Article

B Urban Analytics and City Science

EPB: Urban Analytics and City Science
2023, Vol. 50(3) 709–722

DOI: 10.1177/23998083221129614
journals.sagepub.com/home/epb
SAGE

A method to derive small area estimates of linked commuting trips by mode from open source LODES and ACS data

Kevin Credit
National Centre for Geocomputation, Maynooth University, Ireland

Zander Arnao
The College, University of Chicago, Chicago, IL, USA

Abstract

This paper describes a fully customizable open source method to create linked origin-destination data on commuting flows by mode at the Census tract scale by combining LODES and ACS data from the US Census Bureau. With additional work, the method could be scaled to the entire US (with a small number of exceptions) for every year from 2002 to 2019. For demonstration purposes, the paper applies this method to 2015 commuting flows in Cook County, Illinois. At an aggregate scale, the results of this application show that commuting by all modes is dominated by travel to large regional employment centres. However, the pattern is more localised for the walking mode, and focused along corridors of mode-specific infrastructure investment for the cycling and transit modes, as might be expected. The auto and work from home modes demonstrate the most distributed pattern of travel, revealing more instances of commuting to regional sub-centres than the other modes.

Keywords

Travel behaviour, commuting, big data, transportation modelling, urban analytics

Introduction

Cities face a variety of urban transportation challenges in the 21st century. Managing traffic congestion, commute times and infrastructure investment and maintenance costs have traditionally been the primary focus for many transportation planning agencies and public officials in North America (Rodrigue, 2020). In recent years, growing concern over the myriad climate-related, environmental, health and social externalities from automobile-centred transportation systems has foregrounded the need for transportation planners to better understand (and plan for) non-auto modes (Woodcock et al. 2009; De Nazelle et al. 2011; Neves and Brand 2019; Lee et al. 2017;

Corresponding author:
Kevin Credit, National Centre for Geocomputation, Maynooth University, Maynooth, Ireland.
Email: kevin.credit@mu.ie

Lovasi et al. 2009; Vojnovic and Darden 2013; Frank et al., 2006; Frank et al. 2004; Wang et al. 2016; Yang et al. 2018; Lindstrom, 2008; De Hartog et al. 2010; Mueller et al. 2015). Overlaid on these issues are the *even more recent* shifts in commuting and migration patterns related to the pandemic, including apparent increases in remote working, migration from dense central cities to more sprawling, auto-oriented ones and attendant declines in public transit ridership (Polzin and Choi, 2021).

To make plans to address these challenges, practitioners and researchers require a high volume of spatially and temporally fine-grained data. The standard four-step and newer activity-based travel demand models (TDM) produce estimates of the volume of trips between urban zones by trip purpose and mode (National Academies of Sciences, Engineering, and Medicine, 2012; 2014; Metropolitan Washington Council of Governments, 2018). These methods often rely on expensive household travel surveys that must be deployed at long (e.g. 10-year) intervals, and thus can become quickly outdated (Toole et al., 2015; Cuauhtemoc et al., 2017). In addition, activity-based models (ABM) employ computationally demanding spatial microsimulation methods to create synthetic (disaggregate) populations from which zonal trips are estimated and could benefit from further empirical validation[1] and more lightweight, open source options (Rasouli and Timmermans, 2014; Stabler and Freedman, 2021). Recent approaches have also moved to incorporate 'big' data sources – such as GPS traces, smart card data from public transit systems and user-generated data from apps like Strava and CycleTracks – into the transportation demand modelling process (Pelletier et al., 2011; Toole et al., 2015; Cuauhtemoc et al., 2017; Tu et al., 2018; Zhu et al., 2019). While these data provide automated collection with greater speed, scope and variety of information about transportation patterns, significant issues remain in the application of these data sources (Milne and Watling, 2019). Big data from smartphones have a number of known issues which create challenges for researchers (Zandbergen 2009; Prelipcean and Yamamoto 2018), including detection accuracy (Chen et al., 2018; Harding et al., 2021), variation in service and phone quality (Jariyasunant et al., 2014; Harding et al., 2021), and battery drain (Jariyasunant et al., 2014). Smart cards and open source applications like CycleTracks often feature samples which over-represent younger, economically active and tech-savvy populations (Bagchi and White 2005; Milne and Watling 2019), as well as intensive app users (Chen et al., 2018).

Given these issues, it is apparent that despite the recent methodological advances in estimating travel demand, Cervero's (2006) argument in favour of lightweight sketch planning models that focus *specifically* on representative and reliable measures of non-auto transportation at high spatial resolutions remains relevant. Thus, the purpose of this paper is to develop a reproducible, fully open source method for creating detailed data on commuting flows by mode for small geographic areas. We do this by combining ACS data on (origin-based) commuting by mode with LODES data on residential-employment connections to derive estimates of linked origin-destination (O-D) trips by mode. With some additional work, this method could be applied across the US at the Census tract (neighbourhood) scale for every year that LODES and the ACS/Decennial Census are both available (2002–2019, with a few exceptions), which would provide planners and researchers a vast resource for investigating transportation mode-related spatial and temporal dynamics. Here, we apply the method in Cook County, Illinois (the greater Chicago area) using 2015 data from LODES and 2013 to 2017 5-year estimates on means of transportation to work from the ACS. In this application, we estimate more than 1.7 million commuting trips over nearly 440,000 unique tract-to-tract links for the walking, transit, auto, work from home and cycling modes.

Importantly, we apply distance decay thresholds – derived empirically from National Household Transportation Survey (NHTS) data on travel distance by mode in the region – to the walking and cycling modes to redistribute (unrealistic) long distance trips to nearby tracts. Similarly, we use open source travel times (from the R package r5r) by public transit to estimate and redistribute viable transit trips. Even with these redistributions, the method preserves the total number of trips by the

five modes of interest for the county from LODES, as well as the origin-based mode share percentage from the ACS for each individual tract.

