Generating a synthetic population of the United States*

Abhijin Adiga[a], Aditya Agashe[a], Shaikh Arifuzzaman[a], Christopher L. Barrett[a,c], Richard Beckman[a], Keith Bisset[a], Jiangzhuo Chen[a], Youngyun Chungbaek[a], Stephen Eubank[a,e], Sandeep Gupta[a], Maleq Khan[a], Chris J. Kuhlman[a], Eric Lofgren[a], Bryan Lewis,[a] Achla Marathe[a,b], Madhav V. Marathe[a,c], Henning S. Mortveit[a,d], Eric Nordberg[a], Caitlin Rivers[a], Paula Stretz[a], Samarth Swarup[a], Amanda Wilson[a], Dawen Xie[a]

[a]Network Dynamics and Simulation Science Laboratory;
[b]Department of Agricultural and Applied Economics;
[c]Department of Computer Science;
[d]Department of Mathematics;
[e]Department of Population Health Sciences;
Virginia Tech

January 29, 2015

Abstract

We describe the methodology for generating a synthetic population of the United States.

A synthetic population integrates a variety of databases from commercial and public sources into a common architecture for data exchange. The process preserves the confidentiality of the individuals in the original data sets, yet produces realistic attributes and demographics for the synthetic individuals. The synthetic population is a set of synthetic people and households, located geographically, each associated with demographic variables recorded in the census. Joint demographic distributions are reconstructed from the marginal distributions available in typical census data using an iterative proportional fitting (IPF) technique [5]. Each synthetic individual is placed in a household with other synthetic individuals. Each household is located geographically using land-use data and data pertaining to transportation networks. The process guarantees that a census of our synthetic population is statistically indistinguishable from the original census. The basic process can be extended to assign other personal and behavioral attributes using additional data sources. Table 1 shows some of the data we have used in constructing previous synthetic populations.

The methodology for generating synthetic populations is illustrated in Figure 1. The following list describes the sequence of steps involved:

1. *population synthesis*, in which a synthetic representation of each household in a region is created from Census data. The individuals in the households are endowed with individual and household level characteristics such as age, gender, marital status, household income, household size, and location;

2. *activity assignment*, in which each synthetic person in a household is assigned a set of activities to perform during the day, along with the times when the activities begin and end, as given by activity or time-use survey data;

3. *location choice*, in which an appropriate real location is chosen for each activity for every synthetic person based on a gravity model and data sources such as land use patterns, tax data or commercial location data;

*NDSSL Technical Report 15-009

Table 1: Partial list of datasets we have used in the construction of previous synthetic populations.

Dataset	Description
American Community Survey	U.S. Census data used to build the synthetic set of anonymous people with the aggregate statistics matched.
Land use data	Type of land use e.g. residential, commercial, industrial etc.
National Household Travel Survey (NHTS)	Data on the travel behavior of the American public.
National Health Interview Survey (NHIS)	Provides demographic information of the population such as household size, household income, householders' ages, number of workers, number of cell phones in a household. Also includes a parameter for the importance of each household with respect to national population.
Dun & Bradstreet (D&B)	Describes retail locations, types and the number of employees.
American Time Use Survey (ATUS)	Survey data on individual's activities such as paid work, childcare, volunteering, socializing etc.
HERE (formerly NAVTEQ)	Road Network and transportation map.

4. *contact estimation*, in which each synthetic person is deemed to have made contact with a subset of other synthetic people simultaneously present at a location.

The resulting model is a dynamic representation of human mobility and interaction over the course of a normative day. From this we can also induce a social contact network, which is an interaction-based graph whose vertices are synthetic people, labeled by their demographics, and whose edges represent estimated contacts, labeled by duration of contact and type of activity. [2, 7] This social contact network is specific to a geographic location because of its dependence on "contingent realities" for the area – demographics of people who live there and the distribution of actual activity locations. It provides a plausible, bottom-up mechanism for generating large scale structure without making assumptions about hierarchies. It also makes it possible to model realistic interaction with the built environment, e.g. convenience stores that sell tobacco.

