International Journal of Transportation
Vol.4, No.1 (2016), pp.41-56
http://dx.doi.org/10.14257/ijt.2016.4.1.03

A Two-stage, Fitted Values Approach to Activity Matching

Kristian Lum[1], Youngyun Chungbaek[2], Stephen Eubank[3] and Madhav Marathe[4]

[1,2,3,4]*University of Network Dynamics and Simulation Science Laboratory, Virginia Tech*

[1]*kristianlum@gmail.com*, [2]*ychungba@vbi.vt.edu*, [3]*seubank@vbi.vt.edu*, [4]*mmarathe@vbi.vt.edu*

Abstract

Accurate and rich representations of constituent actor populations are a critical component of agent-based models. Such populations are designed so that demographic, behavioral, procedural, and geographic characteristics of the synthetic population jointly reflect the available information about the true population. Information about the attributes to be mimicked in the synthetic population is often derived from survey samples of the real actors of interest -- such as firms in a market or households in a city. This additional information is then mapped to individual actors in such a way that each actor in the population represents one sample from the joint distribution of all assigned attributes. These actors then interact according to rules, which are functions of their attributes. Thus, accurate attribute matching is necessary to ensure that model outputs are meaningful. In real applications, behavioral surveys often yield complex data types, such as daily activity schedules or action sequences, for which it is difficult to conceive of adequate conditional models of behavior that could be used to generate new behavioral data as a function of covariates. Here we propose a method for assigning behavioral templates to synthetic agents from a set of survey templates. Our method first maps the complex behavioral data to a reduced-dimension Euclidean space, then estimates conditional models in this space. We then make predictions in Euclidean space for synthetic actors and assign them the template schedule that minimizes the distance to the predicted value. By employing a two-step process, we also ensure that within-household dependence structures are maintained in the synthetic population. We illustrate the method with an application to a synthetic representation of households in the state of Israel and demonstrate superior ability to generate accurate joint distributions between demographic characteristics and behavioral activity relative to the standard behavior assignment method.

Keywords: *activity assignment, synthetic populations*

1. Introduction

Synthetic representations of human populations are the backbone of agent-based models in which the actors or agents represent people. Such simulation systems have been used extensively in epidemiology [1, 2], the social sciences [3], transportation sciences [4, 5, 6, 7, 8] and disaster management and planning [9], among other fields. The conclusions drawn from any of these human-based simulation systems all depend critically on the fidelity with which the underlying base population and the interactions among the actors are modeled.

Creating synthetic representations of human populations, henceforth called synthetic populations, has proved to be a challenging endeavor, involving the synthesis of many data sources and types. Although many different methods exist, the template for this procedure typically involves some combination of the following components:

ISSN: 2287-7940 IJT
Copyright ⓒ 2016 SERSC

- *Assignment of demographic characteristics*. Using census data, a base population is created such that the marginal totals of each demographic variable match those of the true population at the finest resolution known. The iterative proportional fitting procedure (IPFP), as in Deming and Stephan [10], Fienberg [11], and [12] is used to model the joint distribution of the demographic variables while respecting the known marginal totals [13, 14].
- *Assignment of activity schedule*. Conditional on the demographic characteristics generated in the previous step, sequences of activities are assigned to each agent. This has been achieved by creating new schedules by sampling from a set of sequentially conditional probabilities, as in Kitamura and Kermanshah [15] and Kitamura [16]. Others, such as Bowman and Ben-Akiva [17] and Ben-Akiva and Bowman [18], have focused on nested logit models to model broad classes of activity templates or tours in addition to small sub-trip refinements. These methods, for the most part, focus on the generation of individual schedules, largely ignoring intra-household dependencies. The more recent work of Gliebe and Koppelman [19, 20], Borgers, Hofman and Timmermans [21], Zhang, Timmermans and Borgers [22], Zhang and Fujiwara [23] and Bradley and Vovsha [24], to name a few, delve into the intricacies of joint modeling of decision making. Vaughn, Speckman and Pas [25] and Speckman, Sun and Vaughn [26] take a different approach. They re-sample from an existing set of activity schedules using CART, classification and regression trees [27]. In doing so, each synthetic household is assigned the activity schedules of a survey household, precluding the need for explicit modeling of within-household dependence.
- *Assignment of activity locations*. Given the daily schedule of activities, agents are assigned locations for each of those activities. This may be done using gravity models in which the probability a location is selected for an activity is inversely proportional to its distance from anchor location(s), such as home or work [28] or trip chaining [29, 30]. This step may be combined with the assignment of activity schedules, if the schedules are created dynamically. A thorough and detailed review of activity and location assignment can be found in Bhat and Koppelman [31]. An overview of matching techniques can be found in Hobeika and Paradkar [32] and Timmermans and Zhang [33].
- Derivation of social/contact networks. Once agents have been assigned activity locations, contact networks may be derived according to the amount of time agents are collocated. That is, if agent *i* and agent*i'* are at location *L* for *m* minutes, then they receive an edge between them with weight *m*. These edges may also be labeled based on where the interaction takes place (*i.e.*, home, work, school, *etc.*,) [28].

Assignment of activity schedules is the focus of this paper. In particular, we focus on extending the re-sampling, Classification and Regression Tree (CART)-based methodology developed in Vaughn, Speckman and Pas [25] and Speckman, Sun and Vaughn [26], which we refer to here as VSP for the authors of the 1997 paper. This particular methodology is the target of this work, as it embedded in many agent-based simulations, including TranSims [7, 34], a popular transportation modeling software that remains widely used among transportation researchers (see *e.g.*, Huang

 Copyright ⓒ 2016 SERSC

[35], Jeihani [36], Ullah [37], Volosin [38], Montz [39], and Isbell [40]) and other transportation simulations [41, 42]. Outside of transportation, this method for activity assignments continues to be used in a wide variety of agent-based models [43, 44, 45]. The VSP method fits a regression tree in which the dependent variable is the total number of minutes a survey household spends outside of the home; independent variables are household-level demographics such as the income of the oldest household member or the number of children present in the household. The output of the algorithm is a decision tree that divides the surveyed households into "leaf nodes" based on the household-level demographic characteristics. Synthetic households are then mapped to the appropriate leaf node, following the fitted decision tree. The activity schedules of one survey household in the corresponding leaf node is assigned at random to the synthetic household. According to Speckman, Sun and Vaughn [26], heads of households are matched with heads of households, adults to adults, and children to children, *etc*.

