Adjustable Attribute Matching in Digital Similars of Populations

Kazi Ashik Islam, S. S. Ravi, Henning S. Mortveit, Samarth Swarup

University of Virginia, Charlottesville VA 22904, USA
{ki5hd,ssravi,henning.mortveit,swarup}@virginia.edu

Abstract. A digital similar (DS) of a population of a region is a common starting point for agent-agent based simulations. Here, an integer linear programming-based algorithm is presented that refines an existing, high-resolution methodology for constructing DSs. The extension consists of constructing a household-to-residence mapping that maximizes the correlation between household income of individual households and residence property values of individual residences. The algorithm is applied to a coastal region of Virginia (US) where we demonstrate that new household-to-residence assignment generates significantly different outcomes than the existing approach which is random assignment at block-group level. Using the context of road inundation and measures such as "time to evacuate" and "time to reach critical care", it is demonstrated significant differences across household income segments with the new method, while no such difference is established with the prior method.

Keywords: Digital similar · societal resilience · synthetic population · population digital twin

1 Introduction

Highly-detailed population models form a basis for many agent-based simulation models and computational modeling across domains such as epidemiology [2,25], disaster preparedness and planning [19, 23], and urban science [5, 12]. Such efforts rely on detailed, individual-level representations of entire populations of the study region, which include relevant demographic information. We refer to these as digital similars, though they are also referred to as synthetic populations or population digital twins. These representations are synthesized by integrating multiple datasets on demographics, activity patterns, and residences and other locations that people visit. The success of such studies depends in large part on the veridicality of the population representation. Data integration is typically done by matching common attributes that are relevant to the purpose. For example, when merging demographic and activity schedule data, matching is done on demographic attributes that are determined to be relevant to predicting activity durations [17, 27]. When such attributes are not available, matching has been done randomly [1, 3].

It is possible that, in the absence of precise data that would allow attribute matching, we might still wish to have a non-random matching. An example is the assignment of households to residence locations, where we might wish to have a correlation between household incomes and residence values. We don't typically expect this correlation to be maximal, as a household might buy a residence and then either have their income change or have the residence value change (increase or decrease). Lacking empirical data on the level of correlation, our approach here is to develop a method that allows us to do adjustable attribute matching. For a concrete example, we use household income and residence value as the two attributes to be matched. Our location data for US digital similars is generated through a combination of modeling, data fusion, and model training based a broad variety of data sources [4, 13, 20, 22]. This data resource has subsequently been augmented with estimates of residential property value from parcel data. Household income data are available through the US Census, which we used in generating the synthetic people and households that constitute our US digital similar [1,3]. In earlier work, these households of the digital similar were assigned residence locations at random within each Census block group of the study region. In our new methodology described here, we use an integer program (IP) formulation to construct an assignment that maximizes the correlation between household income, a PUMS [28] variable, and the detailed estimates of residence values. The IP-formulation also supports construction of an assignment where the Pearson correlation coefficient falls within a prescribed interval as long as the interval's upper bound does not exceed the maximal.

The new method for the household-residence assignment is illustrated for the Eastern Shore, Virginia (ESVA), a region that is exposed to storm surges and flooding. Using our related work on evacuation routing for inundated transportation networks [15, 16] in combination with digital similars for the two counties, Accomack and Northampton, of ESVA, our setup is as follows: inundation data from TideWatch [29] is spatially joined with the road infrastructure [13] and adjusted speed limits are determined. Through routing, we determine the following for each household/residence location on the ESVA:

- The time needed to evacuate to a target destination across the state border with Maryland in the north;
- The time needed to reach urgent care/hospital;
- The time needed for emergency personnel (e.g., a fire truck) to reach the residence location.

The metrics are determined for (a) random assignment of household to residences and for (b) the new IP-based assignment presented in Section 3. In each case, we measure the travel times with and without inundation and assess the fraction of households for which travel times are impacted, for low-, middle-, and high-income households (based on the assigned residences). We find that there are significant differences in travel times only when matching is done in a correlated way. While this is not unexpected, it leads us to believe that any analysis of disaster response, evacuation, etc., that is based on the new matching will show more meaningful patterns by household income and other, correlated,

demographic variables such as race and ethnicity. In related work for construction of digital similars (or synthetic populations) such as [9–11,14,21,26,31], we are not aware of such details being incorporated in the methodology.