Note that it is *impossible* to build such a network by simply conducting a survey; the use of generative models to build such networks is a unique feature of this work [9]. The resulting data set is very rich, and can be applied to many problem domains. For example, we have used synthetic populations as part of a very detailed and extensive simulation of the aftermath of an improvised nuclear detonation [3, 17].

Based on methods and software developed by this team, [5, 6] Research Triangle International (RTI) has produced synthetic populations for the United States and a few other countries. Others are following suit, [14] so much so that the use of a synthetic population is now called "traditional" [20] and has spawned an entire line of research complete with review articles. [19] It is worth noting that most other attempts at synthetic population generation are limited to what we refer to as *baseline population synthesis* (described in section 1 below), which just contains a set of synthetic individuals with realistic demographic attributes. We go beyond this step to add realistic activity patterns and locations for these activities, and to induce contact networks.

Below we describe each of the steps involved in more detail.

1 Baseline population synthesis

In this step we use data from the American Community Survey (ACS) [18] to create a disaggregated set of agents endowed with various demographic variables. The methodology is based on the work of Beckman et al. [5]

The ACS provides data for public use that are resolved to the *block group* level, which is a geographical region containing between 600 and 3000 people.[1] For each block group, tables of distributions of many demographic characteristics – such as age, gender, household income, household size – are provided. We refer to these as *marginal* distributions.

[1] `https://www.census.gov/geo/reference/gtc/gtc_bg.html`

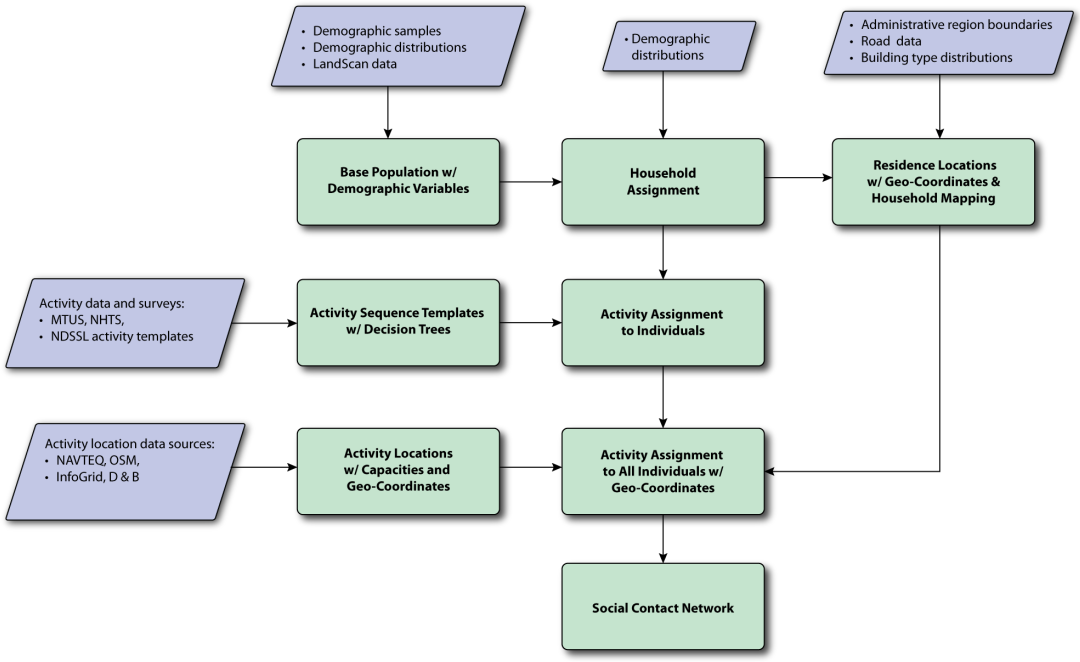


Figure 1: The synthetic population generation pipeline.

To create a synthetic population for the block group, we would like to generate a joint distribution from the given marginal distributions and then sample from it the desired number of times. In the absence of any other information, the only approach available would be to treat the given marginal distributions as independent and to create a joint distribution by multiplying the marginal distributions. Obviously, this would ignore any correlations between the demographic variables.

