# Likeness: a toolkit for connecting the social fabric of place to human dynamics

Joseph V. Tuccillo[‡*], James D. Gaboardi[‡]

**Abstract**—The ability to produce richly-attributed synthetic populations is key for understanding human dynamics, responding to emergencies, and preparing for future events, all while protecting individual privacy. The Likeness toolkit accomplishes these goals with a suite of Python packages: `pymedm`/`pymedm_legacy`, `livelike`, and `actlike`. This production process is initialized in `pymedm` (or `pymedm_legacy`) that utilizes census microdata records as the foundation on which disaggregated spatial allocation matrices are built. The next step, performed by `livelike`, is the generation of a fully autonomous agent population attributed with hundreds of demographic census variables. The agent population synthesized in `livelike` is then attributed with residential coordinates in `actlike` based on block assignment and, finally, allocated to an optimal daytime activity location via the street network. We present a case study in Knox County, Tennessee, synthesizing 30 populations of public K–12 school students & teachers and allocating them to schools. Validation of our results shows they are highly promising by replicating reported school enrollment and teacher capacity with a high degree of fidelity.

**Index Terms**—activity spaces, agent-based modeling, human dynamics, population synthesis

## Introduction

Human security fundamentally involves the functional capacity that individuals possess to withstand adverse circumstances, mediated by the social and physical environments in which they live [Hew97]. Attention to human dynamics is a key piece of the human security puzzle, as it reveals spatial policy interventions most appropriate to the ways in which people within a community behave and interact in daily life. For example, "one size fits all" solutions do not exist for mitigating disease spread, promoting physical activity, or enabling access to healthy food sources. Rather, understanding these outcomes requires examination of processes like residential sorting, mobility, and social transmission.

* *Corresponding author: tuccillojv@ornl.gov*
‡ *Oak Ridge National Laboratory*

*Copyright © 2022 Oak Ridge National Laboratory. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*
*Notice: This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).*

Modeling these processes at scale and with respect to individual privacy is most commonly achieved through agent-based simulations on *synthetic populations* [SEM14]. Synthetic populations consist of individual agents that, when viewed in aggregate, closely recreate the makeup of an area's observed population [HHSB12], [TMKD17]. Modeling human dynamics with synthetic populations is common across research areas including spatial epidemiology [DKA+08], [BBE+08], [HNB+11], [NCA13], [RSF+21], [SNGJ+09], public health [BCD+06], [BFH+17], [SPH11], [TCR08], [MCB+08], and transportation [BBM96], [ZFJ14]. However, a persistent limitation across these applications is that synthetic populations often do not capture a wide enough range of individual characteristics to assess how human dynamics are linked to human security problems (e.g., how a person's age, limited transportation access, and linguistic isolation may interact with their housing situation in a flood evacuation emergency).

In this paper, we introduce Likeness [TG22], a Python toolkit for connecting the social fabric of place to human dynamics via models that support increased spatial, temporal, and demographic fidelity. Likeness is an extension of the UrbanPop framework developed at Oak Ridge National Laboratory (ORNL) that embraces a new paradigm of "vivid" synthetic populations [TM21], [Tuc21], in which individual agents may be attributed in potentially hundreds of ways, across subjects spanning demographics, socioeconomic status, housing, and health. Vivid synthetic populations benefit human dynamics research both by enabling more precise geolocation of population segments, as well as providing a deeper understanding of how individual and neighborhood characteristics are coupled. UrbanPop's early development was motivated by linking models of residential sorting and worker commute behaviors [MNP+17], [MPN+17], [ANM+18]. Likeness expands upon the UrbanPop approach by providing a novel integrated model that pairs vivid residential synthetic populations with an activity simulation model on real-world transportation networks, with travel destinations based on points of interest (POIs) curated from location services and federal critical facilities data.

We first provide an overview of Likeness' capabilities, then provide a more detailed walkthrough of its central workflow with respect to `livelike`, a package for population synthesis and residential characterization, and `actlike` a package for activity allocation. We provide preliminary usage examples for Likeness based on 1) social contact networks in POIs 2) 24-hour POI occupancy characteristics. Finally, we discuss existing limitations and the outlook for future development.

### Overview of Core Capabilities and Workflow

UrbanPop initially combined the vivid synthetic populations produced from the American Community Survey (ACS) using the *Penalized-Maximum Entropy Dasymetric Modeling* (P-MEDM) method, which is detailed later, with a commute model based on origin-destination flows, to generate a detailed dataset of daytime and nighttime synthetic populations across the United States [MPN+17]. Our development of Likeness is motivated by extending the existing capabilities of UrbanPop to routing libraries available in Python like `osmnx`[1] and `pandana`[2] [Boe17], [FW12]. In doing so, we are able to simulate travel to regular daytime activities (work and school) based on real-world transportation networks. Likeness continues to use the P-MEDM approach, but is fully integrated with the U.S. Census Bureau's ACS Summary File (SF) and Census Microdata APIs, enabling the production of activity models on-the-fly.

Likeness features three core capabilities supporting activity simulation with vivid synthetic populations (Figure 1). The first, spatial allocation, is provided by the `pymedm` and `pmedm_legacy` packages and uses Iterative Proportional Fitting (IPF) to downscale census microdata records to small neighborhood areas, providing a basis for population synthesis. Baseline residential synthetic populations are then created and stratified into agent segments (e.g., grade 10 students, hospitality workers) using the `livelike` package. Finally, the `actlike` package models travel across agent segments of interest to POIs outside places of residence at varying times of day.

### Spatial Allocation: the `pymedm` & `pmedm_legacy` packages

Synthetic populations are typically generated from census microdata, which consists of a sample of publicly available longform responses to official statistical surveys. To preserve respondent confidentiality, census microdata is often published at spatial scales the size of a city or larger. Spatial allocation with IPF provides a maximum-likelihood estimator for microdata responses in small (e.g., neighborhood) areas based on aggregate data published about those areas (known as "constraints"), resulting in a baseline for population synthesis [WCC+09], [BBM96], [TMKD17]. UrbanPop is built upon a regularized implementation of IPF, the P-MEDM method, that permits many more input census variables than traditional approaches [LNB13], [NBLS14]. The P-MEDM objective function (Eq. 1) is written as:

$$\max -\sum_{it} \frac{n}{N} \frac{w_{it}}{d_{it}} \log \frac{w_{it}}{d_{it}} - \sum_{k} \frac{e_k^2}{2\sigma_k^2} \tag{1}$$

where $w_{it}$ is the estimate of variable $i$ in zone $t$, $d_{it}$ is the synthetic estimate of variable $i$ in location $t$, $n$ is the number of microdata responses, and $N$ is the total population size. Uncertainty in variable estimates is handled by adding an error term to the allocation $\sum_k \frac{e_k^2}{2\sigma_k^2}$, where $e_k$ is the error between the synthetic and published estimate of ACS variable $k$ and $\sigma_k$ is the ACS standard error for the estimate of variable $k$. This is accomplished by leveraging the uncertainty in the input variables: the "tighter" the margins of error on the estimate of variable $k$ in place $t$, the more leverage it holds upon the solution [NBLS14].

