielski@und.edu (M.A. Niedzielski), okelly.1@osu.edu (M.E. O'Kelly), eric.boschmann@du.edu (E.E. Boschmann).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.09.004
0198-9715/

patterns. This shortcoming is overcome by place-based models that seek to identify aggregated experiences of socio-spatial deprivation.

Because fully enumerated and fine-grained data such as the Irish POWCAR commuting dataset (O'Kelly et al., 2012) are rare, researchers face a trade-off when working with place-based models based on zonally aggregated journey-to-work data such the U.S. Census Transportation Planning Package (CTPP) and the Longitudinal Employer-Household Dynamics (LEHD) data. Despite these excellent datasets from the U.S. Census, there are often gaps between the data that are available and ideal detailed commuting data suited to models and data analysis. These datasets provide socio-economic and journey-to-work data at three levels: focused at home, at work, and on the commuting flow between them. Data by origins/destinations are available by many *single* socio-economic classes (e.g. race, age, income) but also tabulated for a limited number of *combinations* of them (e.g. race and income). However, flow data are available only for generic workers (i.e. not disaggregated) or a limited set of *single* classes (poverty status, age, earnings, or industry). The implication of this coarse flow data is that researchers are limited in their ability to understand detailed socio-spatial patterns of accessibility, spatial mismatch, segregation, commuting efficiency, and jobs–housing balance. At base, the data are not quite sufficient to tackle the kinds of interesting social science questions that require very detailed cross tabulations.

Given such data constraints, very little research has been able to comparatively analyze actual employment flows of workers based upon race/ethnicity and income simultaneously. While vast complexities do exist, the dominant findings from the research literature suggest that racial and ethnic minorities have more restricted job access (Grengs, 2012; Horner & Mefford, 2007; Holzer, Ihlanfeldt, & Sjoquist, 1994), as do lower income workers (Grengs, 2012; Hu, 2014; Immergluck, 1998; Kawabata, 2003; Sanchez, Shen, & Zhong-ren, 2004). In employment access literatures such as spatial mismatch, one debate of particular interest is the meaning of distance in the home-work commuting linkage (Hanson & Pratt, 1988, 1995). While strict dualisms may be problematic, commute distance can represent a 'choice', or a 'constraint', for a worker. In some instances, a longer commute between home and work may represent 'choice': more job opportunities to choose from, willingness to travel further for higher paying jobs, many housing options and high mobility, or choosing to trade off longer commutes for better housing options. Shorter distances between home and work can represent 'constraint': limited transportation options constrain possible job opportunity locations, limited housing options, unemployment due to limited nearby jobs, only low paying jobs nearby, or bearing the burden of longer commutes especially to low paying jobs. With more recent trends of reurbanism, gentrification, and the desirability of shorter commutes (Fishman, 2005), the meanings of commute distance have become more complex. What is needed is a methodology that allows simultaneous examination of actual commuting outcomes based on race/ethnicity and income allowing integrated illumination to see if the effects of distance are consistent when measured together.

In one promising approach, O'Kelly and Lee (2005) proposed a computational method [hereafter OKL] to synthesize disaggregated commutes and their approach has been used to synthesize trips by occupation (Lee, 2012), race/ethnicity (Jang & Yao, 2014), and gender and occupation (Kim, Sang, Chun, & Lee, 2012; Sang, O'Kelly, & Kwan, 2011). The approach is undergoing extensive development in the application to large problem instances (Farber, O'Kelly, Miller, & Neutens, 2015). The need for synthesizing *actual* commutes between zones such as block groups, tracts or traffic analysis zones exists for several reasons. First, highly socially and spatially disaggregate flows would provide insight into actual worker travel behavior. Currently, such data are unavailable and researchers typically use proxies such as potential travel behavior based on what it could be rather that what it actually is to compute, for example, probabilities of securing employment or reducing segregation. Alternatively, they use socially detailed but spatially coarse data from the American Community Survey Public Use Microdata Samples dataset. Second, LEHD and CTPP contain observed flows for only a limited number of socio-economic classes and these may be used for validation tests. But they also contain observed home/work zonal counts for many single and double socio-economic classes ideal for synthesizing detailed commuting flows. Third, highly disaggregate flow data can provide for computations of explanatory variables (parameters, coefficients, balancing factors, shadow prices depending on the context) allowing inferences about factors contributing to push/pull in movement of specific socio-economic categories of workers.

Table 1
Socio-economic variables for which LEHD origin, destination, and flow data are available.

O_{ik}, D_{jk}, X_{ijk}	O_{ik}, D_{jk} only
Age (3)	NAICS sector (20)
Earnings (3)	Race (7)
Industry sector (3)	Ethnicity (2)
	Educational attainment (4)
	Sex (2)
	Age: & Earnings (9); & NAICS sector (60); & Race (21); & Ethnicity (6); & Educational Attainment (12); & Sex (6)
	Earnings: & NAICS sector (60); **& Race (21);** & Ethnicity (6); & Educational Attainment (12); & Sex (6)
	Industry sector: & Age (9); & Earnings (9) & NAICS sector (60); & Race (21); & Ethnicity (6); & Educational Attainment (12); & Sex (6)

Number in parenthesis indicates number of categories for each class. Bold indicates the classes chosen for the validation test and empirical study, respectively, in this paper.

Given the increasing need for synthetic flow data, and its general usefulness for urban research, what is unknown is how well OKL actually models real work trip distributions. In the absence of such tests, researchers have assumed the fit between predicted and observed trips is good. Using Wichita, Kansas data, the purpose of this paper is to test this assumption and validate the OKL technique using LEHD income-disaggregated commutes and to synthesize income–race commutes to illustrate the effectiveness of this approach for urban research.

2. Synthetic commuting flows

The essential idea in most micro-simulation or synthesis problems is to take some high level observation and split it using rates or proportions that reflect shares. The more accurately the data can provide information about the macro states, the more capable will be the fitted model. To this information theoretic perspective, we add an idea from the field of constraint programming. If we can narrow the options for the fitted data by adding control totals, there is a good chance that the resultant will (a) fit the known facts, i.e. the observed work trip distribution, and (b) make unbiased assumptions about the missing details, because the fitted data "entirely reflects the information contained in the constraints" (Shipley, 2009, 158)

The OKL approach synthesizes the unknown disaggregated (i.e. micro state) commuting flows, X_{ijk}, using two types of known macro state constraints, the aggregate commuting flows, X_{ij} (the high level observation), and the aggregate and disaggregate origin and destination totals, O_i/O_{ik} and D_j/D_{jk} (the rates that reflect shares), where

i = index of origin or home locations, $i = 1\ldots n$

j = index of destination or work locations, $j = 1\ldots n$

k = index of worker categories for a given socio-economic class, $k = 1\ldots K$

O_i = number of workers living in origin i

O_{ik} = number of workers of type k living in origin i

D_j = number of workers employed in destination j

D_{jk} = number of workers of type k employed in destination j

X_{ij}^{obs} = observed number of workers commuting from origin i to destination j