In reality, however, there are strong correlations between demographic variables, e.g., between age and household income. Generating a synthetic population by ignoring these correlations would clearly be a poor strategy. Fortunately, the ACS also provides a 5% representative sample for each region.

This sample is generated from a larger region known as a Public Use Microdata Area (PUMA), which consists of at least 100,000 people.[2] The sample is known as a Public Use Microdata Sample (PUMS). The PUMAs are contiguous and each block group is included in only one PUMA. A PUMS record is essentially a complete Census record, after uniquely identifying information such as names and addresses have been eliminated.

We incorporate the PUMS information into the inference of the joint distribution from the marginal distributions using a statistical procedure known Iterative Proportional Fitting (IPF). [8] IPF is a rigorous way of combining the information from the marginal distributions and the sample data. It has been shown to preserve important properties of the data: it gives a constrained maximum entropy estimate of the true joint distribution, [13] and it preserves the odds ratios given in the sample in the absence of any marginal information to the contrary. [15]

With the use of the IPF procedure, combined with the marginal and sample data, we thus obtain a joint distribution over a selected set of marginal variables for each block group, e.g. age of householder, household income, and household size. We sample this joint distribution and find a matching record (an entire household) from the PUMS data. This matching record is copied into the synthetic population. This procedure is repeated until the synthetic population size matches the true population size for the block group.

Since the entire record is copied over from the PUMS, correlations between variables within the record are maintained. This results in a synthetic population that closely matches the structure of the true population.

Validation: It is important to note, as mentioned above, that this procedure actually offers some theoretical guarantees.

[2] `https://www.census.gov/geo/reference/puma.html`

The generated synthetic population is guaranteed to match the marginal distributions of the true population for the selected demographic variables. It is also guaranteed to match the correlations between various demographic variables as far as the sample is representative (the representativeness of the sample is guaranteed by the ACS).

Additional validation is carried out by comparing the distributions of variables that are not included in the IPF step with their distributions as given by the ACS. For example, the ACS provides the distribution of the number of workers in the family for each block group. This variable is not included in the IPF step but is included in the sample data, and is therefore carried along into the synthetic population when we copy the records from the sample data to create the synthetic population. We can therefore compare the distribution of the number of workers in the family in the synthetic population with the true distribution given by the ACS tables. This validation procedure consistently shows a close match between the synthetic population and the true distributions from the ACS (Beckman et al. show one such example [5]).

2 Activity assignment

In this step we use data from the National Household Travel Survey (NHTS)[3] to assign a daily activity sequence to each agent in the synthetic population constructed in the previous step. The NHTS contains detailed information on individuals' movements and activities over the course of a normative day. [23]

The activity patterns for different members of a household are typically strongly dependent on each other. For instance, if there is a child under twelve years of age in the household, then an adult will likely be present in the home with the child whenever the child is at home. Any method for assigning activity sequences to agents must be able to take into account these dependencies.

NHTS surveys households, not individuals. At NDSSL, we have developed a method that exploits this by assigning activity sequences one household at a time, thereby preserving within-household activity correlations. This method, known as the Fitted-Values Method or FVM, [16] proceeds in three steps:

1. For each synthetic household, select the survey household to which it is most similar.
2. For each synthetic individual in the household, find the individual in the survey household who is most similar.
3. Assign the activity sequence of the matching survey individual to the corresponding synthetic individual.

In step 1, the similarity between synthetic and survey households is judged using the (asymmetric) Hausdorff distance. This is done as follows. For each person in the synthetic household, we use a person-person distance measure (D_{PP}, explained in Section 2 below) to find the closest matching person in the survey household. The distance between two households is then taken to be the maximum of the person-person distances. Since the worst of the person-person distances is treated as the household distance, when we find the best matching household, we can be sure that no person will be assigned an ill-fitting activity sequence.

Two stage fitted-value approach

The person-person distance is evaluated on the basis of only those demographics relevant to activity sequences as follows. First, note that a naïve measure like the Euclidean distance or Mahalanobis distance between demographics would not be a good choice for D_{PP}. This is because what we actually care about is the activity sequence of the survey individual. Demographic variables that are not relevant for determining the activity sequence can distort a Euclidean or Mahalanobis distance, and thereby result in poor matching of activity sequences to synthetic individuals.