Activity assignment affects many aspects of the simulation model. For example, in epidemiological models for disease spread, the disease diffuses along contact networks. As described, these contact networks are derived directly from co-location patterns, which are a function of both the assigned activity schedule and the location at which the activity was assigned to take place. Thus, insufficiently modeling activity schedules can result in non-representative contact networks. For example, if agents are assigned activity schedules without regard to age, then the model may erroneously co-locate elderly individuals and young individuals at a university, resulting in unrealistic synthetic contact networks. Thus it is important that activity schedules be assigned in a way that is plausible given each actor's demographic attributes. Of course, activity assignment is important in other application areas as well. In transportation models, activity schedules determine the timing, origin, and destination of trips, with family members often sharing rides and transportation. Thus, it is not only desirable that an activity schedule be a good fit to a synthetic individual marginally, it is also important that household members are jointly assigned activity schedules that produce realistic household-level activity patterns.

This suggests the need for a conditional model for activity sequences given demographic covariates. Activity schedules are high-dimensional categorical time series with complex dependence structures, making it difficult to build meaningful conditional models. For example, if one activity is vacuuming and the household owns only one vacuum, then there is strongly negative dependence between the activity schedules of each household member with respect to vacuuming. Moreover, it is impossible for an individual to perform household tasks while at work or in the car, and so forth, creating dependence between activity and location. The nature of dependence over time between activities and between activities across different household members is sufficiently complex to make the specification of realistic probability models on activity sequences rather daunting. While several methods have been suggested to generate entirely new activity sequences, validation of these methods is difficult because it is challenging to even produce an exhaustive list of all of the a priori rules that must be obeyed by an activity sequence. Our method abrogates any concerns about the realism of activity sequences, and this is one of its major strengths.

Our approach simplifies activity assignment by first restricting the possible activity schedules to those sampled in the survey as in VSP. We further simplify activity assignment by modeling the activity sequences in a reduced space-- we take summary statistics of the activity sequence rather than treating them as a time series. We perform matching in this reduced space, using a novel, two-stage fitted-values matching approach. We account for intra-household dependence structure by

Copyright ⓒ 2016 SERSC

assigning all synthetic household members schedules from among those of a single survey household. Section 2 gives an outline of the notation that is to be used throughout this manuscript. Section 3 describes our activity assignment method in detail, including specific distance metrics and models used, and an example. Section 4 shows the result of our methodology as applied to activity assignment for a synthetic population of Israel. Here, we compare this method to the VSP method. Section 5 concludes and offers generalizations and avenues for possible further research.

2. Notation

We begin by clarifying the notation that will be used throughout this document. The goal is to match activity sequences, A^*_{i*j*}--categorical valued stochastic processes, from the survey respondents to synthetic individuals. We will refer to the j^{th} synthetic household as H_j and the i^{th} synthetic person in the j^{th} synthetic household as P_{ij}. For each of these individuals, we have individual demographic information, such as age, sex, income, education level and household demographic information, such as the number of children in the household or the income of the oldest household member. All of the household- and individual-level variables are contained in the variable X, with X_{ij} referring to the individual- and household-level demographics of P_{ij}. Analogous variables for the population of survey respondents are represented similarly with asterisks. A lower dimensional representation of $A^*_i{}^*_j{}^*$, the sequence of activities of $P^*_i{}^*_j{}^*$, is contained in the variable $Y^*_i{}^*_j{}^* = [\, Y^*_i{}^*_j{}^* *1, Y^*_i{}^*_j{}^* *2, \ldots Y^*_i{}^*_j{}^*_K]$-- the total time duration spent doing each of K=6 activity types-- work, home, shop, other, college, travel. Table 1 gives a summary of the notation.

Table 1. Summary of Notation

Variable name	Synthetic population	Survey
Household subscript	i	i*
Individual subscript	j	j*
Household	H_j	H^*_{j*}
Person in household	P_{ij}	P^*_{i*j*}
Demographic information	X_{ij}	X^*_{i*j*}
Activity duration information	NA	Y^*_{i*j*k}

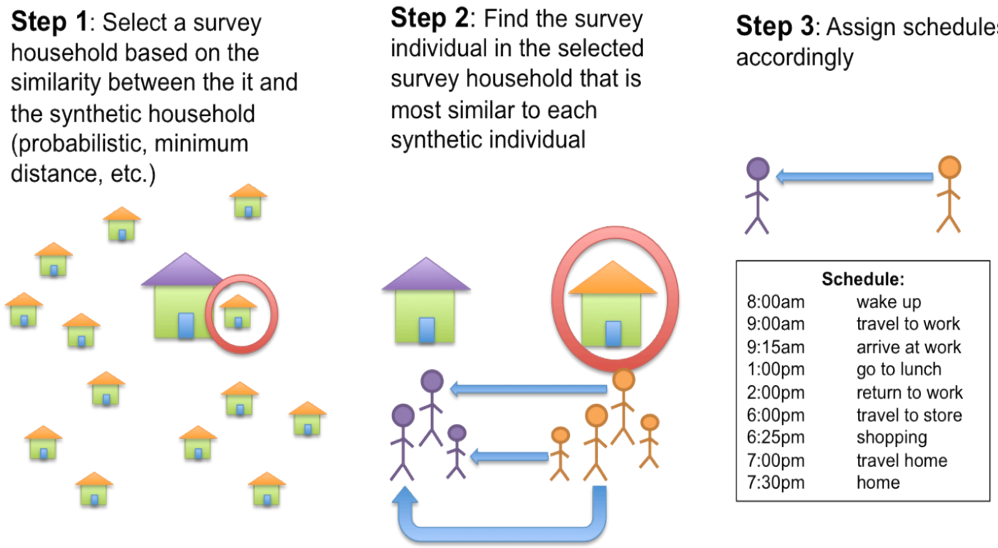
3. Method

3.1. Overview

Our method for assigning activity schedules can be broken down into three steps. This process is repeated (and can be done in parallel) for each synthetic household. For each synthetic household, in the first step we select the survey household to which it is most similar. In this case, similarity is defined by a distance metric, DHH(Hj, H*j*). Intuitively, in order for this distance metric to represent demographic similarity, it should have the property that if Hj is demographically indistinguishable from H*j*, the distance is zero. If Hj is very similar to household H*j*, this distance will be small, and conversely, if Hj is very unlike H*j* in terms of its demographics, this distance should be large. Once the synthetic household is assigned a survey household, individuals within the synthetic household are compared to individuals within the selected survey household based upon the

 Copyright ⓒ 2016 SERSC

pairwise between-individual distance metric, DPP(Pij, P*i*j*). This distance metric should have properties similar to those of the household distance metric:



demographically similar individuals should be separated by short distances while individuals who are very different should be far apart. In the third step, each synthetic individual is assigned the activity schedule of the survey individual most similar (closest) to them in the chosen survey household. This procedure is shown summarized in Figure 1.

