The Journal of Transport and Land Use jtlu.org
Vol. 18 No. 1 [2025] pp. 197–220

JTLU

Synthesizing activity locations in the context of integrated activity-based models

Natalia Zuniga-Garcia (corresponding author)
Argonne National Laboratory
nzuniga@anl.gov

Pedro Veiga de Camargo
Argonne National Laboratory
c@margo.co

Abstract: Activity-based models are a powerful tool for transportation analysis and represent the future of the industry in terms of modeling techniques. However, the data-hungry aspect of these models makes them difficult and slow to build. This paper presents a set of methodologies to synthesize activity locations for U.S. cities, providing estimates of locations by land-use type in areas with limited available data. The methodology includes a regression method to estimate the number of locations by land-use type complemented by selective use of open data. Detailed information from the entire Southern California Association of Governments (SCAG) area, comprising more than 100,000 km^2, is used to calibrate the model. A zero-inflated negative binomial (ZINB) regression is proposed to tackle the excess of zeros in the dataset. The model is estimated using a Bayesian approach that quantifies the coefficients' variability, uses information regarding prior beliefs, and estimates zero-inflated probabilities by zone. The main results suggest that the proposed methodological framework can be used to estimate locations in a fast and efficient way without the need for detailed land-use information. Transportation planners and policymakers can use the results and methods provided in this research to approximate activity location distributions in activity-based models.

Keywords: Location synthesis, land use, activity-based models, zero-inflated negative binomial, Bayesian regression

Article history:
Received: October 18, 2022
Received in revised form:
August 22, 2024
Accepted: December 20, 2024
Available online: March 13, 2025

1 Introduction

Activity-based transport models (ABM) have become a powerful tool for analyzing complex scenarios and answering nuanced policy questions related to urban and regional transportation systems. Although many ABM frameworks exist, some key challenges to their development and use in practice remain, one of which is related to input data.

It is well known that model results for an ABM are particularly sensitive to their disaggregate input data (Kagho et al., 2020). However, the spatial location of activities in different activity-based travel demand models varies greatly. Zone-based methods are commonly used in ABM, with location choice models usually given at the traffic analysis zone (TAZ)-level. Examples include CUSTOM (comprehensive utility-based system of activity-travel scheduling options modeling (Habib, 2017), ActivitySim (Galli et al., 2009), DaySim (Bradley et al., 2010), and BEAM (Behavior, Energy, Autonomy, and

https://doi.org/10.5198/jtlu.2025.2291
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

Mobility) modeling framework (Laarabi et al., 2023). Point-based methods are rare. Models that use coordinate-level locations usually utilize location points that are specified at the zone level, and then coordinates are randomly generated within the zones, as in CEMDAP (comprehensive econometric micro-simulator for daily activity-travel patterns) (Ziemke et al., 2015). While for MATSim (Multi-Agent Transport Simulation) (Horni et al., 2016), the generation process for coordinates differs for different synthetic travel demand generation scenarios. In some cases, it is random, and in other cases, facility locations are sampled following a probability distribution of distances to different land-use facilities already provided at a discrete scale for a scenario region, as explained in Chapter 61 of Horni et al. (2016). Although limited, point-based methods can provide greater detail for routing decisions and a higher precision on the vehicle-miles travel estimates in ABM, which can be useful for detailed energy and emission analysis. However, point-based methods require more data (and are limited to data availability) and increase the model complexity, which can increase computational time.

This research focuses on developing point-based activity locations for Polaris (Auld et al., 2016), a high-performance, open-source agent-based framework designed for simulating large-scale ABM transportation systems. One key example of disaggregated data used by Polaris is the exact position and characterization of activity *locations*. Unlike more traditional models, trips in Polaris are modeled from point to point in the network and not from zone to zone. Figure 1 shows an example of some of the *locations* used by Polaris in the Chicago network and illustrates the representation of *location* point coordinates in the context of Polaris ABM. This paper defines ***location* points** as discrete geographical points with specific and unique land use within a TAZ and used as generation-attraction areas in the ABM.

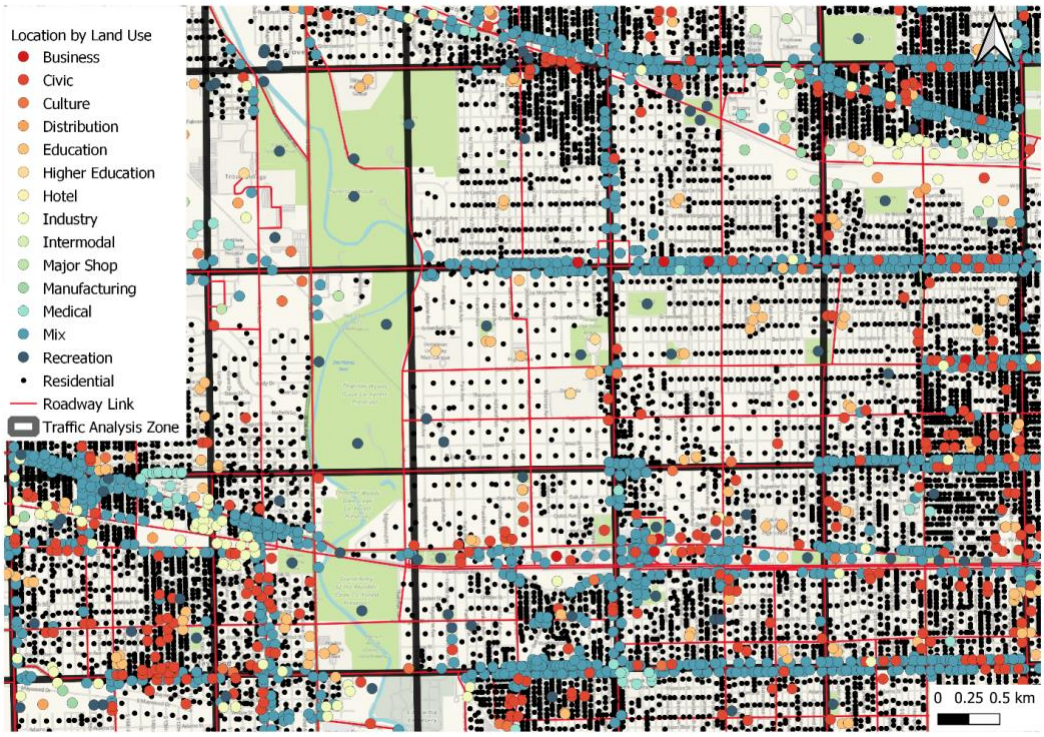


Figure 1. Example of Polaris locations in Chicago network

Identifying relevant *locations* for entire modeling regions is a challenging task and varies in difficulty depending on data availability for different cities. While residential *locations* can be inferred from multiple sources (e.g., Census), other general land-use types are more restricted (e.g., business and recreation). Without an authoritative source of such *locations* that could feed into Polaris models and ABM more generally, it would be convenient to have a solution for synthesizing *locations* consistent with zone-level socio-economic variables and modeling network, as well as reasonably distributed across the zone and along network links.

In considering solutions for such synthetization, it is relevant to consider that not all *locations* are created equal, as some are widely present throughout the network and highly correlated with the local population, employment, and network density (e.g., residential and retail *locations*), while others exist in much smaller numbers and are located in a fashion that is somewhat independent of the characteristics of its immediate surroundings (zone) and might have their physical address determined as a function of other similar *locations* (e.g., hospitals, police stations, government buildings).

1.1 Points of interest research

The general use of points of interest (PoI) data is well-established in the literature and provides a wide variety of techniques for data manipulation (Eckardt & Mateu, 2017; Kashian et al., 2019; Liu & Long, 2016; Yang et al., 2015). PoI data has long been used in transportation surveys to facilitate data collection (Bellemans et al., 2010), and there is also substantial literature on the use of PoIs in land-use and accessibility models (Adhvaryu et al., 2019; Fatima et al., 2019; Jiang et al., 2015; Zhang et al., 2018). Although its use in transport modeling, particularly ABM, is newer and more sparse (Klinkhardt et al., 2021; Niu & Li, 2019), the explosion of agent-based transportation models (Kagho et al., 2020) may result in substantial research in this field as model developers search for effective solutions for building transport models in a more automated manner.

Research on PoI use in modeling urban dynamics highlights the necessity of implementing robust validation procedures for PoI data (Yang et al., 2015), particularly crowdsource data (Hochmair et al., 2018; Kashian et al., 2019; Yeow et al., 2021). However, studies validating proprietary PoI data sources are limited due to a lack of ground-truth data to validate such sources. Therefore, the work we present here is relevant to the literature since it provides a methodology to generate PoI *locations* when the data is limited. We also provide a comparison of different sources of PoI data and test a validation framework that can be used when PoI data is needed.

1.2 Objectives and contributions

The main objective of this research is to provide a methodological framework for the estimation of activity *locations* in data-limited contexts, such as where parcel-level data and land-use information are not available. To this end, we proposed a two-component method. The first part consists of regression modeling for *locations* that are difficult to synthesize from open-source locations. A zero-inflated negative binomial (ZINB) regression model is proposed to tackle the excess of zeros across the zones. An efficient Bayesian approach with an advanced sampling methodology (Hoffman & Gelman, 2014) is proposed to quantify the coefficients' uncertainty, include priors beliefs, and provide metrics to quantify the zero-inflated probability across different zones. The second part provides a description of the estimation of *locations* that can be approximated using open-source PoI data. Through the proposed method, the expected outcome

consists of a method to approximate the count of location points by land use aggregated at a TAZ level. Specifically, we want to be able to replicate Figure 1 for any city in which data availability is challenging.