What we would like to have is a means of weighting the demographic variables according to their importance in determining the activity sequence. We form summary statistics of the activity sequences – such as hours spent at home, at work, or at school – and explain them as a function of the individuals' demographics in a regression model. Then we define D_{PP} to be the Mahalanobis distance between the fitted values of the summary statistics for the survey household and the predicted values of the summary statistics for the synthetic household. The benefit of using the fitted values instead of the true values for the survey household is that it guarantees that individuals that are demographically identical will be evaluated to have a distance of zero.

[3] http://nhts.ornl.gov/

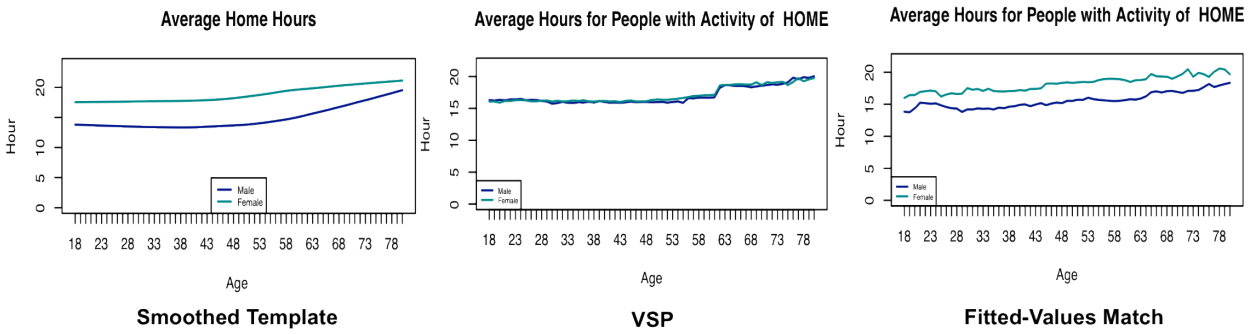


Figure 2: A comparison of the fitted-values method (right) with the data (left) and the earlier VSP method (middle). FVM captures the difference between male and female average hours at home, while the VSP method fails to do so [16].

Once the person-person distance has been calculated, we can calculate the between-household distances and identify the most similar survey household. Once the most similar survey household is identified, *individuals* within the synthetic household are compared to *individuals* within the selected survey household based upon D_{PP}. Finally each synthetic individual is assigned the attributes of the survey individual most similar (closest) to them in the chosen survey household.

Validation: We validate the assignment by comparing the average time spent in various activities broken down by age and gender between the survey data and the synthetic population. The comparison for average hours spent at home is shown in Figure 2. [16] The summary statistic generated by an earlier method due to Vaughn et al., [24] which we refer to as the VSP method after the initials of the authors. We see that while both methods capture the slight increase with age of average hours at home, FVM also manages to reproduce the difference between males and females while the VSP method does not. A similar comparison for various other activities is presented by Lum et al. [16]

3 Location choice

In this step we use data from multiple sources to assign a location for each activity for each agent.

Assigning home locations: Cartographic boundary shapefiles, defining the boundary of each block group, are available from the U.S. Census.[4] The Census and ACS also provide housing unit distributions for each block group, which contain counts of the numbers of buildings with various numbers of housing units (1, 2, 3-4, 5-9, ..., 50+). To create geographical locations for these residential buildings, we use road network data from HERE (formerly NAVTEQ).[5]

A geometric intersection of the boundary polygon with the road network segments gives the section of the road network that falls within the block group. Residential locations are allocated to streets based on the street type (e.g. smaller streets are more likely to have single-family homes), and the distribution of building types. Finally, households are allocated home locations by proceeding iteratively through the list of households and locations.

Assigning locations for other activities: Once home locations have been assigned, locations for other activities are assigned using a gravity model. Data on business and other activity locations are obtained from Dun & Bradstreet and data on school locations are obtained from the National Center for Education Statistics (NCES). [21, 22]

The general idea behind the gravity model is to choose the next activity location for a person probabilistically, given their current activity location. Probabilities are assigned to be proportional to location capacities and inversely proportional to location distances (from the person's current location). Thus, nearby large locations are more likely to be chosen than far-away small locations.