The P-MEDM procedure outputs an *allocation matrix* that estimates the probability of individuals matching responses from the ACS Public-Use Microdata Sample (PUMS) at the scale of census block groups (typically 300–6000 people) or tracts (1200–8000 people), depending upon the use-case.

Downscaling the PUMS from the Public-Use Microdata Area (PUMA) level at which it is offered (100,000 or more people) to these neighborhood scales then enables us to produce synthetic populations (the `livelike` package) and simulate their travel to POIs (the `actlike` package) in an integrated model. This approach provides a new means of modeling population mobility and activity spaces with respect to real-world transportation networks and POIs, in turn enabling investigation of social processes from the atomic (e.g., person) level in human systems.

Likeness offers two implementations of P-MEDM. The first, the `pymedm` package, is written natively in Python based on `scipy.optimize.minimize`, and while fully operational remains in development and is currently suitable for one-off simulations. The second, the `pmedm_legacy` package, uses `rpy2` as a bridge to [NBLS14]'s original implementation of P-MEDM[3] in R/C++ and is currently more stable and scalable. We offer `conda` environments specific to each package, based on user preferences.

Each package's functionality centers around a `PMEDM` class, which contains information required to solve the P-MEDM problem:

- The individual (household) level constraints based on ACS PUMS. To preserve households from the PUMS in the synthetic population, the person-level constraints describing household members are aggregated to the household level and merged with household-level constraints.
- PUMS household sample weights.
- The target (e.g., block group) and aggregate (e.g., tract) zone constraints based on population-level estimates available in the ACS SF.
- The target/aggregate zone 90% margins of error and associated standard errors ($SE = 1.645 \times MOE$).

The `PMEDM` classes feature a `solve()` method that returns an optimized P-MEDM solution and allocation matrix. Through a `diagnostics` module, users may then evaluate a P-MEDM solution based on the proportion of published 90% MOEs from the summary-level ACS data preserved at the target (allocation) scale.

### Population Synthesis: the `livelike` package

The `livelike` package generates baseline residential synthetic populations and performs agent segmentation for activity simulation.

#### *Specifying and Solving Spatial Allocation Problems*

The `livelike` workflow is oriented around a user-specified *constraints* file containing all of the information necessary to specify a P-MEDM problem for a PUMA of interest. "Constraints" are variables from the ACS common among people/households (PUMS) and populations (SF) that are used as both model inputs and descriptors. The constraints file includes information for bridging PUMS variable definitions with those from the SF using helper functions provided by the `livelike.pums` module, including table IDs, sampling universe (person/household), and tags for the range of ACS vintages (years) for which the variables are relevant.

1. https://github.com/gboeing/osmnx
2. https://github.com/UDST/pandana
3. https://bitbucket.org/nnnagle/pmedmrcpp

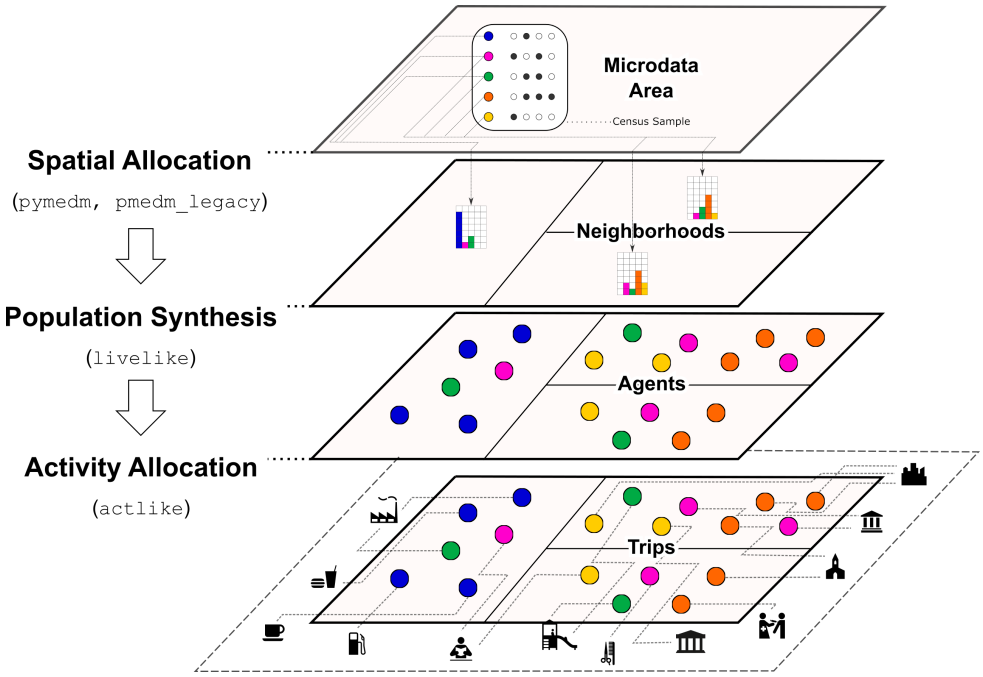


***Fig. 1:*** *Core capabilities and workflow of Likeness.*

The primary `livelike` class is the `acs.puma`, which stores information about a single PUMA necessary for spatial allocation of the PUMS data to block groups/tracts with P-MEDM. The process of creating an `acs.puma` is integrated with the U.S. Census Bureau's ACS SF and Census Microdata 5-Year Estimates (5YE) APIs[4]. This enables generation of an `acs.puma` class with a high-level call involving just a few parameters: 1) the PUMA's Federal Information Processing Standard (FIPS) code 2) the constraints file, loaded as a `pandas.DataFrame` and 3) the target ACS vintage (year). An example call to build an `acs.puma` for the Knoxville City, TN PUMA (FIPS 4701603) using the ACS 2015–2019 5-Year Estimates is:

```
acs.puma(
    fips="4701603",
    constraints=constraints,
    year=2019
)
```

The `censusdata` package[5] is used internally to fetch population-level (SF) constraints, standard errors, and MOEs from the ACS 5YE API, while the `acs.extract_pums_constraints` function is used to fetch individual-level constraints and weights from the Census Microdata 5YE API.

Spatial allocation is then carried out by passing the `acs.puma` attributes to a `pymedm.PMEDM` or `pmedm_legacy.PMEDM` (depending on user preference).

### *Population Synthesis*

The `homesim` module provides support for population synthesis on the spatial allocation matrix within a solved `P-MEDM` object. The population synthesis procedure involves converting the fractional estimates from the allocation matrix ($n$ household IDs by $m$ zones) to integer representation such that whole people/households are preserved. This `homesim` module features an implementation of [LB13]'s "Truncate, Replicate, Sample" (TRS) method. TRS works by separating each cell of the allocation matrix into whole-number (integer) and fractional components, then incrementing the whole-number estimates by a random sample of unit weights performed with sampling probabilities based on the fractional component. Because TRS is stochastic, the `homesim.hsim()` function generates multiple (default 30) realizations of the residential population. The results are provided as a `pandas.DataFrame` in long format, attributed by:

- PUMS Household ID (`h_id`)
- Simulation number (`sim`)
- Target zone FIPS code (`geoid`)
- Household count (`count`)

Since household and person-level attributes are combined when creating the `acs.puma` class, person-level records from the PUMS are assumed to be joined to the synthesized household IDs many-to-one. For example, if two people, $A01$ and $A03$, in household $A$ have some attribute of interest, and there are 3 households of type $A$ in zone $G$, then we estimate that a total of 6 people with that attribute from household $A$ reside in zone $G$.