The comprehensive, timely, and spatially granular estimates of travel flows by mode produced by this method can inform important future work in both transportation planning practice and research. Public planning agencies can use this technique to study aggregate travel patterns by mode without commissioning expensive travel surveys, and as an option to validate synthetically derived activity-based models. Researchers can use these data as inputs to discrete choice or spatial interaction models in order to better understand the contextual and built environment-related determinants of non-auto travel at a regional or national scale. The temporal dimension of the LODES and ACS datasets also allow for time series analysis of the determinants or changes in modal flows at the neighbourhood scale over time.

Modelling travel demand

The conventional travel demand modelling (TDM) process consists of four steps as follows: trip generation, trip distribution, mode choice and route assignment (National Academies of Sciences, Engineering, and Medicine, 2012; 2014; Metropolitan Washington Council of Governments, 2018). In the simplest version of this framework, the region is split into areal traffic analysis zones (TAZ), which are slightly larger than Census tracts. The number of trip 'productions' and 'attractions' for a given trip purpose in each TAZ are estimated based on the characteristics of households gathered from a large transportation survey, such as the NHTS, and Census data on the number of households and/or employees in each TAZ. These trip productions and attractions are then put into a spatial interaction model, along with the distance from each (origin) TAZ to every other (destination) TAZ, to produce estimates of the number of linked O-D trips (by purpose). These linked trips are classified by mode using multinomial or nested logit models, where mode choice is estimated as a function of level of service, traveller characteristics and area characteristics. Once the number of trips by mode between every origin and destination has been estimated, they are applied to the street (or transit) network using algorithms that take into account shortest paths and capacity constraints to determine the actual volume of travel on each component of the travel network (National Academies of Sciences, Engineering, and Medicine, 2012).

While these models have been widely used in transportation planning for more than half a century, the conventional approach has a number of serious issues. First, the data required to estimate trip distribution (i.e. trips between origins and destinations) and mode share is extensive and not completely available from public sources at fine-grained temporal intervals. Household travel surveys – including the federal NHTS – are expensive to deploy and generally only updated every 10 years (or more) (Toole et al., 2015; Cuauhtemoc et al., 2017). This means that traditional TDMs do not have the ability to monitor or react to fast-changing conditions or events. Similarly, TAZs are relatively large spatial units that mask fine-grained travel patterns, particularly for nonmotorized modes (Cervero, 2006). Of course, travel by nonmotorized modes – and sometimes even transit – is often not considered at all in conventional TDMs, which means that we have critically little knowledge about regional-level transportation patterns by walking and cycling (National Academies of Sciences, Engineering, and Medicine, 2012; Cervero, 2006).

More complex activity-based models (ABM) have been developed in recent years to address some of these issues (Bhat et al., 2002; National Academies of Sciences, Engineering, and Medicine, 2014; Rasouli and Timmermans, 2014; Stabler and Freedman, 2021). These models use spatial microsimulation methods to generate a dataset of synthetic individuals (Tanton, 2014). Travel is then modelled at the individual level, based on each synthetic person's individual characteristics, household characteristics, time constraints and activities. This provides the ability to understand much finer-grained spatial and temporal characteristics of travel, including trip chains

(National Academies of Sciences, Engineering, and Medicine, 2014). However, these models are computationally intensive and even more complex to implement than conventional TDMs, with few open source options (Stabler and Freedman, 2021) or ways to empirically validate the results, since the outputs occur fundamentally at the individual level (Rasouli and Timmermans, 2014). They also rely on the same expensive, intermittent travel surveys as conventional TDMs (Tajaddini et al., 2020).

Recently, the availability of new, digitally generated 'big' sources of data on mobility – including smartphone GPS traces, smart card ingress and egress data from transit agencies and user-created information from exercise apps such as Strava – have offered the potential to overcome the problem of costly individual-level data collection (Pelletier et al., 2011; Toole et al., 2015; Cuauhtemoc et al., 2017; Tu et al., 2018; Zhu et al., 2019; Tajaddini et al., 2020). The collection of individual-level *empirical* data on mobility – rather than producing a synthetic estimate – theoretically simplifies the ABM process and removes concerns over 'ground-truth' validation. However, smartphone-derived data have several issues which increase the difficulty of using them in TDM applications (Zanbergen 2009; Preplipcean and Yamamoto 2018). A number of barriers, for example, urban canyons, user behaviour and variation in service and phone quality (Jariyasunant et al., 2014; Harding et al., 2021), make it difficult for smartphone GPS to consistently detect user location with accuracy (Harding et al., 2021), which necessitates supplementary data processing to remove errors (Chen et al., 2018). GPS tracking, especially from specialized apps created by researchers, also leads to drain of battery life, which is a strong concern for users and limits tracking time (Jariyasunant et al., 2014). It is also not possible to ascertain which travel mode is being used simply from the trajectory of GPS traces alone.

Similar problems beset other sources of big transportation data like public transit smart cards. These are credit card-sized devices that store and process data for trip fare collection systems in public transit (Pelletier et al., 2011). Much research has employed smart card data to study public transit (Tu et al., 2018). These data can suffer from issues of representativeness and inaccuracy due to lack of random sampling and failure in technology (Bagchi and White 2005). Crucial context like trip purpose and demographics are often missing because such information is anonymized or not recorded at all (Cuauhtemoc et al., 2017; Milne and Watling 2019). These problems are characteristic of broader issues with big transportation data (Milne and Watling 2019). And, while data from workout trackers like CycleTracks and Strava provide concrete information on mode, as with many internet applications, their users are often younger, more affluent and generally more comfortable with technology than the general population (Milne and Watling 2019). These features may correlate with their displayed mode choice preferences, which makes it somewhat difficult to generalize patterns derived from these data to the wider population.

