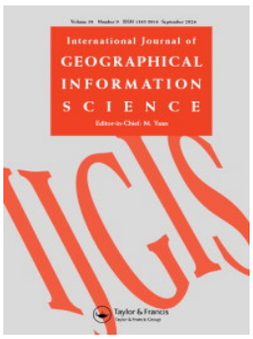

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis20

Detecting synthetic population bias using a spatially-oriented framework and independent validation data

Jessica Embury, Atsushi Nara, Sergio Rey, Ming-Hsiang Tsou & Sahar Ghanipoor Machiani

To cite this article: Jessica Embury, Atsushi Nara, Sergio Rey, Ming-Hsiang Tsou & Sahar Ghanipoor Machiani (2024) Detecting synthetic population bias using a spatially-oriented framework and independent validation data, International Journal of Geographical Information Science, 38:9, 1912-1938, DOI: 10.1080/13658816.2024.2358399

To link to this article: https://doi.org/10.1080/13658816.2024.2358399

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2024, VOL. 38, NO. 9, 1912–1938
https://doi.org/10.1080/13658816.2024.2358399

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Detecting synthetic population bias using a spatially-oriented framework and independent validation data

Jessica Embury[a], Atsushi Nara[a], Sergio Rey[a], Ming-Hsiang Tsou[a] and Sahar Ghanipoor Machiani[b]

[a]Department of Geography, San Diego State University, San Diego, CA, USA; [b]Department of Civil, Construction, and Environmental Engineering, San Diego State University, San Diego, CA, USA

ABSTRACT

Models of human mobility can be broadly applied to find solutions addressing diverse topics such as public health policy, transportation management, emergency management, and urban development. However, many mobility models require individual-level data that is limited in availability and accessibility. Synthetic populations are commonly used as the foundation for mobility models because they provide detailed individual-level data representing the different types and characteristics of people in a study area. Thorough evaluation of synthetic populations is required to detect data biases before the prejudices are transferred to subsequent applications. Although synthetic populations are commonly used for modeling mobility, they are conventionally validated by their sociodemographic characteristics, rather than mobility attributes. Mobility microdata provides an opportunity to independently/externally validate the mobility attributes of synthetic populations. This study demonstrates a spatially-oriented data validation framework and independent data validation to assess the mobility attributes of two synthetic populations at different spatial granularities. Validation using independent data (SafeGraph) and the validation framework replicated the spatial distribution of errors detected using source data (LODES) and total absolute error. Spatial clusters of error exposed the locations of underrepresented and overrepresented communities. This information can guide bias mitigation efforts to generate a more representative synthetic population.

ARTICLE HISTORY
Received 13 October 2023
Accepted 17 May 2024

KEYWORDS
Population synthesis; data validation; data bias; activity-based model; agent-based model

1. Introduction

Models of human mobility are used to address complex scenarios that cannot, or should not, be readily replicated in the real world. Through the simulation of individual-based mobility, we can test scenarios related to viral transmission (Silva *et al*. 2020, Kerr *et al*. 2021, Truszkowska *et al*. 2021, 2022), public health policy (Epstein, 2009, Tracy, et al. 2018), emergency evacuation strategies (Torrens, 2018,

CONTACT Jessica Embury jembury8568@sdsu.edu

Trivedi and Rao, 2018), transportation management (Benenson *et al*. 2008, Scherr *et al*. 2020), and urban development (Batty, 2005, Ligmann-Zielinska and Jankowski, 2007, Torrens and Nara, 2012). However, access to the individual-level data required to build individual-based mobility models is limited in its availability and accessibility (Crooks *et al*. 2008, Anderson and Dragićević, 2020, Heppenstall *et al*. 2020).

Population synthesis can be used to generate a synthetic (i.e. artificial) population of individuals for applications which require individual-level data that is not available elsewhere. Synthetic populations represent the different types and various characteristics of individuals in a study area's population. Depending on the application, individuals in a synthetic population may be assigned to a household with additional household attributes. The aggregate characteristics of the individuals, and their households, should be representative of the entire study area as well as the smaller spatial units where individuals reside. While recently developed population synthesis tools (e.g. Chapuis *et al*. 2021, Salat *et al*. 2023) can ease the process of generating a synthetic population, their internal validation processes for aggregate populations may not detect errors at finer scales, or for individuals.

Thorough evaluation of synthetic populations is crucial for detecting and minimizing pre-existing, technical, and emergent data biases: pre-existing biases exist in a model's source data and can be transferred to its output data, technical biases are the result of the modeling process (e.g. overfitting), and emergent biases arise as the output data are used and depend on the context of the application (e.g. transportation planning). Synthetic population biases must be addressed because they 'systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others' (Friedman and Nissenbaum, 1996, p. 332).

Despite their common use as a foundation for mobility modeling, synthetic populations are conventionally evaluated by their sociodemographic characteristics rather than their mobility attributes. While there are positive associations between sociodemographics and mobility (Lenormand *et al*. 2015), the evaluation of a synthetic population's mobility attributes has the potential to directly improve the accuracy of individual-based mobility models.

Mobility microdata – large, fine-resolution data sets with detailed location information collected from mobile applications – provide the means to validate the mobility attributes of synthetic populations. In addition, mobility microdata can be used as external/independent validation data, not otherwise used during population synthesis. The use of independent data during validation can improve the detection of data biases (Cambridge Systematics, Inc., 2010). Despite this novel opportunity, mobility microdata also contain data biases that underrepresent underserved and/or vulnerable individuals (Rodriguez-Carrion *et al*. 2018, Schlosser *et al*. 2021, Sourbati and Behrendt, 2021). Awareness of data bias within the validation data is important to consider when interpreting the results of synthetic population evaluation.

For our study, we used a population synthesis technique (i.e. iterative proportional updating) to generate synthetic populations at two spatial granularities (Census Block Groups and Sub-Regional Areas). Then, we developed a framework for

spatially-oriented validation using independent mobility microdata to detect and characterize biases. Our study was guided by two research questions:

1. Is the usage of mobility microdata suitable for validating the mobility attributes of synthetic populations?
2. Does synthetic population validation using a spatial framework and diverse data sources add value to the model and its subsequent applications?

This paper addresses a gap in the literature and demonstrates that the validation of mobility attributes complements existing methods of synthetic population evaluation. The demonstrated data validation method has the potential to improve the accuracy of synthetic populations, following further calibration, and the subsequent realism of individual-based mobility models.

2. Related work

2.1. Pre-existing data bias

Pre-existing data biases are present in a synthetic population's source and validation data in the form of data generation bias, which is introduced during data collection and/or compilation. Ethical concerns about data generation biases in mobility microdata and other forms of geospatial microdata are well documented in the literature (Wesolowski *et al*. 2013, Rodriguez-Carrion *et al*. 2018, Coston *et al*. 2021, Schlosser *et al*. 2021, Sourbati and Behrendt, 2021). For instance, mobility microdata collected from mobile devices is only representative of mobile device owners, and ownership varies across different geographies and sociodemographic characteristics (Wesolowski *et al*. 2013). Older individuals are particularly underrepresented by mobility microdata (Sourbati and Behrendt, 2021); in contrast, higher income individuals tend to be overrepresented (Schlosser *et al*. 2021). Administrative data has been used to test mobility microdata for bias and reliability (Coston *et al*. 2021). Although strategies exist for mitigating data generation bias (Estabrooks and Japkowicz, 2001, Rodriguez-Carrion *et al*. 2018, Mohammed *et al*. 2020, Schlosser *et al*. 2021), debiased data are only estimates, at best. The development of strategies to mitigate data generation bias is an ongoing research effort.

In addition to data generation biases, synthetic populations are susceptible to statistical data biases stemming from the spatial aggregation of their source data. Spatially aggregated data is subject to the modifiable areal unit problem (MAUP), a source of statistical bias caused by the data's geographic scale and zonal boundaries (Openshaw and Taylor, 1979). In synthetic populations, the MAUP tends to manifest in the form of increased error at finer spatial granularities (Harland *et al*. 2012). Despite this widespread source of data bias, examples of multiscale population synthesis studies are limited.

2.2. Technical data bias

The process of population synthesis can introduce technical biases into synthetic populations. The oldest method of population synthesis is thought to be iterative

proportional fitting (Deming and Stephan, 1940, Beckman *et al*. 1996); this technique estimates individual attribute values, known as marginal controls, in contingency tables (i.e. sampling distributions) using aggregate attribute values as constraints (i.e. targets) (Castiglione *et al*. 2014, Lomax and Norman, 2016). Iterative proportional fitting was the dominant method of population synthesis described in the literature until the mid-2000s and remains a valid option for applications which do not require individuals to have household membership. However, because iterative proportional fitting is ill-equipped to assign attributes to both individuals and households, technical bias favoring either individual or household characteristics is introduced into synthetic populations in which individuals are household members (Guo and Bhat, 2007, Zhu and Ferreira, 2014).

Subsequent population synthesis methods, such as optimization and probabilistic approaches, have advanced upon iterative proportional fitting to generate more accurate synthetic populations. Optimization approaches, such as iterative proportional updating (IPU), address iterative proportional fitting's primary source of technical bias by using optimization models to minimize the differences between summed attribute values and aggregate constraints for both individual and household characteristics (e.g. Ye *et al*. 2009, Abraham *et al*. 2012). However, synthetic populations generated using an optimization approach tend to reflect the pre-existing biases of their source data (Ramadan and Sisiopiku, 2019). On the other hand, probabilistic approaches produce synthetic populations that are not as likely to share the source data's pre-existing biases, but can introduce technical bias from model overfitting (e.g. Farooq *et al*. 2013, Sun and Erath, 2015, Fournier *et al*. 2021, Kukić and Bierlaire, 2022, Zhou *et al*. 2022).

Machine learning approaches to population synthesis offer new ways to manage and analyze increasingly large data sets with high dimensionality (e.g. Saadi *et al*. 2016, Borysov *et al*. 2019, Alonso-Betanzos *et al*. 2021), yet their limited explainability, due to 'black box' decision-making, may introduce hidden technical biases. Like emergent data biases, examples of technical bias originating from machine learning methods of population synthesis can be difficult to detect and are not well documented.