All the activities, at an agent-level scale, would have a *location* point (with unique land use) as the destination. This would provide static *locations* as a disaggregated set of feasible points for each agent. Agent activities are assigned to these static *locations* using a destination choice model, where TAZ areas are categorized based on their attraction potential given by the respective *location* points and land-use characteristics. A destination choice model, developed by (Auld & Mohammadian, 2011), selects a TAZ accounting for competition and agglomeration effects. For runtime purposes, it employs stratified sampling based on travel time and employment, with four strata and 200 alternatives each. Then, the choice of individual *location* points in a zone is assigned randomly based on the land use and activity type. *Location* points are connected to the roadway network, as shown in Figure 1, where secondary virtual connector links are defined to unify the roadway network with each individual point. These virtual connectors are akin to centroid connectors in traditional models.

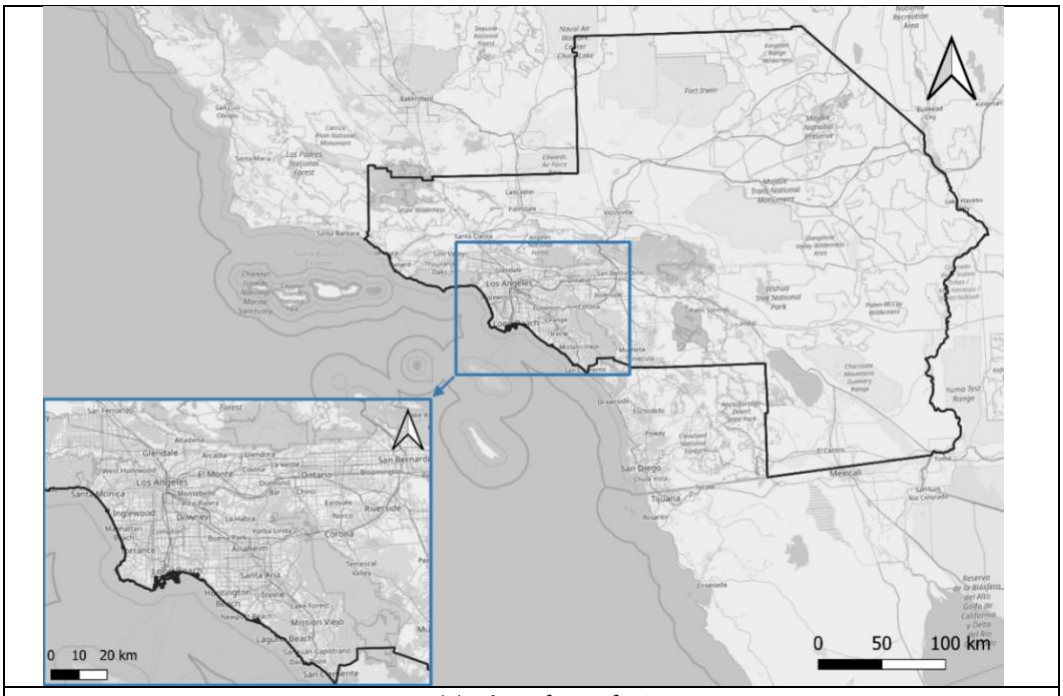
The principal contributions of this paper include (i) a data-driven approach to facilitate the determination of activity *locations* for ABM frameworks; (ii) the application of a Bayesian regression method in the context of *location* estimates that can be used to model similar data; (iii) a validation methodology for the use of crowdsourced PoI data. The remaining sections of this paper are organized as follows. The next section provides an overview of the data used in the analysis. A methodology section provides a description of the proposed methods. The results section presents the main findings and discusses the modeling results. Finally, the summary and conclusion section provide final remarks, limitations, and future work.

2 Data description

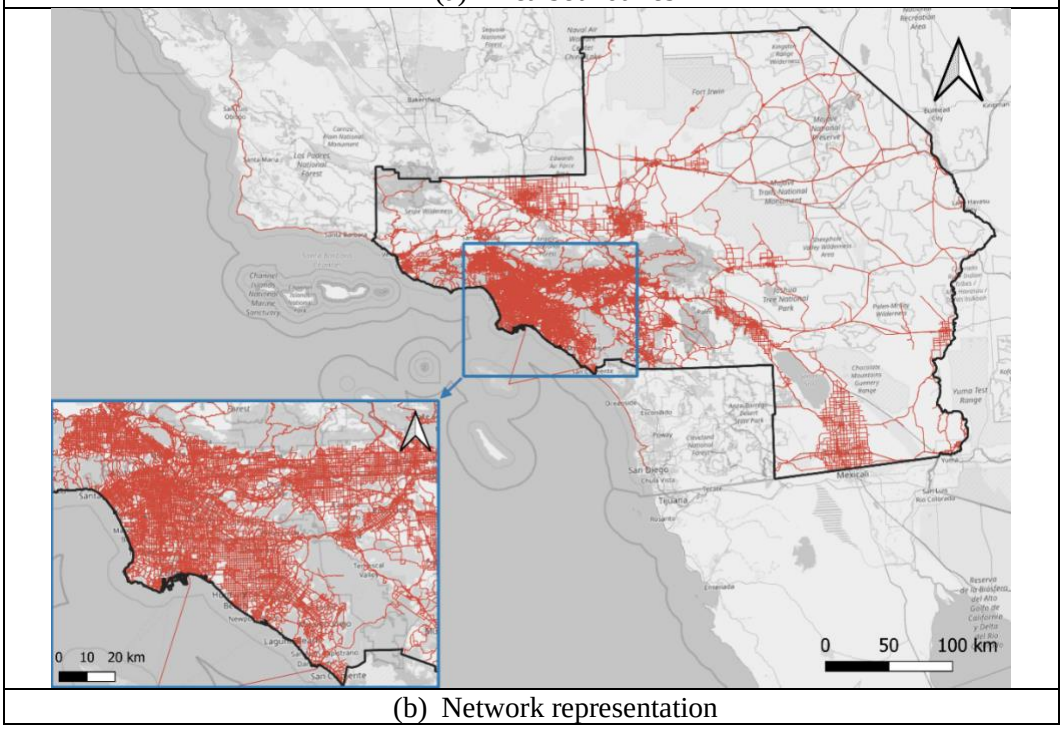
2.1 Data sources

We used parcel-level zoning data for the entire Southern California Association of Governments (SCAG) modeling area, available on its Regional Data Platform website (SCAG, 2022). The SCAG modeling area covers a total area of 103,000 km^2 or over 39,800 sq mi and nearly 43,000 parcels, as displayed in Figure 2a. It consists of six counties: Imperial (with 11% of the area and 5% of the parcels), Los Angeles (with 12% of the area and 13% of the parcels), Orange (with 2% of the area and 4% of the parcels), Riverside (with 18% of the area and 14% of the parcels), San Bernardino (with 51% of the area and 63% of the parcels), and Ventura (with 6% of the area and 2% of the parcels). Each parcel in the SCAG area is classified into one of 23 groups and 136 categories, detailing specific uses such as single dwelling units and commercial storage and more ill-defined uses such as mixed commercial and industrial land use.

In the SCAG region, there are 4,108 TAZs. This information, also obtained from the Regional Data Platform, is used to summarize land-use and parcel data. Other SCAG information used in the analysis includes demographic and employment information (summarized by TAZ) and the transportation roadway network information, as shown in Figure 2b. The roadway network provides information about the infrastructure type, such as freeway, heavy rail, principal road, and arterials, among others.



(a) Area boundaries



(b) Network representation

Figure 2. Southern California Association of Governments (SCAG) modeling area

2.2 Data processing

The parcel data was used to obtain land-use classifications summarized by zone. In this study, we focused on ten land-use classifications: (1) business, (2) recreation (e.g., playgrounds, parks, nature trails, hunting grounds), (3) civic (publicly or privately owned and managed facilities for meetings, conventions or exhibitions and other community,

social and multi-purpose uses), (4) manufacturing, (5) mix (*locations* can be used for multiple diverse activities), (6) multimodal (transport terminals, such as intercity bus/rail stations, and airports), (7) education (primary and secondary), (8) higher education, (9) medical, and (10) commercial hotels/motels. An overview of the distribution of the *location* count across the 4,000+ zones is shown in Figure 3. The data shows a significant number of zones with zero counts for certain classifications, with non-zero data ranging from 80% (business) to 0.2% (medical). This data distribution is relevant because it highlights the need to model data using a zero-inflated model that accommodates an excess of zeros in the data.

Furthermore, Figure 3 shows that zones with extreme values can be classified as outliers compared to the data distribution, such as the recreation type with a mean of 8.2 *locations* per zone and a maximum value of more than 8,000 counts. These outliers can be due to multiple reasons, such as the size of the zone (and therefore, the count by zone could be very big) or specialized areas (many *location* points in the same area), and do not represent the average distribution of the data for modeling purposes. Therefore, it is necessary to process the data to remove extreme/outlier values (defined as a data point that differs significantly from other observations) that can challenge the modeling process.

There are *location* types that are present in only a very small number of zones due to specific land use (hospitals, universities, hotels, schools) and, therefore, cannot be synthesized directly at the zone level. Assigning them to zones would be problematic, as the physical location of some *location* types depends on the physical locations of other *locations* of the same type and other types as well. (e.g., Hospitals will normally be spread around large metropolitan areas and not be far from residential agglomerations, but they may also be built in small clusters when they have complementary specializations.) However, the specificity of these "problematic" *location* types is an important asset that allows for the use of PoI data directly, as search parameters can be narrowed down to match each *location* type. The *location* types that can be better approximated using the PoI method are (1) education (primary and secondary), (2) higher education, (3) medical, and (4) commercial hotels/motels.