Given these issues with 'big' data on non-auto travel, and the limitations associated with estimating TDM and ABM in general, our method fills a number of gaps. We use open source ACS and LODES data, so the method can be easily implemented without a purpose-built travel survey. Since LODES contains data on the linked location of every individual employee-workplace at the Census block scale across the country (for every year), this approach also provides high levels of spatial and temporal granularity without the associated complexity of deriving synthetic populations (as in ABM) or relying on potentially unrepresentative samples (as in digital sources of data or smartphone GPS traces). We also directly estimate the modal split of each O-D link from ACS data, foregrounding the ability to study nonmotorized travel patterns within a regional context. The data produced by this method can also be easily combined with other (Census or locally derived) datasets to study the role of demographics, the built environment, or other neighbourhood characteristics on the volume of travel by mode. And, perhaps most importantly, the method is relatively lightweight to implement, which means it can be more easily used by citizens, transportation activists, policymakers and land use planners. The lightweight nature of the application can also be useful for

transportation planners in sketch planning contexts, for specific studies of nonmotorized travel, and also as a part of the empirical validation of ABM results.

Methods

Data

This method described in this paper is based on two primary open datasets prepared by the US Census Bureau: (1) the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Statistics (known as LODES) and (2) the American Community Survey (ACS) 5-year estimates (Manson et al., 2020). LODES is collected primarily from the unemployment insurance reporting system and delineates residential and workplace locations from these (and other federal administrative) records. It is important to note that since the employment location is reported by employers, in some cases it may not be the exact location where a given employee physically works (Graham et al. 2014), for example, in the case of large institutional employers, which Murikami (2007) also identifies as a particular concern for using the early LEHD data for block-level transportation analysis. LODES also does not denote whether a given origin-destination link is actually taken as a trip. However, the method presented in this paper helps deal with that uncertainty by combining these employee–employer links with data on commuting by mode (including, implicitly, the percentage working from home), thus providing an estimate for actual trips taken across each O-D link.

Counts of the individual employee–employer links are aggregated to the Census block level for every year from 2002 to 2019[2] to create the basic data product (Graham et al. 2014). Within LODES there are three primary datasets as follows: (1) the Origin-Destination (OD) file, which contains job totals associated with a linked residence → workplace flow; (2) the Workplace Area Characteristics (WAC) file, which contains job totals summed by the workplace block and (3) the Residence Area Characteristics (RAC) file, which contains job totals summed by residence block. Each of these files contains more detailed breakdowns of the job counts by various demographic characteristics. Since this method is concerned primarily with creating linked origin-destination data on commuting by mode, the OD file is used, which (due to its specificity) contains only coarse demographic information, grouping jobs into large-scale age, earnings and industry categories. For this paper's application, we used the total number of jobs for each linked origin-destination Census block (field 'S000') from the LODES 2015 'JT00' (all jobs) OD file. We summed these block-level flows within their nesting Census tracts to match the scale of analysis of the ACS mode share data.

Unlike LODES, which relies fundamentally on administrative records, the ACS is a sample survey that is conducted at regular intervals. In order to produce a spatially representative estimate, responses for a given geographical unit are averaged over time; the smaller the unit, the longer the timespan needed to produce a reasonable estimate. At the Census tract scale, 5-year estimates (averages) are used to provide figures with the lowest possible margin of error, although error estimates can still be quite large (see the Margins of Error section in the Supplementary Material for more detail). The ACS collects a variety of information on demographics, housing conditions, and employment, but our interest here is solely on commuting information. The ACS questionnaire asks respondents to describe the means of transportation used to travel to work 'in the last week[3]' (Graham et al. 2014). For this application, we used the percentage of workers 16 and over who use each of these five means of transportation to work – 'car, truck, or van', 'public transportation (excluding taxicab)', 'bicycle', 'worked at home' and 'walked' – by Census tract from the 2013 to 2017 ACS.

Since the survey is administered to people at their place of residence, the ACS counts of commuting by mode are tabulated at the place of residence only, with no information on the

destination of these trips. Thus, the purpose of this paper is to combine this origin-based information on commuting by mode from the ACS with the total linked origin-destination flows from LODES to create estimates of commuting flows by mode over each individual O-D link.

Naive method - equal distribution

The most straightforward approach to applying origin-based rates of commuting by mode to the linked O-D flows from LODES – which we will call the 'naive' method here – is to multiply each mode's ACS commuting share (S) for a given tract (i) across all flows leaving that tract (T) to obtain an 'expected' count of flows (E) by mode m (Equation (1)).

$$E_{im} = S_{im}T_i \tag{1}$$

While this method produces estimates of flows by mode that preserves the correct share of commuting by mode leaving a tract, it unfortunately assumes that this share is *equally distributed* across all destinations (j). This means that for a tract a in a suburban location (e.g. Arlington Heights) with 2% walk commute mode share and a large number of flows to the central city (e.g. the Loop), this method will estimate that 2% of the commuting flows on this particular link (and, in fact, all links leaving tract a) are walking flows, despite a very long distance between tract a and the Loop (e.g. 35–40 km). Given a sufficiently large number of observed flows on a given link, especially in the context of a traditional suburban commuting pattern, this misestimate could be quite large. To solve this problem, and to produce more realistic estimates of commuting by mode, we developed the 'weighted' method described below.

Weighted method – distance and travel time redistribution

The primary intuition of the weighted method is to weight estimated flows (E_{ijm}) by some distance decay parameter based on the length of a given ij link and then to redistribute the 'excess' (i.e. unrealistic) flows across more realistic (i.e. nearby) links. To do this effectively, two different methods must be employed for the walking/cycling and transit modes due to their different characteristics[4].