2.3. Synthetic population evaluation

Synthetic population evaluation is performed through verification, calibration, and validation; the inclusion of each assessment is crucial to the detection and minimization of bias. Verification, frequently performed before population synthesis, evaluates the soundness of the method's conceptual, logical, and physical models to correct errors and limit technical biases (Crooks *et al*. 2008). Calibration, conducted throughout population synthesis, refers to the processes of quantifying the model's uncertainty, as well as the fine-tuning of the physical model and its parameters to improve the synthetic population's alignment to its source data, especially at finer resolutions (Castiglione et al. 2014, Crooks *et al*. 2008). While calibration might effectively reduce technical bias, it can encourage the transference of pre-existing bias to the synthetic population. Validation, which takes place after population synthesis, has a critical role in gauging the extent to which technical and pre-existing bias transference has occurred.

The purpose of validation is to determine how well a synthetic population represents its real-world counterpart (Manson *et al*. 2012). The use of independent/external validation data identifies data biases that reduce accuracy (Cambridge Systematics, Inc., 2010). In the best-case scenario, synthetic population validation would use 'ground truth' data collected from all individuals in the study area. However, the collection of ground truth validation data is not feasible for most, if not all, synthetic populations. Historically, the lack of available fine-scale and individual data essentially prevented the independent/external validation of synthetic populations (Crooks *et al*. 2008, Heppenstall *et al*. 2020). Adapted validation techniques that use source data (i.e. internal data validation) are normalized and broadly accepted for synthetic population validation.

Measures regularly used to validate a synthetic population include total absolute error, standard root mean square error, Pearson correlation coefficient, and coefficient of determination (Harland *et al*. 2012, Lovelace and Dumont, 2016, Borysov *et al*. 2019, Prédhumeau and Manley, 2023). For all methods, aggregated sociodemographic characteristics of the synthetic population are compared to administrative data to gauge error in the synthetic population. None of these methods consider the spatial variation or spatial autocorrelation of errors. Typical validation data is either untabulated microdata, like the US Census Bureau's Public Use Microdata Sample, or an aggregated data overview, such as the US Census Bureau's American Community Survey.

Total absolute error directly compares the synthetic population to the validation data and provides the number of synthetic individuals that have been misclassified for each attribute, or all attributes at once (Harland *et al*. 2012, Wu *et al*. 2022, Prédhumeau and Manley, 2023); this measure is the most straightforward representation of error. Standard root mean square error builds upon total absolute error but emphasizes large errors with higher values (Müller and Axhausen, 2011, Sun and Erath, 2015, Borysov *et al*. 2019, Wu *et al*. 2022).

The Pearson correlation coefficient is a measure of the linear relationship strength between two datasets, but it does not provide a head-to-head data comparison (Gartlehner and Moore, 2008, Niroumand *et al*. 2013). The Pearson correlation coefficient is typically used to complement and build confidence in other validation measures (Borysov *et al*. 2019, Prédhumeau and Manley, 2023). The coefficient of determination is the square of the Pearson correlation coefficient, representing the variance of error in the synthetic population (Borysov *et al*. 2019, Wu *et al*. 2022). Similar to the Pearson correlation coefficient, the coefficient of determination is best used in tandem with other validation measures (Renaud and Victoria-Feser, 2010).

Since synthetic populations are commonly used to understand transportation and movement (e.g. Guo and Bhat, 2007, Bradley *et al*. 2010, Zhu and Ferreira, 2014, Trivedi and Rao, 2018, Scherr *et al*. 2020, Kianersi *et al*. 2021, Wang *et al*. 2022), the validation of mobility characteristics, in addition to sociodemographic attributes, would provide valuable information about biases that are likely to emerge during applications. Although not addressed by the existing literature, mobility microdata presents an opportunity to independently validate a synthetic population's mobility characteristics. Large trajectory data sets have already been shown to explain observed mobility patterns (Jin *et al*. 2023). Mobility microdata provides information about a

greater proportion of individuals in the study area than travel surveys that are often used for model development. Because it is more comprehensive than the source data, mobility microdata may serve as a substitute for ground truth data during synthetic population validation; further, the increasing availability of mobility microdata makes it a practical alternative. While independent data validation with mobility microdata might result in more representative synthetic populations, we must consider pre-existing biases in the validation data (Section 2.1).

3. Materials and methods

3.1. Study area

The study area was San Diego County, CA. San Diego County contains 1794 Census Block Groups (CBGs) and 41 Sub-Regional Areas (SRAs), which are Census Tract aggregations delineating the county's larger neighborhood regions. The county's residents are demographically diverse and reside in urban, suburban, and rural communities. Moreover, San Diego County's presence along the US-Mexico border further complicates its residents' mobility and activity dynamics.

This study focused on synthesizing populations of residents living in moderate-density to high-density communities. Since most of San Diego County's residents live in urban or suburban communities, rural areas with population densities below 100 residents per square mile were omitted. In addition, due to their unique attributes, two military bases were removed. The filtering process created a mostly contiguous study area of 1756 CBGs and 34 SRAs. The study area had a 'fuzzy' eastern boundary because the selected SRAs extended slightly beyond the selected CBGs.

3.2. Study data

This study used six data sets to generate and evaluate synthetic populations at two spatial granularities (CBGs and SRAs) (Table 1). The boundaries for CBGs and SRAs (San Diego Association of Governments, 2015a, 2015b) were spatially joined to the study data in order to perform spatial aggregations by CBG and SRA, and spatial analysis (Table 1, Rows 1–2). Administrative data sets used for population synthesis, calibration, and internal data validation included a travel survey (individual sociodemographic and mobility data) (State of California, 2018), a community survey (fine-scale sociodemographic data) (US Census Bureau, 2022a) and commuter data (fine-scale origin-destinations and industry data) (US Census Bureau, 2022b) (Table 1, Rows 3–5); while generally considered reliable, these data sets include pre-existing nonresponse bias (Shapiro, 2001).

The sixth data set was obtained from a private data company, SafeGraph (SafeGraph, 2023), and was used for independent/external data validation (Table 1, Row 6). We accessed SafeGraph's 'Monthly Patterns' data for 2019 using the Dewey API (https://www.deweydata.io/). The 'Monthly Patterns' data sets contain monthly aggregated information about the number of devices (i.e. device counts), recorded continuously (24 hours per day, 7 days per week), at individual points of interest. We compiled the data to create a data set containing the total device counts for 2019 for

Table 1. The study datasets and their sources, years, original temporal and geographic extents, and descriptions.

Dataset	Source	Year	Original temporal extent	Original geographic extent	Description
Census Block Groups	US Census Bureau's TIGER/Line shapefiles, accessed through San Diego Association of Governments	2010	N/A	Census Block Groups (CBGs)	Spatial dataset containing the 2010 Census Block Group boundaries for the study area. 2010 administrative boundaries are valid for the temporal extent of the study data.
Subregional Areas	San Diego Association of Governments	2010	N/A	Sub-Regional Areas (SRAs)	Spatial dataset containing the 2010 Sub-Regional Area boundaries for the study area. 2010 administrative boundaries are valid for the temporal extent of the study data.
National Household Travel Survey geocoded data (NHTS)	Department of Transportation, State of California	2017	Daily	Individuals	Travel survey data that includes tables for geocoded trips, household characteristics, individual characteristics, and daily travel diaries. The tables are joined by unique identifiers assigned to households and individuals.
American Community Survey 5-Year Estimates (ACS)	US Census Bureau	2017	Annual	CBGs	Comprehensive socio-demographic data that is derived from a survey of US residents.
LEHD Origin-Destination Employment Statistics Residence Area Characteristics (LODES)	Longitudinal Employer-Household Dynamics, US Census Bureau	2017	Annual	Census Blocks	Commuter data, compiled from administrative records and survey data, that specifies where workers in different industry sectors (i.e. NAICS) live.
SafeGraph Monthly Patterns (SafeGraph)[a]	SafeGraph[a]	2019	Monthly	CBGs	Detailed information about visits to an extensive list of places of interest, described using top categories. SafeGraph does not differentiate between work and non-work visits to places of interest. 2017 data was unavailable.

[a]SafeGraph (https://www.safegraph.com) is a private data company that aggregates anonymized location data from numerous mobile applications in order to provide insights about physical places, via the SafeGraph Community (i.e. data users conducting research and analysis). To enhance privacy, SafeGraph excludes information if fewer than five devices visited an establishment in a month from a given CBG.

all points of interest in the study area. SafeGraph data is biased by data suppression performed when fewer than five device counts from an origin CBG were recorded at a point of interest during a given month. SafeGraph data also shares data generation biases that are common to most mobility microdata including the underrepresentation of older individuals who may not own or carry a mobile device, and the overrepresentation of affluent individuals who are more likely to carry multiple devices (Schlosser *et al*. 2021, Sourbati and Behrendt, 2021). In this study, we assume that one device count equals a trip by a single individual.

3.3. Population synthesis and calibration

The marginal controls for population synthesis were selected using the administrative data. An attribute was eligible for selection if it was present in the travel survey and either the community survey or commuter data, and if the attribute did not include suppressed data. The following household and individual marginal controls were used for the study:

- Household: household size, annual household income, number of workers, number of vehicles, and housing status (i.e. own or rent)
- Individual: age, sex, racial identity, Hispanic identity, work status, work category, and education completed.

From the community survey, 174 attributes related to the marginal controls were selected and combined to match the less detailed format of the travel survey (e.g. the total number of different age groups was reduced). From the commuter data, the number of workers across 23 industry sectors, combined into four work categories, were used for the work status and work category. The processed community survey and commuter data provided the target values for population synthesis.

Next, Iterative Proportional Updating (IPU) and the *ipfr* R package (Ward, 2020) were used to create a multidimensional sampling distribution, a contingency table with a 'dimension' for each marginal control. IPU takes an optimization approach that iteratively balances the household and individual marginal controls and returns an optimal solution with minimal differences between the sampling distribution and target values; below is the formula for the IPU optimization model (Equation 1) (Ye *et al*. 2009):

$$\begin{aligned} &\textit{Minimize } \sum_j \left[\sum_i \left(d_{i,j}w_i - c_j\right)/c_j\right]^2 \textit{ or } \sum_j \left[\left(\sum_i d_{i,j}w_i\right)^2/c_j\right] \textit{ or } \sum_j \left[\left|\left(\sum_i d_{i,j}w_i - c_j\right)\right|/c_j\right] \\ &\quad \textit{Subject to } w_i \geq 0, \\ &\quad \textit{where } i \textit{ denotes a household } (i = 1, 2, \ldots, n), \\ &\qquad j \textit{ denotes the constraint or population characteristic of interest } (j = 1, 2, \ldots, m), \\ &\qquad d_{i,j} \textit{ represents the frequency of the population characteristic(household or person type)} j \textit{ in household}, \\ &\qquad w_i \textit{ is the weight attributed to the ith household, and} \\ &\qquad c_j \textit{ is the value of the population characteristic } j \end{aligned} \tag{1}$$

Then, households from the travel survey, along with their individual members, were randomly sampled based on the sampling distribution. Households ($n = 2666$) were

included during sampling if one or more of that household's individuals ($n = 5749$) made a geocoded trip within the study area.