The remaining *location* types have between 8% and 80% of non-zero data and can be approximated using a regression method. These *locations* are (1) business, (2) recreation, (3) civic, (4) manufacturing, (5) mix, and (6) multimodal. For this data, the cleaning process consisted of identifying those values considered as extreme or outliers based on the data distribution. Specifically, we removed data points that differed significantly from other observations, and we verified that the removed outliers were located more than three standard deviations from the mean. Approximately 16% of the data was classified as containing an excessive number of counts per zone. Figure 4 shows the distribution of the location counts after the cleaning process. A total of 3,441 zones are used for the model.

Finally, roadway infrastructure information was obtained using a geographic overlay between the SCAG network and the TAZ zones to obtain the approximate number of centerline kilometers of link types in each zone. A summary of these estimates, as well as population and employment data per zone are shown in Table 1. This table also describes the location counts and corresponds to the data used in the modeling section.

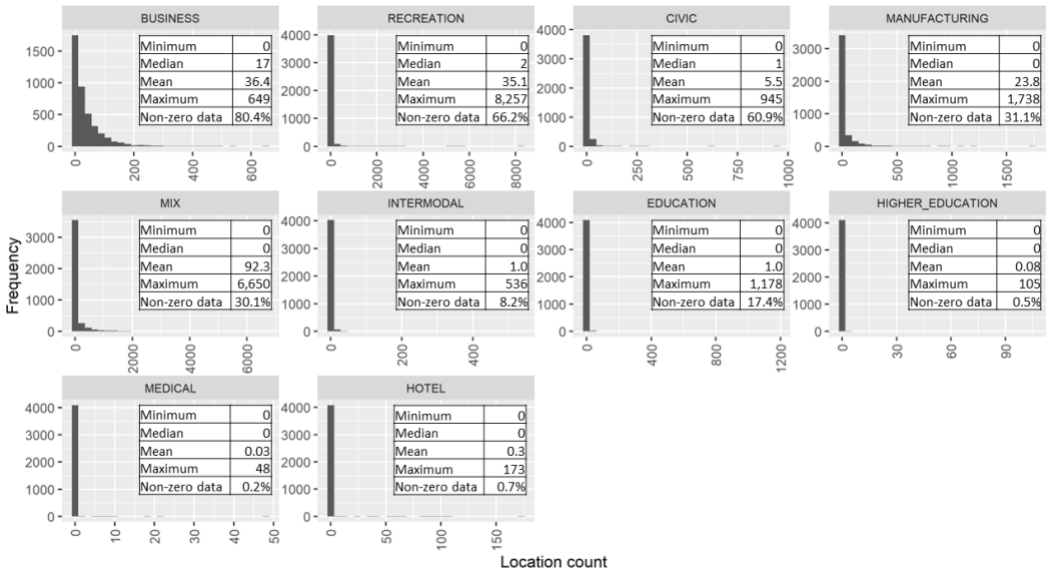


Figure 3. Raw data description of location counts

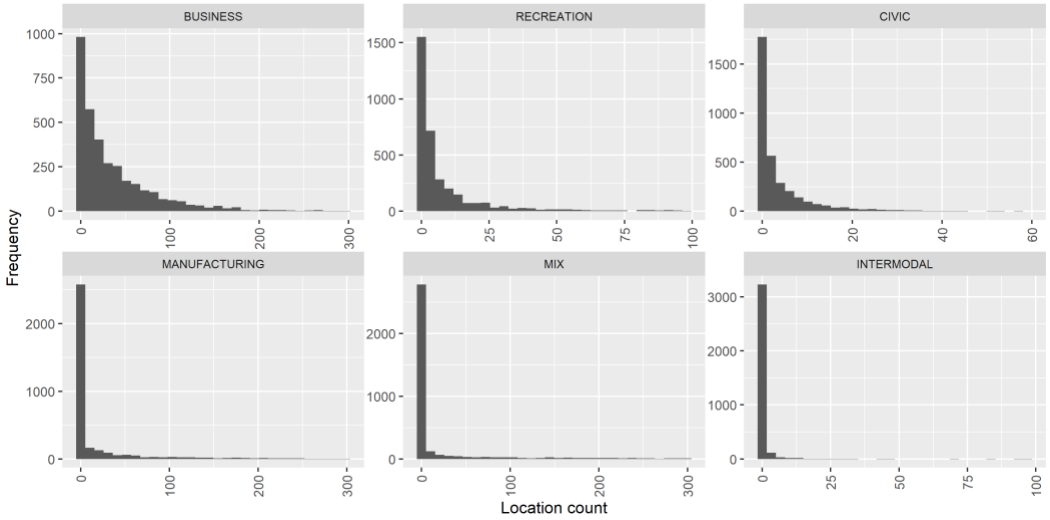


Figure 4. Distribution of location counts after the cleaning process

Table 1. Model data description

Variable	Units	Summary (sample = 3,441 zones)				
		Min	Median	Mean	Max	SD
Locations						
Business	Count	0.000	19.000	36.190	297.000	45.873
Recreation	Count	0.000	2.000	8.341	98.000	15.506
Civic	Count	0.000	1.000	4.229	59.000	7.358
Manufacturing	Count	0.000	0.000	17.630	299.000	44.895
Mix	Count	0.000	0.000	17.700	299.000	50.350
Intermodal	Count	0.000	0.000	0.639	97.000	4.307
Population						
Population density	1,000 / km^2	0.000	2.988	3.847	38.724	3.823
Household density	1,000 / km^2	0.000	0.950	1.284	23.109	1.500
Employment density						
Agricultural	1,000 / km^2	0.000	0.000	0.005	0.544	0.030
Construction	1,000 / km^2	0.000	0.027	0.065	2.529	0.145
Retail	1,000 / km^2	0.000	0.055	0.153	11.394	0.436
Manufacturing	1,000 / km^2	0.000	0.008	0.079	6.655	0.254
Wholesale	1,000 / km^2	0.000	0.008	0.056	6.814	0.227
Transportation	1,000 / km^2	0.000	0.011	0.059	6.452	0.188
Infrastructure						
Freeway	10,000 km	0.000	0.000	0.130	3.133	0.277
Heavy rail	10,000 km	0.000	0.000	0.026	0.815	0.076
Principal road	10,000 km	0.000	0.104	0.140	1.308	0.150
Local road	10,000 km	0.000	0.141	0.273	10.024	0.563
Minor road	10,000 km	0.000	0.161	0.223	3.650	0.247

Min = minimum, Max = maximum, SD = standard deviation.

3 Methodology

Our methodology is divided into two main sections. The first section corresponds to the count model description used to estimate the first set of location types. The second part consists of a PoI method for those *location* points with specific land use (hospitals, universities, hotels, schools) with a small count and located sparsely in the dataset; therefore, a statistical model would not be appropriate.

3.1 Regression model for the location counts

3.1.1 Model description

The regression model proposed in this work corresponds to a ZINB model aiming to address the excess of zeros in the data. The model is estimated under the Bayesian framework using the No-U-turn sampling (NUTS) method (Hoffman & Gelman, 2014), an extension of the Hamiltonian Monte Carlo (HMC) (Brooks et al., 2011) that reduces the need for a large number of iterations. This section provides a description of the method.

The negative binomial (NB) model is given by:

$$p(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)y!}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{y} \quad (1)$$

With mean μ and over-dispersion parameter θ with $\mu, \theta > 0$ and $\Gamma(.)$ is the gamma function. The implied variance by θ is $Var(y) - \mu\left(1 + \frac{\mu}{\theta}\right)$, thus smaller values of θ results in very high variance and $\theta \to \infty$ is equivalent to the Poisson model. Generally, a $\theta > 1.1$ is considered high and a Poison distribution is preferred over a NB one. The NB relates the parameter μ with covariates as follows:

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

Where $x_1, x_2, \ldots, x_n$ is a set of n observables covariates.

The ZINB is a two-component mixture model assigning a mass of p_i to the extra zeros and a mass of $(1 - p_i)$ to an NB count distribution, where $0 \leq p_i \leq 1$ and $i \in \{1,2, \ldots, m\}$ corresponds to the location type. Since we are modeling six *location* types, $m = 6$. For this m-dimensional set of location counts $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$, the ZINB distribution can be written as:

$$p(y_i: \mu_i, \theta_i) = \begin{cases} p_i + (1 - p_i)\left(1 + \frac{\mu_i}{\theta_i}\right)^{-\theta_i} & if\ y_i = 0, \\ (1 - p_i)\frac{\Gamma(y_i + \theta_i)}{\Gamma(\theta_i) y_i!}\left(1 + \frac{\mu_i}{\theta_i}\right)^{-\theta_i}\left(1 + \frac{\theta_i}{\mu_i}\right)^{-y_i} & if\ y_i > 0 \end{cases} \tag{3}$$

The ZINB regression model relates μ_i and p_i to the set of covariates $x_1, x_2, \ldots, x_n$ of size n as follows:

$$\log(\mu_i) = \delta_{i_0} + \delta_{i\ 1} x_1 + \delta_{i_2} x_2 + \cdots + \delta_{i_n} x_n \tag{4}$$

$$\log(p_i) = \zeta_{\mathrm{i}_0} + \zeta_{\mathrm{i}_1} x_1 + \zeta_{\mathrm{i}_2} x_2 + \cdots + \zeta_{\mathrm{i}_n} x_n \tag{5}$$

In our case, these covariates correspond to population, employment, and infrastructure information for each zone. The sample size n is the total number of zones after the cleaning process, corresponding to 3,441. Equation 4 correspond to the NB model and Equation 5 corresponds to the zero-inflated (ZI) model. The goal of the regression is to estimate the set of parameters of Equations 4 and 5, $\delta_{i1}, \delta_{i2}, \ldots, \delta_{in}$ and $\zeta_{i1}, \zeta_{i2}, \ldots, \zeta_{in}$ that allow the estimation of the distribution mean (μ_i) and the ZI probability (p_i). Thus, $p(y_i: \mu_i, \theta_i)$ could be used to predict estimations of location counts $\widehat{y_i}$ in new areas and p_i can be used to describe the probability that the area count is zero or ZI probability.