Weighted method for walking and cycling. For walking and cycling trips, we first calculate the street network distance (in km) d_{ij} between tracts for all observed links in LODES based on the latitude and longitude coordinates of tract centroids snapped to the closest street segment from OpenStreetMap. This is calculated in R (using the r5r package[5]) for the nearly 440,000 unique O-D pairs in the dataset. A constant value of 0.05 km is used for intra-tract (i.e. on-diagonal) flows to avoid numerical issues with multiplying or dividing by 0. Then we find the weighted modal flows on each link (W_{ijm}) based on a mode-specific distance decay parameter (β_m) according to equation (2).

$$W_{ijm} = E_{ijm}d_{ij}^{-\beta_m} \tag{2}$$

The calculation of W_{ijm} in equation (2) is also subject to two constraints as follows: (1) a minimum threshold υ below which the raw number of estimated flows by mode (E_{ijm}) remains unweighted and (2) a maximum threshold μ above which the weighted number of flows = 0. We discuss empirical derivations for υ and μ for walking and cycling in Section 2.3.2. The full expression of the if-else statement defining the calculation of W_{ijm} according to these constraints is shown in equation (3).

$$\left(d_{ij} < \upsilon \Rightarrow W_{ijm} = E_{ijm}\right) \wedge \left(d_{ij} > \mu \Rightarrow W_{ijm} = 0\right) \wedge \left(\upsilon < d_{ij} < \mu \Rightarrow W_{ijm} = E_{ijm} d_{ij}^{-\beta_m}\right) \tag{3}$$

This method necessarily produces fewer flows than are observed in the original LODES data because long flows are removed and medium-distance flows are discounted by β. To maintain consistency with the original data, we need to redistribute these 'excess' flows (X_{ijm}) to nearby tracts. To do this, we start by finding excess flows by subtracting W_{ijm} from E_{ijm} (Equation (4)) to find excess flows on each link.

$$X_{ijm} = E_{ijm} - W_{ijm} \tag{4}$$

Now, we find the set of 'nearby' tracts N_{im} for each tract i (for each mode) by filtering the data to find all ij links within μ. This set serves as the realistic set of (nearby) possible destinations for redistributing the excess flows. We sum the total number of flows to all j tracts within this set for each i (Equation (5)) to find the total number of flows to nearby tracts (TN_i), which will serve as a denominator for redistribution.

$$TN_i = \sum_{j=1}^{N_{im}} T_{ij} \tag{5}$$

We also sum X_{ijm} for each tract i (across all observed ij links) to find the total excess flows by origin (TX_{im}) (Equation (6)):

$$TX_{im} = \sum_{j=1}^{n_{ijm}} X_{ijm} \tag{6}$$

Now, we divide TX_{im} across all nearby tracts (N_{im}) *weighted by* the ratio of flows on an individual link (T_{ij}) to the total number of flows from tract i to all nearby destinations (TN_i) to find each individual link's proportional redistributed flows by mode (ADD_{ijm}) (Equation (7)):

$$ADD_{ijm} = TX_{im}\left(T_{ij}/TN_i\right) \tag{7}$$

This is added to the weighted number of flows found in equation (3) to produce the final redistributed weighted estimate of modal flows on each link (F_{ijm}) (Equation (8)):

$$F_{ijm} = W_{ijm} + ADD_{ijm} \tag{8}$$

This approach has a couple of key advantages. First, excess trips are not *evenly* redistributed to nearby tracts, which would be unrealistic. Instead, the redistributions follow the pattern of the full LODES commuting data. This means that if a given ij link has a particularly high proportion of observed flows, the redistribution of excess walking and cycling flows is *weighted* by that overall proportion, so redistributions stack onto real commuting patterns appropriately. Second, this method preserves the total number of LODES flows leaving any given tract[6] (E_{im}), as well as the ACS commute mode share, while still providing a much more realistic estimate of flows by mode based on the distance of a given link than the naive method. For clarity, Figure S1 in the Supplementary Material provides a simplified visual example of the methodology described in Equations (1)–(8).

Defining parameters for weighting walking and cycling flows. How do we obtain estimates for the β, μ and υ parameters by mode? While our method for estimating walking and cycling flows is completely flexible[7] and allows these parameters to be customised for a given geographic area or use case; in this application, we have defined them based on empirical data from the 2009 National

Household Transportation Survey. First, we selected home-based work trips (from the 'Trips' file) for Metropolitan Statistical Areas (MSA) of 1 million or more with heavy rail in the East North Central Census Division. We then extracted all individual walking and cycling trip distances from this subset. We plotted the percentage of trips in 1-km distance bands and fit them with an exponential curve, as shown in Figure S2 in the Supplementary Material.

The slope term for these curves were used as the β (distance decay) parameters for our modelling purposes. In this case, $\beta_{WALK} = -.714$ and $\beta_{CYCLE} = -.329$. The μ parameter was determined by the maximum observed trip distance[8] in the subset. In this case, $\mu_{WALK} = 3.5$ and $\mu_{CYCLE} = 6.8$. Finally, υ was chosen based on trial and error according to the observed β coefficients. Given the relatively steep value for β_{WALK} and the high proportion of observed walking trips within 1 km, $\upsilon_{WALK} = 1$. Essentially, if we were to set υ_{WALK} to a smaller value, it would heavily discount a large proportion of very short (i.e. intra-tract) walking trips whose excess values would then be redistributed across the larger μ_{WALK}, which seemed inappropriate. Likewise, υ_{CYCLE} was also set to 1.

Weighted method for transit times using r5r. For the transit mode, discount and redistribution of excess flows by *distance* alone is not appropriate, as transit travel occurs at particular nodes and along particular networks. In this case, we need to know which links represent 'realistic' potential transit flows in order to create the transit-relevant set of 'nearby' tracts ($N_{iTRANSIT}$). Excess flows, then, are those leaving a given tract i on links where transit trips are not viable. These are redistributed according to each individual link's proportional flows by mode as described in Equations (7) and (8). In this case, $\upsilon_{TRANSIT}$ = links with 1–3 rides. Our assumption is that transit is a viable option only when a trip is actually taken (>0 rides), but trips with more than 3 transfers are also extremely unlikely to actually be taken for commuting purposes. No distance decay parameter was assigned to transit trips.