Figure 1. Overview of the Activity-matching Process

3.2. Defining Distances

We define the between household distance using the (asymmetric) Hausdorff distance:

$$D_{HH}(H_j, H^*_{j*}) = \max_{P_{ij} \subset H_j} \{D_{PH}(P_{ij}, H^*_{j*})\} \quad (1)$$

$$D_{PH}(P_{ij}, H^*_{j*}) = \min_{P^*_{i*j*} \subset H^*_{j*}} \{D_{PP}(P_{ij}, P^*_{i*j*})\} \quad (2)$$

This between-household distance metric is a minimax over a pairwise, between-individual distance metric, D_{PP}. This is most easily thought of algorithmically. Consider synthetic household H_j and survey household H^*_{j*}. For each $P_{ij} \subset H_j$, find the survey individual, $P^*_{i*j*} \subset H^*_{j*}$ with the shortest distance to P_{ij}. This distance is defined to be the distance between person P_{ij} and household H^*_{j*}, $D_{PH}(P_{ij}, H^*_{j*})$. The distance between the two households is then the maximum of all of these person-to-household distances. This distance metric is represented in Figure 2 using Euclidean distance as the between-person distance, $D_{PP}(P_{ij}, P^*_{i*j*})$; red edges indicate the minima defined by, $D_{PH}(P_1, H^*)$ and $D_{PH}(P_2, H^*)$.

The Hausdorff distance enjoys a long history of use in matching algorithms, particularly in regard to matching images and objects [46, 47, 48, 49, 50, 51]. In the context of activity sequence matching, this distance metric is intuitively appealing. Members of each synthetic household are assigned the activity schedules from among those of one survey household; this distance metric finds the best match for each individual within a household. The worst of these best matches is considered the total distance for the household. In this way, the Hausdorff distance protects against any synthetic individual being assigned a very ill-fitting activity sequence.

Copyright ⓒ 2016 SERSC

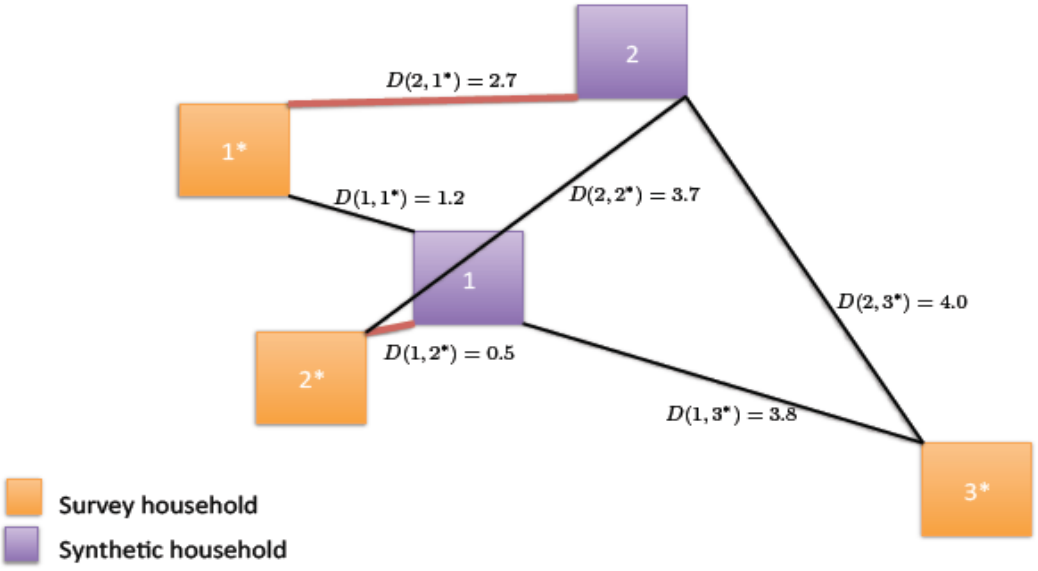


Figure 2. Illustration of the Hausdorff Distance

The Hausdorff distance is defined in terms of D_{pp} for all pairs P_{ij} and $P^*_i{}^*_j{}^*$, and thus to fully specify the Hausdorff distance, one must also define DPP. DPP is meant as a metric for the similarity between individuals, information about which is contained in the demographic covariates, X. Below we discuss the advantages and disadvantages of three candidate pairwise distance metrics: Euclidean distance, Mahalanobis distance, and a fitted-values distance.

One naive distance metric for DPP(P_{ij}, $P^*_i{}^*_j{}^*$) is the Euclidean distance,

$$D_{PP}(P_{ij}, P^*_{i^*j^*}) = \| \mathbf{X}_{ij} - \mathbf{X}^*_{i^*j^*} \| \tag{3}$$

And the related generalization, the Mahalanobis distance,

$$D_{PP}(P_{ij}, P^*_{i^*j^*}) = (\mathbf{X}_{ij} - \mathbf{X}^*_{i^*j^*})^T \mathbf{S}^{-1} (\mathbf{X}_{ij} - \mathbf{X}^*_{i^*j^*}) \tag{4}$$

Where S is the shared covariance matrix of X_{ij} and $X^*_i{}^*_j{}^*$, which are assumed to follow the same multivariate distribution. Note that both Equations 3 and 4 define valid metrics on the space of $X_{ij} \subset \chi$, where χ is R_p if all demographic covariates are continuous. If X_{ij} contains all binary variables, for example, χ is $\{0,1\}^p$.Further investigation of the properties of these candidate metrics is required. In the case of Equation 3, those covariates with the largest variance contribute most to the distance on expectation, regardless of their relationship to the activity sequence A. That is, assume that each of the p covariates in $X_{ij} = [x_{ij}\ 1, \ldots, x_{ijp}]$ have means $\mu = [\mu_1, \ldots, \mu_p]$ and variances $\sigma^2 = [\sigma^2_1, \ldots, \sigma^2_p]$. Following the requirement of the Mahalanobis distance, also assume that X and X* share the same distribution. Equation 1 can be re-expressed as a sum, where each summand represents the contribution of one covariate to the overall distance: $\boldsymbol{D_{PP} = \sum_{l=1}^{p} (x_{ijl} - x^*_{i^*j^*l})^2}$. Then, the expected contribution of the l^{th} covariate is $E[(x_{ij1} - x^*_{i^*j^*1})^2] = 2\,\sigma_l^2$. Thus, covariates that have little relationship with A can still have a strong effect in determining the inter-person distance. By a similar argument, if we use the Mahalanobis distance of Equation 4, the contribution of each covariate is some function of its variance and covariance with the other covariates. If all are uncorrelated (*i.e.*, S is diagonal), then

Copyright ⓒ 2016 SERSC

all covariates contribute equally, regardless of the relationship between the l^{th} covariate and the activity sequence A. For both of these distances, A has no bearing on the distance between X and X^*.