### *Agent Generation*

The synthetic populations can then be segmented into different groups of agents (e.g., workers by industry, students by grade) for activity modeling with the `actlike` package. Agent segments may be identified in several ways:

- Using `acs.extract_pums_segment_ids()` to fetch the person IDs (household serial number + person line number) from the Census Microdata API matching some criteria of interest (e.g., public school students in $10^{\text{th}}$ grade).
- Using `acs.extract_pums_descriptors()` to fetch criteria that may be queried from the Census Microdata API. This is useful when dealing with criteria

4. https://www.census.gov/data/developers/data-sets.html
5. https://pypi.org/project/CensusData

more specific than can be directly controlled for in the P-MEDM problem (e.g., detailed NAICS code of worker, exact number of hours worked).

The function `est.tabulate_by_serial()` is then used to tabulate agents by target zone and simulation by appending them to the synthetic population based on household ID, then aggregating the person-level counts. This routine is flexible in that a user can use any set of criteria available from the PUMS to define customized agents for mobility modeling purposes.

### *Other Capabilities*

*Population Statistics*: In addition to agent creation, the `livelike.est` module also supports the creation of population statistics. This can be used to estimate the compositional characteristics of small neighborhood areas and POIs, for example to simulate social contact networks (see Students). To accomplish this, the results of `est.tabulate_by_serial` (see Agent Generation) are converted to proportional estimates to facilitate POIs (`est.to_prop()`), then averaged across simulations to produce Monte Carlo estimates and errors `est.monte_carlo_estimate()`).

*Multiple ACS Vintages and PUMAs*: The `multi` module extends the capabilities of `livelike` to multiple ACS 5YE vintages (dating back to 2016), as well as multiple PUMAs (e.g., a metropolitan area) via the `multi` module. Using `multi.make_pumas()` or `multi.make_multiyear_pumas()`, multiple PUMAs/multiple years may be stored in a `dict` that enables iterative runs for spatial allocation (`multi.make_pmedm_problems()`), population synthesis (`multi.homesim()`), and agent creation (`multi.extract_pums_segment_ids()`, `multi.extract_pums_segment_ids_multiyear()`, `multi.extract_pums_descriptors()`, and `multi.extract_pums_descriptors_multiyear()`). This functionality is currently available for `pmedm_legacy` only.

## Activity Allocation: the `actlike` package

The `actlike` package [GT22] allocates agents from synthetic populations generated by `livelike` POI, like schools and workplaces, based on optimal allocation about transportation networks derived from `osmnx` and `pandana` [Boe17], [FW12]. Solutions are the product of a modified integer program (Transportation Problem [Hit41], [Koo49], [MS01], [MS15]) modeled in `pulp` or `mip` [MOD11], [ST20], whereby supply (students/workers) are "shipped" to demand locations (schools/workplaces), with potentially relaxed minimum and maximum capacity constraints at demand locations. Impedance from nighttime to daytime locations (Origin-Destination [OD] pairs) can be modeled by either network distance or network travel time.

### *Location Synthesis*

Following the generation of synthetic households for the study universe, locations for all households across the 30 default simulations must be created. In order to intelligently site pseudo-neighborhood clusters of random points, we adopt a dasymetric [QC13] approach, which we term *intelligent block-based* (IBB) allocation, whereby household locations are only placed within blocks known to have been populated at a particular period in time and are placed with a greater frequency proportional to reported household density [LB13]. We employ population and housing counts within 2010 Decennial Census blocks to formulate a modified Variable Size Bin Packing Problem [FL86], [CGSdG08] for each populated block group, which allows for an optimal placement of household points and is accomplished by the `actlike.block_denisty_allocation()` function that creates and solves an `actlike.block_allocation.BinPack` instance.

### *Activity Allocation*

Once household location attribution is complete, individual agents must be allocated from households (nighttime locations) to probable activity spaces (daytime locations). This is achieved through spatial network modeling over the streets within a study area via OpenStreetMap[6] utilizing `osmnx` for network extraction & preprocessing and `pandana` for shortest path and route calculations. The underlying impedance metric for shortest path calculation, handled in `actlike.calc_cost_mtx()` and associated internal functions, can either take the form of distance or travel time. Moreover, household and activity locations must be connected to nearby network edges for realistic representations within network space [GFH20].

With a cost matrix from all residences to daytime locations calculated, the simulated population can then be "sent" to the likely activity spaces by utilizing an instance of `actlike.ActivityAllocation` to generate an adapted Transportation Problem. This mixed integer program, solved using the `solve()` method, optimally associates all population within an activity space with the objective of minimizing the total cost of impedance (Eq. 2), being subject to potentially relaxed minimum and maximum capacity constraints (Eq. 4 & 5). Each decision variable ($x_{ij}$) represents a potential allocation from origin $i$ to destination $j$ that must be an integer greater than or equal to zero (Eq. 6 & 7). The problem is formulated as follows:

$$\min \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} \tag{2}$$

$$\text{s.t.} \quad \sum_{j \in J} x_{ij} = O_i \quad \forall i \in I; \tag{3}$$

$$\text{s.t.} \quad \sum_{i \in I} x_{ij} \geq minD_j \quad \forall j \in J; \tag{4}$$

$$\text{s.t.} \quad \sum_{i \in I} x_{ij} \leq maxD_j \quad \forall j \in J; \tag{5}$$

$$\text{s.t.} \quad x_{ij} \geq 0 \quad \forall i \in I \quad \forall j \in J; \tag{6}$$

$$\text{s.t.} \quad x_{ij} \in \mathbb{Z} \quad \forall i \in I \quad \forall j \in J. \tag{7}$$

*where*

$i \in I$ = each household in the set of origins
$j \in J$ = each school in the set of destinations
$x_{ij}$ = allocation decision from $i \in I$ to $j \in J$
$c_{ij}$ = cost between all $i, j$ pairs
$O_i$ = population in origin $i$ for $i \in I$
$minD_j$ = minimum capacity $j$ for $j \in J$
$maxD_j$ = maximum capacity $j$ for $j \in J$

6. https://www.openstreetmap.org/about

The key to this adapted formulation of the classic Transportation Problem is the utilization of minimum and maximum capacity thresholds that are generated endogenously within `actlike.ActivityAllocation` and are tuned to reflect the uncertainty of both the population estimates generated by `livelike` and the reported (or predicted) capacities at activity locations. Moreover, network impedance from origins to destinations ($c_{ij}$) can be randomly reduced through an internal process by passing in an integer value to the `reduce_seed` keyword argument. By triggering this functionality, the count and magnitude of reduction is determined algorithmically. A random reduction of this nature is beneficial in generating dispersed solutions that do not resemble compact clusters, with an example being the replication of a private school's student body that does not adhere to public school attendance zones.

