ORIGINAL PAPER

Open Access

A method to create a synthetic population with social networks for geographically-explicit agent-based models

Na Jiang^{1*} , Andrew T. Crooks² , Hamdi Kavak¹ , Annetta Burger³ and William G. Kennedy¹

Abstract

Geographically-explicit simulations have become crucial in understanding cities and are playing an important role in Urban Science. One such approach is that of agent-based modeling which allows us to explore how agents interact with the environment and each other (e.g., social networks), and how through such interactions aggregate patterns emerge (e.g., disease outbreaks, traffic jams). While the use of agent-based modeling has grown, one challenge remains, that of creating realistic, geographically-explicit, synthetic populations which incorporate social networks. To address this challenge, this paper presents a novel method to create a synthetic population which incorporates social networks using the New York Metro Area as a test area. To demonstrate the generalizability of our synthetic population method and data to initialize models, three different types of agent-based models are introduced to explore a variety of urban problems: traffic, disaster response, and the spread of disease. These use cases not only demonstrate how our geographically-explicit synthetic population can be easily utilized for initializing agent populations which can explore a variety of urban problems, but also show how social networks can be integrated into such populations and large-scale simulations.

Keywords: Synthetic population generation, Agent-based modeling, New York, Traffic dynamics, Disease, Disaster

1 Introduction

Over the last two decades, computational modeling approaches, and especially geographically-explicit simulations have become crucial in understanding the complex nature of cities and are playing an important role in Urban Science (Batty, 2007; Heppenstall et al., 2011) ranging from the micro-movement of pedestrians to that of macro patterns of urban growth (e.g., Crooks et al., 2015; Xie et al., 2007). Coinciding with this growth is proliferation of data from various sources (e.g., cell phones, social media, GPS logs) which are being used in geographical urban simulations in order to capture reality better (e.g., Ye et al., 2021; Huang et al., 2020; Manley et al., 2015). At the same time, there are growing demands for obtaining realistic

individual-level population data to simulate and capture more population characteristics or behaviors within the urban environment. Hence, researchers have placed considerable efforts to create artificial or synthetic populations (e.g., Barrett et al., 2009; Müller & Axhausen, 2011; Wise 2014; Gallagher et al., 2018; Lim, 2020). Besides academic researchers, research organizations, such as, US Census, Research Triangle Institute (RTI) International, are also attempting to build and share synthetic population datasets to support various research activities including computational urban modeling (Renardy et al., 2020).

However, the majority of existing synthetic population datasets share two common limitations. One is that they miss social networks, the other is that they can not be reused in different modeling applications. The lack of social networks is a concern as such networks play a critical role in various urban situations from that of crisis

*Correspondence: njiang8@gmu.edu

¹Department of Computational and Data Sciences George Mason University, Fairfax, VA, United States

Full list of author information is available at the end of the article

(e.g., Ersing & Kost, 2012), to that of disease spread (e.g., Pescosolido & Levy, 2002), and human mobility (e.g., Hu et al., 2021). As such, modelers are increasingly using social networks in urban simulations, especially those utilizing agent-based modeling (e.g., Hespenthal et al., 2021; Batty, 2013; Kim et al., 2020) and we would argue that such synthetic populations should also include social networks. The rationale for creating social networks is that they allow us to study the connections among various agents within different contexts (e.g., family, social group) and how such connections lead to aggregate patterns emerging.

As discussed, not only are social networks missing from existing publicly available synthetic populations, most of synthetic population datasets are often built to meet specific research requirements, such as, traffic simulation in a specific area (e.g., Wise et al., 2017) or disease spread in a specific city (e.g., Eubank et al., 2004) which makes it challenging to reuse such a specific synthetic population for different research questions or study areas. To overcome these limitations, the purpose of this paper is to introduce a new method that creates a synthetic population dataset incorporating social networks and share both of the method and resulting dataset with the scientific community to enable reuse of them as researchers see fit.

To illustrate the novelty of our proposed method and how it can be reused for different urban applications, we demonstrate three agent-based models which tackle three different types of problems all of which require realistic synthetic populations (i.e., traffic dynamics, disaster responses, and disease outbreaks). These three models are built in different programming languages (i.e., Java and Python), but they all use the synthetic population dataset generated by using the method introduced in Section 3. Our motivation for this is twofold. First to demonstrate the re-usability and generalizability of our synthetic population in different modeling languages and platforms. Second, by sharing the synthetic population and models, we provide a basic test environment to allow readers not only to replicate the results presented in this paper, but also to extend the models to their own application areas. In the remainder of this paper, we provide a brief background into synthetic population generation (Section 2) before introducing our methodology in Section 3. We then show the results in Section 4 before turning to illustrate how our dataset can be used to explore traffic dynamics, disaster responses, and disease outbreaks (Section 5). Finally in Section 6, we provide a summary of the paper and identify areas of further work.

2 Background

The vast majority of population synthesis methods utilized in agent-based models originate from microsimulation techniques (Orcutt, 1957) and involve a two-

step process of fitting a population to a set of relevant attributes and constraints and then generating individual units on the fitted population (Birkin & Wu, 2012; Müller & Axhausen, 2011; Wise, 2014). Traditionally, population synthesis methods can be broken into two main approaches: 1) synthetic reconstruction (SR); and 2) combinatorial optimization (CO) or re-weighting (Barthelemy & Toint, 2013; Pritchard & Miller, 2012). Also, in the last few years, we have witnessed an increasing number of studies using statistical (Sun & Erath, 2015; Saadi et al., 2016) and machine learning approaches (Borysov et al., 2019). However, such approaches are still in their infancy and not widely used. The SR method involves obtaining the joint distributions of relevant attributes and using Iterative Proportional Fitting (IPF) with the sample population to create a fitted population and generate individual households from that population (Deming & Stephan, 1940). While, CO involves creating a population and modifying it with the sample population until it meets a threshold of required constraints, for example a specific age distribution (Huang & Williamson, 2002; Wise, 2014). Both methods have their advantages and disadvantages. For example, CO can minimize errors by using constraints extracted from disaggregated datasets, such as, Public Use Microdata Areas (PUMAs) or Samples of Anonymized Records (SAR) (Farooq et al., 2013; Lim 2020; Wheaton et al., 2009). However, a disadvantage is that it is computationally intensive while at the same time needs detailed disaggregated data which is not always available. Such disaggregated data is not needed for SR.