Last, the model was fine-tuned by calibrating the parameters (Table 2). The model used minimum and maximum ratio parameters to set how many times a household record could be sampled, a maximum iteration parameter, and a secondary importance parameter that balanced optimization priorities between household and individual marginal controls. For minimum ratio, maximum ratio, and secondary importance, the selected parameter value was a threshold where further fine-tuning no longer produced a detectable reduction in overall differences between the synthetic populations' aggregated attribute values and the target values. For maximum iteration, there was little discernible difference between the tested parameter values, so the lowest value was selected.

The population synthesis process described in this section was performed for both CBGs and SRAs in order to assess synthetic populations at different spatial granularities. Figure 1 illustrates the population synthesis and calibration procedure.

3.4. Synthetic population validation

The synthetic populations were validated based on two types of mobility attributes. The first group of attributes included the number of commutes originating from each CBG/SRA, by work category and in total. The study's four work categories were clerical or administrative, labor (i.e. manufacturing, construction, maintenance, and farming), professional (i.e. professional, management, and technical), and sales or service. These attributes were internally validated by the commuter data. These mobility attributes can be compared head-to-head using the synthetic population and commuter data.

Table 2. Parameter values tested during calibration of the population synthesis model.

Parameter	Minimum allowed	Maximum allowed	Minimum tested	Maximum tested	Step size	Selected
Minimum ratio	0	1	0.05	0.2	0.05	0.1
Maximum ratio	0	1	5	20	5	10
Maximum iteration	1	N/A	100	500	100	100
Secondary importance	0	1	0.7	1	0.1	0.8

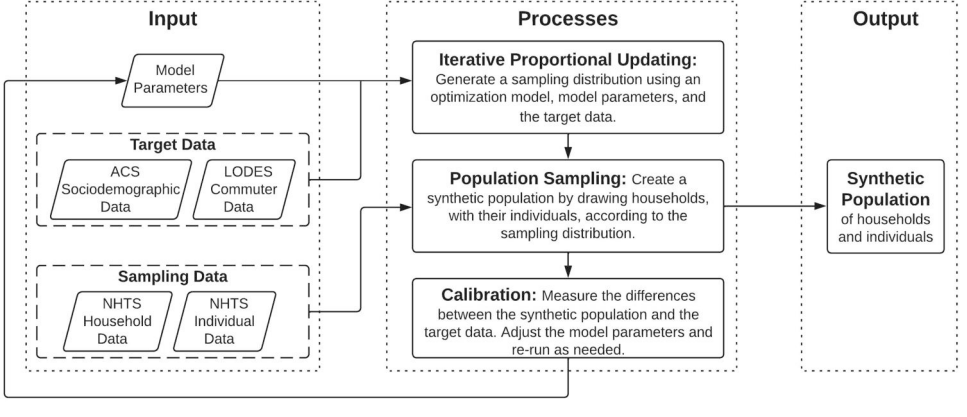

Figure 1. The methodological framework for population synthesis and model calibration.

The second group of attributes included the number of trips originating from each CBG/SRA, by trip purpose and in total. The eight trip purposes (i.e. activity types) selected from the travel survey were attending school as a student, buying goods, buying meals, buying services, health care visits, other general errands, recreation, and religious or community activities. These attributes were independently/externally validated using SafeGraph data. Because SafeGraph data provides the destinations of trips (i.e. points of interest), rather than trip purposes, we made a crosswalk to join the synthetic populations' trip purposes to the destinations' industry classifications. Additionally, since SafeGraph records data for work and non-work trips, we created a second crosswalk to match the synthetic populations' work categories to the destinations' industry classifications. For example, the 'restaurants and other eating places' industry classification was matched to the 'buying meals' trip purpose and the 'food services' work trip. If multiple industry classifications were matched to a work category, then the number of commutes were proportionally divided among them. Using the crosswalks, the number of work and non-work trips originating from each CBG/SRA were compared to the number of SafeGraph trips made to destinations with matched industry classifications. For these mobility attributes, we must use indirect comparison techniques, rather than a direct one-to-one comparison, because we are validating one day of the synthetic populations' trips to a year of aggregated SafeGraph data. Further, SafeGraph does not capture data for all individuals in the study area. The assessment of SafeGraph data's suitability for independent/external data validation was one of the study's two objectives.

Three methods were used to validate the synthetic populations. The first two methods are well-established for synthetic population validation, while the third is a novel spatially-oriented validation framework. For the first method, we determined the total absolute error (TAE) in the number of commutes originating from each CBG/SRA of the synthetic populations, using the commuter data. For the second method, we calculated Pearson correlation coefficients to evaluate the strength of the linear relationship between the synthetic populations' mobility attributes and the validation data. Since the Pearson correlation analysis is not a direct head-to-head comparison, both groups of mobility attributes were validated.

The third method follows a three-step spatially-oriented validation framework. First, we calculated the mobility attribute differences between the synthetic populations and the validation data. Percentage differences between the synthetic populations and commuter data were used to directly validate the number of commutes originating from each CBG/SRA (Equation 2). Meanwhile, percentile differences between the synthetic populations and the SafeGraph data were used as a normalized measure to indirectly validate the number of trips originating from each CBG/SRA (Equation 3). Since we assume that validation data is less biased than the synthetic populations, these mobility differences represent error in the synthetic populations.

$$Percentage\ Difference_i = \frac{NWSP_i - NWL_i}{NWL_i} \times 100$$

$$where\ NWSP_i\ is\ the\ number\ of\ workers\ estimated\ by\ the\ synthetic\ population\ at\ a\ geographic\ region\ i, and$$

$$NWL_i\ is\ the\ number\ of\ workers\ obtained\ from\ LODES\ at\ a\ geographic\ region\ i \quad (2)$$

$$Percentile\ Difference_i = \frac{RSP_i}{NSP} \times 100 - \frac{RSG_i}{NSG} \times 100$$

where RSP_i *is the rank number of trips estimated by the synthetic population at a geographic region* i,
NSP is the total number of geographic regions in the synthetic population,
RSG_i *is the rank number of trips obtained from the SafeGraph data at a geographic region* i, *and*
NSG is the total number of geographic regions in the SafeGraph data

(3)

Next, the mobility differences were spatially analyzed to characterize the distribution of error in the synthetic populations. Using the Global Moran's I, we determined whether the percentage/percentile differences displayed spatial autocorrelation. After confirming global spatial autocorrelation, hot and cold spots were identified using local spatial autocorrelation (Anselin, 1995). We assigned cluster designations (i.e. high-high or low-low) to spatial units if their Local Moran's I was statistically significant at a confidence level of 95% ($p <= 0.05$). A high-high designation indicated that both the spatial unit and its neighbors have high values (i.e. positive mobility differences). Similarly, a low-low designation indicated that the spatial unit and its neighbors have low values (i.e. negative mobility differences). Clusters of high-high ($p <= 0.05$) or low-low ($p <= 0.05$) spatial units were merged using a dissolve operation to identify regions (i.e. dissolved clusters) where positive spatial autocorrelation occurred.

To visualize the spatial units' percentage/percentile differences, we mapped their standard deviations from the mean. Positive deviations from the mean indicated that the synthetic population's values were higher than those in the validation data (i.e. overrepresentation), whereas negative deviations indicated that the synthetic population's values were lower than the validation data (i.e. underrepresentation). A difference of zero meant there was no difference between the synthetic population and validation data. The dissolved clusters were added to the map to reinforce the locations of clusters. Last, by interpreting the spatial distributions of percentage/percentile differences, we characterized bias within the synthetic populations.

The first two validation methods built a foundation for the spatially-oriented validation framework. The results of the first method served as a baseline for assessment, and the findings from the second method justified the usage of mobility differences as a validation measure. The establishment of a spatial validation framework that supports direct and indirect comparisons to validation data was the study's second objective. Figure 2 summarizes the entire three-method validation process.

4. Results

4.1. Synthetic population composition

This section compares the total size and overall sociodemographic compositions of the synthetic populations to administrative data.

The synthetic population generated for SRAs (SP-SRA) had a total population of 3.09 million individuals in 1.13 million households, and the synthetic population generated for CBGs (SP-CBG) had 2.85 million individuals in 1.09 million households. The

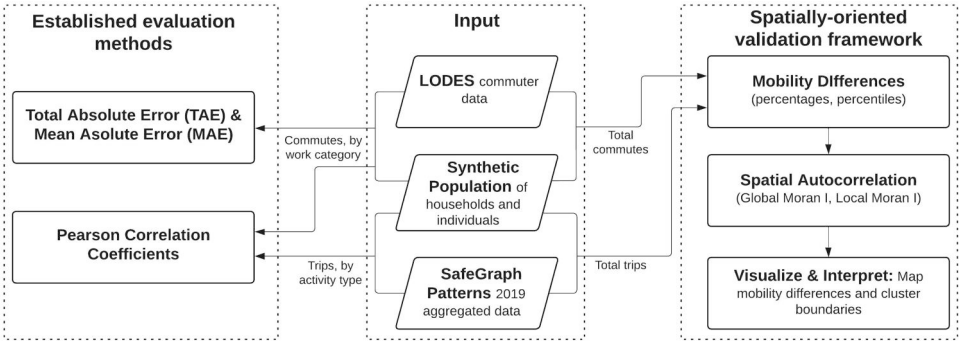

Figure 2. The process for validating the synthetic populations' worker origins with internal validation data (LODES), and the synthetic populations' trip origins and activities with independent/external validation data (SafeGraph), using two established validation methods (left) and a spatially-oriented validation framework (right).

total populations of SP-SRA and SP-CBG were 4.7% and 12.0% smaller than the administrative data's population size ($n = 3{,}237{,}526$), respectively. There were 2.6% more households in SP-SRA ($n = 1{,}125{,}462$) and 0.6% fewer households in SP-CBG ($n = 1{,}090{,}164$) than in the administrative data ($n = 1{,}096{,}758$). In addition, when compared to the administrative data's working population size ($n = 1{,}392{,}686$), SP-SRA ($n = 1{,}391{,}458$) had 0.1% fewer workers while SP-CBG ($n = 1{,}441{,}229$) had 3.5% more workers.