How do we obtain estimates of realistic transit links? The R package r5r (version 0.6.0) is used to query transit travel times to and from each tract centroid in the study area. This package, which stands for 'Rapid Realistic Routing with R5 in R' (Pereira et al., 2021), takes as inputs the General Transit Feed Specification (GTFS) and OpenStreetMap (OSM) data for a particular region and uses that data to calculate transit travel times from a set of origins to destinations directly in R (similar to OpenTripPlanner, but operable entirely in R). The user specifies a range of parameters, including the maximum walking distance (i.e. to a transit station) threshold, the maximum total travel time threshold, the start date/time and a fuzzy travel time window[9] for departure times. In this case, the outputs are based on the trip whose waiting times are closest to average within the departure window. Importantly, the 'breakdown' parameter provides a detailed output for each link, including whether or not transit was actually used for a particular link (i.e. if the calculated transit travel + waiting + transfer times ≤ walking time for a particular link) and the number of transit rides required to get to a given destination. Parameters used in this r5r application can be found in Table S1 in the Supplementary Material.

Results

In order to gauge (1) how well the method is preserving the data inputs and (2) the extent to which the results generally match other available sources of modal commuting data, we employ several internal and external validation approaches, which can be accessed in the Validation section of the Supplementary Material. Beyond validation, we can also assess the characteristics of the estimated flows by mode in order to better understand travel patterns in Cook County. Figure 1 shows the distribution of estimated flows by distance band and mode. Given its small overall mode share percentage (4.27%), walking makes up a fairly considerable proportion of trips within 5 km: 28.87%. Cycling trips are most heavily concentrated in the 2–5 km band, making up 3.62% of all

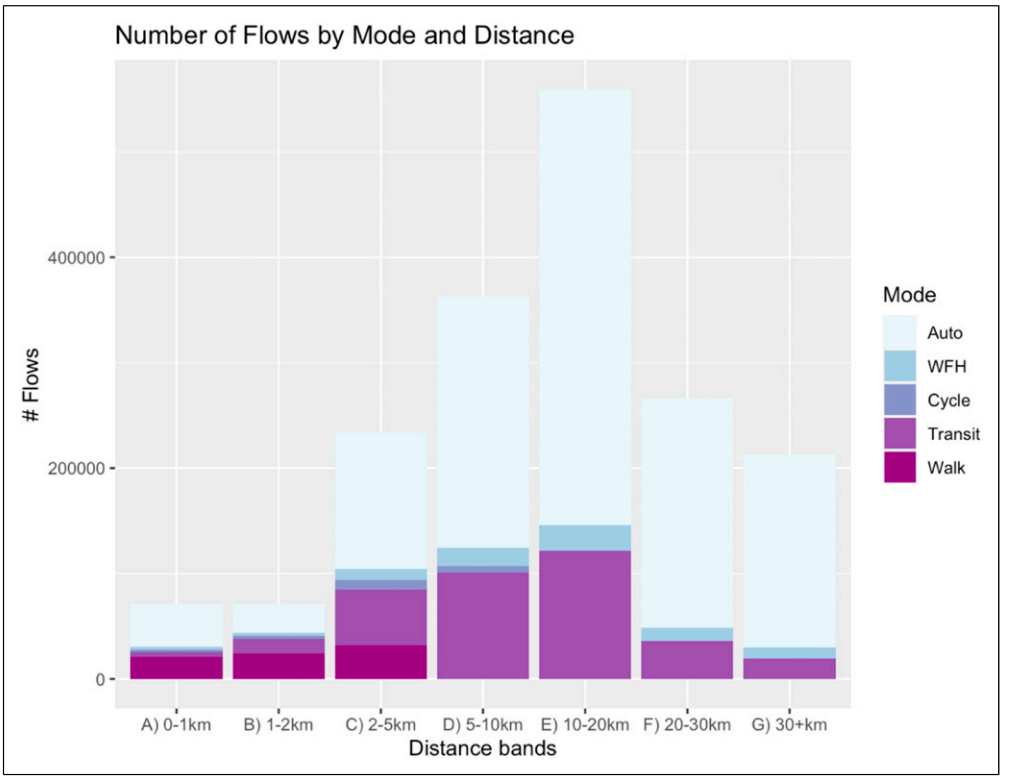

Figure 1. Histogram showing distribution of weighted trips by mode and distance.

trips at that distance compared to a 1.07% overall mode share. Transit trips, which aren't as heavily structured by distance, are most concentrated in the 1–20 km range. Transit demonstrates a 23.25% mode share for trips of these distances compared to an 19.40% mode share overall. Interestingly, across all distances – even within 1 km – auto travel outcompetes all other modes.

Beyond these descriptive characteristics, the most useful application of this method is to examine the spatial patterns of linked commuting flows by mode. Figure 2 shows maps of the weighted flows for links with the largest 1000 trips (to increase visual clarity of the key patterns) for each of the five modes of interest, created using the 'FlowMapper' tool (flowmapper.org). Commuting by auto is the most spatially dispersed, which is not surprising given that it is the most common mode of travel (both in Cook County and in the United States overall) and that auto infrastructure provides for rapid travel over long distances and is widely distributed and heavily invested in. Activity concentrates on the Loop, with the largest flows coming from neighbourhoods/suburbs such as Mount Greenwood, Edison Park, Midway and the South Loop. At the same time, we can observe a more widespread distribution of employment (in-commuting) sub-centres in the northwest portion of the county, for example, Schaumburg.