Turning to applications, synthetic populations have been utilized to study demography, ecology, epidemiology, transportation, to name a few (e.g., Geard et al., 2013; Müller & Axhausen, 2011; Birkin & Wu, 2012). One methodology that has grown over the last two decades to capture the complexity of urban systems, is that of agent-based modeling (Crooks et al., 2019). At the core of agent-based models are individual agents which form a synthetic population, who interact with each other, and it is through these interactions aggregate patterns emerge (e.g., individuals driving cars can lead to traffic jams). Within agent-based modeling, synthetic populations are generally considered to be just the artificial society of the agents and we would argue that little attention has been given to how to incorporate realistic social networks for a given population based on actual real world demographic information. Even though there are many agent-based models that utilize social networks (e.g., Amblard et al., 2015), most of these networks were grown during the simulation (e.g., Pires & Crooks 2017; Kim et al., 2020), use stylized networks (e.g., Alizadeh et al., 2017), or simply assume adjacent agents are part of the same network (e.g., Agrawal et al., 2013). Others have created synthetic populations with realistic social networks (e.g., Eubank et al., 2004;

Barrett et al., 2009), however, the agents operate within a network and are not geographically-explicit. Others have created geographically-explicit synthetic populations with social networks, but in such works, the agents' social networks are based on social media connections and not family or workplace relationships (Wise, 2014). In the following section, we address these limitations and introduce a new method that creates a geographically-explicit synthetic population with realistic social networks.

3 Methodology

3.1 Study area

Before presenting our method, we first introduce our study area as shown in Fig. 1: New York City and its surrounding area (the state of Connecticut and parts of New Jersey, New York state, and Pennsylvania). The study area covers a 262 x 234 km area, home to 23,004,272 people based on the 2010 US Census. Our rationale for choosing this study area, is that it is slightly larger than the official metropolitan statistical area allowing us to capture high-density urban, suburban and rural areas, and thus it provides a more heterogeneous synthetic population. In addition, using this larger area allows potentially capturing more patterns of life, such as commuting patterns in and out the New York City.

3.2 Method details

Figure 2 shows the workflow of our synthetic population generation method along with the data and code packages used in each step. This method is derived from the

works of (Barthelemy & Toint, 2013) and (Wise, 2014), but includes extensions such as the creation of social networks based on real locations and the addition of educational sites as shown in Table 1. Before proceeding into the workflow, data preprocessing was carried out, which included ensuring the road network was topologically connected and no duplicate records (i.e., edges) existed. The first step was to create spaces on the road network to place homes and businesses based on the road type, for example, primary and secondary roads where used to constrain the location of homes and work spaces. We then created individuals using synthetic reconstruction based on the 2010 US Census data and grouped them to households. In step 2, individuals in households were then assigned workplaces consistent with the data from the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) dataset sourced from (US Census Bureau, 2019). Furthermore, we assigned younger household members to the closest daycare/schools based on their ages (e.g., daycare, elementary, middle, and high school) whose locations were sourced from the US Environmental Protection Agency (EPA) Office of Environmental Information (OEI) collected by (USEPA Office of Environmental Information, 2015). Lastly, three types of social networks were created based on (1) being in the same household, (2) working in the same workplace, or (3) attending the same educational institute. Household networks are fully connected, while their family members' school and work networks are based on interactions with individuals in these locations

Fig. 1 Study Area

Fig. 2 Workflow for Generation of Synthetic Population and Networks

based on concepts from small world social networks (e.g., Newman & Watts, 1999; Dunbar, 1998). Our synthetic population method was coded in Python. we made the source code, source data, and resulting synthetic data publicly available which allows others to repeat the processes presented in this paper or utilize the generated data (Code: <https://bit.ly/SynPopABM>; Source and resulting synthetic population data: <https://osf.io/3vsaj/>)

3.2.1 Step 1: create individuals and spaces

To create individuals and spaces in Step 1 (Fig. 2), there are two predominant ways to synthesize populations, as noted in in Section 2. While we could have chosen combinatorial optimization, we did not choose it because this method requires disaggregate level data to minimize the error during the population synthesis process as discussed in Section 2. Such data is not always available and as we wanted a generic method that could be applied to multiple areas (or countries) we chose synthetic reconstruction, as this method only requires data at the census tract level. Using synthetic reconstruction, we created individuals

to represent every person within every census tract and assign their sex and age based on information from the 2010 US Census data (US Census Bureau, 2010c). Specifically, we managed to create individuals under each age group of male and female. However, we did not consider ethnicity or race in our methods because the focus of this method is on locations and networks more than detailed demographic characteristics (e.g. race). Within this step, after creating the individuals, our method grouped them into households based on the household types within a tract using a uniform distribution, which is inspired by (Barthelemy & Toint, 2013) and (Wise, 2014). However, we also added some constraints related to age differences among the members under the same household to minimize the error between the synthetic dataset and US Census data. These constraints only applied to household types with children under 18 and were derived through trial and error with respect to population synthesis. The specifics of the constraints are: 1) husband's age – wife's age between (-4, +9); 2) father's age - child's age less than or equal to 50; 3) mother's age – child's age less than or

Table 1 Various synthetic population methods: adaptations and differences

	Geographically-explicit	Households Types	Daytime Locations	Social Networks
Barthelemy and Toint (2013)	No	6	No	No
Wise (2014)	Yes	12	Work Sites	Ego and Social Media Networks
Our Method	Yes	12	Work and Educational Sites	Small-World Network

equal to 40. The US Census categorizes households into 11 types (such as, husband-and-wife families, single childless adult households, households with a child less than 18, and male and female single householders over 65). However, based on our analysis of the data, we added one more type, that of people living in group quarters, which can be institutional (e.g., correctional facilities for adults, juvenile facilities, nursing facilities/skilled-nursing facilities) or non-institutional (e.g., college/university student housing, or military quarters). We further assumed that for each tract, there's only one example of group quarters and those who belong to a group quarters all live in this location. Hence, there are total of 12 types households in our synthetic population.