The model estimation is done in R software using the Bayesian regression models in Stan (brms) library (Bürkner, 2017). The estimation used 4 chains of 2,000 iterations each and 1,000 warm up iterations for a total of 4,000 sampling iterations. The variable selection is performed based on the coefficients' 95% confidence intervals (CI). The computational time of the final model is approximately 60 minutes on a Windows laptop with a 3 GHz Intel® processor (11th generation Core i7-1185G7) and 128 GB of memory. The priors used are:

$$\theta_i \sim gamma(0.01,\ 0.01) \tag{6}$$

$$\delta_{i_0} \sim student_t(3.0, -2.0,\ 2.5) \tag{7}$$

$$\zeta_{i_0} \sim logistic(0,\ 1) \tag{8}$$

And δ_{i_d}, ζ_{i_d}follow an improper flat prior with $d \in \{1,2, \dots, n\}$.

3.1.2 Using the model results

Although predicting the number of *locations* for a zone is the most complex phase in synthesizing a set of *locations* for an integrated ABM framework, it is also necessary to physically allocate each one of these *locations* to the zone by defining its geographic coordinates. Our approach to this task, which goes beyond the goals of this paper, consists of randomly selecting positions alongside roads in the network to place each one of the *locations* generated. Controlling the probability of choosing each link based on its network hierarchy (i.e., link type) and the distance from the centerline that should be chosen are both aspects that are currently being investigated and might be the topic of a future paper.

3.2 Point-of-Interest method

3.2.1 Description and data sources

As mentioned before, PoI data was used to generate *location* data for very specific facility types, namely education (primary and secondary schools), higher education (universities), medical (hospitals), and commercial hotels/motels. Although there is no shortage of data providers selling PoI data in the United States in 2022, the use of Open-Street Maps (OSM) would be preferred for both cost savings and expedience, as one can obtain the data on demand instead of waiting for specific procurement processes.

This use of OSM PoI data for characterizing the activity side of transportation models is not new (Klinkhardt et al., 2021), and other related uses, such as the modeling of an urban change, have also been explored (Zhang & Pfoser, 2019). Validating the OSM PoI data itself is still a substantial concern for those using it (Hochmair et al., 2018; Kashian et al., 2019; Liu & Long, 2016; Zhou et al., 2022), although performing this validation is not trivial in the absence of a reference *ground-truth*, which Yeow et al., (2021) (Yeow et al., 2021) has recently explored in detail. One difficulty in using the most common public APIs for data validation frequently omitted in the literature (Yeow et al., 2021) is the restrictive user terms imposed by these services, often precluding users from storing query results for subsequent analysis, making it virtually impossible to replicate results at a future point in time precisely.

In the face of these restrictions, we have chosen the HERE data API (HERE, 2022) as a reference dataset, as it provides a more generous free API use tier, and designed the validation process to run all necessary comparisons against the OSM data at runtime and dispose of the HERE data. One noteworthy aspect of the HERE data download, which is common to most public APIs providing PoI data, is that the number of return values for queries is limited to a rather small number of results (99 in HERE's case), so it is necessary to divide the search area into small areas for which we can guarantee (and monitor) that the number of PoIs returned is smaller than the 99 limit.

3.2.2 Validation methodology

We have used the validation procedures described in the literature (Yeow et al., 2021) as the basis for the procedures implemented in this analysis, but a few changes and

additions were made in order to look into some of the differences we located in the data. In a nutshell, the validation metrics used can be described as follows:

- **Mutual proximity** – Percentage of OSM *locations* within a certain distance from a HERE *location* and vice-versa. This is an important metric when multiple records for the same facility may exist (e.g., multiple buildings in a university).
- **Zones with *locations*** – The total number of TAZs with *locations* for each data source and a Pearson correlation coefficient between the vector totals for each data source.
- **Hexbins with *locations*** – The total number of Hexbins with *locations* for each data source and a Pearson correlation coefficient between the vector totals for each data source.

A Hexbin mesh with sides measuring 1 km covering the entire model area was created for this measure. Since deserts and national/state parks cover a large portion of the modeling area, we have limited the analysis to those hexbins for which there is at least one building in the Microsoft US Building Footprint dataset (Microsoft, 2022), taking that as a reasonable indication of human activity in the area, which resulted in approximately 50% of all hexbins to be removed from the analysis, and therefore from the data querying process.

4 Results and discussion

4.1 Count model estimate

The modeling results are summarized in Table 2, where the NB parameters (Equation 4) are shown first, followed by the ZI parameters (Equation 5). The models were first estimated using all the variables, and the variable selection procedure consisted of iteratively removing those variables that covered a significant area near the value 0 in the 95% CIs because this would mean that there is a high possibility that the true value is 0.

The estimates for the mean of the over-dispersion parameter θ are all under 1, suggesting that a NB model is preferred over a Poisson one. We can observe that as the number of zones with non-zero *location* counts decreases, the model shows a smaller over-dispersion parameter value and, therefore, a higher variance. Intermodal *locations*, with less than 10% of non-zero data, have an over-dispersion parameter θ of 0.066, while business *locations* (with more than 80% of non-zero data) show a $\theta = 0.797$.

The models' parameters in Table 2 (NB part) indicate that business, civic, and manufacturing *locations* are likely directly proportional to the population density, while recreation, mix, and intermodal points are predominantly in zones with lower population, as indicated by the coefficient sign. This is expected as the number of businesses is expected to be higher in areas where the population is high, which is the case of civic (public buildings or institutions owned and operated by governmental or other public agencies) and manufacturing *locations*. In contrast, recreational areas (such as parks, fields, and camping grounds) in California are typically in remote zones, similar to mixed-use and intermodal (transport facilities) areas.

The business *location* counts increase as the number of jobs in the retail sector increases, while it is lower where manufacturing jobs are predominant. The inverse effect is observed with the manufacturing counts. This result matches the expectation that a higher number of businesses are present where the number of retail employment is higher and, similarly, more manufacturing points are expected in areas with high manufacturing jobs. Also, it is observed that the count of recreational points decreases as the

employment density increases since, as discussed previously, most recreational *locations* in California are located away from urban areas. Employment density variables are not very significant for the mix and intermodal uses.

In terms of infrastructure types, although the number of heavy rail centerline kilometers is low, it was significant for most of the land use/types, and the effect is highest for the intermodal facilities, followed by manufacturing areas. As the number of principal roads' centerline length increases, more location counts are expected for business, recreation, civic, and manufacturing use, with the highest impact on business and manufacturing *locations*. Meanwhile, local and minor roads, such as centerline kilometers, also positively affect the number of *location* counts. The freeway centerline variable did not contribute significantly to the model. Infrastructure centerline length variables are insignificant when determining the count of mixed-use points.

Table 2. ZINB model estimation, NB part

	Business				Recreation				Civic			
Parameter (δ)	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**
Intercept	2.898	0.062	2.772	3.019	2.697	0.059	2.581	2.813	1.153	0.076	1.003	1.304
Population												
Population density	0.069	0.008	0.055	0.084	-0.208	0.012	-0.231	-0.186	0.087	0.023	0.044	0.131
Household density	–	–	–	–	–	–	–	–	-0.258	0.058	-0.371	-0.149
Employment												
Agricultural	1.815	0.783	0.391	3.467	-5.237	1.022	-7.110	-3.194	–	–	–	–
Construction	–	–	–	–	–	–	–	–	–	–	–	–
Retail	0.500	0.087	0.334	0.675	-0.369	0.080	-0.532	-0.217	-0.170	0.069	-0.300	-0.024
Manufacturing	-0.415	0.098	-0.598	-0.215	-0.612	0.151	-0.908	-0.320	–	–	–	–
Wholesale	–	–	–	–	-1.183	0.268	-1.713	-0.645	–	–	–	–
Transportation	–	–	–	–	–	–	–	–	0.538	0.225	0.108	1.012
Infrastructure												
Freeway	–	–	–	–	–	–	–	–	–	–	–	–
Heavy rail	0.989	0.292	0.435	1.561	–	–	–	–	1.951	0.384	1.211	2.718
Principal road	0.787	0.157	0.489	1.089	0.510	0.169	0.185	0.841	0.657	0.204	0.251	1.057
Local road	0.299	0.066	0.173	0.433	0.174	0.059	0.063	0.291	0.199	0.064	0.080	0.335
Minor road	0.943	0.100	0.752	1.137	–	–	–	–	0.356	0.124	0.117	0.604
Family Specific												
Dispersion (θ)	0.797	0.028	0.742	0.853	0.474	0.021	0.435	0.518	0.442	0.015	0.414	0.472
	Manufacturing				**Mix**				**Intermodal**			
Parameter	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**
Intercept	3.298	0.120	3.061	3.534	4.081	0.092	3.899	4.258	0.744	0.274	0.218	1.278
Population												
Population density	0.093	0.035	0.024	0.164	-0.068	0.057	-0.183	0.045	-0.289	0.039	-0.364	-0.212
Household density	-0.570	0.091	-0.749	-0.390	0.309	0.176	-0.024	0.669	–	–	–	–
Employment												
Agricultural	–	–	–	–	–	–	–	–	-5.683	5.785	-16.029	7.463
Construction	0.966	0.373	0.274	1.726	–	–	–	–	–	–	–	–
Retail	-0.235	0.102	-0.410	-0.011	–	–	–	–	–	–	–	–
Manufacturing	0.552	0.211	0.154	0.995	–	–	–	–	–	–	–	–
Wholesale	0.912	0.300	0.332	1.506	–	–	–	–	–	–	–	–
Transportation	0.720	0.258	0.262	1.268	0.292	0.486	-0.553	1.362	–	–	–	–
Infrastructure												
Freeway	–	–	–	–	-0.308	0.178	-0.644	0.053	-0.760	0.384	-1.459	0.054
Heavy rail	1.119	0.415	0.342	1.956	–	–	–	–	2.118	1.193	-0.048	4.594
Principal road	0.974	0.277	0.424	1.522	–	–	–	–	–	–	–	–
Local road	–	–	–	–	–	–	–	–	1.096	0.472	0.199	2.043
Minor road	0.493	0.191	0.126	0.876	–	–	–	–	–	–	–	–
Family Specific												
Dispersion (θ)	0.473	0.033	0.411	0.539	0.395	0.032	0.332	0.458	0.066	0.011	0.049	0.091