The pattern is generally similar for working from home, with a much heavier concentration of activity from near northside neighbourhoods like Lincoln Park to the Loop. Substantial 'flows' to Evanston (Northwestern University) are also visible. This makes sense given the relatively high concentration of professional, technical and creative occupations in these neighbourhoods. For transit, the overall pattern is relatively dispersed, but the Loop remains the dominant destination,

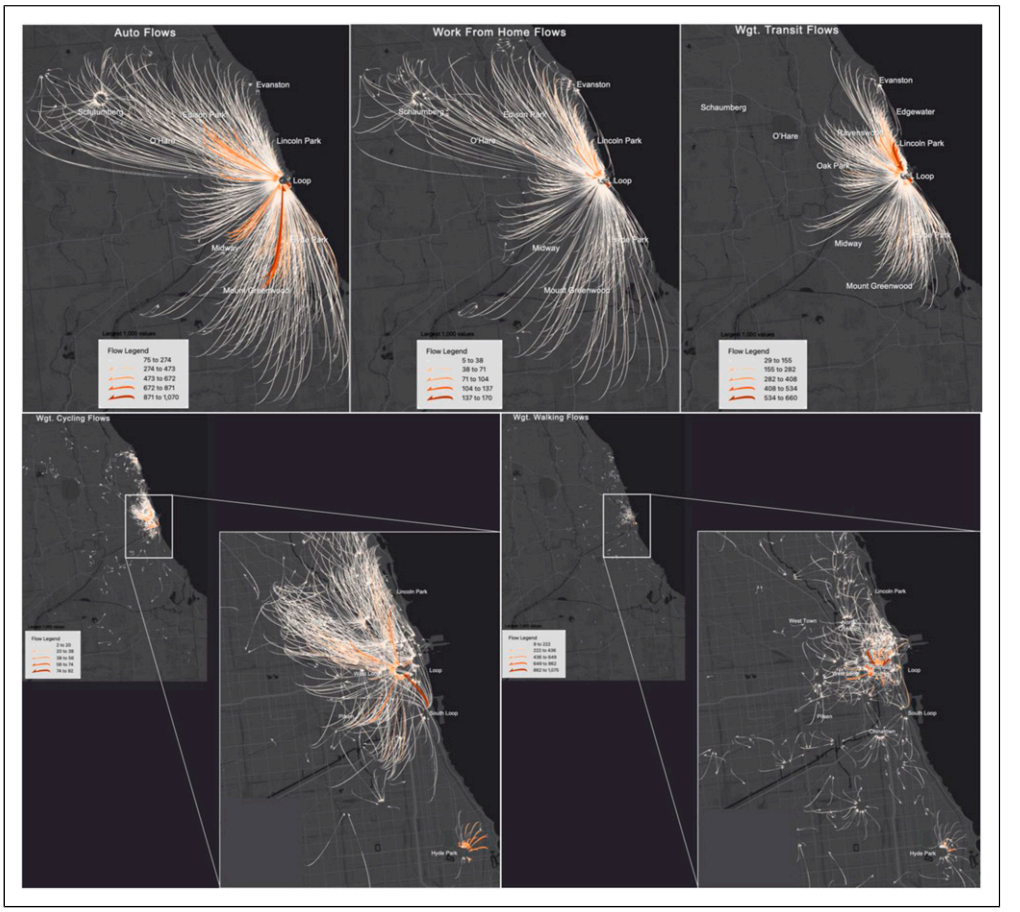

Figure 2. Five panel map showing flows for links with largest 1000 trips for auto (top left), work from home (top middle), transit (top right), cycling (bottom left) and walking (bottom right). Arrows denote direction of the flows.

with the largest flows coming from areas well-serviced by bus – including the South Loop and the near north, for example, Lincoln Park and Lakeview – as well as areas served by specific CTA 'L' lines such as Edgewater (Red Line), West Town, Oak Park (Blue Line) and Ravenswood (Brown Line). Evanston, on the CTA's Purple Line, also shows up as a prominent commuting *destination*.

For cycling, the pattern displays some inhibition distance, which makes sense given the method's parameters as well as the basic logic of the cycling mode, that is, walking is likely preferable for *very* short trips. We see the major university/employment centres, that is, the Loop (centre city), Hyde Park (University of Chicago) and Evanston (Northwestern University), show up strongly as cycling commuting destinations, dominated by commuting trips to the Loop. The largest of these flows come from gentrifying, young neighbourhoods such as Pilsen, Old Town, Lincoln Park, Lakeview, West Town and the South Loop, which matches our intuition about cyclists. These areas have also received some of the highest profile investments in cycling infrastructure in the city, including the Lakefront Trail, protected bicycle lanes on the Milwaukee Avenue bicycle corridor, and the 606/ Bloomingdale Trail, a previously abandoned elevated train line that has been renovated into a new pedestrian and cycling trail.

Finally, for the walking mode we see the largest concentrations of activity in major employment centres that are also located in walkable neighbourhood contexts, for example, in Hyde Park and the Loop. Interestingly, smaller (more walkable) sub-centres in nearby neighbourhoods are also evident, including the West Loop, Chinatown, South Loop and the North/Clybourn retail corridor. The overwhelming regional pattern, however, consists of nearby residential neighbourhoods commuting to the Loop (in particular, the financial district) by walking.

Discussion and conclusions

In this paper, we have developed a fully customizable method for estimating the number of commuting flows by the walking, cycling, transit, work from home and auto modes to and from every individual Census tract in Cook County, IL by combining origin-based information on commuting by mode from the ACS with the linked origin-destination flows from LODES. This method – which is implementable through the supplementary code available (https://github.com/kcredit/LODES-ACS-commuting-flows) – provides a useful resource for transportation planners, active transportation advocates and researchers to study spatial and temporal variation in auto and non-auto travel patterns.

Indeed, the resulting spatial patterns of travel by mode reveal a number of interesting characteristics. As expected, we observe high volumes of trips by all modes to regional employment centres such as the Loop, Evanston and Schaumberg. The auto and work from home modes provide the most distributed patterns of commuting, with prominent sub-centres in more outlying areas of the county (like Schaumberg). Outside of these areas, additional mode-specific patterns also emerge. The unique impact of cycling infrastructure is apparent in the Milwaukee Avenue bicycle corridor, while the pattern of transit trips tends to follow the largest Chicago Transit Agency (CTA) 'L' lines, including the Red, Blue, Purple and Brown Lines. A high volume of transit commuting to the Loop is also concentrated in the well-connected near northside neighbourhoods, which are served by a large number of frequent bus lines.