After the optimal solution is found for an `actlike.ActivityAllocation` instance, selected decisions are isolated from non-zero decision variables with the `realized_allocations()` method. These allocations are then used to generate solution routes with the `network_routes()` function that represent the shortest path along the network traversed from residential locations to assigned activity spaces. Solutions can be further validated with Canonical Correlation Analysis, in instances where the agent segments are stratified, and simple linear regression for those where a single segment of agents is used. Validation is discussed further in Validation & Diagnostics.

### Case Study: K–12 Public Schools in Knox County, TN

To illustrate Likeness' capability to simulate POI travel among specific population segments, we provide a case study of travel to POIs, in this case K–12 schools, in Knox County, TN. Our choice of K–12 schools was motivated by several factors. First, they serve as common destinations for the two major groups—workers and students—expected to consistently travel on a typical business day [RWM+17]. Second, a complete inventory of public school locations, as well as faculty and enrollment sizes, is available publicly through federal open data sources. In this case, we obtained school locations and faculty sizes from the Homeland Infrastructure Foundation-Level Database (HIFLD)[7] and student enrollment sizes by grade from the National Center for Education Statistics (NCES) Common Core of Data[8].

We chose the Knox County School District, which coincides with Knox County's boundaries, as our study area. We used the `livelike` package to create 30 synthetic populations for the Knoxville Core-Based Statistical Area (CBSA), then for each simulation we:

- Isolated agent segments from the synthetic population. K–12 educators consist of full-time workers employed as primary and secondary education teachers (2018 Standard Occupation Classification System codes 2300–2320) in elementary and secondary schools (NAICS 6111). We separated out student agents by public schools and by grade level (Kindergarten through Grade 12).
- Performed *IBB* allocation to simulate the household locations of workers and students. Our selection of household locations for workers and students varied geographically. Because school attendance in Knox County is restricted by district boundaries, we only placed student households in the PUMAs intersecting with the district (FIPS 4701601, 4701602, 4701603, 4701604). However, because educators may live outside school district boundaries, we simulated their household locations throughout the Knoxville CBSA.
- Used `actlike` to perform optimal allocation of workers and students about road networks in Knox County/Knoxville CBSA. Across the 30 simulations and 14 segments identified, we produced a total of 420 travel simulations. Network impedance was measured in geographic distance for all student simulations and travel time for all educator simulations.

Figure 2 demonstrates the optimal allocations, routing, and network space for a single simulation of 10th grade public school students in Knox County, TN. Students, shown in households as small black dots, are associated with schools, represented by transparent colored circles sized according to reported enrollment. The network space connecting student residential locations to assigned schools is displayed in a matching color. Further, the inset in Figure 2 provides the pseudo-school attendance zone for 10th graders at one school in central Knoxville and demonstrates the adherence to network space.

#### *Students*

Our study of K–12 students examines social contact networks with respect to potentially underserved student populations via the compositional characteristics of POIs (schools).

We characterized each school's student body by identifying student profiles based on several criteria: minority race/ethnicity, poverty status, single caregiver households, and unemployed caregiver households (householder and/or spouse/parnter). We defined 6 student profiles using an implementation of the density-based K-Modes clustering algorithm [CLB09] with a distance heuristic designed to optimize cluster separation [NLHH07] available through the `kmodes` package[9] [dV21]. Student profile labels were appended to the student travel simulation results, then used to produce Monte Carlo proportional estimates of profiles by school.

The results in Figure 3 reveal strong dissimilarities in student makeup between schools on the periphery of Knox County and those nearer to Knoxville's downtown core in the center of the county. We estimate that the former are largely composed of students in married families, above poverty, and with employed caregivers, whereas the latter are characterized more strongly by single caregiver living arrangements and, particularly in areas north of the downtown core, economic distress (pop-out map).

#### *Workers (Educators)*

We evaluated the results of our K–12 educator simulations with respect to POI occupancy characteristics, as informed by commute and work statistics obtained from the PUMS. Specifically, we used work arrival times associated with each synthetic worker (PUMS *JWAP*) to timestamp the start of each work day, and incremented this by daily hours worked (derived from PUMS *WKHP*) to create a second timestamp for work departure. The estimated departure time assumes that each educator travels to the school for a typical 5-day workweek, and is estimated as $JWAP + \frac{WKHP}{5}$.

7. https://hifld-geoplatform.opendata.arcgis.com
8. https://nces.ed.gov/ccd/files.asp
9. https://pypi.org/project/kmodes

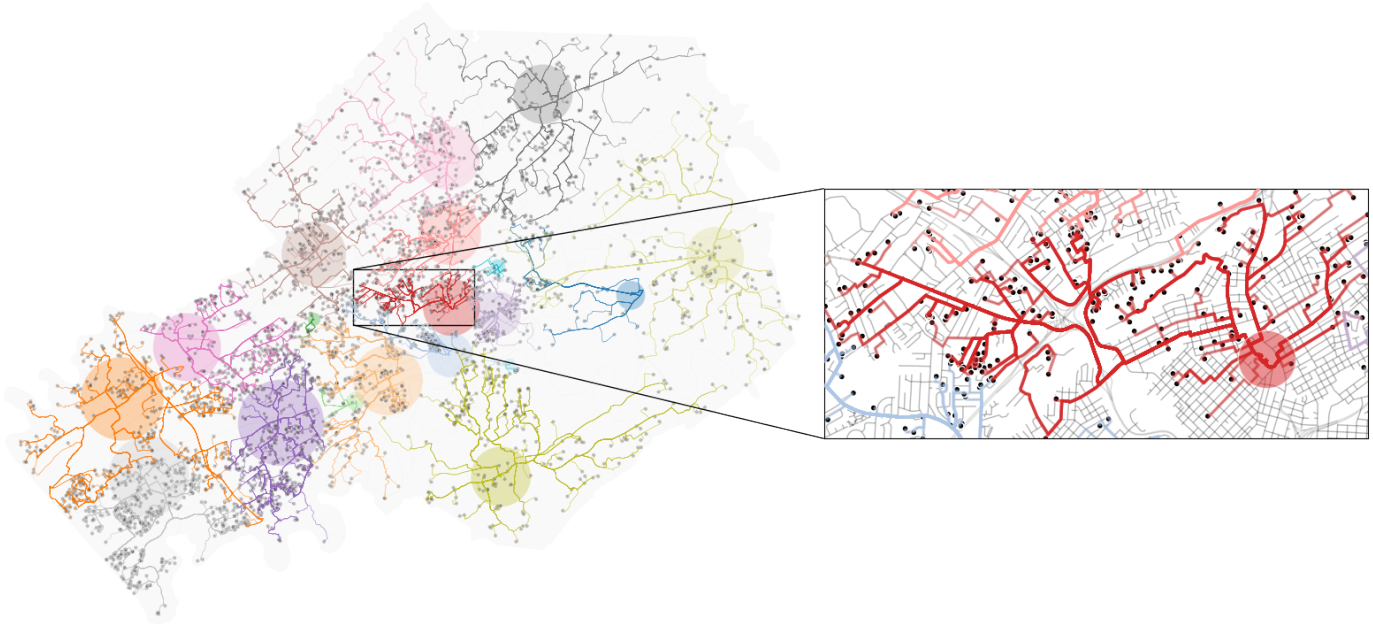

***Fig. 2:*** *Optimal allocations for one simulation of $10^{th}$ grade public school students in Knox County, TN.*


***Fig. 3:*** *Compositional characteristics of K–12 public schools in Knox County, TN based on 6 student profiles. Glyph plot methodolgy adapted from [GLC+15].*