Home, work, and school (including college) activities are defined to be *anchor* activities, in the sense that the locations

[4] `https://www.census.gov/geo/maps-data/data/cbf/cbf_blkgrp.html`
[5] `https://www.here.com/navteq-redirect/?lang=en-US`

for these activities are chosen first, and the locations for other activities (e.g. shopping) are chosen relative to these anchor activity locations.

If the home location is i, then work location j is chosen according to the following distribution [1].

$$Pr(j|i) \propto w_j e^{bT_{ij}}, \tag{1}$$

where w_j is the capacity of location j for the work activity, b is a negative constant, and T_{ij} is the travel time from location i to location j. Note that a location can have different capacities for different activities. For example, a school can have a large capacity for school activities but a relatively smaller capacity for work activities.

For assigning other activity locations, such as shopping, we take into account the locations for the anchor activities both before and after the current activity. For example, if a synthetic individual has a shopping activity between work and home activities (i.e., she goes shopping on the way home from work), the probability distribution over shopping locations is defined as, [1]

$$Pr(k|j,i) \propto s_k e^{bT_{jk}+aT_{ki}}, \tag{2}$$

where j is the work location, i is the home location, and s_k is the capacity of location k for the shopping activity. T_{jk} and T_{ki} are the travel times as before.

To avoid having to compute a distribution over the entire set of locations, activity assignment is done in two steps. First the region is divided into Traffic Analysis Zones (TAZs) and the centroid of each TAZ is assigned a capacity which is the sum of the capacities of all locations within the TAZ. We first assign locations at the level of the TAZs, and then, in the second step, use the same model over all the locations within the chosen TAZ. Alternatively, a distance threshold can be imposed where we only consider candidate locations within, say, 60 miles of the current location.

Validation: Validation of this step is discussed extensively by Beckman et al. [4] They extended the synthetic population to model cell phone traffic by integrating it with a model of the cellular infrastructure, and showed that they are able to match various measures of cellular traffic. This indirectly shows that the mobility model gives a good representation of actual population densities across the region over time. They also showed that the distribution of distances traveled is a power law, which is consistent with what has been reported in the literature [12]. The mobility patterns have also been validated using various measures of traffic such as trip counts between various zones. [10]

4 Contact estimation

Once each person has been assigned a location for each activity during the day, we can induce a contact network as follows.

Activity locations and activity times are combined to construct a time-indexed map between persons and locations. This is a bipartite graph, as illustrated in Figure 3a. From this, we can induce a person-person contact network by assuming a *mixing model*.

The simplest mixing model assumes that two people in the same location at the same time are in contact, as shown in Figure 3b. This is a "complete mixing" model within each location. Large locations can be divided into *sub-locations*. For example, a school is divided into classrooms. The size or capacity of a sub-location depends on the location of which it is a part. For example, classrooms might be assigned a greater capacity than rooms in an office building. Instead of assuming full mixing, some other mixing model such an Erdős-Rényi random graph can also be used.

At the end of this step, we obtain a contact network for the entire population of the region, which can be quite large. For example, the network for the city of Los Angeles contains 16.23 million nodes and 459 million edges. [26]

Validation: There are no empirical data about contact networks at this scale, therefore we validate these networks by evaluating various metrics on the constructed networks. Network structural measures, such as degree distributions (i.e., the distribution of numbers of contacts per person), show expected patterns. Since the network is dynamic, we show the temporal degree for a few randomly chosen locations in Figures 4a, 4b, and 4c. We see that the degrees of the home locations drop during the day, while the degrees of the work and school locations increase during the day.