3.2.2 Step2: assign daytime locations

Daytime locations (i.e., workplaces and daycare/schools) were assigned to every agent generated in Step 1. Previous methods such as (Barthelemy & Toint, 2013) did not include such characteristics and (Wise, 2014) only considered workplaces. Our method also placed home and

work locations along the road network to make the synthetic population geographically-explicit. Such a process was not carried out in (Barthelemy & Toint, 2013) synthetic population work and hence their population is not geographically-explicit. While our approach potentially requires pre-processing of geographic maps compared to procedurally-generated cities (Kim et al. 2018), it is more realistic to represent geographic areas. Furthermore, to mimic the “real” world we also took into consideration general zoning, where we restricted businesses to secondary roads with the exception of religious centers and daycare/schools that may be located on residential roads. No businesses were allowed to be placed on primary roads (i.e., highways or motorways) as these are divided and have limited-access (US Census Bureau, 2010a). Without detailed land parcel information and the wish to keep the method simple, houses were placed on local roads at point locations at least 50m apart (see Fig. 3 (c)) or on top of each other when population density is high (e.g., representing apartment complexes, see Fig. 3 (d)) based on the number of occupied housing units in each census

tract. Workplaces were randomly placed either onto secondary roads, approximately 20m apart or at local road intersections. The number of workplaces in each census tract is disaggregated from County level business establishment counts from the US Census Bureau's County Business Patterns (US Census Bureau, 2010b). To determine the size of the population in each workplace, we used a lognormal distribution within census tracts based on findings that job size distributions in U.S. cities are lognormal (Eeckhout, 2004). The information of commuting patterns related to the work places was derived from the LEHDs LODS dataset based on (US Census Bureau, 2019). After we aggregated the information to the tract-level, we assigned work-age agents to a random workplace location within a tract based on the origin-destination statistics. Data for school locations were extracted from the Educational Institution dataset retrieved from the EPA OEI (USEPA Office of Environmental Information, 2015). The dataset contains geographic coordinates of educational institutions, enrollments, grade levels, and start age and end age of each institution. We assigned school-age agents to the nearest available school location within a tract. School-age agents were sorted into schools based

on grade and enrollment levels. Figure 3(a) and (b) displays representative suburban and high density urban area results of this step, an example of household, workplace and education locations for one census tract within our study area.

3.2.3 Step 3: create social networks

Finally, as Step 3, we created three types of networks, specifically: household, work, and educational (which are subdivided into school and daycare networks). In the work of (Barthelemy & Toint, 2013), social networks were not generated, while (Wise, 2014) only generated social networks based on a ego network from the social media data. Our approach is different: individuals received a link to each agent based on living in the same household and either working in the same workplace or attending the same daycare/education institute. Small-world networks are one of the most implemented networks in social simulations (Amblard et al., 2015). Hence, if the group size of a household, work or education site was greater than 5, a (Newman & Watts, 1999) small-world network was generated where the number 5 was chosen based on the work of (Dunbar, 1998). While for people who live in the same household, if the group size was less than 5, everyone was

Table 2 Summary of the generated synthetic datasets (n differs in the size depending on the network, $n \in [0, 14]$)

Dataset	Size	Format	Dimensions (rows $\times$ columns)	Description
Synthetic Population	2.53 GB	.csv	22,921,302 $\times$ 9	Synthetic individuals and their demographic, location, and work or educational information
Household Network	1.4 GB	.csv	22,921,302 $\times$ n	Individuals and their members live in the same household
Work Network	990 MB	.csv	9,730,906 $\times$ n	Individuals and their members in synthetic work networks
School Network	435 MB	.csv	4,123,393 $\times$ n	Individuals and their members in synthetic school networks
Daycare Network	94 MB	.csv	882,603 $\times$ n	Individuals and their members in synthetic daycare networks

connected. Figure 4 shows the process to create the social networks. Figure 4(a) shows selected population from our dataset, as discussed above people under same household are fully connected (4(b)). Work and educational network were created based on the worksite location and educational location (4(c)). Figure 4 (d) is an example of the multi-layer network within a home including an individual's family ties and proximity ties to people at work and school.

4 Results from the synthetic population generation

After providing the details about our method in Section 3, this section illustrates the resulting datasets. Five different datasets were generated, including one synthetic

population dataset and four social network datasets (links between the network datasets and the population dataset was via the unique ID of the synthetic individual). Table 2 shows the basic information of each dataset. To display the structure of the resulting datasets, Fig. 5(a) shows a sample record extracted from the synthetic population dataset and Fig. 5(b) shows the sample of the synthetic social networks extracted from household network. Furthermore, it is also important to verify and validate the results, which is what we turn to in section 4.1 and 4.2. To verify our method, basic statistical analyses were carried out which are presented below. For validation, our results are compared to a benchmark dataset (i.e., RTI data) to test the robustness of our method.

4.1 Verification

Verification of our synthetic population was done by comparing selected measures from the official US Census data to our results as shown in Table 3. During this process, we found the number of individuals in our synthetic population was not identical to the 2010 US Census data that showed 23,004,272 individuals lived in the study area. Looking into this issue, we identified 116 problematic census tracts out of a total 5,500 tracts. These problematic tracts had inconsistent counts with respect to total numbers of males and females or no data was given for the number of individuals under 18 years old from the official US Census data. Over the whole dataset, there were 82,970 individuals living in those problematic tracts. Our method was not able to generate the individuals in these problematic tracts because of the inconsistencies discussed above. However, the percentage of the individuals living in problematic tracts is only 0.36% of the total population of the study area. We decided to leave these tracts out of our synthetic population. In the end, our synthetic population method generates 22,921,302

	id	age	sex	hhold	htype	wp	long	lat	RID
0	36005000200i1	76.0	f	36005000200h0	10	36005000200h0	-73.855405	40.807544	66517
1	36005000200i9	89.0	f	36005000200h1	10	36005000200h1	-73.857143	40.812771	657987
2	36005000200i207	72.0	f	36005000200h10	10	36005000200h10	-73.854977	40.806293	411083
3	36005000200i108	28.0	f	36005000200h100	9	36061008800w1199	-73.861698	40.811852	764495
4	36005000200i2996	4.0	m	36005000200h1000	1	36005000400s0	-73.855311	40.807093	655841

(a)

	FromNode	A	B	C	D
0	09001030300i1	09001030300i56	09001030400i2431	09001030300i70	09001030400i2427
1	09001030300i3	09001030300i23	09001030200i2428	09001030300i67	09001030200i2427
2	09001030300i4	09001030300i13	09001044500i2292	09001030300i28	09001044500i2283
3	09001030300i8	09001030300i14	09001030400i2443	09001030300i147	09001030400i2442
4	09001030300i13	09001030300i4	09001030300i28	09001030300i36	09001044500i2292

(b)

Fig. 5 Sample Records: (a) Synthetic Population; (b) Synthetic Social Network

Table 3 Population counts based on multiple sources

	Original Census Tract (5506 Tracts)	Census without Problematic Tracts (5390 Tracts)	Synthetic Population
Individuals	23,004,272	22,921,302	22,921,302
Households	8,468,450	8,453,097	8,457,710
Male	11,108,508	11,053,104	11,053,104
Female	11,895,764	11,868,198	11,868,198

individuals, including 17,697,433 adults (age ≥ 18) and 5,223,869 children (Age < 18), which were grouped into 8,457,710 households. However, the total number of households was slightly greater than the number from US Census records, because we generated households for people living in group quarters (as discussed in Section 3). If we excluded this population we would have the identical household amount to the official US Census data (i.e., 8,453,097).