Paper organization. In Section 2 and 3 we present the new algorithm with proofs. Following this, we describe our approach to scaling which is through a spatial decomposition of the study region followed by a multi-pass process for each blockgroup. In Section 4 we demonstrate the use of the new digital similar in the context of flooding for the Eastern Shore, Virginia (ESVA). Specifically, we demonstrate how road inundation causes quite different impacts to the ESVA population when broken down by income compared to when a random household-to-residence mapping is used. The measures considered were time-to-evacuate, time-to-reach-urgent-care, and emergency-response-time. We conclude with a summary in Section 5.

2 Approach

2.1 Generalized Pearson Correlation Coefficient

Let $\mathbb{R}$ denote the set of real numbers and $[n]$ denote the set $\{1, 2, \ldots, n\}$. Consider two sets of n variables, say $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, where each variable takes on a value from $\mathbb{R}$. In addition, there is a function f that maps $X \times Y$ to $\mathbb{R}$. Thus, for each pair of variables $x_i \in X$ and $y_j \in Y$, the value $f(x_i, y_j)$ is in $\mathbb{R}$.

A **perfect matching** between X and Y is a permutation π of $[n]$ such that x_i is matched with $y_{\pi(i)}$. For a given perfect matching π between X and Y, the value of the **Generalized Pearson Correlation Coefficient**, denoted by $\mathrm{GPCC}(X, Y, \pi)$, is defined as follows:

$$\mathrm{GPCC}(X, Y, \pi) \;=\; \sum_{i=1}^{n} f(x_i, y_{\pi(i)})\,. \tag{1}$$

A special case of this is the common definition of the Pearson Correlation Coefficient (PCC), where π is the identity permutation (i.e., each x_i gets matched with y_i) and the function f is defined by

$$f(x_i, y_i) \;=\; \frac{1}{n-1}\left(\frac{x_i - \mu(X)}{\sigma(X)}\right)\left(\frac{y_i - \mu(Y)}{\sigma(Y)}\right). \tag{2}$$

In the above equation, $\mu(X)$ and $\mu(Y)$ are respectively the sample means of X and Y and $\sigma(X)$ and $\sigma(Y)$ are respectively the sample standard deviations of X and Y.

Graph Theoretic Definitions We will use a few standard definitions from graph theory. These definitions can be found in many texts [30, e.g.]. A bipartite graph $G(V_1, V_2, E)$ has two disjoint sets of nodes V_1 and V_2, and each edge in E has one node from V_1 and the other from V_2. A **matching** M in G is a subset of edges such that no two edge of M are incident on the same node. The size

of a matching M is the number of edges in M. When there is a weight $w(e)$ associated with each edge $e \in E$, the weight of a matching M is the sum of the weights of the edges in M.

A bipartite graph $G(V_1, V_2, E)$ is **balanced** if $|V_1| = |V_2|$. For a balanced bipartite graph, with $|V_1| = |V_2| = n$, a **perfect matching** of G is a matching of size n. Consider a balanced bipartite graph $G(V_1, V_2, E)$ which has a perfect matching. Suppose $V_1 = \{v_1, v_2, \ldots v_n\}$ and $V_2 = \{w_1, w_2, \ldots w_n\}$. Now, any perfect matching of G represents a one-to-one correspondence between V_1 and V_2. If the nodes in V_1 are ordered as $\langle v_1, v_2, \ldots v_n \rangle$, then a perfect matching M can be thought of as a permutation π of $[n]$. In other words, M is the set of edges given by $\{v_i, w_{\pi(i)} : 1 \leq i \leq n\}$. This view of a perfect matching in a balanced bipartite graph allows us to formulate the problem of maximizing GPCC as that of constructing an appropriate perfect matching in such a bipartite graph.

When there are edge weights, a **maximum weight perfect matching** of $G(V_1, V_2, E)$ is a perfect matching whose weight is a *maximum* among all the perfect matchings of G. It is well known that if a balanced bipartite graph $G(V_1, V_2, E)$, where $|V_1| = |V_2| = n$, has a perfect matching, then such a matching of maximum weight can be computed in time $O(n|E|)$, see [6].

2.2 Maximizing Generalized Pearson Correlation Coefficient

Given the definition of GPCC by Equation (1), it is of interest to consider the problem of finding a permutation π that *maximizes* the GPCC value. From the discussion in Section 2.1, it can be seen that this maximization problem can be solved by a simple reduction to the **maximum weight perfect matching** (MWPM) problem on balanced bipartite graphs. Since the MWPM problem can be solved efficiently [6–8], it follows that the problem of finding a permutation that maximizes the GPCC value can also be solved efficiently.