The property that the expected contribution to the distance is unrelated to A is undesirable. It allows for the possibility that the effect of irrelevant covariates is large relative to the effect of highly relevant covariates. Consider the case in which irrelevant covariates are increasingly included, *i.e.*, $x^*_i{}^*_j{}^*_1$ independent of A for increasingly many $x^*_i{}^*_j{}^*_1$. The distance functions of equations 1 and 2 increasingly become dominated by contributions from covariates that are unrelated to A. In practice, this is a problem because one does not know a priori which covariates to include in the distance function or how to weight them in terms of their explanatory capacity with respect to A. Ideally, one would be able to include as many covariates as are available and automatically weight them with respect to their importance in determining A.

A is unweildy-- it is complex and high dimensional. Defining a meaningful conditional model for A, from which to extract such weights, is challenging. Rather than model A directly, we define a conditional model on summary statistics of A, in this case, the total number of hours spent doing each of K activities. This information is contained in the variable $Y = [Y^*_i{}^*_j{}^*_1,\ldots,Y^*_i{}^*_j{}^*_K]$. We model Y using a simple model, $Y_k = f_k(X) + \varepsilon$, where ε is idiosyncratic error. In this case, we choose $f_k(X) = X\beta_k$, a simple linear regression model. Although this model is likely inadequate as a generative model of Y (it does not preclude negative values for Y, for example), we use it only as a tool to automatically generate weights for our distance function,

$$D_{PP}(P_{ij}, P^*_{i^*j^*}) = (\hat{Y}_{ij} - \hat{Y}^*_{i^*j^*})^T \mathbf{S}^{-1} (\hat{Y}_{ij} - \hat{Y}^*_{i^*j^*}) \quad (5)$$

Where $\hat{Y}_{ij} = \{\hat{Y}_{ijk}\}$ for $\hat{Y}_{ijk} = \mathbf{X}_{ij}\hat{\beta}_k$. $\hat{Y}$ are the predicted values of Y for synthetic individuals from the fitted model. Similarly, $\hat{Y}^*_k = \mathbf{X}^*\hat{\beta}_k$ are the model's fitted values, and S = diag (σ_1^2 , …, σ_K^2) contains the residual variance of each regression on the diagonal. This is the Mahalanobis distance between the fitted values of Y for $P^*_{i^*j^*}$ and the predicted values of Y for P_{ij}. In practice, for computational expediency, we assume S is diagonal and estimate it from the survey data. The following properties hold for our distance, D_{PP}:

- D_{PP} defines a valid metric on the space of fitted and predicted values, each k x 1 vectors in R^p.
- Regardless of β, two individuals that are demographically indistinguishable (*i.e.*, $X_{ij} = X^*_{i^*j^*}$) are at a distance of zero apart. Note that if we had used the true values for the Y^*s instead of the fitted values this would not be the case.
- Individuals that are demographically similar with respect to Y have a smaller distance than those that are different.
- This scheme inherits the additional property from the linear model literature that if the l^{th} covariate is uncorrelated with Y_k for k=1…K, as the number of surveyed individuals increases, the l^{th} coefficient of each regression approaches zero. Thus, whether two individuals are similar along the l^{th} dimension of X has no effect on the distance between them, at least asymptotically. At the asymptote, demographic covariates that are uncorrelated with Y are given a weight of zero.

From $D_{PP}(P_{ij}, P^*_{i^*j^*})$, $D_{HH}(H_j, H^*_{j^*})$ inherits some desirable properties. If $H_j$is demographically indistinguishable from H*j* (*i.e.*, each individual in H_j is

Copyright ⓒ 2016 SERSC

demographically identical to an individual in H^*_{j*}), then $D_{HH}(H_j, H^*_{j*})=0$. This is a consequence of the fact that $X_{ij} = X^*_{i*j*}$ implies that $D_{PP}(P_{ij}, P^*_{i*j*}) = 0$. Furthermore, if Hjis demographically similar to H^*_{j*}, then $D_{HH}(H_j, H^*_{j*})$ is small. Here demographic similarity is small whenever the maximum amount of dissimilarity between matched individuals in the two households is small.

3.3. Matching

Having calculated all of the between household distances, $D_{HH}(H_j, H^*_{j*})$, matching is simple. We assign H^*_{m*} as a match to H_j if $D_{HH}(H_j, H^*_{m*}) = \min_{j*} D_{HH}(H_j, H^*_{j*})$; the m^{*th} survey household is matched to the j^{th} synthetic household. In order to assign schedules to individuals $P_{ij} \subset H_j$, we find the survey individual who best matches each synthetic individual within the selected m^{*th} household. That is, $P_{ij}$is assigned the activity schedule of the survey individual $P_{n*m*} \subset H^*_{m*}$ such that $D_{PP}(P_{ij}, P^*_{n*m*}) = \min_{i*} D_{PP}(P_{ij}, P^*_{i*m*})$. In the case that there are multiple minima, we select the least frequently assigned schedule. For large number of synthetic households, this is essentially equivalent to random assignment from the set of households that achieve the minimum distance.

3.4. Example

Here we provide a concrete example of the matching process. Suppose we are given the synthetic household described in Table 2. The personal income field refers to the income decile, number of children is the number of persons under 18 years of age in the household (the same for all members), and the predicted values are the expected number of hours spent performing activities related to home, work, shopping, other, and college respectively under independent regression models for each activity type. For example, the18 year old male in this household is expected to spend, on average, 10.27 hours at home and 1.19 hours at college each day under the fitted regression model.

Table 2. Information About Synthetic Individuals

HH	Person	Sex	Age	Personal Income	Children	Predicted Values
1	1	male	18	4	2	(10.27, 1.28, 0.34, 8.59, 1.19)
1	2	male	45	9	2	(6.36, 7.28, 0.46, 6.22, 0.05)
1	3	female	45	8	2	(9.03, 5.60, 0.40, 5.73, -0.50)

Similarly, suppose the pool of survey households from which we will select a match for the above listed synthetic household consists of the following three households:

 Copyright ⓒ 2016 SERSC

Table 3. Information about Survey Individuals

HH	Person	Sex	Age	Personal Income	Children	Predicted Values
1^*	1^*	female	37	7	3	(9.80, 5.05, 0.48, 6.02, -0.39)
1^*	2^*	male	39	10	3	(7.54, 7.47, 0.40, 5.71, 0.41)
2^*	1^*	male	20	4	2	(10.05, 1.28, 0.22, 8.59, 1.41)
2^*	2^*	female	44	6	2	(9.27, 4.50, 0.52, 5.68, 0.25)
2^*	3^*	male	52	4	2	(10.05, 1.28, 0.70, 6.93, 0.41)
3^*	1^*	male	20	4	1	(10.61, 1.28, 0.37, 8.20, 1.39)
3^*	2^*	female	46	8	1	(9.37, 5.60, 0.43, 5.33, -0.30)

Assume that in this case the residual variance of each of the regressions is one. We calculate $D_{PP}(P_{ij}, P^*_{i*j*})$, shown in Table 4. For example, $D_{PP}(P_{11}, P^*_{1*1*}) = (10.27 - 9.80)^2 + \ldots + (1.19 + 0.39)^2$.