SE = standard error, L-95% CI = lower 95% confidence interval, U-95% CI = upper 95% confidence interval.

Table 2 (continuation). ZINB model estimation, ZI part

	Business				Recreation				Civic			
Parameter (ζ)	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**
Intercept	-1.735	0.116	-1.973	-1.517	-2.874	0.456	-3.900	-2.112	-1.035	0.583	-2.256	0.040
Population												
Population density	–	–	–	–	0.215	0.046	0.132	0.306	0.187	0.045	0.104	0.282
Household density	–	–	–	–	–	–	–	–	–	–	–	–
Employment												
Agricultural	-2.821	4.524	-15.125	2.490	–	–	–	–	–	–	–	–
Construction	–	–	–	–	-1.025	2.876	-7.463	1.489	–	–	–	–
Retail	–	–	–	–	-0.533	0.910	-3.136	0.361	-1.290	1.728	-5.691	0.674
Manufacturing	–	–	–	–	–	–	–	–	–	–	–	–
Wholesale	–	–	–	–	–	–	–	–	–	–	–	–
Transportation	–	–	–	–	–	–	–	–	–	–	–	–
Infrastructure												
Freeway	-0.744	0.368	-1.536	-0.099	-6.304	5.258	-20.872	-0.716	–	–	–	–
Heavy rail	–	–	–	–	–	–	–	–	-31.762	23.990	-88.862	-0.828
Principal road	-2.195	0.796	-3.908	-0.805	–	–	–	–	-9.064	4.992	-20.424	-1.136
Local road	–	–	–	–	-0.505	0.834	-2.735	0.433	-5.236	2.642	-11.331	-1.045
Minor road	0.124	0.244	-0.386	0.577	-3.408	2.191	-8.569	0.028	-15.784	5.221	-27.272	-6.722
	Manufacturing				**Mix**				**Intermodal**			
Parameter (ζ)	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**	**Mean**	**SE**	**L-95% CI**	**U-95% CI**
Intercept	1.454	0.129	1.200	1.706	0.233	0.099	0.037	0.425	0.870	0.275	0.298	1.388
Population												
Population density	–	–	–	–	0.092	0.041	0.013	0.173	–	–	–	–
Household density	0.612	0.075	0.467	0.758	0.273	0.125	0.035	0.523	–	–	–	–
Employment												
Agricultural	-6.481	5.260	-22.166	-0.807	–	–	–	–				
Construction	–	–	–	–	–	–	–	–	-11.568	5.511	-24.846	-3.523
Retail	1.823	0.341	1.198	2.500	0.600	0.239	0.174	1.091	–	–	–	–
Manufacturing	-20.447	2.717	-26.040	-15.500	–	–	–	–	-3.936	3.134	-11.735	0.316
Wholesale	-11.068	2.525	-16.176	-6.305	3.352	0.647	2.135	4.700	–	–	–	–
Transportation	-2.874	0.811	-4.523	-1.364	–	–	–	–	–	–	–	–
Infrastructure												
Freeway	-0.807	0.190	-1.181	-0.444	–	–	–	–	-1.729	0.893	-3.845	-0.342
Heavy rail	-4.550	0.865	-6.317	-2.962	–	–	–	–	-31.946	27.268	-110.673	-5.439
Principal road	-0.734	0.332	-1.396	-0.059	–	–	–	–	–	–	–	–
Local road	-0.313	0.121	-0.572	-0.100	–	–	–	–	0.645	0.305	0.114	1.295
Minor road	-0.405	0.214	-0.834	0.008	-0.430	0.171	-0.776	-0.102	–	–	–	–

SE = standard error, L-95% CI = lower 95% confidence interval, U-95% CI = upper 95% confidence interval.

The ZI model's parameter in Table 2 helps explain the excess of zeros in the TAZs. For example, for the *location* counts in recreation, mix, and civic points, an increase in the population density results in a higher probability that the area has zero *locations* of these types. Generally, as the number of centerline kilometers of the different infrastructure types analyzed decreases, the probability of zero counts in the zones increases. This result indicates that the roadway network present has a high impact on the *locations*' points since areas with no infrastructure tend to have lower to no *location* counts.

The Bayesian framework allows the estimation of the parameter distribution, and therefore, the coefficients' variability can be described. Figure 5 shows examples of parameters' distribution and trace plots for the NB part of the model. The parameter

distribution of the intercept of business, with a mean of 2.898 and [2.772, 3.019] as the 95% CIs (Table 2), shows a distribution that suggests that the parameter is likely non-zero. This is relevant because it gives confidence in the estimate and provides a measure of variability, unlike a frequentist estimate, where the parameter is given as a point value. The trace plots show the convergence of the chains in the sampling process. Four chains were used with 2,000 iterations each and 1,000 warm-up iterations. Figure 5 shows that the estimates reached stability before the warm-up iterations, which shows that the estimation reached convergency.

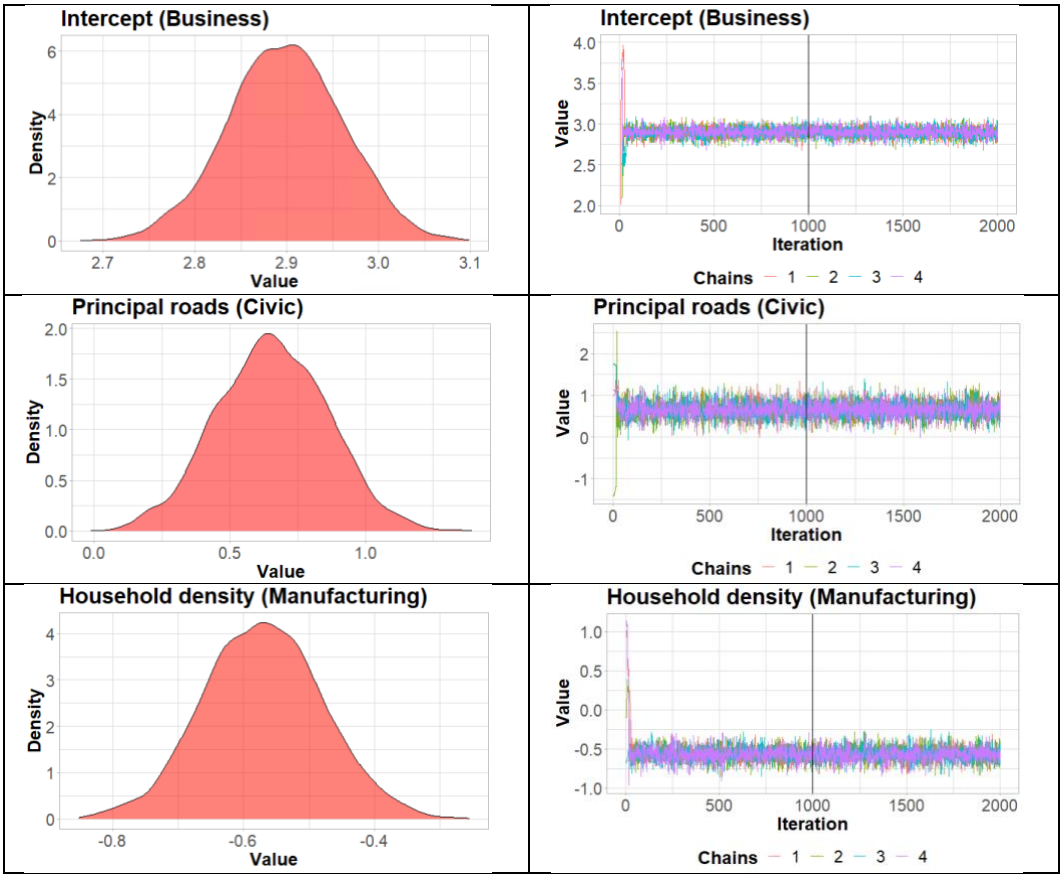


Figure 5. Parameter distribution and trace plots for the NB part of the ZINB model

Twenty posterior distributions were estimated and compared to the data distribution. Figure 6 shows the posterior distributions (dashed lines) and the training data distribution (shaded area) for each of the six *location* points models. The figure shows that the estimated posterior distributions are very similar to the data, suggesting that the estimate can be used to model the data accurately, although some variation can be observed. The *locations* with the higher number of zeros in the training data, and therefore, a significantly smaller number of zones with *location* counts, show a higher peak in the number of predicted zeros (a high-density concentration is around zero). These models also have the highest variance since the over-dispersion parameter θ is significantly lower than other *location* types. Therefore, it is likely that these models do not perform as well for these *location* types because they might overestimate zeros due to the lack of non-zero data available to provide confident *locations* counts estimates.