Next, we assessed the synthetic populations' sociodemographic compositions for each of the marginal controls. For the household attributes (household size, annual income, number of workers, number of vehicles, and housing status [i.e. own or rent]), the compositions of the synthetic populations were similar to the administrative data, with two notable exceptions. First, both synthetic populations had fewer households with one worker and more households with two workers. Second, SP-CBG had relatively more households living in owned housing than either SP-SRA or the administrative data.

There were more discrepancies for the individual attributes (age, sex, racial identity, Hispanic identity, work status, work category, and education completed). For the age category, SP-SRA and SP-CBG had fewer individuals that were 18–34 years old than the administrative data, with corresponding discrepancies for the education completed (lower number of N/A values) and work (lower number of N/A values) categories. Also, SP-CBG had fewer individuals of Hispanic ethnicity than either SP-SRA or the administrative data. Table 3 contains the complete results of the composition analysis.

4.2. Validation measure 1: total absolute errors

To obtain a baseline with which to compare the results of validation using the spatially-oriented validation framework, we calculated the total absolute error (TAE) in the number of commuters originating from each spatial unit, using the commuter data set. We also calculated summary statistics for the absolute errors to reveal variation across the spatial units.

Table 3. Comparison of the household and individual sociodemographic compositions of SP-SRA and SP-CBG to the administrative source data.

Attribute			Administrative Data[a]	(%)	SP-SRA	(%)	SP-CBG	(%)
Household attributes	*Household size (number of household members)*	1	263,096	24.0	240,919	21.4	255,479	23.4
		2	356,193	32.5	346,707	30.8	354,626	32.5
		3	188,719	17.2	215,050	19.1	201,676	18.5
		4	161,145	14.7	181,620	16.1	166,815	15.3
		5+	127,605	11.6	141,166	12.5	111,568	10.2
	Annual household income ($)	0–25,000	179,735	16.4	199,928	17.8	184,423	16.9
		25,001–50,000	212,233	19.4	213,812	19.0	199,314	18.3
		50,001–75,000	184,832	16.9	190,154	16.9	186,895	17.1
		75,001–100,000	141,287	12.9	143,094	12.7	140,218	12.9
		100,001–150,000	186,349	17.0	187,704	16.7	189,725	17.4
		150,001–200,000	89,625	8.2	88,174	7.8	87,991	8.1
		200,001+	102,697	9.4	102,596	9.1	101,598	9.3
	Number of household workers[b]	0	186,757	17.0	183,925	16.3	204,947	18.8
		1	514,342	46.9	402,093	35.7	411,002	37.7
		2	308,643	28.1	446,167	39.6	404,615	37.1
		3+	87,019	7.9	93,277	8.3	69,600	6.4
	Number of household vehicles	0	61,951	5.6	69,040	6.1	56,405	5.2
		1	341,368	31.1	351,969	31.3	339,973	31.2
		2	438,430	40.0	445,924	39.6	442,474	40.6
		3	170,880	15.6	170,237	15.1	165,261	15.2
		4+	84,129	7.7	88,292	7.8	86,051	7.9
	Housing status (owned or rented residence)	own	577,378	52.6	609,163	54.1	629,706	57.8
		rent	519,380	47.4	516,299	45.9	460,458	42.2
Individual attributes	*Age (years)*	0–17	719,059	22.2	842,101	27.3	725,026	25.4
		18–34	883,392	27.3	719,984	23.3	596,508	20.9
		35–64	1,218,436	37.6	1,134,855	36.8	1,122,617	39.4
		65+	416,639	12.9	388,919	12.6	405,737	14.2
	Sex	female	1,611,745	49.8	1,544,991	50.1	1,441,326	50.6
		male	1,625,781	50.2	1,540,868	49.9	1,408,562	49.4
	Racial identity	American Indian or Alaska Native	17,707	0.5	17,544	0.6	11,784	0.4
		Asian	383,009	11.8	351,275	11.4	308,989	10.8
		Black or African American	162,943	5.0	121,455	3.9	95,754	3.4
		Native Hawaiian or Pacific Islander	14,017	0.4	10,958	0.4	8429	0.3
		White	2,289,287	70.7	2,183,891	70.8	2,066,282	72.5
		multiple	164,953	5.1	177,123	5.7	172,961	6.1
		other	205,612	6.4	223,613	7.2	185,689	6.5
	Hispanic identity	no	2,154,841	66.6	2,114,737	68.5	2,059,721	72.3
		yes	1,082,685	33.4	971,122	31.5	790,167	27.7
	Work status	no	1,645,812	50.8	1,498,663	48.6	1,408,659	49.4
		yes	1,591,714	49.2	1,587,196	51.4	1,441,229	50.6
	Work category	clerical or administrative	182,529	5.6	178,701	5.8	160,782	5.6
		labor[c]	243,278	7.5	237,059	7.7	194,587	6.8
		professional[d]	629,395	19.4	696,274	22.6	671,212	23.6
		sales or service	464,096	14.3	471,254	15.3	412,382	14.5
		N/A (not working)	1,718,228	53.1	1,502,571	48.7	1,410,925	49.5
	Education completed	less than high school	287,410	8.9	230,055	7.5	155,192	5.4
		high school	400,465	12.4	435,021	14.1	402,587	14.1
		some college	664,202	20.5	629,570	20.4	630,880	22.1
		Bachelor's degree	500,425	15.5	570,802	18.5	562,015	19.7
		graduate degree	312,236	9.6	383,201	12.4	379,590	13.3
		N/A (<25 years old)	1,072,788	33.1	837,210	27.1	719,624	25.3

[a]The 'Work status' and 'Work category' attributes are sourced from LODES data; all other attributes are sourced from ACS data.
[b]Number of household workers is an estimate based on the number of workers per family unit.
[c]Labor: Manufacturing, construction, maintenance, and farming.
[d]Professional: Professional, management, and technical.

SP-SRA had an absolute error of 190,439 commuters across all work categories, accounting for 12.0% of the total workforce in SP-SRA. All work categories except for the 'sales or service' category had positive error, indicating more commuters in the synthetic population than the validation data. The 'professional' work category (i.e. professional, management, and technical industries) had the greatest absolute error (TAE = 540,682 commuters) among the four work categories as well as the most variation across spatial units (SD = 11,763.5). Meanwhile, the 'labor' work category (i.e. manufacturing, construction, maintenance, and farming industries) had the lowest absolute error (TAE = 2293 commuters) and the 'sales or service' category had the least variation ($SD = 7{,}041.1$). For all work categories, scatterplots of the absolute error and the total number of commuters (validation data) revealed that error became greater, in either the positive or negative direction, as the number of commuters increased.

SP-CBG, the synthetic population with higher spatial granularity, had an absolute error of 33,741 commuters, making up only 2.3% of SP-CBG's total workforce. Despite SP-CBG's lower total absolute error, the individual work categories had high absolute error, which surpassed the errors in SP-SRA's labor, professional, and sales or service categories. The 'clerical or administrative' and 'professional' work categories had positive errors, while the 'labor' and 'sales or service' errors were negative. Like SP-SRA, the 'professional' work category had the greatest total absolute error (TAE = 514,433 commuters) and variation ($SD = 341.2$) in SP-CBG. Again, the 'sales or service' category had the least variation ($SD = 183.1$) but the 'clerical or administrative' work category had the least error (TAE = 26,707 commuters). Table 4 lists the full results of the error analysis.

The analysis revealed general trends in the synthetic populations. For instance, population synthesis overestimated the number of commuters in the 'professional' work category while underestimating the number of commuters in the 'sales or service' category. Therefore, it stands to reason that spatial units that are home to a lot of 'professional' commuters will have a greater positive error while spatial units that are home to many 'sales or service' commuters will have a greater negative error. Figure 3 illustrates the spatial distribution of total error, which manifests in positive and negative clusters. Spatial autocorrelation was confirmed using the Global Moran's

Table 4. Total error and summary statistics for the number of commuters residing in each spatial unit, by work category, for SP-SRA and SP-CBG.

Work Category		Total error	Minimum error	Maximum error	Mean absolute error	Standard deviation
SP-SRA	All work categories	190,439	−13,744	25,584	5601.1	9374.9
	Clerical or administrative	45,908	−1055	4477	1350.2	1098.9
	Labor[a]	2293	−3279	2958	67.4	1407.6
	Professional[b]	540,682	1113	43,442	15,902.4	11,763.5
	Sales or service	−398,444	−24,959	−52	−11,718.9	7041.1
SP-CBG	All work categories	33,741	−2814	2,462	19.2	303.2
	Clerical or administrative	26,707	−281	453	15.2	49.9
	Labor[a]	−42,163	−899	275	−24.0	60.9
	Professional[b]	514,433	−77	4551	292.6	341.2
	Sales or service	−465,236	−2280	96	−264.6	183.1

[a]Labor: Manufacturing, construction, maintenance, and farming.
[b]Professional: Professional, management, and technical.