Table 4. Between-individual Distances

	P^*_{1*1*}	P^*_{1*2*}	P^*_{2*1*}	P^*_{2*2*}	P^*_{2*3*}	P^*_{3*1*}	P^*_{3*2*}
P_{11}	23.60	55.47	0.16	20.86	3.59	0.31	32.34
P_{12}	17.08	1.70	57.33	16.66	50.51	59.92	12.85
P_{13}	0.99	6.03	31.63	1.90	22.14	30.83	0.31

Based upon these inter-individual distances, we find the best match within each survey household for each of the synthetic individuals; the corresponding distance is shown in parentheses. These are shown in Table 5.

Table 5. Best within-household matches

	H^*_{1*}	H^*_{2*}	H^*_{3*}
P_{11}	P^*_{1*1*} (23.60)	P^*_{2*1*} (0.16)	P^*_{3*1*}(0.31)
P_{12}	P^*_{1*2*} (1.70)	P^*_{2*2*} (16.66)	P^*_{3*2*} (12.85)
P_{13}	P^*_{1*1*} (0.99)	P^*_{2*2*} (1.90)	P^*_{3*2*} (0.31)
$D_{HH}(H_j, H^*_{j*})$	23.60	16.66	12.85

The maximum within each household of the minimum between-person distance (the column-wise maximum) is the Hausdorff distance. This is shown in the bottom row of Table 5. We select the third survey household, H^*_{3*}, as the best match for H_1, as this household achieves the minimum distance from H_1. The synthetic individual is assigned the activity schedule of the survey individual in H^*_{3*}, that is the best match. P_{11} is assigned the activity schedule of P^*_{3*1*}; both P_{12} and P_{13} are assigned the activity schedule of P^*_{3*2*}.

4. Results

We apply our household activity matching methodology to assigning activity schedules to a synthetic population of Israel, consisting of approximately 7 million synthetic individuals. We include in $\boldsymbol{X}$ the personal covariates age, sex, and income. We also include several household-level covariates in $\boldsymbol{X}$ for each individual: the number of adults in the household, the number of children in the household, the income of the highest earning individual in the household, and the age of the oldest household member. We note that it may be advantageous to include more covariates

Copyright ⓒ 2016 SERSC

in $\boldsymbol{X}$, particularly those relating to geographic location or more detailed indicators of socio-economic class. In this case, this information was not present in the data we have obtained; however, we note that this model is quite general and adding that information, should it become available, is as easy as adding additional covariates to $\boldsymbol{X}$. For each of the $\boldsymbol{K}$ activity types, we fit a linear regression model, $\boldsymbol{Y}_k = \boldsymbol{X}\beta_k + \varepsilon$, using the covariates described, where $\boldsymbol{Y}_k$ is the number of minutes spent doing activity k. We use an AIC-based variable selection technique, stepwise selection, to obtain a parsimonious model. In order to avoid enforcing a linear trend in age or income, we define categorical income variables representing income deciles and five-year age block categorical variables for age. The variable selection procedure selected each of the candidate variables in at least one of the regression models. The results of this regression are given in Table 6. Columns correspond to activity type and rows correspond to variables. Variables that the variable selection procedure excluded from the model are shown with a dot. These results are mostly unsurprising, as they conform to traditional gender role stereotypes and general intuition. For example, the coefficient on sex is positive for the number of minutes spent at home, meaning females tend to spend more time in the home; it is negative for work. Average number of minutes per day spent at work increases with increasing income group (income 1 indicates the group with the lowest income and 7 indicates the group with the highest income). Here, we also see that the average number of minutes spent at college is decreasing with age. The youngest group, aged 18-24 is the baseline group, and as such, its effect is subsumed by the intercept.

Table 6. Regression Results

Coefficient	Home	Work	Shopping	College	Travel	Other
Intercept	16879	27026	740	6605	11609	25013
Sex (female)	6161	-5179	.	-1161	.	.
Income 1	10496	-17237	888	2245	.	1919
Income 2	7208	-5015	399	2026	.	-5610
Income 3	1516	-474	322	1762	.	-3913
Income 4	679	1527	64	223	.	-3172
Income 5	-130	3492	-35	261	.	-3693
Income 6	-3601	4392	192	1063	.	-1909
Income 7	-1291	5062	-234	613	.	-4296
age(24,29]	.	.	322	-527	.	-3338
age(29,34]	.	.	398	-2057	.	-2064
age(34,39]	.	.	1182	-2267	.	-3197
age(39,44]	.	.	1665	-2511	.	-4674
age(44,49]	.	.	1144	-2939	.	-4711
age(49,54]	.	.	1740	-3579	.	-5961
age(54,59]	.	.	1461	-3538	.	-5415
age(59,64]	.	.	1410	-4481	.	-1705
age(64,100]	.	.	885	-5096	.	-717
household size	-1819	.	-96	-220	.	1188
Oldest Age	93	.	-12	.	16	.
Highest Income	482	.	71	-254	-153	.
# Children	2027	.	.	.	195	-967

 Copyright ⓒ 2016 SERSC

We compare our results to the original survey data and the results produced using the VSP algorithm. These results are shown in Figure 3. We present results for each activity type, disaggregated by gender and age. The column marked as smoothed template refers only to the data within the survey. These were smoothed using a kernel smoother to give a better sense of the general trend in the data. The other columns, labeled as VSP and Fitted Values Match are made from applying the two activity assignment methods to the synthetic population. Qualitatively, one would hope to see similar activity profiles by age and gender in the synthetic population as appeared in the survey.

The top row, labeled Home, shows the average number of hours spent at home by sex and age. In the survey data, we see a marked difference between the sexes, with men spending fewer hours than women at home on average. There is a slight uptick in the number of hours spent at home in the older ages. Using VSP, this trend is largely lost. On average both sexes spend the same amount of time at home. The increase in the older ages, however is preserved. Using our method, labeled as Fitted-Values Match, we are able to maintain the difference between the sexes. The women spend more time at home on average. We also are able to reproduce the increase for those older synthetic individuals.

The row labeled Work shows a similar comparison: the proportion of synthetic individuals with a work activity by age and gender. We again see that the VSP method fails to reproduce a difference in the sexes in terms of the average number of hours spent at work; this effect is seen under our method, though its magnitude appears smaller than in the survey data.