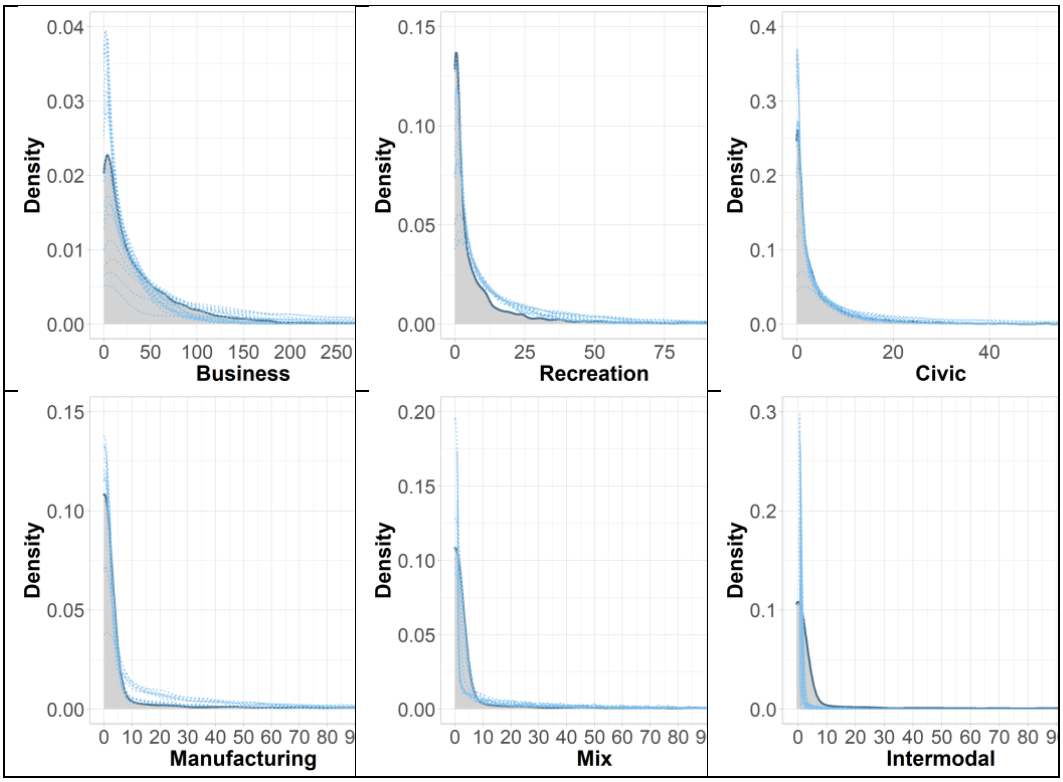


Figure 6. Posterior predictions (dotted lines) comparison with data distribution (shaded area)

The model estimates can be used to calculate the probability of having an area with zero *location* points of a specific use type by calculating the ZI probability. This estimation is useful because it can provide insights into the spatial distribution of the *location* points with a metric that can describe the probability of not having points in the area. For example, Figure 7 shows a comparison of the zonal ZI probabilities for business and recreational *locations*. It is expected that the number of business points is higher than the number of recreational points in the area. We can observe that the map shows a higher number of areas where recreational points are not expected (many areas have 95%-100% probabilities) while the business ZI probability is generally lower, meaning there are higher chances that these areas have non-zero business *locations* points.

4.2 PoI method

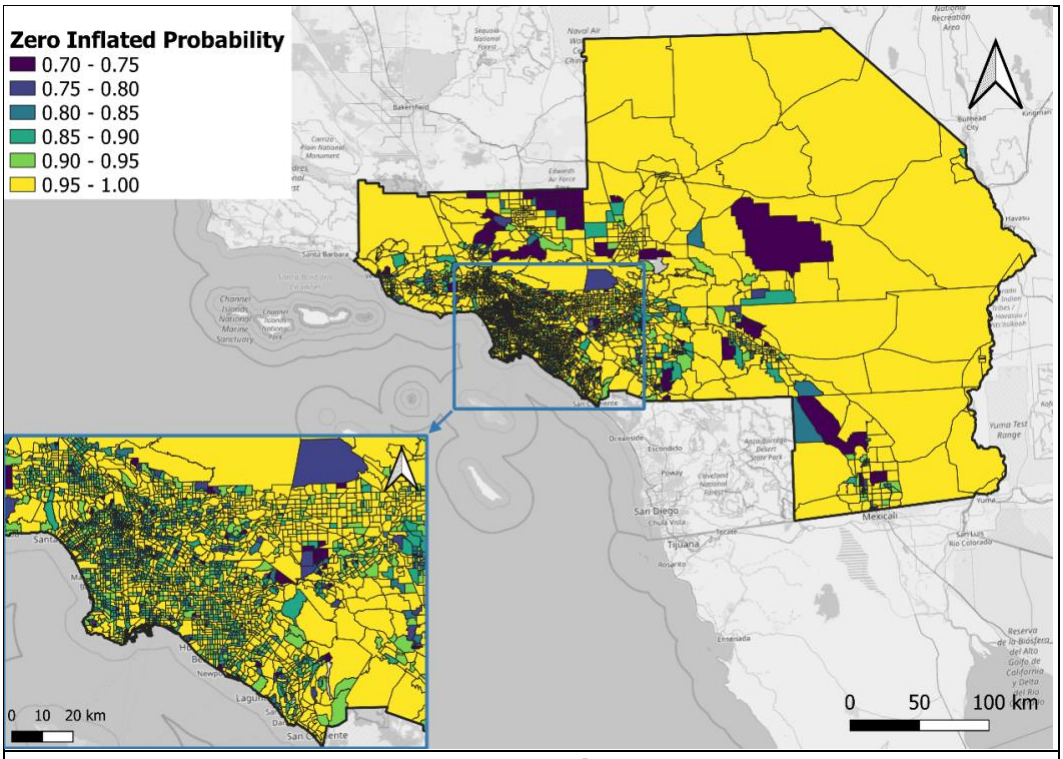
As the focus of the PoI methodology was on the validation of crowdsourced (OSM) data, the first step was to define the specific query parameters for both OSM and HERE and query both APIs to perform an initial comparison between both datasets for a point in time in early 2022. The code used in this section is available through the following GitHub repository: *link*.

The initial comparison showed that discrepancies between OSM and HERE are of different natures for each of the *location* types, as shown in Table 3, and therefore specific analysis for each *location* type is necessary.

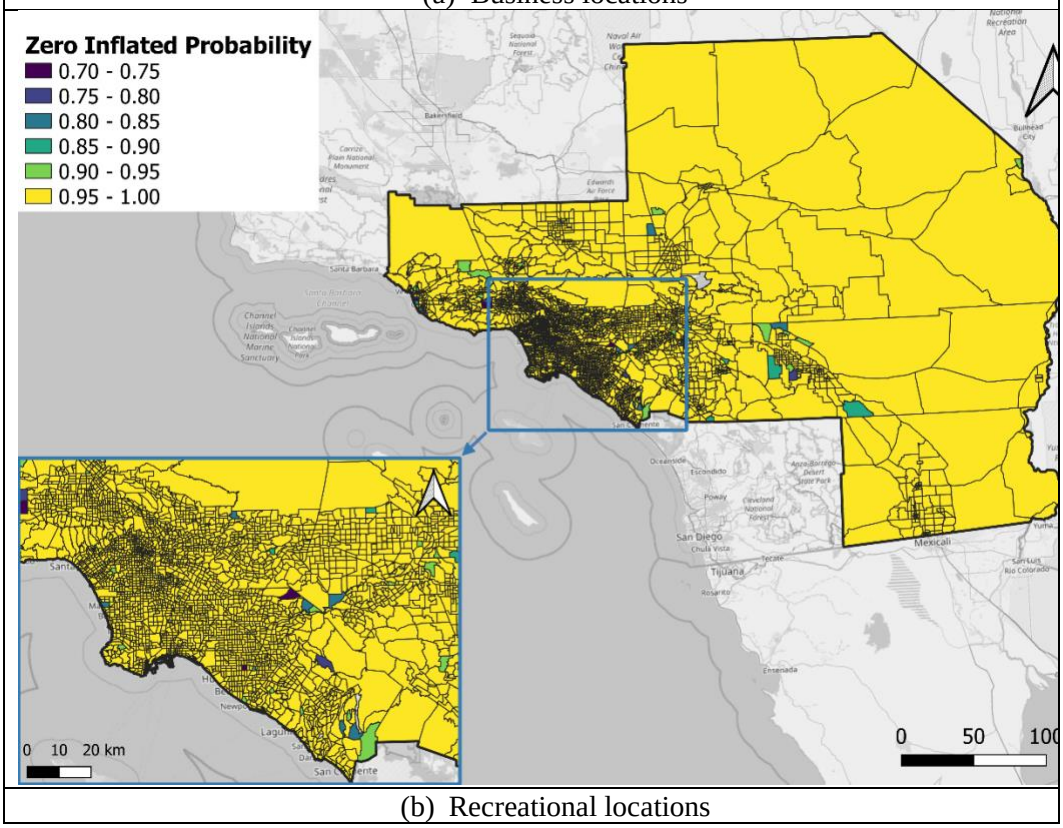
Table 3. Total PoI counts and validation metrics