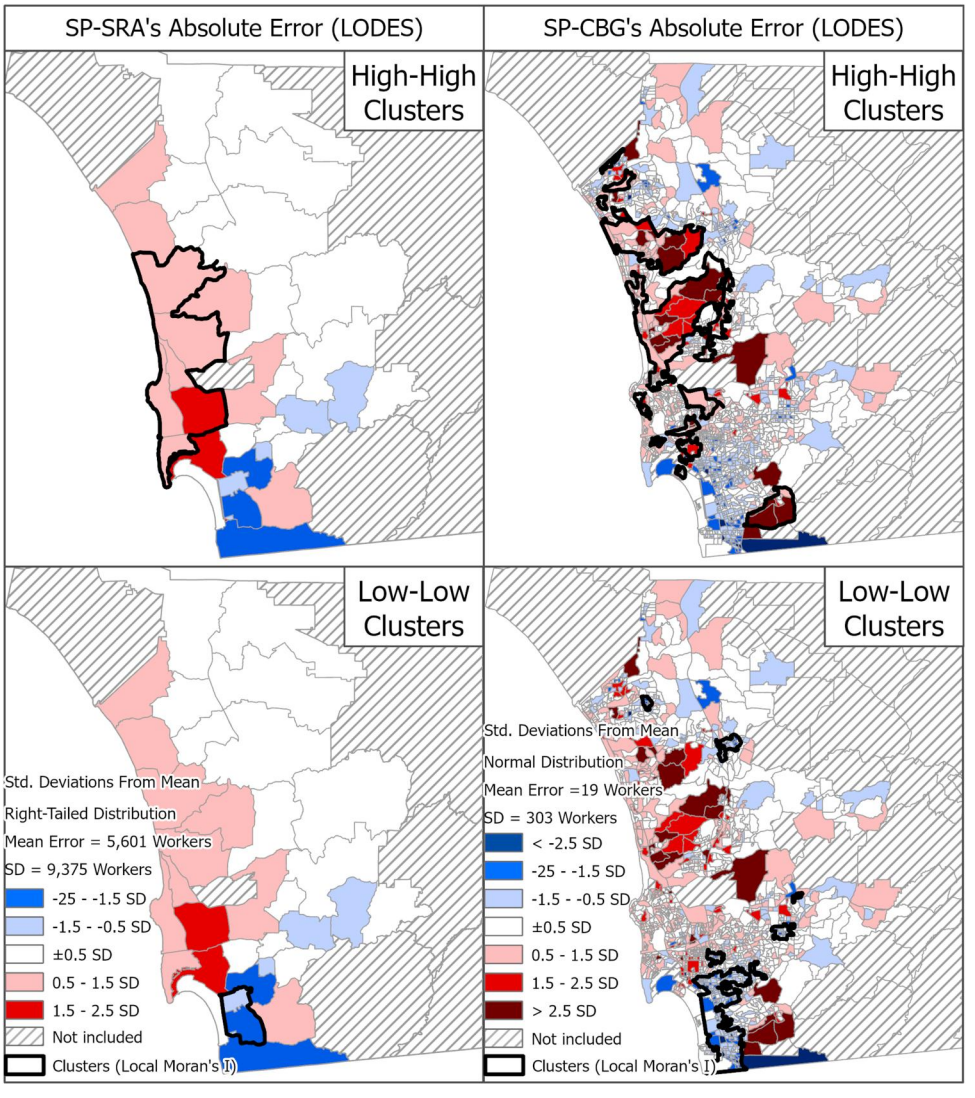

Figure 3. The spatial distribution of total error in the total number of commuters residing in each spatial unit in SP-SRA (left) and SP-CBG (right), symbolized using standard deviations from the mean. Based on the Local Moran's I, the bold black lines in the top maps outline dissolved clusters highlighting statistically significant spatial clusters of positive standard deviation (high-high), whereas those in the bottom maps outline dissolved clusters of negative standard deviation (low-low).

I (SP-SRA: $I = 0.301$, $p = 0.002$; SP-CBG: $I = 0.368$, $p \leq 0.001$). The maps of both synthetic populations display dissolved clusters of positive error in the central and northwest parts of the study area, while a dissolved cluster of negative error is located in the southern part of the study area. In Section 4.4, we compare these results to those derived using the spatially-oriented validation framework.

4.3. Validation measure 2: Pearson correlation coefficients

In addition to absolute error, we calculated Pearson correlation coefficients for the number of commuters residing in each spatial unit to determine the linear relationship of this attribute in the synthetic populations and the commuter data. Across all work categories, high correlation coefficients in SP-SRA ($r \geq 0.942$, $p \leq 0.001$) and SP-CBG ($r \geq 0.796$, $p \leq 0.001$) indicated strong linear relationships between the synthetic populations and commuter data. The correlation coefficients for SP-SRA were higher than those for SP-CBG for all work categories. Additionally, the correlation coefficients were higher for work categories with more total commuters (i.e. 'professional', 'sales or service') and were lower for work categories with fewer total commuters (i.e. 'labor', 'clerical or administrative') (Table 5).

We also calculated the Pearson correlation coefficient values for the number of trips originating from each spatial unit by activity. In general, the values of the Pearson correlation coefficients for the number of trips (SafeGraph) were lower than those for the number of commutes (commuter data), implying weaker linear relationships between the synthetic populations and SafeGraph data. However, the correlation coefficients remained greater for SP-SRA than SP-CBG. The highest Pearson correlation coefficients were for the total number of trips originating from each spatial unit (SP-SRA: $r = 0.942$, $p \leq 0.001$; SP-CBG: $r = 0.914$, $p \leq 0.001$), rather than the number of trips related to a specific activity. Also, the correlation coefficients were higher for activities with a clear association with a point-of-interest industry classification (e.g. buying meals and restaurants/other eating places); this phenomenon is more prominent in SP-SRA (Table 5).

Since there are strong linear relationships between the mobility attributes in the synthetic populations and the validation datasets, the usage of mobility differences as a validation measure for the spatially-oriented validation framework is reasonable.

Table 5. Pearson correlation coefficients (r) for the total number of workers and the number of workers by industry per spatial unit for SP-SRA and LODES commuter data, and SP-CBG and LODES commuter data.

Attribute		SP-SRA (r)		SP-CBG (r)	
Work category (LODES)	Total workers (all categories)	0.946	***	0.883	***
	Clerical or administrative	0.942	***	0.796	***
	Labor: manufacturing, construction, maintenance, farming	0.951	***	0.801	***
	Professional: professional, management, technical	0.959	***	0.923	***
	Sales or service	0.979	***	0.879	***
Activity (SafeGraph)	Total trips (all activities)	0.942	***	0.914	***
	Attend school as a student	0.906	***	0.890	***
	Buy goods (groceries, clothes, appliances, gas)	0.900	***	0.890	***
	Buy meals (go out for a meal, snack, carry-out)	0.925	***	0.903	***
	Buy services (dry cleaner, banking, car service, pet care)	0.903	***	0.859	***
	Health care visit (medical, dental, therapy)	0.928	***	0.869	***
	Other general errands (post office, library)	0.850	***	0.856	***
	Recreational activities (parks, movies, bars, museums)	0.917	***	0.861	***
	Religious or other community activities	0.904	***	0.818	***

Pearson correlation coefficients (r) for the total number of trips and the number of trips by activity per spatial unit for SP-SRA and SafeGraph data, and SP-CBG and SafeGraph data.
Significance levels:.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

4.4. Validation measure 3: spatially analyzed mobility differences

The spatially-oriented validation framework uses mobility differences as the validation measure. *Percentage* differences are calculated when the validation data is directly comparable to the synthetic populations (e.g. commuter data); otherwise, *percentile* differences are calculated (e.g. SafeGraph data). Because the method supports direct and indirect comparisons of the synthetic populations to validation data, both the commuter and SafeGraph data sets were used for validation.

The commuter data was used to validate the total number of commutes originating from each spatial unit. The mean percentage difference for SP-SRA equaled 15.5% with a standard deviation of 21.4%. SP-CBG had a lower mean percentage difference (3.9%) but greater variability (SD = 35.5%) than SP-SRA. The frequency of percentage differences for SP-SRA and SP-CBG both approached normal distributions.

Upon mapping, the percentage differences for SP-SRA and SP-CBG displayed obvious spatial clustering that was confirmed using the Global Moran's I (SP-SRA: $I = 0.275$, $p = 0.004$; SP-CBG: $I = 0.405$, $p < 0.001$) (Figure 4). The locations of dissolved clusters of positive and negative mobility differences were identified using the Local Moran's I. Dissolved clusters of spatial units with negative standard deviations from the mean highlight neighborhoods where the number of commuters was underestimated. Likewise, dissolved clusters of spatial units with positive standard deviations from the mean identify neighborhoods where the number of commuters was overestimated. Dissolved clusters of positive standard deviation can be found in the central and coastal regions of the study area while a large dissolved cluster of negative standard deviation is located in the southern part of the study area. The spatial distributions of mobility differences (Figure 4) and total error (Figure 3) are strikingly similar, although the dissolved clusters of positive standard deviation extend further east in the maps of total error.

The validation method was repeated using SafeGraph as the validation dataset. The mean difference for SP-SRA equaled 0 percentiles, as expected for the normalized measure, with a standard deviation of 10 percentiles. However, the frequency of SP-SRA's percentile differences was irregular and multimodal. SP-CBG also had a mean of 0 percentiles, but it had a higher standard deviation (22 percentiles). The frequency of percentile differences for SP-CBG approached a normal distribution. When mapped, there were visible clusters of percentile differences for SP-SRA and SP-CBG (Figure 5).

The Global Moran's I confirmed the positive spatial autocorrelation for SP-SRA ($I = 0.338$, $p = 0.001$) and SP-CBG ($I = 0.346$, $p < 0.001$). With the Local Moran's I, we defined the locations of dissolved clusters of positive and negative standard deviation from the mean. Dissolved clusters of positive standard deviation (i.e. overestimated number of trips) were located along the coast, and dissolved clusters of negative standard deviation (i.e. underestimated number of trips) dotted the study area to the south and the east. Of note, the locations of these dissolved clusters (SafeGraph) are similar to, but not the same as, the dissolved clusters for the number of commutes (commuter data). Considering that the commuter data only represents work commutes, not all trips, it makes sense that the distribution of clusters would not align perfectly. However, similarities in the distribution suggest a positive correlation between the total number of commutes and the total number of trips.