Figure 4d shows that the degree distribution of the non-home locations in the union graph follows a power law. The union graph is constructed by ignoring the time-stamps on the edges. Power-law distributions are reported in the literature in similar contexts such as González et al. [12]

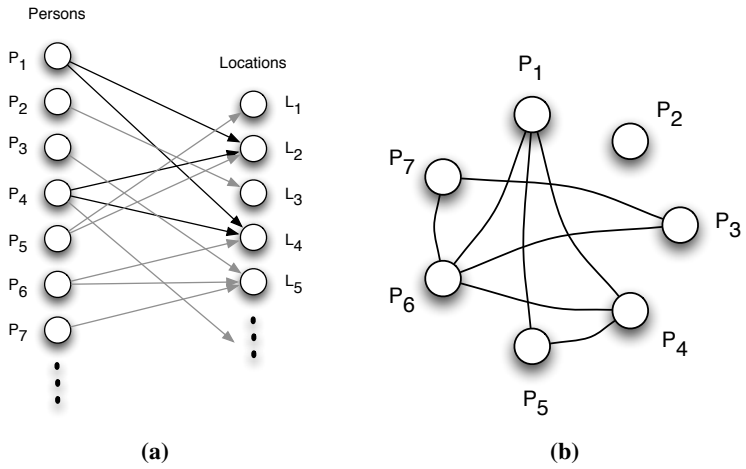
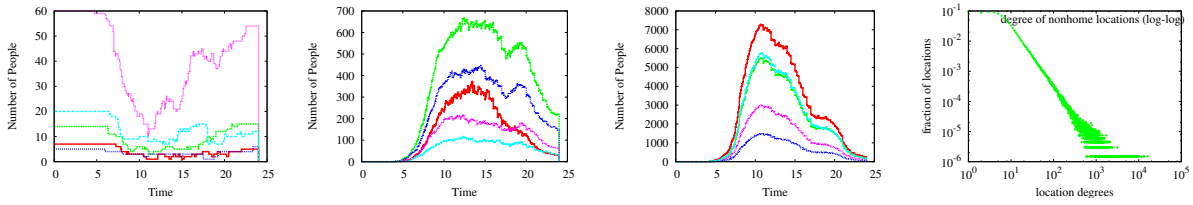


Figure 3: Left: a bipartite person-location graph where edges represent the locations visited by people. Note that two people may interact in more than one location over the course of a day, e.g., P_1 and P_4 meet at locations L_2 and L_4. Right: the induced person-person social contact network. There is an edge between any two people who are at the same location for an overlapping time during the day. If they meet at more than one location, the contact times are added. Figure reproduced from [17].



(a) Home **(b)** Work **(c)** School **(d)** Non-home union

Figure 4: The first three figures show the temporal degree of four randomly chosen locations of each type in the synthetic contact network for Los Angeles. The figure on the right shows the degree distribution of the non-home locations in the "union" graph, which is obtained by ignoring time stamps on the edges. Figures reproduced from [26].

In addition to structural validity, validity with respect to dynamic measures, such as the spread of epidemics, is evaluated and discussed by Eubank et al. [10, 11] They measure various network characteristics, such as clustering coefficient, expansion, and overlap ratio, and show their relevance and realism for epidemic dynamics. Xia et al. [25, 26] compare the synthetic networks for different cities in terms of micro-structure (graphlets), and epidemic intervention efficacy and show that the networks have differences explainable in terms of social and cultural differences between regions.

References

[1] C. Barrett, R. Beckman, K. Bisset, S. Eubank, A. Marathe, M. V. Marathe, and P. Stretz. Building social contact networks. Technical Report NDSSL-TR-06-095, Virginia Bioinformatics Institute, Virginia Tech, 2006.

[2] C. Barrett, K. Bisset, J. Leidig, A. Marathe, and M. Marathe. An integrated modeling environment to study the co-evolution of networks, individual behavior, and epidemics. *AI Magazine*, 31(1):75–87, 2009.

[3] C. Barrett, K. Bisset, S. Chandan, J. Chen, Y. Chungbaek, S. Eubank, Y. Evrenosoğlu, B. Lewis, K. Lum, A. Marathe, M. Marathe, H. Mortveit, N. Parikh, A. Phadke, J. Reed, C. Rivers, S. Saha, P. Stretz, S. Swarup, J. Thorp, A. Vullikanti, and D. Xie. Planning and response in the aftermath of a large crisis: An agent-based informatics framework. In R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, editors, *Proceedings of the 2013 Winter Simulation Conference*, 2013.