The third row similarly presents the results of college attendance by age and gender. The smoothed survey data shows high rates of attendance for younger people. Between 18 and 23 years of age, female college attendance hovers around 25%. Male college attendance in our survey data is quite high at age 18, followed by a noticeable drop until 23 years of age, at which point it exhibits a second peak in attendance, reaching almost 30% at age 27. For both sexes, college attendance remains relatively low from age 35 onward, with a small increase in the late sixties. The VSP method assigns college related activities at a much lower rate to the younger ages than is seen in the survey data. 18 year old males are assigned college activities at about a 10% rate, as compared to nearly 40% in the survey; females are assigned college activities at only a slightly higher rate. From age 25 to the mid-fifties, college attendance under VSP remains relatively constant at slightly greater than 10% for both sexes, with a sharp drop in the late fifties. From the late fifties until the mid-seventies, college attendance is relatively flat at about 7%. Compared to the survey data, this is extremely high. The method we propose, however, seems to reproduce the pattern seen in the survey data relatively well. Although both our method and VSP assign higher college attendance to 18-year-old females than males, our method retains the precipitous peak in the mid-late twenties for the males and the gradual decline in attendance for females starting around age 23.

The last three rows show results for the proportion of individuals that go shopping throughout the day, the proportion that engage in other activity types that do not fit the broad classification system, and the proportion that travel. In each case, slight differences may be seen between the sexes and across age. As a general pattern, VSP tends not to reproduce differences by sex, though differences across age may be present. Our method does typically produce different rates of activity types by gender, though, the magnitude of the gender effect in some cases is smaller than that seen in the survey. The shape of the rate of activity attendance across age is qualitatively smoother under our method, lacking the cliff apparent around the mid-50s seen under the VSP method.

Copyright ⓒ 2016 SERSC

That there were consistent differences by sex across many of the comparisons we showcase is partly due to the fact that our survey data is from Israel. The reason that our method was able to reproduce the differences in activity patterns by gender is because the regression coefficient on the sex variable was estimated to have a large effect. This same effect could be achieved by matching only males survey respondents to synthetic males and vice versa. For surveys from other countries and cultures, gender may not play as important a role in determining activity patterns. Because our method provides automatic weights for the distance function, it is unnecessary to possess substantial culture-specific knowledge of the factors that influence activity patterns to develop reasonable activity sequences for synthetic populations. For example, in other data sets, it may be more important to match people of equivalent income levels rather than genders. Our method learns the appropriate weighting of the various demographics attributes from the data, so these types of determinations, which are generally time-consuming and require in-depth knowledge of the social structures in the population of interest, can be made automatically.

5. Discussion

We have presented a novel method for activity assignment in synthetic populations that shows improvement to the degree of accuracy with which activity schedules are assigned. By developing a technique to incorporate both individual-level covariates and household-level covariates in the household matching process, we have achieved improved matching. Where before, using a single-layer CART allocation method resulted in unrealistic patterns of college attendance by age, work habits by income level and gender, *etc.*, our method has largely resolved these inconsistencies. Future research might consider more complicated functions for f – perhaps the CART model could find a new home embedded within this larger procedure. It would also be useful to formalize hypothesis tests for whether Y is an adequate projection of A. In this paper, we have simply assumed that Y is a meaningful summary.

Although this method has been discussed in the context of activity assignment, it could easily be extended to any situation in which attributes of elements in a set are to be matched to other elements in a set while maintaining within-set dependence. To generalize entirely, the fitted value matching approach could be used for prediction in any setting in which no adequate model exists for generating a valid realization of the dependent variable.

Acknowledgments

We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. Special thanks to Richard Beckman for numerous comments and discussion on this topic. This work has been partially supported by the following grants: DTRA Grant HDTRA1-11-1-0016 NSF NetSE Grant CNS-1011769, NSF SDCI Grant OCI-1032677, NIH MIDAS Grant 2U01GM070694-09, and DTRA CNIMS Contract HDTRA1-11-D-0016-0010.

 Copyright ⓒ 2016 SERSC

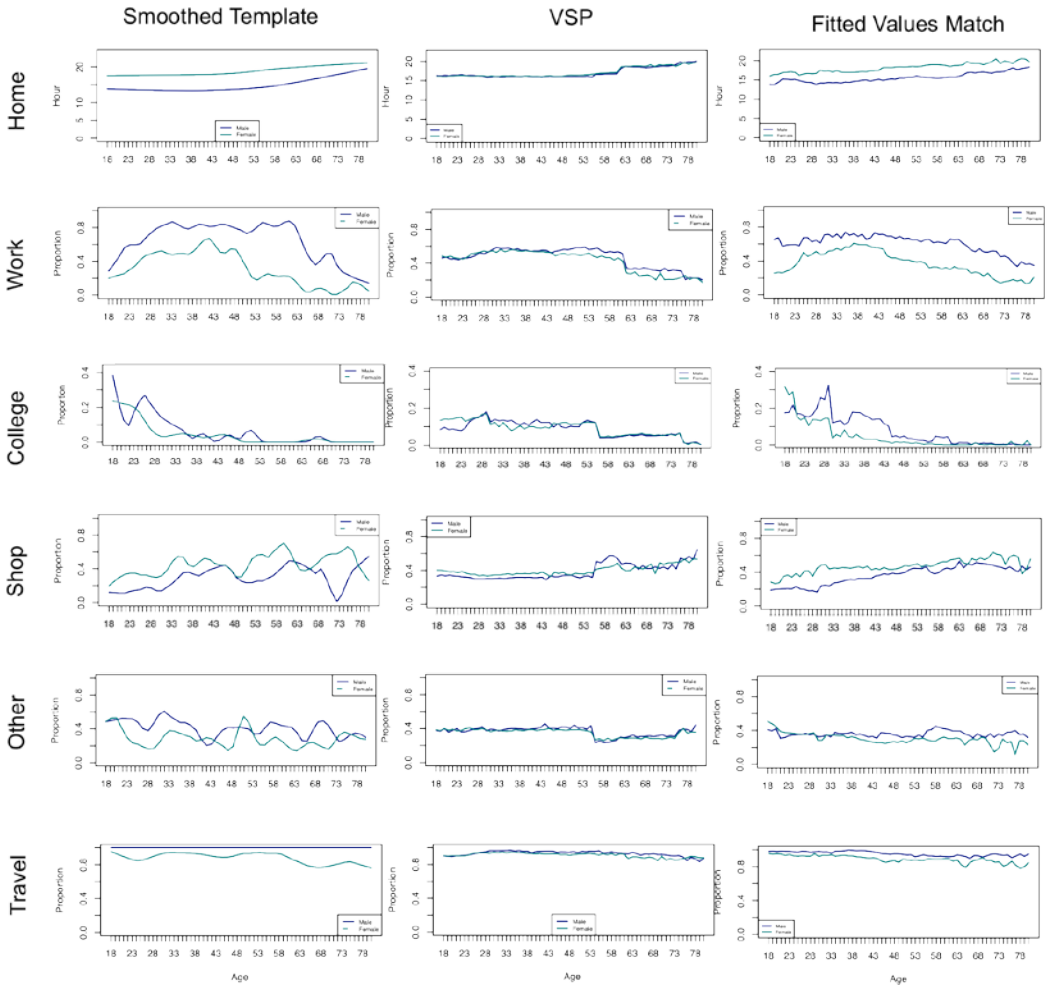


Figure 3. Comparison Activity Profiles from the Survey Data (smoothed template) VSP, and our Proposed Method (Fitted Values Match)

References

[1] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng and M. V. Marathe, "EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks", Proceedings of the 2008 ACM/IEEE conference on Supercomputing, **(2008)**.