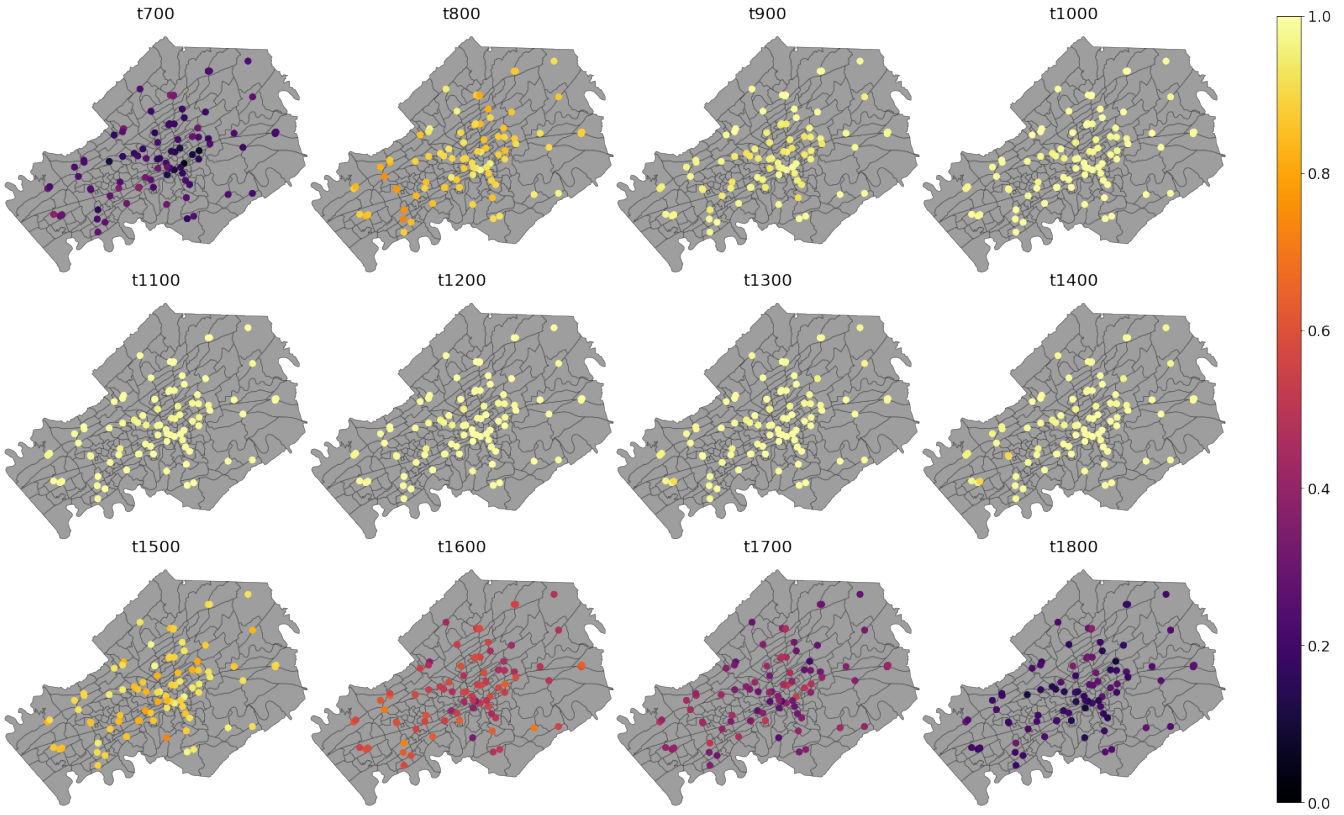


***Fig. 4:*** *Hourly worker occupancy estimates for K–12 schools in Knox County, TN.*

Roughly 50 educator agents per simulation were not attributed with work arrival times, possibly due to the source PUMS respondents being away from their typical workplaces (e.g., on summer or winter break) but still working virtually when they were surveyed. We filled in these unkown arrival times with the modal arrival time observed across all simulations (7:25 AM).

Figure 4 displays the hourly proportion of educators present at each school in Knox County between 7:00 AM (`t700`) and 6:00 PM (`t1800`). Morning worker arrivals occur more rapidly than afternoon departures. Between the hours of 7:00 AM and 9:00 AM (`t700`–`t900`), schools transition from nearly empty of workers to being close to capacity. In the afternoon, workers begin to gradually depart at 3:00 PM (`t1500`) with somewhere between 50%–70% of workers still present by 4:00 PM (`t1600`), then workers begin to depart in earnest at 5:00 PM into 6:00 PM (`t1700`–`t1800`), by which most have returned home.

Geographic differences are also visible and may be a function of (1) a higher concentration of a particular school type (e.g., elementary, middle, high) in this area and (2) staggered starts between these types (to accommodate bus schedules, etc.). This could be due in part to concentrations of different school schedules by grade level, especially elementary schools starting much earlier than middle and high schools[10]. For example, schools near the center of Knox County reach worker capacity more quickly in the morning, starting around 8:00 AM (`t800`), but also empty out more rapidly than schools in surrounding areas beginning around 4:00 PM (`t1600`).

### *Validation & Diagnostics*

A determination of modeling output robustness was needed to validate our results. Specifically, we aimed to ensure the preservation of relative facility size and composition. To perform this validation, we tested the optimal allocations of those generated by Likeness against the maximally adjusted reported enrollment & faculty employment counts. We used the maximum adjusted value to account for scenarios where the population synthesis phase resulted in a total demographic segment greater than reported total facility capacity. We employed Canonical Correlation Analysis (CCA) [Kna78] for the K–12 public school student allocations due to their stratified nature, and an ordinary least squares (OLS) simple linear regression for the educator allocations [PVG+11]. Because CCA is a multivariate measure, it is only a suitable diagnostic for activity allocation when multiple segments (e.g., students by grade) are of interest. For educators, which we treated as a single agent segment without stratification, we used OLS regression instead. The CCA for students was performed in two components: Between-Destination, which measures capacity across *facilities*, and Within-Destination, which measures capacity across *strata*.