At the same time, the approach taken in this paper can be extended in a variety of useful ways. Given sufficient computational power, this analysis could theoretically be scaled to the state- or national-level for most years between 2002 and 2019 in order to better understand intra- and inter-regional travel patterns. Of course, expanding beyond the application here would require some additional work. Our method for estimating transit flows relies on r5r travel time data, which is available only in regions with large-scale transit networks. The procedure used to fit distance decay curves for the walking and cycling modes currently comes from 2009 NHTS data specific to the large cities in the region of the country in which Chicago is located. If this method were expanded to additional regions, region-specific curves would need to be fit, or a more generalised estimate used based on state or national data. Calculating tract-to-tract street network distances at scale could also be quite computationally expensive using r5r, and would require the manual download of area-specific street network data from OSM, so simple great circle distance calculated directly from the latitude and longitude coordinates of tract centroids may be preferable in that case[10].

Beyond characterizing aggregate patterns, this method could also be particularly useful for analysing specific spatially granular corridors and conditions of interest, for example, specific neighbourhood origin-destination relationships. Fine-grained temporal analyses, particularly in terms of changes in the work from home patterns, would also be insightful to examine in forthcoming iterations of the LODES data (2020–on) in order to better understand the changes in commuting due to the pandemic-era shift to remote work. Demographically, we could also look at finer-grained subsets of commuters based on existing breakdowns in the LODES O-D data based on age, income and broad industry classification. Further external validation for the walking and cycling outputs could be done by comparing to other large sources of data such as Strava – although

these data differ in terms of purpose (i.e. commuting vs. recreation), additional validation would be interesting and useful to users of both products.

Finally, the data created by this paper's method could be used as an important input to future research employing spatial interaction models to explicitly study the competing importance of various built environment and demographic 'push' and 'pull' factors on tract-level commuting. The role of built environment factors on commuting by mode, in particular, is one of the most-studied topics in urban and transportation planning, and could benefit from a more comprehensive analysis using this method's outputs.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. As Rasouli and Timmermans (2014, p.51) note in a mostly positive review of activity-based models, 'a lack of empirical evidence that activity-based models indeed perform better than four-step models [has] been the major [reason] for not adopting activity-based models in planning practice'.
2. For all but seven states for particular years: Alaska 2017–2019, Arizona 2002–2003, Arkansas 2002, 2019, DC 2002–2009, Massachusetts 2002–2010, Mississippi 2002–2003, 2019, and New Hampshire 2002 (US Census Bureau, 2021).
3. Important to note that in some cases this may skew modal estimates if travel taken in the last week is not typical for an individual.
4. In other words, what constitutes 'nearby' for walking and cycling trips is not the same as for transit, given the fact that transit trips occur only along specific infrastructural networks.
5. Specifically, we ran a 'walking' mode query with a large enough travel time threshold to capture all origins and destinations in Cook County (3000 h), which took about an hour to run on a standard laptop. Returned travel times are multiplied by the set walking speed parameter (5 kph) to obtain distances in km. More information on r5r can be round in Section 3.3.3.
6. For the modes of interest, based on the ACS commute mode share.
7. As shown in the associated code here: [redacted].
8. One extremely long distance outlier for walking trips (>3.9x larger than the second-largest distance) was removed.
9. From the r5r package documentation: 'Time window in minutes for which r5r will calculate multiple travel time matrices departing each minute. By default, the number of simulations is 5 times the size of "time_window" set by the user. Defaults window size to "1", the function only considers 5 departure times. This parameter is only used with frequency-based GTFS files…The travel_time_matrix function uses an R5-specific extension to the RAPTOR routing algorithm (see Conway et al., 2017). This RAPTOR extension uses a systematic sample of one departure per minute over the time window set by the user in the "time_window" parameter. A detailed description of base RAPTOR can be found in Delling et al. (2015)' (Pereira et al., 2021).
10. In our case, street network and great circle distances were correlated at 0.993.

References

Bagchi M and White PR (2005) The potential of public transport smart card data. *Transport Policy* 12(5): 464–474. DOI: 10.1016/j.tranpol.2005.06.008.

Bhat CR, Srinivasan S and Guo JY (2002) *Activity-Based Travel-Demand Modeling for Metropolitan Areas in Texas: Data Sources, Sample Formation, and Estimation Results*. Texas Department of Transportation. https://ctr.utexas.edu/wp-content/uploads/pubs/4080_3.pdf

Cervero R (2006) Alternative approaches to modeling the travel-demand impacts of smart growth. *Journal of the American Planning Association* 72(3): 285–295.

Chen P, Shen Q and Childress S (2018) A GPS data-based analysis of built environment influences on bicyclist route preferences. *International Journal of Sustainable Transportation* 12(3): 218–231. DOI: 10.1080/15568318.2017.1349222.

Cuauhtemoc A, Erath A and Fourie PJ (2017) Transport modelling in the age of big data. *International Journal of Urban Sciences* 21(51): 19–42. DOI: 10.1080/12265934.2017.1281150.

de Hartog JJ, Boogaard H, Nijland H, et al. (2010) Do the health benefits of cycling outweigh the risks? *Environmental Health Perspectives* 118(8): 1109–1116. DOI: 10.1289/ehp.0901747.

de Nazelle A, Nieuwenhuijsen MJ, Antó JM, et al. (2011) Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. *Environment International* 37(4): 766–777. DOI: 10.1016/j.envint.2011.02.003.

Frank LD, Andresen MA and Schmid TL (2004) Obesity relationships with community design, physical activity, and time spent in cars. *American Journal of Preventive Medicine* 27(2): 87–96.

Frank LD, Sallis JF, Conway TL, et al. (2006) Many pathways from land use to health: associations between neighbourhood walkability and active transportation, body mass index, and air quality. *Journal of the American Planning Association* 72(1): 75–87.

Graham MR, Kutzbach MJ and McKenzie B (2014) *Design Comparison of LODES and ACS Commuting Data Products*. Center for Economic Studies, US Census BUreau. https://www2.census.gov/ces/wp/2014/CES-WP-14-38.pdf

Harding C, Faghih Imani A, Srikukenthiran S, et al. (2021) Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys. *Transportation* 48(5): 2433–2460. DOI: 10.1007/s11116-020-10135-7.