Metric	Measure	Hospitals	Universities	Hotels	Schools[1]
Total PoIs	OSM	148	129	3,602	2,869
	HERE	148	231	3,965	665
Proximity	Isolated OSM locations	8	36	482	2,222
	Isolated HERE locations	8	154	750	179
Zone level	Pearson Correlation coefficient	0.93	0.34	0.53	0.07
	OSM present and HERE absent	7	35	241	1,652
	HERE present and OSM absent	8	141	522	214
Hexbin level	Pearson Correlation coefficient	0.87	0.40	0.60	0.19
	OSM present and HERE absent	16	34	247	1,390
	HERE present and OSM absent	17	137	485	147

[1] Includes primary and secondary schools only



(a) Business locations



(b) Recreational locations

Figure 7. Zero inflated probability by zone

4.2.1 Hospitals

Hospitals are identified with the value *'hospital'* for the *tag 'healthcare'* in OSM. We have found, however, that the tag is used somewhat loosely, so we have limited our OSM search to PoIs that were also tagged to have emergency departments, which is a stronger sign that the facility is what one would normally call a hospital. In the HERE dataset, we have limited our search to HERE's code *800-8000-0325* (HERE, 2021).

The total number of hospitals was virtually identical in both datasets. The spatial distribution of hospitals across the modeling area, however, shows more significant differences between the two datasets, and when looking into mutually close elements in both datasets, we verify that there are eight hospitals in each dataset that are more than 1 km from any other hospital in the other dataset. Similar results are found when looking into TAZs and hexbins that contain hospitals in one dataset but do not so for the other dataset.

4.2.2 Universities

Universities are identified with the value *'university'* for the *tag 'amenity'* in OSM, although building footprints and university dorms are also tagged with the value 'university'. In the HERE dataset, we have limited our search to HERE's code *800-8200-0173* (HERE, 2021).

The differences found for universities were not expected, as these tend to be very few across a metropolitan area while being significant enough that one would expect that all of them to be identified properly in either dataset being utilized in this analysis.

The key to this difference has its best clue in the number of *locations* from the HERE dataset that are "isolated." At close manual inspection, we have been able to identify dozens of points identified as universities in the HERE dataset that are not university campuses but rather offices for universities dedicated to remote learning or physical institutions located abroad. In that sense, the OSM dataset can be considered superior to the HERE dataset.

4.2.3 Hotels

Hotels are precisely identified with the value *'hotel'* for the *tag 'tourism'* in OSM. However, other short-term accommodations, such as hostels and motels, are equally relevant in transport models, and therefore all of them should be included in this dataset. Consequently, the tags *'hotel', 'hostel', 'motel', 'guest house', 'chalet'* and *'apartment'* were all included in the data query/building process. In the HERE dataset, we have limited our search to HERE's codes *500-5000-0000, 500-5000-0053, and 500-5000-0054.*

Once again, the number of *locations* in both datasets is very close, with OSM having less than 10% fewer locations than HERE. However, the spatial correlation between the two datasets could have been much better for such a close number of total *locations*. The most surprising number in this comparison is that there are over 522 zones (just over 12% of the total zones) for which HERE has at least one hotel record and for which OSM has none, in contrast with 241, where the reverse is true.

With this picture, it would seem that the OSM dataset might have more thorough coverage in some areas, whereas HERE seems to have more consistent coverage across the whole modeling area.

4.2.4 Schools

Schools were the most interesting case of all when we tried to identify only primary and secondary schools only. The HERE API has specific codes for those two types of facilities, respectively *800-8250-0287* and *800-8250-0288* (HERE, 2021), while the OSM data has to be parsed by looking at other tags such as names to identify only those two classes of schools.

As Table 3 shows clearly, the HERE API either misses a large number of schools grossly or is so strict with their classification that most schools in the modeling area would not fit in either of those categories. This led us to explore more lenient school classifications, including all points tagged as schools on OSM and the HERE category *800-8250-0000*, shown in Table 4.

It becomes clear that this is a more adequate comparison between the two datasets, especially when the Pearson correlation coefficient becomes the second highest among all *location* types. Other major differences emerge, especially concerning spatial coverage, made evident by the existence of over 600 zones for which only HERE has any school record, in contrast with just over 100 zones for which only OSM has a school *location* record. As shown in Figure 8, the lack of coverage in the OSM data happens in the densest parts of the modeling area and not in a concentrated manner, with the vast majority of the zones having at least one school *location* record in both OSM and HERE.

Table 4. Total PoIs and validation metrics revisited for the case of schools

Metric	Measure	Schools	All Schools
Total PoIs	OSM	2,869	5,615
	HERE	665	9,180
Proximity	Isolated OSM locations	2,222	149
	Isolated HERE locations	179	878
Zone level	Pearson Correlation coefficient	0.07	0.61
	OSM present and HERE absent	1,652	101
	HERE present and OSM absent	214	616
Hexbin level	Pearson Correlation coefficient	0.19	0.69
	OSM present and HERE absent	1,390	158
	HERE present and OSM absent	147	533

Given these differences, the superior dataset with respect to school location could be either HERE or OSM, depending on the specific use cases, with OSM being strongly preferred in cases where identification of primary and secondary school *locations* is critical.

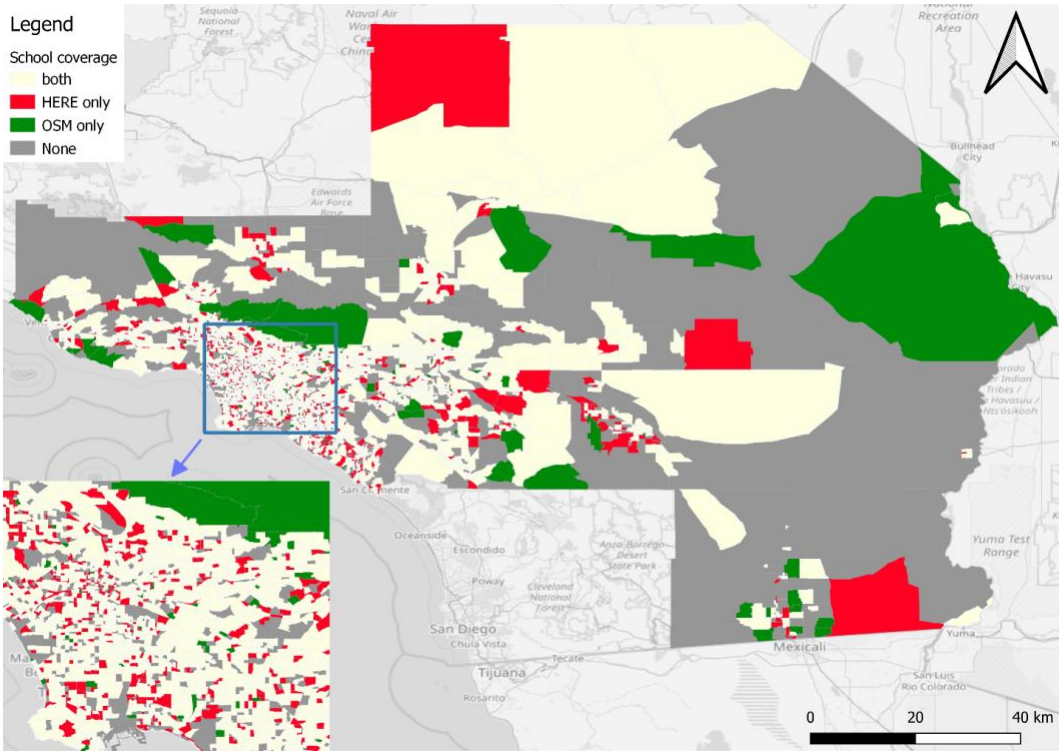


Figure 8. Zonal coverage for school locations

5 Summary and conclusions

This work presented methods to estimate activity *location* points using land-use parcel data from a 100,000+ km^2 area in California's SCAG region. We identified ten activity land-use/types and aggregated the information by TAZ along with data from other sources, such as the roadway network. The analysis of *location* count estimates by TAZ revealed that some land-use classification types have a high number of areas with zero counts and, therefore, cannot be synthesized directly at the zone level, but it brings the opportunity to use PoI data to complement the information. For this reason, our methodology is divided into two main methods. A set of land-use types is modeled using regression methods. A ZINB model is proposed to address the excess of zeros. The second set of land types can be approximated with PoI methods. We estimated ZINB models using a Bayesian framework that allows for the quantification of the coefficients' variability and proposed a PoI method through a comparison of two main sources, OSM and HERE maps, and the use of different validation metrics.