[4] R. Beckman, K. Channakeshava, F. Huang, J. Kim, A. Marathe, M. Marathe, G. Pei, S. Saha, and A. K. S. Vullikanti. Integrated multi-network modeling environment for spectrum management. *IEEE Journal on Selected Areas in Communications*, 31(6):1158–1168, June 2013.

[5] R. J. Beckman, K. A. Baggerly, and M. D. McKay. Creating synthetic base-line populations. *Transportation Research A – Policy and Practice*, 30:415–429, 1996.

[6] R. J. Beckman et al. TRANSIMS-release 1.0: The Dallas-Fort Worth case study. Technical Report LA-UR-97-4502, Los Alamos National Laboratory, 1997.

[7] K. Bisset and M. Marathe. A cyber-environment to support pandemic planning and response. *DOE SciDAC Magazine*, pages 36–47, 2009.

[8] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Annals Math. Stats*, 11(4):427–444, 1940.

[9] J. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press, 2005.

[10] S. Eubank, H. Guclu, V. S. A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, May 2004.

[11] S. Eubank, V. S. A. Kumar, M. Marathe, A. Srinivasan, and N. Wang. Structure of social contact networks and their impact on epidemics. *AMS-MIMACS Special Volume on Epidemiology*, 2006.

[12] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding human mobility patterns. *Nature*, 453: 779–782, June 5 2008.

[13] C. T. Ireland and S. Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, Mar. 1968.

[14] M. Lenormand and G. Deffuant. Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation*, 16(4):12, 2013. ISSN 1460-7425. URL `http://jasss.soc.surrey.ac.uk/16/4/12.html`.

[15] R. J. A. Little and M.-M. Wu. Models for contingency tables with known mmargin when target and sampled populations differ. *J. Amer. Statist. Assoc.*, 86(413):87–95, Mar. 1991.

[16] K. Lum, Y. Chungbaek, S. G. Eubank, and M. V. Marathe. A two-stage, fitted values approach to activity matching. In *Procedia - Social and Behavioral Sciences*, 2013.

[17] M. Marathe, H. Mortveit, N. Parikh, and S. Swarup. Prescriptive analytics using synthetic information. In W. H. Hsu, editor, *Emerging Trends in Predictive Analytics: Risk Management and Decision Making*. IGI Global, 2014.

[18] M. Mather, K. L. Rivers, and L. A. Jacobsen. The american community survey. *Population Bulletin*, 60(3):1–20, 2005. URL `http://www.census.gov/acs/www/`.

[19] K. Müller and K. Axhausen. Population synthesis for microsimulation: State of the art. Technical report, Technical Report August. Swiss Federal Institute of Technology Zurich, 2010.

[20] M.-R. Namazi-Rad, P. Mokhtarian, and P. Perez. Generating a dynamic synthetic population using an age-structured two-sex model for household dynamics. *PLoS ONE*, 9(4):e94761, 04 2014. doi: 10.1371/journal.pone.0094761. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0094761`.

[21] National Center for Education Statistics. Characteristics of public and private elementary and secondary schools in the United States: Results from the 201112 schools and staffing survey. Technical Report NCES 2013312, Department of Education, August 2013.

[22] National Center for Education Statistics. Private school universe survey (PSS): Public-use data file user's manual for school year 2011-12. Technical Report NCES 2014351, Department of Education, September 2014.

[23] A. Santos, N. McGuckin, H. Nakamoto, D. Gray, and S. Liss. Summary of travel trends: 2009 National Household Travel Survey. Technical Report FHW A-PL-ll-022, U.S. Department of Transportation Federal Highway Administration, June 2011.

[24] P. Speckman, K. Vaughn, and E. Pas. Generating household activity-travel patterns (HATPs) for synthetic populations. In *Transportation Research Board Annual Meeting*, 1997.

[25] H. Xia, C. L. Barrett, J. Chen, and M. V. Marathe. Computational methods for testing adequacy and quality of massive synthetic proximity social networks. In *Proc. IEEE International Conference on Big Data Science and Engineering (BDSE)*, 2013.

[26] H. Xia, J. Chen, M. V. Marathe, and S. Swarup. Comparison and validation of synthetic social contact networks for epidemic modeling (extended abstract). In *Proceedings of The Thirteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Paris, France, May 2014.