Descriptive Monte Carlo statistics from the 30 simulations were run on the resultant coefficients of determination ($R^2$), which show a goodness of fit (approaching 1). As seen in Table 1, all models performed exceedingly well, though the Within-Destination CCA performed *slightly* less well than both the Between-Destination CCA and the OLS linear regression. In fact, the global minimum of all $R^2$ scores approaches 0.99 (students – Within-Destination), which demonstrates robust preservation of

10. https://www.knoxschools.org/Page/5553

| K–12 | $R^2$ Type | Min | Median | Mean | Max |
|---|---|---|---|---|---|
| Students (public schools) | Between-Destination CCA | 0.9967 | 0.9974 | 0.9973 | 0.9976 |
| | Within-Destination CCA | 0.9883 | 0.9894 | 0.9896 | 0.9910 |
| Educators (public & private schools) | OLS Linear Regression | 0.9977 | 0.9983 | 0.9983 | 0.9991 |

***TABLE 1:*** *Validating optimal allocations considering reported enrollment at public schools & faculty employment at all schools.*

true capacities in our synthetic activity modeling. Furthermore, a global maximum of greater than 0.999 is seen for educators, which indicates a near perfect replication of relative faculty sizes by school.

## Discussion

Our Case Study demonstrates the twofold benefits of modeling human dynamics with vivid synthetic populations. Using Likeness, we are able to both produce a more reasoned estimate of the neighborhoods in which people reside and interact than existing synthetic population frameworks, as well as support more nuanced characterization of human activities at specific POIs (e.g., social contact networks, occupancy).

The examples provided in the Case Study show how this refined understanding of human dynamics can benefit planning applications. For example, in the event of a localized emergency, the results of Students could be used to examine schools for which rendezvous with caregivers might pose an added challenge towards students (e.g., more students from single caregiver vs. married family households). Additionally, the POI occupancy dynamics demonstrated in Workers (Educators) could be used to assess the times at which worker commutes to/from places of employment might be most sensitive to a nearby disruption. Another application in the public health sphere might be to use occupancy estimates to anticipate the best time of day to reach workers, during a vaccination campaign, for example.

Our case study had several limitations that we plan to overcome in future work. First, we assumed that all travel within our study area occurs along road networks. While road-based travel *is* the dominant means of travel in the Knoxville CBSA, this assumption is not transferable to other urban areas within the United States. Our eventual goal is to build in additional modes of travel like public transit, walk/bike, and ferries by expanding our ingest of OpenStreetMap features.

Second, we do not yet offer direct support for non-traditional schools (e.g., populations with special needs, families on military bases). For example, the Tennessee School for the Deaf falls within our study area, and its compositional estimate could be refined if we reapportioned students more likely in attendance to that location.

Third, we did not account for teachers in virtual schools, which may form a portion of the missing work arrival times discussed in Workers (Educators). Work-from-home populations can be better incorporated into our travel simulations by applying work schedules from time-use surveys to probabilistically assign in-person or remote status based on occupation. We are particularly interested in using this technique with Likeness to better understand changing patterns of life during the COVID-19 pandemic in 2020.

## Conclusion

The Likeness toolkit enhances agent creation for modeling human dynamics through its dual capabilities of high-fidelity ("vivid") agent characterization and travel along real-world transportation networks to POIs. These capabilities benefit planners and urban researchers by providing a richer understanding of how spatial policy interventions can be designed with respect to how people live, move, and interact. Likeness strives to be flexible toward a variety of research applications linked to human security, among them spatial epidemiology, transportation equity, and environmental hazards.

Several ongoing developments will further Likeness' capabilities. First, we plan to expand our support for POIs curated by location services (e.g., Google, Facebook, Here, TomTom, FourSquare) by the ORNL PlanetSense project [TBP+15] by incorporating factors like facility size, hours of operation, and popularity curves to refine the destination capacity estimates required to perform `actlike` simulations. Second, along with multimodal travel, we plan to incorporate multiple trip models based on large-scale human activity datasets like the American Time Use Survey[11] and National Household Travel Survey[12]. Together, these improvements will extend our travel simulations to "non-obligate" population segments traveling to civic, social, and recreational activities [BMWR22]. Third, the current procedure for spatial allocation uses block groups as the target scale for population synthesis. However, there are a limited number of constraining variables available at the block group level. To include a larger volume of constraints (e.g., vehicle access, language), we are exploring an additional tract-level approach. P-MEDM in this case is run on cross-covariances between tracts and "supertract" aggregations created with the Max-*p*-regions problem [DAR12], [WRK21] implemented in PySAL's `spopt` [RA07], [FGK+21], [RAA+21], [FBG+22].

As a final note, the Likeness toolkit is being developed on top of key open source dependencies in the Scientific Python ecosystem, the core of which are, of course, `numpy` [HMvdW+20] and `scipy` [VGO+20]. Although an exhaustive list would be prohibitive, major packages not previously mentioned include `geopandas` [JdBF+21], `matplotlib` [Hun07], `networkx` [HSS08], `pandas` [pdt20], [WM10], and `shapely` [G+ ]. Our goal is contribute to the community with releases of the packages comprising Likeness, but since this is an emerging project its development to date has been limited to researchers at ORNL. However, we plan to provide a fully open-sourced code base within the coming year through GitHub[13].

## Acknowledgements

This material is based upon the work supported by the U.S. Department of Energy under contract no. DE-AC05-00OR22725.

## References

[ANM+18] H.M. Abdul Aziz, Nicholas N. Nagle, April M. Morton, Michael R. Hilliard, Devin A. White, and Robert N. Stew-

11. https://www.bls.gov/tus
12. https://nhts.ornl.gov
13. https://github.com/ORNL

art. Exploring the impact of walk–bike infrastructure, safety perception, and built-environment on active transportation mode choice: a random parameter model using New York City commuter data. *Transportation*, 45(5):1207–1229, 2018. doi:10.1007/s11116-017-9760-8.

[BBE+08] Christopher L. Barrett, Keith R. Bisset, Stephen G. Eubank, Xizhou Feng, and Madhav V. Marathe. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC'08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, pages 1–12. IEEE, 2008. doi:10.1109/SC.2008.5214892.

[BBM96] Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, 1996. doi:10.1016/0965-8564(96)00004-3.

[BCD+06] Dimitris Ballas, Graham Clarke, Danny Dorling, Jan Rigby, and Ben Wheeler. Using geographical information systems and spatial microsimulation for the analysis of health inequalities. *Health Informatics Journal*, 12(1):65–79, 2006. doi:10.1177/1460458206061217.

[BFH+17] Komal Basra, M. Patricia Fabian, Raymond R. Holberger, Robert French, and Jonathan I. Levy. Community-engaged modeling of geographic and demographic patterns of multiple public health risk factors. *International Journal of Environmental Research and Public Health*, 14(7):730, 2017. doi:10.3390/ijerph14070730.

[BMWR22] Christa Brelsford, Jessica J. Moehl, Eric M. Weber, and Amy N. Rose. Segmented Population Models: Improving the LandScan USA Non-Obligate Population Estimate (NOPE). American Association of Geographers 2022 Annual Meeting, 2022.

[Boe17] Geoff Boeing. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, September 2017. doi:10.1016/j.compenvurbsys.2017.05.004.