Our algorithm for maximizing the GPCC value is shown in Figure 1. The following proposition establishes the correctness and the running time of the algorithm.

Proposition 1. *Given values for the variables in the sets $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ and a function f that returns the value $f(x_i, y_j)$ for any pair of inputs x_i and y_j, the algorithm in Figure 1 returns a permutation that maximizes the GPCC value defined by Equation* (1)*. Further, the algorithm runs in polynomial time.*

Proof: Since G is a complete balanced bipartite graph and $|V_x| = |V_y| = n$, G has a perfect matching. (For example, the set of edges $\{\{v_i, w_i\} : 1 \leq i \leq n\}$ is a perfect matching for G.) From the discussion in Section 2.1, it can be seen that every matching of sets X and Y represents a perfect matching in G. Further, for every such matching, from Equation (1), the value of GPCC is the sum of the weights of the edges in the corresponding matching. Thus, a perfect matching

Input: The values of $2n$ variables x_1, x_2, ..., x_n and y_1, y_2, ..., y_n; a function f that returns the value $f(x_i, y_j)$ given the values of any pair of variables x_i and y_j.

Output: A permutation π of $[n]$ that maximizes $\text{GPCC}(X, Y, \pi)$ over all permutations of $[n]$.

Steps of the Algorithm:

1. Construct a weighted balanced *complete* bipartite graph $G(V_x, V_y, E)$ as follows. The node sets $V_x = \{v_1, v_2, \ldots, v_n\}$ and $V_y = \{w_1, w_2, \ldots, w_n\}$ are in one-to-one correspondence with sets X and Y respectively. $E = \{\{v_i, w_j\} : 1 \leq i, j \leq n\}$. For each edge $\{v_i, w_j\} \in E$, the weight $w(v_i, w_j)$ is set to $f(x_i, y_j)$.
2. Compute a maximum weight perfect matching M of G.
3. For each edge $\{v_i, w_j\} \in M$, set $\pi(i) = j$.
4. Return the permutation π.

Fig. 1: Algorithm to Find a Permutation that Maximizes GPCC

in G with the largest total weight indeed provides a matching of X and Y with the largest value of GPCC. This establishes the correctness of the algorithm.

To estimate the running time, we assume that for a given pair of values x_i and y_j, the value $f(x_i, y_j)$ can be computed in $O(1)$ time. Step 1 of the algorithm runs in $O(n^2)$ time since the number of edges in G is n^2 and the weight of each edge can be computed in $O(1)$ time. As mentioned earlier, Step 2 runs in $O(n|E|)$ $= O(n^3)$ time since $|E| = n^2$. Step 3 runs in $O(n)$ time. Thus, the running time of the algorithm is dominated by the time used in Step 2. Hence, the algorithm runs in $O(n^3)$ time. ■

3 Integer Linear Programming Formulations for Matching Problems

Overview We present integer linear programming (ILP) formulations for two versions of the matching problem. In the first version, the goal is to obtain a matching for which the GPCC value is within specified bounds. The second version, the goal is to find a matching that maximizes the GPCC value.

Obtaining a GPCC Value Within Given Bounds A natural question that arises in the context of generating synthetic populations is that of finding a permutation that leads to a given GPCC value. From the previous discussion, it can be seen that this problem corresponds to finding a perfect matching of a specified weight in a balanced weighted bipartite graph. Maalouly [18] presents results that suggest this problem, which he refers to as the **Exact Weight Perfect Matching** problem, is unlikely to be efficiently solvable. Here, we present

a method that uses an integer linear programming (ILP) formulation for a relaxed version of the problem. Specifically, we are given values of $2n$ variables $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ and two real values ℓ and u. The goal is to find a permutation π of $[n]$ such that the GPCC value corresponding to π (given by Equation (1)) satisfies the condition $\ell \leq \text{GPCC}(X, Y, \pi) \leq u$.

Using the discussion in previous sections, we can consider the above problem as that of finding a perfect matching M in a balanced complete bipartite graph $G(V_x, V_y, E)$ such that the weight of M is at least ℓ and at most u. Recall that the weight of each edge $\{v_i, w_j\}$ in G is given by $f(x_i, y_j)$. Our {0,1}-ILP formulation for the problem is as follows.

Variables: There are n^2 variables z_{ij}, $1 \leq i, j \leq n$. Each z_{ij} takes on a value from $\{0, 1\}$. The variable z_{ij} represents edge $\{v_i, w_j\}$. The value of $z_{ij} = 1$ if the edge $\{v_i, w_j\}$ is in the chosen perfect matching; otherwise, the value of z_{ij} is 0.