[2] C. L. Barrett, S. Eubank and M. V. Marathe, "An Interaction-Based Approach to Computational Epidemiology", Proceedings An Interaction-Based Approach to Computational Epidemiology, AAAI, **(2008)**.

[3] J. M. Epstein, "Generative social science: Studies in agent-based computational modeling", Princeton University Press, **(2006)**.

[4] K. G. Goulias and R. Kitamura, "Travel demand forecasting with dynamic microsimulation", **(1992)**.

[5] L. Smith, R. Beckman and K. Baggerly, "TRANSIMS: Transportation analysis and simulation system", Los Alamos National Lab., NM (United States), **(1995)**.

[6] K. Nagel, R. J. Beckman and C. L. Barrett, "TRANSIMS for urban planning", Proceedings TRANSIMS for urban planning, 6th International Conference on Computers in Urban Planning and Urban Management, Venice, Italy, **(1999)**.

[7] C. L. Barrett, R. J. Beckman, K. P. Berkbigler, K. R. Bisset, B. W. Bush, S. Eubank, J. M. Hurford, G. Konjevod, D. A. Kubicek and M. V. Marathe, "TRANSIMS (TRansportation ANalysis SIMulation System). Volume 0: Overview, Volume 2: Software, Part 1: Modules, Volume 2: Software, Part 2: Selectors: Volume 2: Software, Part 3: Test Networks, Volume 2: Software, Part 5: Libraries, Volume 3: Files,Volume 6 Installation", Los Alamos National Lab., Los Alamos, NM, Rep. LA-UR-99-1658, LA-UR-99-2574, LA-UR-99-2575, LA-UR-99-2576, LA-UR-99-2578, LA-UR-99-2579, LA-UR-99-2580, **(1999)** pp.

[8] R. Kitamura, C. Chen, R. M. Pendyala and R. Narayanan, "Micro-simulation of daily activity-travel patterns for travel demand forecasting", Transportation, vol. 27, **(2000)**, pp. 25--51.

Copyright ⓒ 2016 SERSC

[9] N. Parikh, S. Swarup, P. E. Stretz, C. M. Rivers, B. L. Lewis, M. V. Marathe, S. G. Eubank, C. L. Barrett, K. Lum and Y. Chungbaek, "Modeling human behavior in the aftermath of a hypothetical improvised nuclear detonation", Proceedings Modeling human behavior in the aftermath of a hypothetical improvised nuclear detonation, Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems, **(2013)**.
[10] W. E. Deming and F. F. Stephan, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", The Annals of Mathematical Statistics, vol. 11, **(1940)**, pp. 427-444.
[11] S. E. Fienberg, "An iterative procedure for estimation in contingency tables", The Annals of Mathematical Statistics, **(1970)**, pp. 907-917.
[12] S. E. Fienberg, "The analysis of incomplete multi-way contingency tables", Biometrics, **(1972)** pp. 177-202.
[13] R. J. Beckman, K. A. Baggerly and M. D. McKay, "Creating synthetic baseline populations", Transportation Research Part A: Policy and Practice, vol. 30, **(1996)**, pp. 415-429.
[14] J. Y. Guo and C. R. Bhat, "Population synthesis for microsimulating travel behavior", Transportation Research Record: Journal of the Transportation Research Board, 2014 **(2007)**, pp. 92-101.
[15] R. Kitamura and M. Kermanshah, "Identifying time and history dependencies of activity choice", Transportation Research Record, **(1983)**.
[16] R. Kitamura, "Sequential, history-dependent approach to trip-chaining behavior", Transportation Research Record, **(1983)**.
[17] J. L. Bowman and M. E. Ben-Akiva, "Activity-based disaggregate travel demand model system with activity schedules", Transportation Research Part A: Policy and Practice, vol. 35, **(2001)**, pp. 1-28.
[18] M. E. Ben-Akiva and J. L. Bowman, "Activity based disaggregate travel demand model system with daily activity schedules", **(1997)**.
[19] J. P. Gliebe and F. S. Koppelman, "A model of joint activity participation between household members", Transportation, vol. 29, **(2002)**, pp. 49-72.
[20] K. G. Goulias, "Multilevel analysis of daily time use and time allocation to activity types accounting for complex covariance structures using correlated random effects", Transportation, vol. 29, **(2002)**, pp. 31-48.
[21] A. Borgers, F. Hofman and H. Timmermans, "Conditional choice modelling of time allocation among spouses in transport settings", European Journal of Transport and Infrastructure Research, vol. 2, **(2002)**, pp. 5-17.
[22] J. Zhang, H. Timmermans and A. Borgers, "Utility-maximizing model of household time use for independent, shared, and allocated activities incorporating group decision mechanisms", Transportation Research Record: Journal of the Transportation Research Board, vol. 1807, **(2002)**, pp. 1-8.
[23] J. Zhang and A. Fujiwara, "Representing heterogeneous intra-household interaction in the context of time allocation", 83rd Annual Meeting of the Transportation Research Board, Washington, DC, **(2004)**.
[24] M. Bradley and P. Vovsha, "A model for joint choice of daily activity pattern types of household members", Transportation, vol. 32, **(2005)**, pp. 545-571.
[25] K. M. Vaughn, P. Speckman and E. I. Pas, "Generating household activity-travel patterns (HATPs) for synthetic populations", Prepared For Presentation at the 76th Annual Meeting of the Transportation Research Board. Washington DC, **(1997)**.
[26] P. L. Speckman, D. Sun and K. M. Vaughn, "Synthesizing Activity-Travel Patterns: A Resampling Approach", Working Paper, National Institute of Statistical Sciences (NISS), Research Triangle Park, NC, **(1998)**.
[27] L. Breiman, "Classification and regression trees", **(1984)**.
[28] C. L. Barrett, R. J. Beckman, M. Khan, V. Anil Kumar, M. V. Marathe, P. E. Stretz, T. Dutta and B. Lewis, "Generation and analysis of large synthetic social contact networks", Proceedings Generation and analysis of large synthetic social contact networks, Winter Simulation Conference, **(2009)**.
[29] R. Kitamura, "Incorporating trip chaining into analysis of destination choice", Transportation Research Part B: Methodological, vol. 18, **(1984)** pp. 67-81.
[30] T. Adler and M. Ben-Akiva, "A theoretical and empirical model of trip chaining behavior", Transportation Research Part B: Methodological, vol. 13, **(1979)**, pp. 243-257.
[31] C. R. Bhat and F. S. Koppelman, "Activity-based modeling of travel demand", Handbook of transportation Science, Springer, **(1999)**, pp. 35-61.
[32] A. G. Hobeika and R. Paradkar, "Comparative analysis of household activity matching approaches in Transportation Analysis and Simulation System", Journal of transportation engineering, vol. 130, **(2004)**, pp. 706-715.
[33] H. J. P. Timmermans and J. Zhang, "Modeling household activity travel behavior: Examples of state of the art modeling approaches and research agenda", Transportation Research Part B: Methodological, vol. 43, **(2009)**, pp. 187-190.
[34] K, C.a.B. Barrett, L. Smith, V. Loose, R. Beckman, J. Davis, D. Roberts and M. Williams, "An Operational Description of TRANSIMS", Los Alamos National Laboratory Unclassified Report, LA-UR-95-2393, Los Alamos, NM, **(1995)**.