Other than the overall population number comparison with US Census, we also verified the male and female population under each age group by comparing it with US Census data. The rationale for this is that we used the 18 age groups for male and female directly from US Census data to create population in step 2 of our method. In this verification part, we excluded the problematic census tracts identified above. Figure 6 displays the male and female population under each age group from US Census and our population dataset. No differences were identified under all age groups, thus our method was able to generate a identical number of male and female population under each age group.

In addition to general population statistics, as noted in Section 3.2, our synthetic population method also generated geographically-explicit social networks. Social networks play an important role with respect to the spread of information and social capital (in times of disasters), and diseases (e.g., Eubank et al., 2004; Ersing & Kost, 2012;

Burger, 2020; Pires & Crooks, 2017). With respect to the household networks, our networks ranged from individuals in cliques with 0 to 10 ties (connections) with the majority of the population in small groups of 2 to 4 individuals as shown in Fig. 7. While the work networks were more skewed which reflects many small places of work, school and daycare networks have a higher degree of connectivity as shown in Table 4 due to the greater number of people in each location.

4.2 Validation

Turning to validation, we chose two reference datasets to compare our synthetic population against which are shown in Fig. 8. First, we compared our synthetic population to the US Census (US Census Bureau, 2010c) dataset based on the average household size attribute. Our rationale for choosing this attribute was that in our synthetic population generation method (Section 3.2), we did not use this attribute when creating the households. On average, the household size difference between the two datasets ranged from -0.01 to 0, indicating the synthetic population varied only slightly from the US Census. Turning to the second dataset, the RTI International synthetic population (RTI, 2019), our rationale for using this dataset is that it was generated using synthetic reconstruction, which is the same method as ours but does not include social networks (as discussed in Section 1) (Wheaton et al., 2009). In addition, our method only used census tract data to generate the synthetic population, while the generation of RTI used both aggregate and disaggregate data (e.g., the Public Use Microdata Sample (PUMS) data from the US Census). Hence, other than proving our population synthesis method was robust, we would also like to argue that our method was able to generate a robust synthetic population by using only census tract data, which is an advantage when disaggregate data is not available. To demonstrate the robustness of our population, similar to the US Census comparison

Fig. 6 Verification Results: Comparison between Our Population and Census on Each Age Group's Population

above, we compared average household size with RTI's average household size from their synthetic population. Figure 8 shows the difference of household size between our synthetic population and that of the RTI population. Overall the difference is less than 1 (i.e., ranging from -0.22 to 0.78). As this figure shows, our synthetic population compares favorably with that of the RTI Synthetic population.

5 Use cases

As discussed in Section 1, synthetic populations have been utilized by various modeling techniques such as microsimulation and agent-based modeling to explore a wide range of phenomena such as traffic dynamics, disaster response, and disease outbreaks (e.g., Wise et al., 2017;

Eubank et al., 2004; Jiang et al., 2020; Xu et al., 2017). Hence, to demonstrate the utility generalizability of our synthetic population, the generated dataset was applied to three different agent-based modeling use cases exploring a variety of urban phenomena. The first use case simulates traffic dynamics in a high-density urban setting (i.e., Manhattan island and surrounding areas, Section 5.1). The second use case (Section 5.2) builds upon the first use case in the sense that it shares the same study area, but it simulates how people might react in the immediate aftermath of a disaster. The last use case (Section 5.3) simulates a hypothetical disease spread through our synthetic population. Our rationale for choosing these applications, other than to demonstrate how the geographically-explicit synthetic population can be used, also shows how our extensions to previous population synthesis work (i.e., social networks and explicit home and work locations) can be used to gain insight into such phenomena. However, it needs to be noted, that these are simplistic models and are given as pedagogic examples. Our aim is to demonstrate the re-usability of our synthetic population dataset and, as discussed in Section 1, all three models have been made publicly available for readers to explore and extend (See: <https://bit.ly/SynPopABM>).

Table 4 Social networks statistics from our synthetic population method

Social Network Types	Nodes	Edges	Average Degree
Household	22,921,302	29,382,456	1.28
Work	9,730,906	24,417,323	2.50
School	4,123,393	10,800,293	2.62
Daycare	882,603	2,318,523	2.63

5.1 Use case 1: traffic dynamics

Agent-based modeling has been used extensively in urban transportation (Wise et al., 2017) which therefore motivated this application and to demonstrate how our synthetic population could be used in this domain. We used JAVA in MASON platform to build a simple agent-based model to capture traffic congestion in a high-density urban area (Luke et al., 2005). Figure 9 shows

the structure of this traffic model, which comprises three main components: 1) agent; 2) environment; and 3) model visualization. The synthetic population was used to initialize the agent component, which contained the agents' basic attributes (e.g., age, sex, home location, work location, etc.) and behaviors (e.g., path finding and commuting). The Environment component not only represented the model's physical environment (e.g., road

networks, buildings), but also comprised of the model scheduler that runs all agents' behaviors (i.e., find path and commute). Lastly, the visualization component updated the model user interface at each time step, for example displaying the physical environment and updating agent's location. Within this model, we used one time step to represent one minute of "real" time. The rationale to use one minute is that it enables us to capture small-scale movement and traffic congestion such as the morning rush hour by simulating agents commuting from home to work.

Before simulating large scale traffic dynamics, verification experiments were undertaken for one road segment to ensure the model had the ability to capture congestion. Figure 10 shows the commuting speed simulated by the model with various sizes of vehicles (e.g., 1, 10, 100, 1000) for a 2.1 km road segment in Manhattan. As to be expected, with more vehicles, the commute along the road segment took longer due to traffic delays. Building upon this verification example and in order to overcome the computational constraints for running the traffic model at full scale, we used 1 agent to represent 100 or 1000 agents. Figure 10 shows the model can generate the same commute speed when 1 agent represents either 100 and 1000 agents. With these verification steps, we feel confident that our model captured basic traffic dynamics. Therefore, we loaded a 0.01% sample of our synthetic population into the model and scale other parameters accordingly. During the simulation, we capture each individual's geo-location at every time step. Figure 11 shows a heat map of traffic density during the morning commute focusing on Manhattan island NYC and surrounding areas. In the figure, the darkest red color indicates the area suffering from the worst traffic delays. Furthermore, the model demonstrates how home and work locations from the synthetic population can be utilized for modeling traffic dynamics.

5.2 Use case 2: groups formation after a disaster event

Building upon the traffic model discussed in Section 5.1, where we demonstrated how normal patterns of life can

be generated (i.e., traffic congestion) using the synthetic population. In this use case, we extend the traffic model to explore how people react in the immediate aftermath of a disaster. Especially, how an individual's social networks impact their decision making. Our rationale for this is that social networks impact people's decision making during disaster events and new social networks emerge as people react and respond to a disaster (e.g., Burger, 2020; Ersing & Kost, 2012; Jones & Faas, 2016; Dynes, 2006).