[CGSdG08] Isabel Correia, Luís Gouveia, and Francisco Saldanha-da Gama. Solving the variable size bin packing problem with discretized formulations. *Computers & Operations Research*, 35(6):2103–2113, June 2008. doi:10.1016/j.cor.2006.10.014.

[CLB09] Fuyuan Cao, Jiye Liang, and Liang Bai. A new initialization method for categorical data clustering. *Expert Systems with Applications*, 36(7):10223–10228, 2009. doi:10.1016/j.eswa.2009.01.060.

[DAR12] Juan C. Duque, Luc Anselin, and Sergio J. Rey. THE MAX-P-REGIONS PROBLEM*. *Journal of Regional Science*, 52(3):397–419, 2012. doi:10.1111/j.1467-9787.2011.00743.x.

[DKA+08] M. Diaz, J.J. Kim, G. Albero, S. De Sanjose, G. Clifford, F.X. Bosch, and S.J. Goldie. Health and economic impact of HPV 16 and 18 vaccination and cervical cancer screening in India. *British Journal of Cancer*, 99(2):230–238, 2008. doi:10.1038/sj.bjc.6604462.

[dV21] Nelis J. de Vos. kmodes categorical clustering library. https://github.com/nicodv/kmodes, 2015–2021.

[FBG+22] Xin Feng, Germano Barcelos, James D. Gaboardi, Elijah Knaap, Ran Wei, Levi J. Wolf, Qunshan Zhao, and Sergio J. Rey. spopt: a python package for solving spatial optimization problems in PySAL. *Journal of Open Source Software*, 7(74):3330, 2022. doi:10.21105/joss.03330.

[FGK+21] Xin Feng, James D. Gaboardi, Elijah Knaap, Sergio J. Rey, and Ran Wei. pysal/spopt, jan 2021. URL: https://github.com/pysal/spopt, doi:10.5281/zenodo.4444156.

[FL86] D.K. Friesen and M.A. Langston. Variable Sized Bin Packing. *SIAM Journal on Computing*, 15(1):222–230, February 1986. doi:10.1137/0215016.

[FW12] Fletcher Foti and Paul Waddell. A Generalized Computational Framework for Accessibility: From the Pedestrian to the Metropolitan Scale. In *Transportation Research Board Annual Conference*, pages 1–14, 2012. URL: https://onlinepubs.trb.org/onlinepubs/conferences/2012/4thITM/Papers-A/0117-000062.pdf.

[G+ ] Sean Gillies et al. Shapely: manipulation and analysis of geometric objects, 2007–. URL: https://github.com/shapely/shapely.

[GFH20] James D. Gaboardi, David C. Folch, and Mark W. Horner. Connecting Points to Spatial Networks: Effects on Discrete Optimization Models. *Geographical Analysis*, 52(2):299–322, 2020. doi:10.1111/gean.12211.

[GLC+15] Isabella Gollini, Binbin Lu, Martin Charlton, Christopher Brunsdon, and Paul Harris. GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software*, 63(17):1–50, 2015. doi:10.18637/jss.v063.i17.

[GT22] James D. Gaboardi and Joseph V. Tuccillo. Simulating Travel to Points of Interest for Demographically-rich Synthetic Populations, February 2022. American Association of Geographers Annual Meeting. doi:10.5281/zenodo.6335783.

[Hew97] Kenneth Hewitt. Vulnerability Perspectives: the Human Ecology of Endangerment. In *Regions of Risk: A Geographical Introduction to Disasters*, chapter 6, pages 141–164. Addison Wesley Longman, 1997.

[HHSB12] Kirk Harland, Alison Heppenstall, Dianna Smith, and Mark H. Birkin. 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, 2012. doi:10.18564/jasss.1909.

[Hit41] Frank L. Hitchcock. The Distribution of a Product from Several Sources to Numerous Localities. *Journal of Mathematics and Physics*, 20(1-4):224–230, 1941. doi:10.1002/sapm1941201224.

[HMvdW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi:10.1038/s41586-020-2649-2.

[HNB+11] Jan A.C. Hontelez, Nico Nagelkerke, Till Bärnighausen, Roel Bakker, Frank Tanser, Marie-Louise Newell, Mark N. Lurie, Rob Baltussen, and Sake J. de Vlas. The potential impact of RV144-like vaccines in rural South Africa: a study using the STDSIM microsimulation model. *Vaccine*, 29(36):6100–6106, 2011. doi:10.1016/j.vaccine.2011.06.059.

[HSS08] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008. URL: https://www.osti.gov/biblio/960616.

[Hun07] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi:10.1109/MCSE.2007.55.

[JdBF+21] Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, James McBride, Jacob Wasserman, Adrian Garcia Badaracco, Jeffrey Gerard, Alan D. Snow, Jeff Tratner, Matthew Perry, Carson Farmer, Geir Arne Hjelle, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Brendan Ward, Giacomo Caria, Mike Taves, Nick Eubank, sangarshanan, John Flavin, Matt Richards, Sergio Rey, maxalbert, Aleksey Bilogur, Christopher Ren, Dani Arribas-Bel, Daniel Mesejo-León, and Leah Wasser. geopandas/geopandas: v0.10.2, October 2021. doi:10.5281/zenodo.5573592.

[Kna78] Thomas R. Knapp. Canonical Correlation Analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85(2):410–416, 1978. doi:10.1037/0033-2909.85.2.410.

[Koo49] Tjalling C. Koopmans. Optimum Utilization of the Transportation System. *Econometrica*, 17:136–146, 1949. Publisher: [Wiley, Econometric Society]. doi:10.2307/1907301.

[LB13] Robin Lovelace and Dimitris Ballas. 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41:1–11, September 2013. doi:10.1016/j.compenvurbsys.2013.03.004.

[LNB13] Stefan Leyk, Nicholas N. Nagle, and Barbara P. Buttenfield. Maximum Entropy Dasymetric Modeling for Demographic Small Area Estimation. *Geographical Analysis*, 45(3):285–306, July 2013. doi:10.1111/gean.12011.

[MCB+08] Karyn Morrissey, Graham Clarke, Dimitris Ballas, Stephen Hynes, and Cathal O'Donoghue. Examining access to GP services in rural Ireland using microsimulation analysis. *Area*, 40(3):354–364, 2008. doi:10.1111/j.1475-4762.2008.00844.x.

[MNP+17] April M. Morton, Nicholas N. Nagle, Jesse O. Piburn, Robert N. Stewart, and Ryan McManamay. A hybrid dasymetric and machine learning approach to high-resolution residential electricity consumption modeling. In *Advances in Geocomputation*, pages 47–58. Springer, 2017. doi:10.1007/978-3-319-22786-3_5.

[MOD11] Stuart Mitchell, Michael O'Sullivan, and Iain Dunning. PuLP: A Linear Programming Toolkit for Python. Technical report, 2011. URL: https://www.dit.uoi.gr/e-class/modules/document/file.php/216/PAPERS/2011.%20PuLP%20-%20A%20Linear%20Programming%20Toolkit%20for%20Python.pdf.