The regression estimates results showed that *location* counts can be approximated using population, employment density, and roadway network information. Results show that the count increases as the population density increases for business, civic, and manufacturing activities. Meanwhile areas with recreational, mixed, and intermodal use tend to be located in zones with lower population and employment density. The heavy rail centerline length variable had a significant effect on *location* counts despite its own low count, with the effect being the highest for the intermodal facilities, followed by manufacturing areas. As the number of principal roads' centerline length increases, a higher number of *location* counts is expected for business, recreation, civic, and manufacturing use, with the highest impact on business and manufacturing locations. Local and minor road centerline kilometers also have a positive and significant effect on the number of *location* counts. The roadway infrastructure centerline kilometers provided significant insights into the count estimates but were also relevant for the ZI model. Results show that as the number of centerline kilometers decreases, the probability that the zone has zero count increases significantly. The ZINB models can be used to estimate *location* counts but also provide additional information useful for further analysis, such as the ZI probability by area.

With regards to the use of PoI data, it was surprising to find that highly established commercial datasets such as HERE maps would have glaring imprecisions in the identification of categories like primary and secondary schools, allowing us to hypothesize that there may not be an appropriate "ground truth" reference for this type of data one could use. One of the advantages that the HERE data provides when it comes to ABM modeling is the consistently higher spatial coverage verified in all but one PoI category, where we verified that the number of TAZs and hexbins in which *locations* were present was much higher than what we verified for OSM. In this context, we believe that the OSM data is more than adequate as a source of discrete *locations* for discrete simulation transport models such as Polaris (Auld et al., 2016) and its open-data character and ease of access make it a sensible choice.

Although the data used in the analysis provides a complete overview of the area's *locations* and activities, some limitations must be highlighted. The land-use classification can be biased and ground truth data such as the validation using high-resolution satellite imagery or site visits. Also, land-use evolution over time can be an important factor in modeling ABM scenarios, and current data only provides the current estate.

This research can be found useful for the development of ABM *location* synthesis. The main results suggest that the proposed methodological framework can be used to estimate locations for ABM networks in a fast and efficient way, without the need for

detailed land-use information. Transportation planners and policymakers can use the results and methods provided in this research to approximate activity *location* distributions in ABM frameworks. Future research on the topic will be focused on methods to physically allocate each one of these *location* points to the zones by defining geographic coordinates.

Acknowledgments

The work done in this paper was sponsored by the U.S. Department of Energy (DOE) Vehicle Technologies Office (VTO) under the Systems and Modeling for Accelerated Research in Transportation (SMART) Mobility Laboratory Consortium, an initiative of the Energy Efficient Mobility Systems (EEMS) Program. The submitted manuscript has been created by the UChicago Argonne, LLC, Operator of Argonne National Laboratory (Argonne). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

Also, we gratefully acknowledge The University of California, Irvine (UCI) for facilitating the access to the data.

Author contribution

The authors confirm contribution to the paper as follows: writing-original draft preparation: Natalia Zuniga-Garcia and Pedro Veiga de Camargo; conceptualization and design: Pedro Veiga de Camargo; methodology: Natalia Zuniga-Garcia and Pedro Veiga de Camargo; data assembly and analysis: Pedro Veiga de Camargo and Natalia Zuniga-Garcia; writing-reviewing and editing: Pedro Veiga de Camargo and Natalia Zuniga-Garcia. All authors have reviewed the results and approved the final version of the manuscript.

References

Adhvaryu, B., Chopde, A., & Dashora, L. (2019). Mapping public transport accessibility levels (PTAL) in India and its applications: A case study of Surat. *Case Studies on Transport Policy*, *7*(2), 293–300. https://doi.org/10.1016/j.cstp.2019.03.004

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., & Zhang, K. (2016). POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transportation Research Part C: Emerging Technologies*, *64*, 101–116. https://doi.org/10.1016/j.trc.2015.07.017

Auld, J., & Mohammadian, A. (2011). Planning-constrained destination choice in activity-based model: Agent-based dynamic activity planning and travel scheduling. *Transportation Research Record*, *2254*, 170–179. https://doi.org/10.3141/2254-18

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., & Timmermans, H. (2010). Implementation framework and development trajectory of FEATHERS activity-based simulation platform. *Transportation Research Record*, *2175*(1), 111–119. https://doi.org/10.3141/2175-13

Bradley, M., Bowman, J. L., & 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. https://doi.org/10.1016/S1755-5345(13)70027-7

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Boca Raton, FL: CRC Press.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Eckardt, M., & Mateu, J. (2017). Analysing highly complex and highly structured point patterns in space. *Spatial Statistics*, *22*(2), 296–305. https://doi.org/10.1016/j.spasta.2017.04.007

Fatima, K., Moridpour, S., & Saghapour, T. (2019). Comparison of elderly public transport accessibility indices: Time-based methods. Paper presented at the Australasian Transport Research Forum, September 30-October 2, Canberra, Australia.

Galli, E., Cuéllar, L., Eidenbenz, S., Ewers, M., Mniszewski, S., & Teuscher, C. (2009). ActivitySim: Large-scale agent-based activity generation for infrastructure simulation. *Proceedings of the 2009 Spring Simulation Multiconference*, 1–9.

Habib, K. N. (2017). A comprehensive utility-based system of activity-travel scheduling options modelling (CUSTOM) for worker's daily activity scheduling processes. *Transportmetrica A: Transport Science*, *14*(4), 292–315. https://www.tandfonline.com/doi/abs/10.1080/23249935.2017.1385656

HERE. (2021). *HERE Geocoding & search API*. Retrieved from https://developer.here.com/documentation/geocoding-search-api/dev_guide/topics-places/places-category-system-full.html

HERE. (2022). *HERE REST APIs: Portable APIs for maps, routing and more*. Retrieved from https://developer.here.com/develop/rest-apis

Hochmair, H. H., Juhász, L., & Cvetojevic, S. (2018). Data quality of points of interest in selected mapping and social media platforms. In P. Kiefer, H. Huang, N. van de Weghe, & M. Raubal (Eds.), *Progress in location based services 2018* (pp. 293–313). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-71470-7_15

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(47), 1593–1623.

Horni, A., Nagel, K., & Axhausen, K. W. (2016). *The multi-agent transport simulation MATSim*. London: Ubiquity Press. http://www.oapen.org/record/613715

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., & Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land-use classification and disaggregation. *Computers, Environment and Urban Systems*, *53*, 36–46. https://doi.org/10.1016/j.compenvurbsys.2014.12.001

Kagho, G. O., Balac, M., & Axhausen, K. W. (2020). Agent-based models in transport planning: Current state, issues, and expectations. *Procedia Computer Science*, *170*, 726–732. https://doi.org/10.1016/j.procs.2020.03.164

Kashian, A., Rajabifard, A., Richter, K.-F., & Chen, Y. (2019). Automatic analysis of positional plausibility for points of interest in OpenStreetMap using coexistence patterns. *International Journal of Geographical Information Science*, *33*(7), 1420–1443. https://doi.org/10.1080/13658816.2019.1584803

Klinkhardt, C., Woerle, T., Briem, L., Heilig, M., Kagerbauer, M., & Vortisch, P. (2021). Using OpenStreetMap as a data source for attractiveness in travel demand models. *Transportation Research Record*, *2675*(8), 294–303. https://doi.org/10.1177/0361198121997415

Laarabi, H., Needell, Z., Waraich, R., Poliziani, C., & Wenzel, T. (2023). *BEAM: The modeling framework for behavior, energy, autonomy and mobility*. Berkeley, CA: Lawrence Berkeley National Laboratory.

Liu, X., & Long, Y. (2016). Automated identification and characterization of parcels with OpenStreetMap and points of interest. *Environment and Planning B: Planning and Design*, *43*(2), 341–360. https://doi.org/10.1177/0265813515604767

Microsoft. (2022, June 30). *Computer generated building footprints for the United States*. Retrieved from https://github.com/microsoft/USBuildingFootprints. (Original work published 2018.)

Niu, F., & Li, J. (2019). An activity-based integrated land-use transport model for urban spatial distribution simulation. *Environment and Planning B: Urban Analytics and City Science*, *46*(1), 165–178. https://doi.org/10.1177/2399808317705658

SCAG. (2022). *SCAG's regional data platform (RDP)*. Retrieved from https://hub.scag.ca.gov/

Yang, Y., Tang, J., Luo, H., & Law, R. (2015). Hotel location evaluation: A combination of machine learning tools and web GIS. *International Journal of Hospitality Management*, *47*, 14–24. https://doi.org/10.1016/j.ijhm.2015.02.008

Yeow, L. W., Low, R., Tan, Y. X., & Cheah, L. (2021). Point-of-Interest (POI) data validation methods: An urban case study. *ISPRS International Journal of Geo-Information*, *10*(11), 735. https://doi.org/10.3390/ijgi10110735

Zhang, L., & Pfoser, D. (2019). Using OpenStreetMap point-of-interest data to model urban change—A feasibility study. *PLOS One*, *14*(2), e0212606. https://doi.org/10.1371/journal.pone.0212606

Zhang, X., Li, W., Zhang, F., Liu, R., & Du, Z. (2018). Identifying urban functional zones using public bicycle rental records and point-of-interest data. *ISPRS International Journal of Geo-Information*, *7*(12), 459. https://doi.org/10.3390/ijgi7120459

Zhou, Q., Wang, S., & Liu, Y. (2022). Exploring the accuracy and completeness patterns of global land-cover/land-use data in OpenStreetMap. *Applied Geography*, *145*, 102742. https://doi.org/10.1016/j.apgeog.2022.102742

Ziemke, D., Nagel, K., & Bhat, C. (2015). Integrating CEMDAP and MATSIM to increase the transferability of transport demand models. *Transportation Research Record*, *2493*(1), 117–125. https://doi.org/10.3141/2493-13