To capture this, another agent-based model was built by extending both agent and environment components of the traffic model as shown in Fig. 12. Within the Agent component, attributes related to health status were added to represent the effects on individual's health brought by the disaster event. Also, the main difference between this model and the one presented in Section 5.1 is the addition of the three types of social networks (Section 3.2.3) to the agent component, which were initialized from the social networks of our synthetic population (Section 3.2.2). The behaviors of the agents were also altered from the previous model, whereby agents can now respond to a disaster event. The response behaviors were based on observations from previous studies (e.g., Mawson, 2007) and implemented using a fast and frugal decision tree (Kennedy & Bassett, 2011). Each step of the model represents one minute of "real" time, during which each agent decides and acts (i.e., moves towards a goal or stays in place). This time step was chosen because we wanted to represent the fine scale decision making and movements of the population. For simplification, agents only commuted by vehicle; however, immediately after a disaster, some flee on foot. Agents also had different work-day schedules and workplaces, resulting in individualized commuter behavior. After the disaster, agents altered their normal behavior from daily commuting to finding safety for themselves (Mawson, 2007; Maslow, 1943). Depending on the availability of shelter and agent knowledge after the disaster, they either choose to form ad hoc groups (i.e., emergent networks), shelter in place, or flee from the affected area.

Fig. 10 Model Verification Using Selected Road Segment

Fig. 11 Traffic Density During Morning Commute

Once they have achieved relative safety, they search for, locate, and join members of their household (based on their social networks instantiated by the synthetic population dataset). Figure 13 shows the impact of a large explosion in Manhattan and the health status of the agents

in and around the immediate blast area as well as the commuting region 1 minute after the explosion. After the explosion, surviving agents flee damaged buildings and create new social groups in the process of attempting to go home. Figure 14 shows an example of how an agent's social

Fig. 12 Model Component Structure of Population Respond to Disaster event

Fig. 13 Agents' Health Status After 1 Minute of the Disaster Event

network from the synthetic population changed (i.e., new ad hoc group formation) as agents interact with others within the disaster area and flee. Through such experiments, we can explore and characterize the reaction of the population of NYC to a disaster.

5.3 Use case 3: disease dynamics

As discussed in Section 2, one of the fields utilizing synthetic populations is that of epidemic research. To

demonstrate how our synthetic population dataset can be utilized in such a domain, we utilized an agent-based model as the tool to develop a simple susceptible, exposed, infected and recovered (SEIR) compartmentalized disease model in Python. Figure 15 shows the model components. Within agent component, we integrated SEIR into this component to represent agents' health status. Also, the model took our synthetic population dataset including its social network as input parameters for the initialization

Fig. 14 Simulation of Social Networks Changes for a Sample of the Population: (a) Before the Disaster Event; (b) Emergent Social Networks After the Disaster Event

of the Agent component and for guiding agent-to-agent interactions (e.g., members of the same household, or children at the same school or colleagues at the same work site interact with each other). The model simulates how a hypothetical disease could spread through the system and captures the health status at each time step of the model. In addition, the model also captured where this change in health status occurs (e.g., at home or work). While the traffic model demonstrated how researchers could use home and work locations to generate traffic congestion, this example shows the utility of why we included social networks in our synthetic population method. Also, by implementing the model in Python, it shows how our synthetic population dataset can be used in other programming languages.

As the synthetic population is geographically-explicit, it is straightforward to select agents for just one location (e.g., home or work). In this example, we extracted only agents who had home locations in Ulster county, NY. Within this county, there are total of 153,253 people living in 58,094 different households, shown in Fig. 16. In this model, one day is divided into 3 time periods (each being 8 hours) which are analogous with simple daily rhythm of home (either sleeping or getting up), being at work (or going to school), and being back at home. In any of the time periods, the agents interact (i.e., spreads or is infected by the disease) through their networks at that specific time. For example, while at work, agents will only interact with others in their work network. While agents are at home, they interact solely with their household networks. It is through these interactions that a disease can spread. In this example, we start off with only one agent being infected with a hypothetical disease. Figure 17(a) shows the SEIR curve, which indicates that the model utilizing our synthetic population can capture standard epidemiological disease dynamics. In addition, the relative locations of where agents get exposed to the disease can be counted. Figure 17(b) show the cumulative number of agents who get exposed to the disease at specific locations (i.e., home, work, daycare, and schools).

6 Conclusion and discussion

While synthetic population research has been growing and is becoming more and more utilized within the geographically-explicit simulations, little attention has

Fig. 17 Disease Model Results: (a) SEIR Results; (b) Exposed Locations

been given to creating geographically-explicit synthetic populations with realistic social networks. To address this shortcoming, in this paper, we have presented a novel method for generating a synthetic population dataset by utilizing aggregate census data which incorporates not only home and work locations but also social networks. Our verification and validation experiments demonstrated that our synthetic population results were comparable to aggregate census tract level (i.e., Census Data) and individual level (i.e., RTI Data) data showing that our method is robust.

Our method also addresses one of the challenges in population synthesis especially when it comes to agent-based modeling. Specifically, the ability to instantiate social networks at the model initialization rather than during the running of the simulation. We consider this important because social networks are increasingly being used in the field of agent-based modeling but so far such networks have not been incorporated in synthetic population as we noted in Section 2.

To show the re-usability and generalizability of our publicly available dataset, we presented three use cases. The first showed how the synthetic population can generate possible traffic dynamics (Section 5.1 while the other two (i.e., disaster response and a disease outbreak Sections 5.2

and 5.3) explicitly showed how social networks can be used in geographically-explicit settings. By implementing these use cases, we have shown the synthetic population can be applied and reused in different modeling environments and programming languages to meet various research purposes. Our intention in sharing our synthetic population method and data is to demonstrate how our synthetic population can be utilized in situations where fine-scale individual-level data might not be available but could be constructed based on aggregate data. As noted in Section 1, all models built for the use cases are shared online. Our rationale for this is to allow reproduction of anything presented in this paper and to allow researchers to extend and adapt such methods and models in this nascent field of integrating spatial data and social networks into large-scale agent-based modeling simulations.