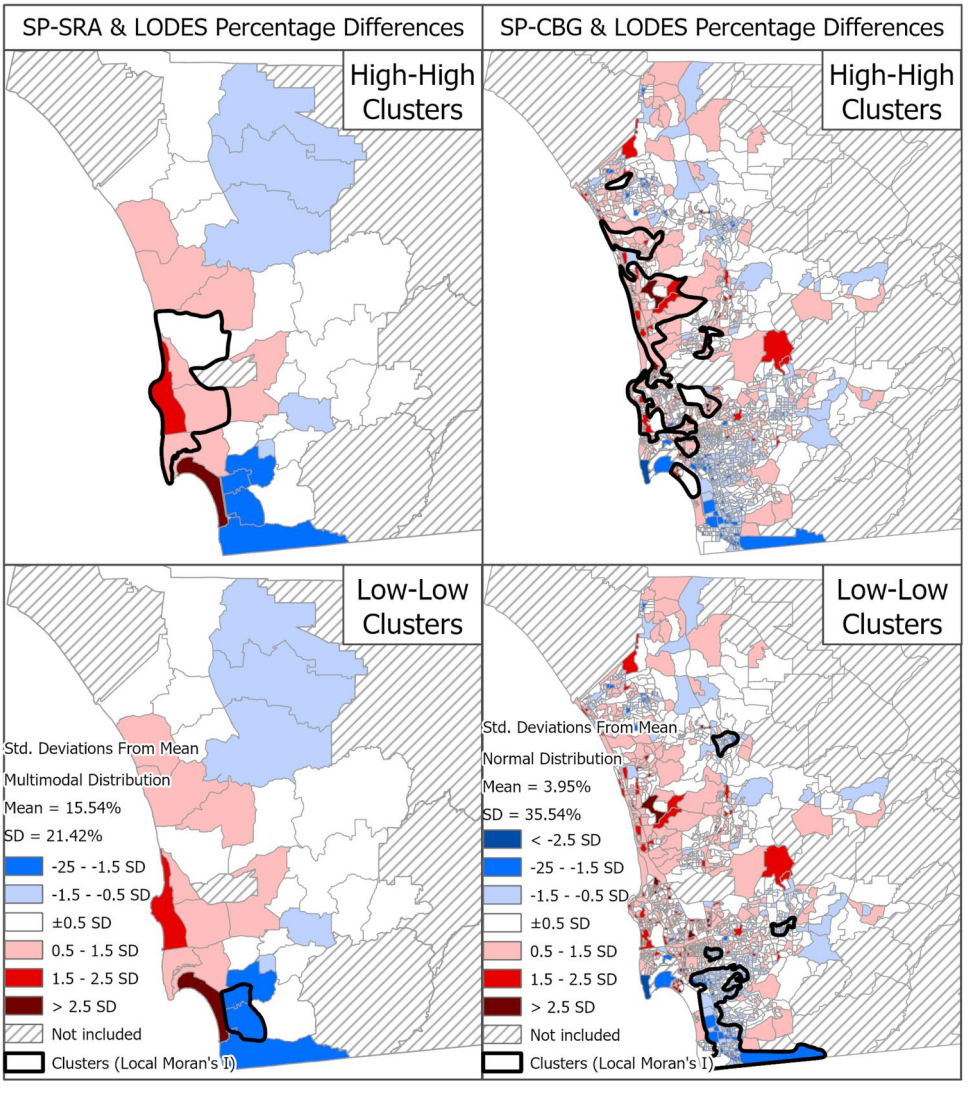

Figure 4. The spatial distribution of percentage differences in the total number of commuters residing in each spatial unit in SP-SRA (left) and SP-CBG (right), symbolized using standard deviations from the mean. Based on the Local Moran's I, the bold black lines in the top maps outline dissolved clusters highlighting statistically significant spatial clusters of positive standard deviation (high-high), whereas those in the bottom maps outline dissolved clusters of negative standard deviation (low-low).

5. Discussion

Our study found pre-existing biases in the travel surveys used for population synthesis. Survey respondents tended to be older, more educated, and more affluent than the average resident in the study area, based on the community survey data. In addition, the proportion of respondents with a Hispanic identity was much lower than the proportion of residents with a Hispanic identity; these discrepancies, along with the

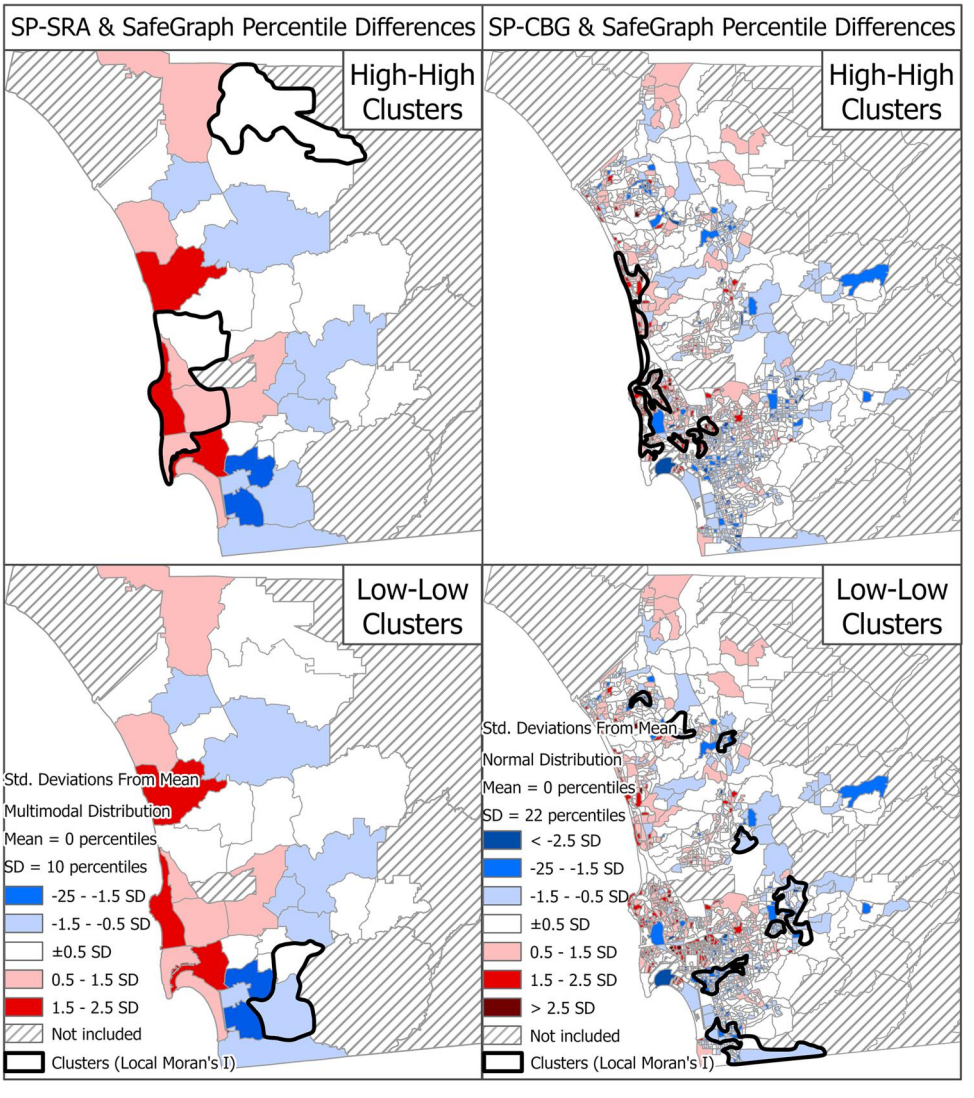

Figure 5. The spatial distribution of percentile differences in the total number of trips originating from each spatial unit in SP-SRA (left) and SP-CBG (right), symbolized using standard deviations from the mean. Based on the Local Moran's I, the bold black lines in the top maps outline dissolved clusters highlighting statistically significant spatial clusters of positive standard deviation (high-high), whereas those in the bottom maps outline dissolved clusters of negative standard deviation (low-low).

combination of households with 5 or more members, are likely explanations for having fewer total individuals in the synthetic populations. When examining the synthetic populations' aggregate attributes for the study area (Section 4.1), these biases were still present, but less severe than in the compiled travel surveys.

Aggregation methods used for the community survey data also introduced errors to the synthetic populations. For instance, the combination of all households with five or more members resulted in synthetic populations with fewer individuals than the study area's target values. Furthermore, there were fewer two-worker households than

in the study area's source data because of data ambiguity (i.e. the number of family units in each household was not specified by the community survey). The first validation measure, total error (Section 4.2), revealed high levels of error and variability for the work category attributes, but the Pearson correlation coefficients for the number of commuters indicated strong linear relationships between the synthetic populations and the commuter data ($r \geq 0.796$, $p \leq 0.001$) (Section 4.3).

The goal of our research was not to entirely eliminate error from our synthetic populations, but to replicate the patterns of error using independent mobility microdata (SafeGraph) and the spatially-oriented validation framework (Section 4.5). The Pearson correlation coefficients for the number of trips using SafeGraph data were statistically significant and generally comparable to those resulting from the correlation analysis with commuter data. Though not identical, the locations of dissolved clusters of total error (Figure 3) and mobility differences (Figure 4), both using commuter data for validation, were analogous to one another in their depictions of large dissolved clusters located in the southern (low-low) and central coastal (high-high) regions of the study area. The mobility differences using SafeGraph validation data (Figure 5) form dissolved clusters of error in the same general regions; some discrepancies are expected in Figure 5 because it includes all mobility while Figures 3 and 4 only include work-related commutes. The shared locations of dissolved clusters for total mobility and the number of commuters is supported by preliminary research that discovered spatial similarity between the SafeGraph and commuter data sets (Embury *et al*. 2022a). These results support the suitability of SafeGraph data for independent data validation and demonstrate the value of using the spatially-oriented validation framework.

Differences in the results for the two synthetic populations at different spatial resolutions (i.e. CBG, SRA) stress the importance of multiscale analysis of human dynamics. While the biases detected by the Pearson correlation analysis were similar for SP-CBG and SP-SRA, SP-CBG had weaker linear relationships with the commuter and SafeGraph validation data than SP-SRA, likely due to its finer spatial granularity (Harland *et al*. 2012). Of note, results using SafeGraph validation data may be heavily influenced by the study's assumption that all synthetic population trips originated from the traveler's residential spatial unit. This assumption is more reasonable for the larger SRA units and likely caused greater error for the smaller CBG units. In these cases, the value of SP-CBG's greater detail is partially diminished by its increased error. However, SP-CBG had a greater number of spatial units and was more useful than SP-SRA for identifying spatial relationships. The reversal of utility in SP-SRA and SP-CBG emphasizes the need for multiscale and spatial evaluation methods.

Despite the implications of the discovered errors and biases, there are still insights to be gleaned from the results. The dissolved clusters of mobility differences identified communities where the total number of commuters and/or the total number of trips were significantly underrepresented (low-low) or overrepresented (high-high). The underrepresented and overrepresented communities match identified regions of high and low COVID-19 vulnerability (Embury et al. 2022b, Tsou *et al*. 2023). The similarities indicate that the synthetic populations underrepresented the study area's marginalized and underserved communities (Tsou *et al*. 2023). The biases in the

synthetic populations, if not mitigated, have the potential to perpetuate harm in these communities.

Perhaps most importantly, the spatially-oriented validation framework demonstrated its value by detecting biases that were not apparent when the synthetic populations' attributes were compared to the administrative source data. Spatial clusters of overrepresentation and underrepresentation can be marked for further investigation and bias mitigation. Several compelling opportunities for bias mitigation research, to be discussed further in Section 6.2, emerged as a result of this study.