 Copyright ⓒ 2016 SERSC

[35] S. Huang, A. W. Sadek, I. Casas and L. Guo, "Calibrating travel demand in large-scale micro-simulation models with genetic algorithms: a TRANSIMS Model Case Study", Transportation Research Board 89th Annual Meeting, 10-2437, **(2010)**.
[36] M. Jeihani, K. Ahn, A. G. Hobeika, H. D. Sherali and H. A. Rakha, "Comparison of TRANSIMS'Light Duty Vehicle Emissions with On-Road Emission Measurements", Journal of the Transportation Research Forum, vol. 45, **(2010)**.
[37] M. S. Ullah, A. Rahman, R. Morocoima-Black and A. Mohideen, "Travel Demand Modeling for a Small MPO Using TRANSIMS", Transportation Research Board 90th Annual Meeting, 11-1183, **(2011)**.
[38] S. S. E. a. P. Volosin, R. M. Pendyala, B. Grady and B. Gardner, "The Application of a Microsimulation Model System to the Analysis of a Light Rail Corridor: Insights from a TRANSIMS Deployment", 91st annual meeting of the Transportation Research Board, Washington, DC, **(2012)**.
[39] T. Montz and Z. Zhang, "Calibration and Validation of a Regional-Level Traffic Model for Hurricane Evacuation", Transportation Research Board 92nd Annual Meeting, 13-2339, **(2013)**.
[40] N. A. Isbell and K. G. Goulias, "Modeling Second-by-Second Traffic Emissions in a Mega-Region", Transportation Research Board 93rd Annual Meeting, 14-2325, **(2014)**.
[41] S. P. Greaves and P. R. Stopher, "Creating a Synthetic Household Travel/Activity Survey – Rationale and Feasibility Analysis", Transportation Research Record 1706, **(2000)**, pp. 82-91.
[42] N. Cetin, K. Nagel, B. Raney and A. Voellmy, "Large-scale multi-agent transportation simulations", Computer Physics Communications, vol. 147, **(2002)**, pp. 559-564.
[43] S. Eubank, H. Guclu, V. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai and N. Wang, "Modelling disease outbreaks in realistic urban social networks", Nature, 429 **(2004)**, pp. 180-184.
[44] J. M. H. G. Chowell, S. Eubank and C. Castillo-Chavez, "Scaling laws for the movement of people between locations in a large city", Physical Review E, vol. 68, **(2003)**, pp. 066102.
[45] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng and M. V. Marathe, "EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks", 2008 ACM/IEEE conference on Supercomputing, Austin, TX, **(2008)** November 15-21.
[46] D. P. Huttenlocher, G. A. Klanderman and W. J. Rucklidge, "Comparing images using the Hausdorff distance", Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 15, **(1993)**, pp. 850-863.
[47] W. Rucklidge, "Efficient visual recognition using the Hausdorff distance", Springer Heidelberg, **(1996)**.
[48] N. Aspert, D. Santa-Cruz and T. Ebrahimi, "Mesh: Measuring errors between surfaces using the hausdorff distance", Proceedings Mesh: Measuring errors between surfaces using the hausdorff distance, Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference, **(2002)**.
[49] D.-G. Sim, O.-K. Kwon and R.-H. Park, "Object matching algorithms using robust Hausdorff distance measures", Image Processing, IEEE Transactions, vol. 8, **(1999)**, pp. 425-429.
[50] O. Jesorsky, K. J. Kirchberg and R. W. Frischholz, "Robust face detection using the hausdorff distance", Proceedings Robust face detection using the hausdorff distance, Audio-and video-based biometric person authentication, **(2001)**.
[51] W.-L. Hung and M.-S. Yang, "Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance", Pattern Recognition Letters, vol. 25, **(2004)**, pp. 1603-1611.

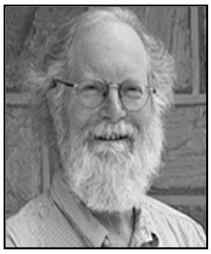
Authors

Kristian Lum, Kristian Lum is the Lead Statistician at the Human Rights Data Analysis Group. Previously, she was a Research Assistant Professor in the Network Dynamics and Simulation Science Laboratory at Virginia Tech. She earned her PhD in Statistics at Duke University in 2010.

Youngyun Chungbaek, Youngyun Chungbaek is a Research Associate in the Network Dynamics and Simulation Science Laboratory at the Virginia Bioinformatics Institute at Virginia Tech. She received her Master's degree in 2003 in Computer Science and her Ph.D. in 2011 in Educational Research and Evaluation at Virginia Tech.

Copyright ⓒ 2016 SERSC

Madhav Marathe,Madhav Marathe is a professor of Computer Science and director of the Network Dynamics and Simulation Science Laboratory. He obtained his Bachelor of Technology degree in 1989 in Computer Science and Engineering from the Indian Institute of Technology, Madras, and his Ph.D. in 1994 in Computer Science from the University at Albany under the supervision of Professors Harry B. Hunt III and Richard E. Stearns. He is a fellow of the IEEE, ACM and AAAS.



Stephen Eubank, Stephen Eubank is deputy director, Network Dynamics and Simulation Science Laboratory, tenured professor, Department of Population Health Sciences, and adjunct professor, Department of Physics. Since arriving at the Biocomplexity Institute at Virginia Tech in January, 2005, he has pursued interests both in developing advanced technology for the study of realistic socio-technical systems and also in understanding how the dynamics of diffusive processes on networks, e.g. disease transmission, are related to the structure of the underlying networks. He is the PI on one of the research groups making up the NIH's MIDAS (Modeling Infectious Disease Agent Study) network.

Copyright ⓒ 2016 SERSC