While our method can create synthetic population datasets along with its social networks, similar to all synthetic population methods, there is always room for improvement. Especially, given the latest US Census data that was available to us was from the 2010. Thus, the synthetic population result will not align with the latest demographics of the area. However, our method was created based on the standard census data, which makes it

possible to modify and replicate the method and generate the population when the 2020 US Census becomes available. This is one reason why we share the code of our method in Section 3. Furthermore, when the 2020 US Census becomes available, we would like to extend our synthetic population dataset to comprise all the US population. Currently, the synthetic population did not comprise financial, employment or educational backgrounds, with such data, the synthetic population could be potentially used by an agent-based modeler wanting to explore issues such as residential location or public health issues (e.g., Jiang et al., 2021; Li et al., 2020). This could be remedied by adding additional variables to the agents from the census dataset.

As discussed in Section 1, the reuse of existing synthetic population datasets has been a challenge within the field of agent-based modeling. One reason behind this is that there are no standard method or agreement to some extent on what attributes should be added to a synthetic population dataset. This is similar to agent-based modeling, more generally, where there are many definitions of such models are in the sense of what is an agent, and also there are many possible attributes that can be assigned to an agent depending on the purpose of the model. Due to these issues, a well built synthetic population should always aim to solve a specific research question and can not be reused to explore other questions. We consider this may undermine the previous efforts made by researchers to the field of utilizing synthetic population datasets in agent-based modeling. To address this challenge caused by these issues, another potential aspect for the future work would be proposing a set of standards that can be referenced by researchers to create, store and describe their own methods and datasets potentially being inspired by initiatives such as the ODD protocol or CityGML and ontological principles which formalize data exchanges (Grimm et al., 2006; Kutzner et al., 2020; Lee et al., 2016). Furthermore, such standards could be beneficial to the field of information management by allowing researchers either to replicate the method or to reuse the resulting datasets.

With respect to the models, we noted in Section 5 that these were only pedagogic in nature and therefore like all models there are limitations and there is always space to improve them. As for the traffic dynamic model (Section 5.1), we only considered people commute from home to work because imitated human behaviors information can be extracted from census data for our synthetic population. Thus, other human activities besides commuting (e.g., shopping, social activities, sporting events) could be added to the model if data was available. Potentially such data could come from multiple data sources, such social media (Padilla et al. 2014; Kavak et al. 2018), urban infrastructure data (e.g., building

types) (Zhang et al., 2021) and human mobility data (e.g., aggregates anonymized location data) (Elarde et al., 2021). By doing this, various types of traffic dynamics could be captured. Similarly, when we simulated the hypothetical disease spread (Section 5.3), we only considered home, work, and educational locations. Such data discussed above could also be added to the disease model to add more types of interaction spaces and locations. In addition, these interactions would generate more types of human activities and could shed light on disease outbreaks from the individual level. For example, infected individuals' trajectory data could be utilized to understand the human mobility during a disease outbreak (e.g., Cheng et al., 2021; Hu et al., 2021) and assist in contact tracing. To expand the disaster model (Section 5.2), more research is needed to form new emergent groups. However, this would require a greater study of post disaster events or other types of emergency field trails for observations, something which is currently not available (Burger, 2020). Even with such limitations and areas of further work, the three models presented in this paper provide a basic experimental setting for agent-based modelers who have their own preferences for programming languages and research areas but want to utilize geographically-explicit synthetic populations at model initialization. This was not the main purpose of the paper, but it illustrates how one could use our novel synthetic population method in a variety of use cases (with or without geographically-explicit social networks) to explore complex geographical systems where agents interact with each other and their environment.

Acknowledgments

This work is made possible by the support of the Center for Social Complexity at George Mason University and the RENEW Institute at the University of Buffalo. The authors would also like to thank two former members of our research team Talha Oz and Xiaoyi Yuan.

Authors' contributions

N.J., A.T.C., A.B., H.K. and W.K. designed the synthetic population method and conceptualised the use case models along with the experiments. N.J. coded the synthetic population method and two use case models (i.e., traffic dynamic and disease outbreak), who also ran the experiment, carried out the analysis and wrote up the results. A.B. coded and ran the experiments of the disaster response use case model, who also carried out the analysis and delivered results. Both A.T.C. and N.J. prepared the initial draft while H.K., A.B. and W.K. provided substantial edits to the paper and all authors approved its content.

Funding

This work was supported in part by the Defense Technology Research Agency (DTRA) under Grant number HDTRA1-16-0043. The opinions, findings, and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the sponsors.

Availability of data and materials

All source and resulting synthetic population data: <https://osf.io/3vsaj/>.

Code availability

Code of synthetic population method and use case models: <https://bit.ly/SynPopABM>.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Author details

¹Department of Computational and Data Sciences George Mason University, Fairfax, VA, United States. ²Department of Geography, University at Buffalo, Buffalo, NY, United States. ³School of Global Public Health, New York University, New York, NY, USA.