6. Conclusion

The two-fold purpose of this study was to assess the suitability of mobility microdata for independent data validation, and to introduce a spatially-oriented data validation framework for synthetic populations. Using IPU, synthetic populations were generated at two spatial granularities (SRAs and CBGs). Both synthetic populations, especially SP-SRA, seemed to have low levels of bias based on their sociodemographic compositions. However, the validation method which measured the total error in the synthetic populations using the commuter data revealed overrepresentation and underrepresentation in the number of commuters in communities across the study area. When mapped, the total errors showed that the synthetic populations underrepresented some of the study area's marginalized communities. These findings were replicated using the spatially-oriented validation framework using both commuter data and SafeGraph data for validation.

6.1. Study limitations

The study and its findings are subject to a number of limitations, several of which are common among spatial and spatiotemporal statistical analyses. First, the study area had a low number of spatial units for the low granularity (SRA) portion of the study. Although the number of spatial units exceeded the minimum ($n > 30$) expected for a Pearson correlation analysis, the low number of SRAs ($n = 34$) limits confidence in the results. Next, the study used irregularly shaped spatial units and data with different temporal resolutions and time periods (i.e. 2017, 2019). As a result, the study is subject to the modifiable areal unit problem (Openshaw and Taylor, 1979) and the modifiable temporal unit problem (Çöltekin *et al*. 2011), which state that results, and their significance, depend on the data's spatial and temporal boundaries. Finally, the study suffered edge effects because mobility into and out of the study area was not considered. The inclusion of inflows and outflows, especially along the US-Mexico border, may have altered the study's findings.

Two of the study's source datasets have considerable biases. The compiled travel surveys had pre-existing biases, discussed in Section 4.1, and the SafeGraph data set has data generation biases, discussed in Sections 2.1 and 3.2. Debiasing the data, as suggested by Coston et al. (2021), may have resulted in more representative synthetic populations and increased confidence in the results of the independent data validation.

The study makes two assumptions that must be recognized. First, all trips made by the synthetic populations' individuals originate from their residential spatial units. This assumption is more problematic for the high granularity (CBG) portion of the study.

The creation of activity schedules for individuals in the synthetic populations will fully address this assumption by defining precise origin and destination locations for every trip (Bradley *et al*. 2010, Drchal *et al*. 2019, Luo *et al*. 2024). Second, the crosswalks used to compare travel surveys and synthetic populations contain generalizations that may have affected the results of the study. For example, SafeGraph's industry classifications were only given one activity purpose, although, in reality, there may be several appropriate activities. The impact of this assumption will also be reduced by activity scheduling and the assignment of trip destinations. While these assumptions may affect the results of this study, they can be addressed in future research.

6.2. Future research directions

This study inspired several focus areas for future research. To start, the incorporation of uncertainty measures in the community survey source data (Wei *et al*. 2023) can provide a probabilistic grounding for validation that would enhance our interpretation of the results. Next, debiased SafeGraph data can be used to validate the synthetic populations. Discrepancies between the validation results can be analyzed to better understand the debiasing process. Similar to the independent data validation performed in this study, the SafeGraph data can be used to independently/externally calibrate the population synthesis model. The inclusion of external data during calibration can improve results and increase overall confidence in the model.

Then, overrepresented and underrepresented communities will be subjected to individual examination and bias mitigation. The established bias mitigation procedures will be compared across the communities and tested for the entire study area. Lastly, activity scheduling will be performed to address the study's assumptions about trip origins. The spatially-oriented data validation framework can be expanded to support the validation of activity schedules. Ultimately, the activities will be simulated by an agent-based model and, once again, the data validation framework can be expanded to introduce parallel evaluation methods fit for agent-based modeling contexts.

On their own, synthetic populations provide valuable insight into the activities and dynamics of individuals. The value of synthetic populations is amplified when they are used for individual-based mobility modeling. Accordingly, close attention to synthetic population validity is critical in advancing realism in individual-based mobility models and preventing the potential perpetuation of harm caused by undetected bias.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This material is based on work supported by the National Science Foundation under Grant No. 2031407. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Notes on contributors

Jessica Embury is a doctoral student in the Departments of Geography at San Diego State University and the University of California Santa Barbara, and a researcher at the Center for Human Dynamics in the Mobile Age. Embury's research focuses on spatial modeling, spatiotemporal data analysis, and the integration of big geospatial data into geographic applications. Embury has experience conducting geographic research and producing spatial models related to social equity issues such as food access, pollution burden, and disease vulnerability.

Atushi Nara is an Associate Professor of Geography at San Diego State University and the Associate Director of the Center for Human Dynamics in the Mobile Age. His research interests include spatial data science, spatiotemporal data analysis and modeling, human dynamics and movement behaviors, complex adaptive systems, and geocomputation education. Nara has received funding from various grant programs, including the National Science Foundation's Infrastructure Management and Extreme Events (IMEE) and Computer Science for All (CSforAll: Research and RPPs) programs.

Sergio Rey is a Professor of Geography at San Diego State University. Rey's research focuses on spatial data science, geocomputation, spatial inequality dynamics, regional science, and open science. He is a Fellow of the American Association for the Advancement of Science, the Spatial Econometrics Association, and the Regional Science Association International. He is the co-founder and lead developer of the open source Python Spatial Analysis Library (PySAL), and has taught workshops on PySAL throughout the world.

Ming-Hsiang Tsou is a Professor of Geography, San Diego State University (SDSU) and the Director of Center for Human Dynamics in the Mobile Age (HDMA). His research interests are in Human Dynamics, Social Media, Big Data, Visualization, Internet Mapping, Web GIS, Mobile GIS, Cartography, and K-12 GIS education. In Spring 2014, Tsou established a new research center, The Center for Human Dynamics in the Mobile Age, a transdisciplinary research area of excellence at San Diego State University to integrate research works from GIScience, Public Health, Social Science, Sociology, and Communication.

Sahar Ghanipoor Machiani (PhD, Virginia Tech) is an Associate Professor of Civil, Construction, and Environmental Engineering at San Diego State University and a Co-director of SDSU Smart Transportation Analytics Research (STAR) Lab. She previously served as an Associate Director (SDSU Director) at the Safe-D National University Transportation Center (UTC). Her contributions have been recognized through honors such as the WTS Technology for Transportation Award, the Western District ITE Transportation Achievement Award, and an SDSU Center for Teaching and Learning Grant Award.

ORCID

Jessica Embury http://orcid.org/0000-0002-6677-9980
Atsushi Nara http://orcid.org/0000-0003-4173-7773
Sergio Rey http://orcid.org/0000-0001-5857-9762
Ming-Hsiang Tsou http://orcid.org/0000-0003-3421-486X
Sahar Ghanipoor Machiani http://orcid.org/0000-0002-7314-2689

Data availability statement

The data and codes that support the findings of this study are available on figshare.com with the link: https://doi.org/10.6084/m9.figshare.24664647

Data were derived from the following resources:

San Diego Association of Governments. (2015a). *CENSUS_BLOCKGROUPTIGER2010.zip* [Data set]. https://rdw.sandag.org/Account/gisdtview?dir=Census

San Diego Association of Governments. (2015b). *Subregional_Areas_2010.zip* [Data set]. https://rdw.sandag.org/Account/gisdtview?dir=Census

SafeGraph. (2023). *Monthly Patterns - Historic Data (2019)* [Data set]. https://marketplace.deweydata.io/#/products/safegraph_mp_*_r_0/documentation

State of California. (2018). *National Household Travel Survey 2017 California Geocoded (Spatial) Data* [Data set]. https://nhts.dot.ca.gov/

United States Census Bureau. (2022a). *American Community Survey Data 2017 5-year estimates* [Data set]. https://www.census.gov/programs-surveys/acs/data.html

United States Census Bureau. (2022b). *LEHD Origin-Destination Employment Statistics Data (2002-2020)* (Version 7.0) [Data set]. https://lehd.ces.census.gov/data/#lodes

References

Abraham, J.E., Stefan, K.J., and Hunt, J.D., 2012. Population synthesis using combinatorial optimization at multiple levels. *Transportation Research Board 91st Annual Meeting*, 12, 3383. https://trid.trb.org/view/1130260

Alonso-Betanzos, A., *et al*., 2021. Generating a synthetic population of agents through decision trees and socio demographic data. *In:* I. Rojas, G. Joya, and A. Català, eds. *Advances in computational intelligence*. Cham, Switzerland: Springer International Publishing, 128–140.

Anderson, T., and Dragićević, S., 2020. NEAT approach for testing and validation of geospatial network agent-based model processes: case study of influenza spread. *International Journal of Geographical Information Science*, 34 (9), 1792–1821.

Anselin, L., 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27 (2), 93–115.

Batty, M., 2005. *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. Cambridge, MA: The MIT Press.

Beckman, R.J., Baggerly, K.A., and McKay, M.D., 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30 (6), 415–429.

Benenson, I., Martens, K., and Birfir, S., 2008. PARKAGENT: An agent-based model of parking in the city. *Computers, Environment and Urban Systems*, 32 (6), 431–439.

Borysov, S.S., Rich, J., and Pereira, F.C., 2019. How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies*, 106, 73–97.

Bradley, M., Bowman, J.L., and Griesenbeck, B., 2010. SACSIM: an applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling*, 3 (1), 5–31.

Cambridge Systematics, Inc, 2010. *Travel model validation and reasonableness checking manual*. 2nd ed. Cambridge, MA: North Carolina Department of Transportation. https://connect.ncdot.gov/projects/planning/tpb%20training%20presentations/fhwa%20model%20validation%20handbook.pdf

Castiglione, J., Bradley, M., and Gliebe, J., 2014. *Activity-based travel demand models: a primer*. Washington, DC: Transportation Research Board.

Chapuis, K., *et al*., 2021. Gen*: an integrated tool for realistic agent population synthesis. *In*: P. Ahrweiler and M. Neumann, eds. *Advances in social simulation. ESSA 2019. Springer Proceedings in Complexity*. Cham: Springer.

Crooks, A., Castle, C., and Batty, M., 2008. Key challenges in agent-based modelling for geo-spatial simulation. *Computers, Environment and Urban Systems*, 32 (6), 417–430.

Coston, A., *et al*., 2021. Leveraging administrative data for bias audits: assessing disparate coverage with mobility data for COVID-19 policy. *In*: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery, 173–184.