Jariyasunant J, Sengupta R and Walker JL (2014) *Overcoming Battery Life Problems of Smartphones When Creating Automated Travel Diaries*. University of California Transportation Center. *UCTC-FR-2014-05*. https://trid.trb.org/view/1323139

Lee J, Vojnovic I and Grady S (2017) The 'transportation disadvantaged': urban form, gender and automobile versus non-automobile travel in the detroit region. *Urban Studies* 55(11): 2470–2498.

Lindstrom M (2008) Means of transportation to work and overweight and obesity: a population-based study in Southern Sweden. *Preventive Medicine* 46(1): 22–28.

Manson S, Schroeder J, Van Riper D, et al. (2020) *IPUMS National Historical Geographic Information System:*. Minneapolis, MN: IPUMS. [dataset] DOI: 10.18128/D050.V15.0. Version 15.0.

Metropolitan Washington Council of Governments (COG). (2018) *User's Guide for the COG/TPB Travel Demand Forecasting Model*. Washington, DC: National Capital Region Transportation Planning Board (TPB). Version 2.3.75 https://www.mwcog.org/transportation/data-and-tools/modeling/model-documentation/

Milne D and Watling D (2019) Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography* 76(April): 235–244. DOI: 10.1016/j.jtrangeo.2017.11.004.

Mueller N, Rojas-Rueda D, Cole-Hunter T, et al. (2015) Health impact assessment of active transportation: a systematic review. *Preventative Medicine* 76: 103–114. DOI: 10.1016/j.ypmed.2015.04.010.

Murakami E (2007) *Longitudinal Employment and Household Dynamics (LEHD): Understanding LEHD and Synthetic Home to Work Flows in 'ON the MAP.'* Federal Highway Administration. http://www.fhwa.dot.gov/planning/census_issues/lehd/

National Academies of Sciences. (2014) Engineering, and medicine. *Activity-Based Travel Demand Models: A Primer*. Washington, DC: The National Academies Press. DOI: 10.17226/22357.

National Academies of Sciences. (2012) Engineering, and medicine. *Travel Demand Forecasting: Parameters and Techniques*. Washington, DC: The National Academies Press. DOI: 10.17226/14665.

Neves A and Brand C (2019) Assessing the potential for carbon emissions savings from replacing short car trips with walking and cycling using a mixed GPS-travel diary approach. *Transportation Research Part A: Policy and Practice* 123: 130–146. DOI: 10.1016/j.tra.2018.08.022.

Pelletier MP, Trepanier M and Morency C (2011) Smart card data use in public transit: a literature review. *Transportation Research Part C-Emerging Technologies* 19(4): 557–568. DOI: 10.1016/j.trc.2010.12.003.

Pereira RHM, Saraiva M, Herszenhut D, et al. (2021) *Intro to r5r: Rapid Realistic Routing with R5 in R.* CRAN. https://cran.r-project.org/web/packages/r5r/vignettes/intro_to_r5r.html

Polzin S and Choi T (2021) *COVID-19's Effects on the Future of Transportation*. United States Department of Transportation, Office of the Assistant Secretary for Research and Technology. DOI: 10.21949/1520705.

Prelipcean AC and Yamamoto T (2018) Workshop synthesis: new developments in travel diary collection systems based on smartphones and GPS receivers. *Transportation Research Procedia* 32: 119–125. DOI: 10.1016/j.trpro.2018.10.023.

Rasouli S and Timmermans H (2014) Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences* 18(1): 31–60. DOI: 10.1080/12265934.2013.835118.

Rodrigue JP (2020) *The Geography of Transport Systems*. 5th ed. New York: Routledge.

Stabler B and Freedman J (2021) *ActivitySim: ActivityBased Travel Demand Modeling Built by and for Users*. Resource Systems Group. https://rsginc.com/

Tajaddini A, Rose G, Kockelman KM, et al. (2020) Recent progress in activity-based travel demand modeling: rising data and applicability. In: de Luca S, Di Pace R and Fiori C (eds), *Models and Technologies for Smart, Sustainable and Safe Transportation Systems*. https://www.intechopen.com/chapters/73240

Tanton R (2014) A review of spatial microsimulation methods. *International Journal of Microsimulation* 7(1): 4–25.

Toole JL, Colak S, Sturt B, et al. (2015) The path most traveled: travel demand estimation using big data resources. *Transportation Research Part C* 58(B): 162–177.

Tu W, Cao R, Yue Y, et al. (2018) Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *Journal of Transport Geography* 69(May): 45–57. DOI: 10.1016/j.jtrangeo.2018.04.013.

US Census Bureau. (2021) *LEHD Origin-Destination Employment Statistics (LODES) Dataset Structure Format Version 7.5*. US Census Bureau. https://lehd.ces.census.gov/data/lodes/LODES7/LODESTechDoc7.5.pdf

Vojnovic I and Darden JT (2013) Class/racial conflict, intolerance, and distortions in urban form: lessons for sustainability from the detroit region. *Ecological Economics* 96(C): 88–98.

Wang Y, Chau CK, Ng WY, et al. (2016) A review on the effects of physical built environment attributes on enhancing walking and cycling activity levels within residential neighbourhoods. *Cities* 50: 1–15.

Woodcock J, Edwards P, Tonne C, et al. (2009) Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *Lancet* 374(9705): 1930–1943. DOI: 10.1016/S0140-6736(09)61714-1.

Yang Y, Wang C, Liu W, et al. (2018) Understanding the determinants of travel mode choice of residents and its carbon mitigation potential. *Energy Policy* 115: 486–493.

Zandbergen PA (2009) Accuracy of iPhone locations: a comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS* 13(s1): 5–25.

Zhu L, Yu FR, Wang Y, et al. (2019) Big data analytics in intelligent transportation systems: a survey. *Ieee Transactions on Intelligent Transportation Systems* 20(1): 383–398. DOI: 10.1109/TITS.2018.2815678.