Received: 3 August 2021 Accepted: 18 January 2022

Published online: 14 February 2022

References

- Agrawal, A., Brown, D.G., Rao, G., Riolo, R., Robinson, D.T., & Bommarito, M. (2013). Interactions between organizations and networks in common-pool resource governance. *Environmental Science & Policy*, 25, 138–146. <https://doi.org/10.1016/j.envsci.2012.08.004>.
- Alizadeh, M., Cioffi-Revilla, C., & Crooks, A. (2017). Generating and analyzing spatial social networks. *Computational and Mathematical Organization Theory*, 23(3), 362–390. <https://doi.org/10.1007/s10588-016-9232-2>.
- Amblard, F., Bouadjio Boulic, A., Sureda Gutierrez, C., & Gaudou, B. (2015). Which models are used in social simulation to generate social networks? A review of 17 years of publications in JASSS. In *Proceedings of the 2015 Winter Simulation Conference* (pp. 4021–4032): Institute of Electrical and Electronics Engineers, Inc. <https://doi.org/10.1109/WSC.2015.7408556>.
- Barrett, C.L., Beckman, R.J., Khan, M., Kumar, V.S.A., Marathe, M.V., Stretz, P.E., ... Lewis, B. (2009). Generation and analysis of large synthetic social contact networks. In *Proceedings of the 2009 Winter Simulation Conference* (pp. 1003–1014): Institute of Electrical and Electronics Engineers, Inc. <https://doi.org/10.1109/WSC.2009.5429425>.
- Barthelemy, J., & Toint, P.L. (2013). Synthetic population generation without a sample. *Transportation Science*, 47(2), 266–279. <https://doi.org/10.1287/trsc.1120.0408>.
- Batty, M. (2007). *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-based Models, and Fractals*. Cambridge: MIT press.
- Batty, M. (2013). *The New Science of Cities*. Cambridge: MIT press.
- Birkin, M., & Wu, B. (2012). A review of microsimulation and hybrid agent-based approaches. In *Agent-Based Models of Geographical Systems* (pp. 51–68): Springer. https://doi.org/10.1007/978-90-481-8927-4_3.
- Borysov, S.S., Rich, J., & Pereira, F.C. (2019). How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies*, 106, 73–97. <https://doi.org/10.1016/j.trc.2019.07.006>.
- Burger, A. (2020). *Disaster through the lens of complex adaptive systems: Exploring emergent groups utilizing agent based modeling and social networks*. PhD Dissertation. Fairfax: George Mason University.
- Cheng, T., Lu, T., Liu, Y., Gao, X., & Zhang, X. (2021). Revealing spatiotemporal transmission patterns and stages of COVID-19 in china using individual patients' trajectory data. *Computational Urban Science*, 1(1), 9. <https://doi.org/10.1007/s43762-021-00009-8>.
- Crooks, A., Croitoru, A., Lu, X., Wise, S., Irvine, J., & Stefanidis, A. (2015). Walk this way: Improving pedestrian agent-based models through scene activity analysis. *ISPRS International Journal of Geo-Information*, 4, 1627–1656. <https://doi.org/10.3390/ijgi4031627>.
- Crooks, A., Malleon, N., Manley, E., & Heppenstall, A. (2019). *Agent-Based Modeling and Geographical Information Systems: A Practical Primer*. London: Sage.
- Deming, W.E., & Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4), 427–444. <https://doi.org/10.1214/aoms/117731829>.
- Dunbar, R.I.M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5), 178–190. [https://doi.org/10.1002/\(SICI\)1520-6505\(1998\)6:5<178::AID-EVAN5>3.0.CO;2-8](https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8).
- Dynes, R. (2006). Social capital: Dealing with community emergencies. *Homeland Security Affairs*, 2(2), 1–26.
- Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, 94(5), 1429–1451. <https://doi.org/10.1257/0002828043052303>.
- Elarde, J., Kim, J.-S., Kavak, H., Züfle, A., & Anderson, T. (2021). Change of human mobility during COVID-19: A United States case study. *PLoS ONE*, 16(11), e0259031.
- Ersing, R.L., & Kost, K.A. (2012). *Surviving Disaster: The Role of Social Networks*. Chicago: Lyceum Books.
- Eubank, S., Guclu, H., Kumar, V.S.A., Marathe, M.V., Srinivasan, A., Toroczkai, Z., ... Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988), 180–184. <https://doi.org/10.1038/nature02541>.
- Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58(C), 243–263. <https://doi.org/10.1016/j.trb.2013.09.012>.
- Gallagher, S., Richardson, L.F., Ventura, S.L., & Eddy, W.F. (2018). SPEW: Synthetic Populations and Ecosystems of the World. *Journal of Computational and Graphical Statistics*, 27(4), 773–784. <https://doi.org/10.1080/10618600.2018.1442342>.
- Geard, N., McCaw, J.M., Dorin, A., Korb, K.B., & McVernon, J. (2013). Synthetic population dynamics: A model of household demography. *Journal of Artificial Societies and Social Simulation*, 16(1), 8. <https://doi.org/10.18564/jasss.2098>.
- Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., ... DeAngelis, D.L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198(1–2), 115–126. <https://doi.org/10.1016/j.ecolmodel.2006.04.023>.
- Heppenstall, A.J., Crooks, A.T., See, L.M., & Batty, M. (2011). *Agent-based Models of Geographical Systems*. Dordrecht: Springer.
- Heppenstall, A., Crooks, A., Malleon, N., Manley, E., Ge, J., & Batty, M. (2021). Future developments in geographical agent-based models: Challenges and opportunities. *Geographical Analysis*, 53(1), 76–91. <https://doi.org/10.1111/gean.12267>.
- Hu, T., Wang, S., She, B., Zhang, M., Huang, X., Cui, Y., ... Li, Z. (2021). Human mobility data in the COVID-19 pandemic: Characteristics, applications, and challenges. *International Journal of Digital Earth*, 14(9), 1126–1147. <https://doi.org/10.1080/17538947.2021.1952324>.
- Huang, X., Li, Z., Jiang, Y., Li, X., & Porter, D. (2020). Twitter reveals human mobility dynamics during the COVID-19 pandemic. *PLOS ONE*, 15(11), 0241957. <https://doi.org/10.1371/journal.pone.0241957>.
- Huang, Z., & Williamson, P. (2002). A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata. Retrieved September 9, 2019, from <https://www.semanticscholar.org/paper/A-COMPARISON-OF-SYNTHETIC-RECONSTRUCTION-AND-TO-THE-Huang/fab88a8ffaa97b81bf36ea53bcfae5f9c719da4>.
- Jiang, N., Burger, A., Crooks, A.T., & Kennedy, W.G. (2020). Integrating social networks into large-scale urban simulations for disaster responses. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoSpatial Simulation* (pp. 52–55): Association for Computing Machinery. <https://doi.org/10.1145/3423335.3428168>.
- Jiang, N., Crooks, A., Wang, W., & Xie, Y. (2021). Simulating urban shrinkage in detroit via agent-based modeling. *Sustainability*, 13(4). <https://doi.org/10.3390/su13042283>.
- Jones, E.C., & Faas, A. (2016). *Social Network Analysis of Disaster Response, Recovery, and Adaptation*. Oxford: Elsevier.
- Kavak, H., Padilla, J.J., Lynch, C.J., & Diallo, S.Y. (2018). Big data, agents, and machinelearning: towards a data-driven agent-based modeling approach. In *Proceedings of the Annual Simulation Symposium, Baltimore, MD* (p. 1.12).
- Kennedy, W.G., & Bassett, J.K. (2011). Implementing a “Fast and Frugal” Cognitive Model within a Computational Social Simulation. In *The Computational Social Science Society of Americas Conference* (2011). Santa Fe, NM (pp. 9–12).
- Kim, J.S., Kavak, H., & Crooks, A. (2018). Procedural city generation beyond game development. *SIGSPATIAL Special*, 10(2), 34.41. <https://doi.org/10.1145/3292390.3292397>.
- Kim, J.-S., Jin, H., Kavak, H., Rouly, O.C., Crooks, A., Pfoser, D., ... Züfle, A. (2020). Location-based social network data generation based on patterns of life. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)* (pp. 158–167): IEEE. <https://doi.org/10.1109/MDM48529.2020.00038>.
- Kutzner, T., Chaturvedi, K., & Kolbe, T.H. (2020). CityGML 3.0: New Functions Open Up New Applications. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88(1), 43–61. <https://doi.org/10.1007/s41064-020-00095-z>.