Çöltekin, A., *et al.*, 2011. Modifiable temporal unit problem. *In: ISPRS/ICA workshop "Persistent problems in geographic visualization"* (ICC2011). Paris, France: ICC2011 Workshop. https://www.zora.uzh.ch/id/eprint/54263/1/2011_C%C3%B6ltekinA_coltekin-etal-ica2011-geovis-workshop.pdf

Deming, W.E., and Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11 (4), 427–444.

Drchal, J., Čertický, M., and Jakob, M., 2019. Data-driven activity scheduler for agent-based mobility models. *Transportation Research Part C: Emerging Technologies*, 98, 370–390.

Embury, J., Nara, A., and Jin, C., 2022a. Spatially weighted structural similarity index: A multiscale comparison tool for diverse sources of mobility data. *In*: *HANIMOB'22: The 2nd ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility Proceedings*, 1 November 2022, Seattle, WA. New York, NY: Association for Computing Machinery.

Embury, J., *et al.*, 2022b. A spatio-demographic perspective on the role of social determinants of health and chronic disease in determining a population's vulnerability to COVID-19. *Preventing Chronic Disease*, 19, E38.

Epstein, J.M., 2009. Modelling to contain pandemics. *Nature*, 460 (7256), 687–687.

Estabrooks, A., and Japkowicz, N., 2001. A mixture-of-experts framework for learning from imbalanced data sets. *In*: F. Hoffmann, D. J. Hand, N. Adams, D. Fisher, and G. Guimaraes, eds., *Advances in intelligent data analysis*. Heidelberg, Germany: Springer, 34–43.

Farooq, B., *et al.*, 2013. Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58 (C), 243–263.

Fournier, N., *et al.*, 2021. Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation*, 48 (2), 1061–1087.

Friedman, B., and Nissenbaum, H., 1996. Bias in computer systems. *ACM Transactions on Information Systems*, 14 (3), 330–347.

Gartlehner, G., and Moore, C.G., 2008. Direct versus indirect comparisons: a summary of the evidence. *International Journal of Technology Assessment in Health Care*, 24 (2), 170–177.

Guo, J.Y., and Bhat, C.R., 2007. Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2014 (1), 92–101.

Harland, K., *et al.*, 2012. Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, 15 (1), 1.

Heppenstall, A., *et al.*, 2020. Future developments in geographical agent-based models: challenges and opportunities. *Geographical Analysis*, 53 (1), 76–91.

Jin, C., *et al.*, 2023. Predicting households' residential mobility trajectories with geographically localized interpretable model-agnostic explanation (GLIME). *International Journal of Geographical Information Science*, 37 (12), 2597–2619.

Kerr, C.C., *et al.*, 2021. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLoS Computational Biology*, 17 (7), e1009149.

Kianersi, D., *et al.*, 2021. Agent-based simulation of human mobility using high-resolution foot-traffic data. *Journal of Student-Scientists' Research*, 3.

Kukić, M., and Bierlaire, M., 2022. One-step simulator for synthetic households generation. *In*: *22nd Swiss Transport Research Conference,* Ascona, Switzerland. https://transp-or.epfl.ch/documents/proceedings/KukicBierlaire_STRC2022.pdf

Lenormand, M., *et al.*, 2015. Influence of sociodemographic characteristics on human mobility. *Scientific Reports*, 5 (1), 10075.

Ligmann-Zielinska, A., and Jankowski, P., 2007. Agent-based models as laboratories for spatially explicit planning policies. *Environment and Planning B: Planning and Design*, 34 (2), 316–335.

Lomax, N., and Norman, P., 2016. Estimating population attribute values in a table: "get me started in" iterative proportional fitting. *The Professional Geographer*, 68 (3), 451–461.

Lovelace, R., and Dumont, M., 2016. *Spatial microsimulation with R*. 1st ed. New York, NY: Chapman and Hall/CRC.

Luo, N., *et al.*, 2024. An integration modeling framework for individual-scale daily mobility estimation. *Travel Behaviour and Society*, 34, 100650.

Manson, S.M., Sun, S., and Bonsal, D., 2012. Agent-based modeling and complexity. *In:* A.J. Heppenstall, A.T. Crooks, L.M. See, and M. Batty, eds., *Agent-based models of geographical systems*. Dordrecht, Netherlands: Springer, 125–139.

Mohammed, R., Rawashdeh, J., and Abdullah, M., 2020. Machine learning with oversampling and undersampling techniques: overview study and experimental results. *In: 2020 11th International Conference on Information and Communication Systems (ICICS)*. New York, NY: Institute for Electrical and Electronics Engineers (IEEE), 243–248.

Müller, K., and Axhausen, K., 2011. Population synthesis for microsimulation: state of the art. *In*: *Transportation Research Board 90th Annual Meeting*. Ascona, Switzerland: 10th Swiss Transport Research Conference. https://www.strc.ch/2010/Mueller.pdf

Niroumand, H., Zain, M.F.M., and Jamil, M., 2013. Statistical methods for comparison of data sets of construction methods and building evaluation. *Procedia - Social and Behavioral Sciences*, 89, 218– 221.

Openshaw, S., and Taylor, P.J., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *In*: N. Wrigley, ed. *Statistical applications in spatial sciences*. London, UK: Pion, 127–144.

Prédhumeau, M., and Manley, E., 2023. A synthetic population for agent-based modelling in Canada. *Scientific Data*, 10 (1), 148.

Ramadan, O., and Sisiopiku, V., 2019. *A critical review on population synthesis for activity- and agent-based transportation models*. Rijeka, Croatia: IntechOpen. https://www.intechopen.com/chapters/67163

Renaud, O., and Victoria-Feser, M.-P., 2010. A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140 (7), 1852–1862.

Rodriguez-Carrion, A., Garcia-Rubio, C., and Campo, C., 2018. Detecting and reducing biases in cellular-based mobility data sets. *Entropy (Basel, Switzerland)*, 20 (10), 736.

Saadi, I., *et al.*, 2016. Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, 90, 1–21.

Salat, H., *et al.*, 2023. Synthetic population Catalyst: a micro-simulated population of England with circadian activities. *Environment and Planning B: Urban Analytics and City Science*, 50 (8), 2309–2316.

San Diego Association of Governments 2015a. *CENSUS_BLOCKGROUPTIGER2010.zip*. https://rdw.sandag.org/Account/gisdtview?dir=Census

San Diego Association of Governments 2015b. Subregional_Areas_2010.zip. https://rdw.sandag.org/Account/gisdtview?dir=Census

SafeGraph 2023. Monthly Patterns - Historic Data (2019). https://marketplace.deweydata.io/#/products/safegraph_mp_*_r_0/documentation

Scherr, W., *et al.*, 2020. Towards agent-based travel demand simulation across all mobility choices – the role of balancing preferences and constraints. *European Journal of Transport and Infrastructure Research*, 20 (4), 4.

Schlosser, F., *et al.*, 2021. Biases in human mobility data impact epidemic modeling (arXiv: 2112.12521). arXiv.

Shapiro, R.Y., 2001. Polling. In: *International encyclopedia of the social & behavioral sciences*. Oxford, UK: Elsevier, 11719–11723.

Silva, P.C.L., *et al.*, 2020. COVID-ABS: an agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons, and Fractals*, 139, 110088.

Sourbati, M., and Behrendt, F., 2021. Smart mobility, age and data justice. *New Media & Society*, 23 (6), 1398–1414.

State of California 2018. National Household Travel Survey 2017 California Geocoded (Spatial) Data. https://nhts.dot.ca.gov/

Sun, L., and Erath, A., 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61, 49–62.
Torrens, P.M., and Nara, A., 2012. Polyspatial agents for multi-Scale urban simulation and regional policy analysis*. *Regional Science Policy & Practice*, 4 (4), 419–445.
Torrens, P.M., 2018. A computational sandbox with human automata for exploring perceived egress safety in urban damage scenarios. *International Journal of Digital Earth*, 11 (4), 369–396.
Tracy, M., Cerdá, M., and Keyes, K.M., 2018. Agent-based modeling in public health: current applications and future directions. *Annual Review of Public Health*, 39 (1), 77–94.
Trivedi, A., and Rao, S., 2018. Agent-based modeling of emergency evacuations considering human panic behavior. *IEEE Transactions on Computational Social Systems*, 5 (1), 277–288.
Truszkowska, A., *et al.*, 2021. Designing the safe reopening of US Towns Through High-Resolution Agent-Based Modeling. *Advanced Theory and Simulations*, 4 (9), 2100157.
Truszkowska, A., *et al.*, 2022. Predicting the effects of waning vaccine immunity against COVID-19 through high-resolution agent-based modeling. *Advanced Theory and Simulations*, 5 (6), 2100521.
Tsou, M.-H., *et al.*, 2023. Analyzing Spatial-temporal impacts of neighborhood socioeconomic status variables on COVID-19 outbreaks as potential social determinants of health. *Annals of the American Association of Geographers*, 113 (4), 891–912.
United States Census Bureau 2022a. American Community Survey Data 2017 5-year estimates. https://www.census.gov/programs-surveys/acs/data.html
United States Census Bureau. 2022b. LEHD *origin-destination employment statistics data* (2002-2020) (Version 7.0). https://lehd.ces.census.gov/data/#lodes
Wang, Y., Hao, H., and Wang, C., 2022. Preparing urban curbside for increasing mobility-on-demand using data-driven agent-based simulation: case study of city of Gainesville, Florida. *Journal of Management in Engineering*, 38 (3), 05022004.
Ward, K., 2020. ipfr: List *balancing for reweighting and population synthesis* (R package version 1.0.2) [Computer software]. https://CRAN.R-project.org/package=ipfr
Wei, R., Knaap, E., and Rey, S.J., 2023. American Community Survey (ACS) data uncertainty and the analysis of segregation dynamics. *Population Research and Policy Review*, 42 (1), 5.
Wesolowski, A., *et al.*, 2013. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society, Interface*, 10 (81), 20120986.
Wu, G., *et al.*, 2022. A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. *Scientific Data*, 9 (1), 19.
Ye, X., *et al.*, 2009. *Methodology to match distributions of both household and person attributes in generation of synthetic populations*. Washington, DC: 88th Annual Meeting of the Transportation Research Board (TRB).
Zhou, M., *et al.*, 2022. Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation. *Computers, Environment and Urban Systems*, 91, 101717.
Zhu, Y., and Ferreira, J., 2014. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, 2429 (1), 168–177.