[MPN+17] April M. Morton, Jesse O. Piburn, Nicholas N. Nagle, H.M. Aziz, Samantha E. Duchscherer, and Robert N. Stewart. A simulation approach for modeling high-resolution daytime commuter travel flows and distributions of worker subpopulations. In *GeoComputation 2017, Leeds, UK*, pages 1–5, 2017. URL: http://www.geocomputation.org/2017/papers/44.pdf.

[MS01] Harvey J. Miller and Shih-Lung Shaw. *Geographic Information Systems for Transportation: Principles and Applications*. Oxford University Press, New York, 2001.

[MS15] Harvey J. Miller and Shih-Lung Shaw. Geographic Information Systems for Transportation in the 21st Century. *Geography Compass*, 9(4):180–189, 2015. doi:10.1111/gec3.12204.

[NBLS14] Nicholas N. Nagle, Barbara P. Buttenfield, Stefan Leyk, and Seth Spielman. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1):80–95, 2014. doi:10.1080/00045608.2013.843439.

[NCA13] Markku Nurhonen, Allen C. Cheng, and Kari Auranen. Pneumococcal transmission and disease in silico: a microsimulation model of the indirect effects of vaccination. *PloS one*, 8(2):e56079, 2013. doi:10.1371/journal.pone.0056079.

[NLHH07] Michael K. Ng, Mark Junjie Li, Joshua Zhexue Huang, and Zengyou He. On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):503–507, 2007. doi:10.1109/TPAMI.2007.53.

[pdt20] The pandas development team. pandas-dev/pandas: Pandas, February 2020. doi:10.5281/zenodo.3509134.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL: https://www.jmlr.org/papers/v12/pedregosa11a.html.

[QC13] Fang Qiu and Robert Cromley. Areal Interpolation and Dasymetric Modeling: Areal Interpolation and Dasymetric Modeling. *Geographical Analysis*, 45(3):213–215, July 2013. doi:10.1111/gean.12016.

[RA07] Sergio J. Rey and Luc Anselin. PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, 37(1):5–27, 2007. URL: https://rrs.scholasticahq.com/article/8285.pdf, doi:10.52324/001c.8285.

[RAA+21] Sergio J. Rey, Luc Anselin, Pedro Amaral, Dani Arribas-Bel, Renan Xavier Cortes, James David Gaboardi, Wei Kang, Elijah Knaap, Ziqi Li, Stefanie Lumnitz, Taylor M. Oshan, Hu Shao, and Levi John Wolf. The PySAL Ecosystem: Philosophy and Implementation. *Geographical Analysis*, 2021. doi:10.1111/gean.12276.

[RSF+21] Krishna P. Reddy, Fatma M. Shebl, Julia H.A. Foote, Guy Harling, Justine A. Scott, Christopher Panella, Kieran P. Fitzmaurice, Clare Flanagan, Emily P. Hyle, Anne M. Neilan, et al. Cost-effectiveness of public health strategies for COVID-19 epidemic control in South Africa: a microsimulation modelling study. *The Lancet Global Health*, 9(2):e120–e129, 2021. doi:10.1016/S2214-109X(20)30452-6.

[RWM+17] Amy N. Rose, Eric M. Weber, Jessica J. Moehl, Melanie L. Laverdiere, Hsiu-Han Yang, Matthew C. Whitehead, Kelly M. Sims, Nathan E. Trombley, and Budhendra L. Bhaduri. LandScan USA 2016 [Data set]. Technical report, Oak Ridge National Laboratory, 2017. doi:10.48690/1523377.

[SEM14] Samarth Swarup, Stephen G. Eubank, and Madhav V. Marathe. Computational epidemiology as a challenge domain for multi-agent systems. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1173–1176, 2014. URL: https://www.ifaamas.org/AAMAS/aamas2014/proceedings/aamas/p1173.pdf.

[SNGJ+09] Beate Sander, Azhar Nizam, Louis P. Garrison Jr., Maarten J. Postma, M. Elizabeth Halloran, and Ira M. Longini Jr. Economic evaluation of influenza pandemic mitigation strategies in the United States using a stochastic microsimulation transmission model. *Value in Health*, 12(2):226–233, 2009. doi:10.1111/j.1524-4733.2008.00437.x.

[SPH11] Dianna M. Smith, Jamie R. Pearce, and Kirk Harland. Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health & Place*, 17(2):618–624, 2011. doi:10.1016/j.healthplace.2011.01.001.

[ST20] Haroldo G. Santos and Túlio A.M. Toffolo. Mixed Integer Linear Programming with Python. Technical report, 2020. URL: https://python-mip.readthedocs.io/_/downloads/en/latest/pdf/.

[TBP+15] Gautam S. Thakur, Budhendra L. Bhaduri, Jesse O. Piburn, Kelly M. Sims, Robert N. Stewart, and Marie L. Urban. PlanetSense: a real-time streaming and spatio-temporal analytics platform for gathering geo-spatial intelligence from open source data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–4, 2015. doi:10.1145/2820783.2820882.

[TCR08] Melanie N. Tomintz, Graham P. Clarke, and Janette E. Rigby. The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services. *Area*, 40(3):341–353, 2008. doi:10.1111/j.1475-4762.2008.00837.x.

[TG22] Joseph V. Tuccillo and James D. Gaboardi. Connecting Vivid Population Data to Human Dynamics, June 2022. Distilling Diversity by Tapping High-Resolution Population and Survey Data. doi:10.5281/zenodo.6607533.

[TM21] Joseph V. Tuccillo and Jessica Moehl. An Individual-Oriented Typology of Social Areas in the United States, May 2021. 2021 ACS Data Users Conference. doi:10.5281/zenodo.6672291.

[TMKD17] Matthias Templ, Bernhard Meindl, Alexander Kowarik, and Olivier Dupriez. Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software*, 79:1–38, 2017. doi:10.18637/jss.v079.i10.

[Tuc21] Joseph V. Tuccillo. An Individual-Centered Approach for Geodemographic Classification. In *11th International Conference on Geographic Information Science 2021 Short Paper Proceedings*, pages 1–6, 2021. doi:10.25436/E2H59M.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C.J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi:10.1038/s41592-019-0686-2.

[WCC+09] William D. Wheaton, James C. Cajka, Bernadette M. Chasteen, Diane K. Wagener, Philip C. Cooley, Laxminarayana Ganapathi, Douglas J. Roberts, and Justine L. Allpress. Synthesized population databases: A US geospatial database for agent-based models. *Methods report (RTI Press)*, 2009(10):905, 2009. doi:10.3768/rtipress.2009.mr.0010.0905.

[WM10] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi:10.25080/Majora-92bf1922-00a.

[WRK21] Ran Wei, Sergio J. Rey, and Elijah Knaap. Efficient re-

gionalization for spatially explicit neighborhood delineation. *International Journal of Geographical Information Science*, 35(1):135–151, 2021. doi:10.1080/13658816.2020.1759806.

[ZFJ14] Yi Zhu and Joseph Ferreira Jr. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record*, 2429(1):168–177, 2014. doi:10.3141/2429-18.