- Lee, Y.-C., Eastman, C.M., & Solihin, W. (2016). An ontology-based approach for developing data exchange requirements and model views of building information modeling. *Advanced Engineering Informatics*, 30(3), 354–367. <https://doi.org/10.1016/j.aei.2016.04.008>.
- Li, Y., Hyder, A., Southerland, L.T., Hammond, G., Porr, A., & Miller, H.J. (2020). 311 service requests as indicators of neighborhood distress and opioid use disorder. *Scientific Reports*, 10(1), 19579. <https://doi.org/10.1038/s41598-020-76685-z>.
- Lim, P.P. (2020). *Population synthesis for travel demand modelling in Australian capital cities*. PhD thesis. The University of Queensland. <https://doi.org/10.14264/uql.2020.822>.
- Luke, S., Cioffi-Revilla, C., Panait, L., Sullivan, K., & Balan, G. (2005). MASON: A multiagent simulation environment. *Simulation*, 81(7), 517–527. <https://doi.org/10.1177/0037549705058073>.
- Manley, E., Orr, S.W., & Cheng, T. (2015). A heuristic model of bounded route choice in urban areas. *Transportation Research Part C: Emerging Technologies*, 56, 195–209. <https://doi.org/10.1016/j.trc.2015.03.020>.
- Maslow, A. (1943). A Theory of Human Motivation. *Psychological Review*, 50(4), 370–396. <https://doi.org/10.1037/h0054346>.
- Mawson, A.R. (2007). *Mass Panic and Social Attachment: The Dynamics of Human Behavior*. Aldershot: Ashgate.
- Müller, K., & Axhausen, K.W. (2011). Population synthesis for microsimulation: State of the art. *Arbeitsberichte Verkehrs- und Raumplanung*, 638, 16. <https://doi.org/10.3929/ethz-a-006127782>.
- Newman, M.E., & Watts, D.J. (1999). Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4–6), 341–346. [https://doi.org/10.1016/S0375-9601\(99\)00757-4](https://doi.org/10.1016/S0375-9601(99)00757-4).
- Orcutt, G.H. (1957). A new type of socio-economic system. *Review of Economics and Statistics*, 39(2), 116–123. <https://doi.org/10.2307/1928528>.
- Padilla, J.J., Diallo, S.Y., Kavak, H., Sahin, O., & Nicholson, B. (2014). Leveraging social media data in agent-based simulations. In *SpringSim (ANSS)*, Tampa, FL (p. 17).
- Pescosolido, B.A., & Levy, J.A. (2002). The role of social networks in health, illness, disease and healing: the accepting present, the forgotten past, and the dangerous potential for a complacent future. In *Social Networks and Health*, vol 8 (pp. 3–25): Emerald Group Publishing Limited. [https://doi.org/10.1016/S1057-6290\(02\)80019-5](https://doi.org/10.1016/S1057-6290(02)80019-5).
- Pires, B., & Crooks, A.T. (2017). Modeling the emergence of riots: A geosimulation approach. *Computers, Environment and Urban Systems*, 61, 66–80. <https://doi.org/10.1016/j.compenvurbysys.2016.09.003>.
- Pritchard, D.R., & Miller, E.J. (2012). Advances in population synthesis: Fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, 39, 685–704. <https://doi.org/10.1007/s11116-011-9367-4>.
- Renardy, M., Eisenberg, M., & Kirschner, D. (2020). Predicting the second wave of COVID-19 in washtenaw county, MI. *Journal of Theoretical Biology*, 507, 110461. <https://doi.org/10.1016/j.jtbi.2020.110461>.
- RTI (2019). RTI U.S. Synthetic Household Population. Retrieved September 9, 2019, from <https://www.rti.org/impact/rti-us-synthetic-household-population>.
- Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016). Hidden markov model-based population synthesis. *Transportation Research Part B: Methodological*, 90, 1–21. <https://doi.org/10.1016/j.trb.2016.04.007>.
- Sun, L., & Erath, A. (2015). A bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61, 49–62. <https://doi.org/10.1016/j.trc.2015.10.010>.
- US Census Bureau (2010a). 2010 TIGER/Line Shapefiles: Roads. Retrieved September 9, 2019, from <https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Roads>.
- US Census Bureau (2010b). County Business Patterns: 2010. Retrieved September 9, 2019, from <https://www.census.gov/data/datasets/2010/econ/cbp/2010-cbp.html>.
- US Census Bureau (2010c). Decennial Census (2010, 2000). Retrieved September 9, 2019, from <https://www.census.gov/data/developers/datasets/decennial-census.html>.
- US Census Bureau (2019). US Census Bureau Center for Economic Studies Publications And Reports Page. Retrieved September 9, 2019, from <https://lehd.ces.census.gov/announcements.html>.
- USEPA Office of Environmental Information (2015). Educational Institutions, US, 2015, ORNL, SEGS. Retrieved September 9, 2019, from <https://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid=%7B9C49AE4B-F175-43D0-BCC6-A928FF54C329%7D>.
- Wheaton, W.D., Cajka, J.C., Chasteen, B.M., Wagener, D.K., Cooley, P.C., Ganapathi, L., ... Allpress, J.L. (2009). Synthesized population databases: A US geospatial database for agent-based models. *Methods report (RTI Press)*, 2009(10), 905. <https://doi.org/10.3768/rtipress.2009.mr.0010.0905>.
- Wise, S. (2014). *Using social media content to inform agent-based models for humanitarian crisis response*. PhD Dissertation. Fairfax: George Mason University.
- Wise, S., Crooks, A., & Batty, M. (2017). Transportation in agent-based urban modelling. In *Agent Based Modelling of Urban Systems, ABMUS 2016* (pp. 129–148): Springer. https://doi.org/10.1007/978-3-319-51957-9_8, 10051.
- Xie, Y., Batty, M., & Zhao, K. (2007). Simulating emergent urban form using agent-based modeling: Desakota in the suzhou-wuxian region in china. *Annals of the Association of American Geographers*, 97(3), 477–495. <https://doi.org/10.1111/j.1467-8306.2007.00559.x>.
- Xu, Z., Glass, K., Lau, C.L., Geard, N., Graves, P., & Clements, A. (2017). A synthetic population for modelling the dynamics of infectious disease transmission in American Samoa. *Scientific Reports*, 7(1), 16725. <https://doi.org/10.1038/s41598-017-17093-8>.
- Ye, X., Li, S., & Peng, Q. (2021). Measuring interaction among cities in china: A geographical awareness approach with social media data. *Cities*, 109, 103041. <https://doi.org/10.1016/j.cities.2020.103041>.
- Zhang, Z., Yin, D., Virrantaus, K., Ye, X., & Wang, S. (2021). Modeling human activity dynamics: an object-class oriented space–time composite model based on social media and urban infrastructure data. *Computational Urban Science*, 1(1), 7. <https://doi.org/10.1007/s43762-021-00006-x>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.