Objective: No optimization objective is needed here.

Constraints:

1. For each node v_i, *exactly* one edge from the chosen matching should be incident on v_i. This leads to the following set of n constraints:
$$\sum_{j=1}^{n} z_{ij} \;=\; 1, \quad 1 \leq i \leq n.$$
2. For each node w_j, *exactly* one edge from the chosen matching should be incident on w_j. This leads to the following set of n constraints:
$$\sum_{i=1}^{n} z_{ij} \;=\; 1, \quad 1 \leq j \leq n.$$
3. The weight of the chosen matching must satisfy the specified upper and lower bounds. This leads to the following two constraints:
$$\sum_{i=1}^{n}\sum_{j=1}^{n} f(x_i, y_j)\, z_{ij} \;\geq\; \ell \;\text{ and}$$
$$\sum_{i=1}^{n}\sum_{j=1}^{n} f(x_i, y_j)\, z_{ij} \;\leq\; u\,.$$
4. Each z_{ij} must take on a value from $\{0, 1\}$:
$$z_{ij} \in \{0, 1\}, \quad 1 \leq i, j \leq n\,.$$

Recovering a matching: When there is a solution, for each z_{ij} that has value 1, we match x_i with y_j.

3.1 Maximizing the GPCC Value

We now present an ILP formulation for finding a permutation that maximizes the GPCC value. As pointed out in Section 2.2, this problem can be solved efficiently by a reduction to the maximum weight perfect matching problem in

bipartite graphs. However, an ILP formulation for the problem is convenient in practice since search heuristics built into ILP solvers such as Gurobi are generally able to generate solutions quickly even for reasonably large problem instances. The ILP formulation presented here is obtained by a minor modification to the formulation presented in Section 3.

Variables: There are n^2 variables z_{ij}, $1 \leq i, j \leq n$. Each z_{ij} takes on a value from $\{0, 1\}$. The variable z_{ij} represents edge $\{v_i, w_j\}$. The value of $z_{ij} = 1$ if the edge $\{v_i, w_j\}$ is in the chosen perfect matching; otherwise, the value of z_{ij} is 0.

Objective: Maximize $\sum_{i=1}^{n} \sum_{j=1}^{n} f(x_i, y_j)\, z_{ij}$.

Constraints:

1. For each node v_i, *exactly* one edge from the chosen matching should be incident on v_i. This leads to the following set of n constraints:
$$\sum_{j=1}^{n} z_{ij} \;=\; 1, \quad 1 \leq i \leq n.$$
2. For each node w_j, *exactly* one edge from the chosen matching should be incident on w_j. This leads to the following set of n constraints:
$$\sum_{i=1}^{n} z_{ij} \;=\; 1, \quad 1 \leq j \leq n.$$
3. Each z_{ij} must take on a value from $\{0, 1\}$:
$$z_{ij} \in \{0, 1\}, \quad 1 \leq i, j \leq n\,.$$

Recovering a matching: When there is a solution, for each z_{ij} that has value 1, we match x_i with y_j.

3.2 Adaptation for Use with Digital Similars

The IP-based algorithm for constructing household-to-residence assignment is applied independently at blockgroup resolution. For a blockgroup, there will typically be disparity between the number of households $|H|$ and the number of residence locations $|R|$. The two cases to consider are (i) $|R| \geq |H|$ and (ii) $|R| < |H|$. For the first case, we apply the IP-based algorithm to the set of households H and a randomly selected subset $R' \subset R$ with $|R'| = |H|$. For the second case, we construct a partition $\mathcal{H} = \{H_1, H_2, \ldots, H_{k+1}$ of H such that $|H_1| = |H_2| = \cdots = H_k = |R|$ and apply the algorithm to the pairs (H_i, R) with $1 \leq i \leq k$. The remaining set H_{k+1} is handled as in the first case.

4 Results

The algorithm was applied to the digital similar of Accomack and Northampton, Virginia, the counties that constitute the Eastern Shore. We compare three methods for household-to-residence assignment: (1) IP-based assignment with

synthetic population P_{IP}, (2) sorted assignment with population P_{sorted}, and (3) random assignment with population P_{random}. In addition, we split the populations into the following income-based sub-demographics:

- **Low income:** $[0, \$55,000]$; 11,375 households
- **Mid income:** $[\$55,000, \$120,000]$; 5,533 households
- **High income:** $> \$120,000$; 1,736 households

The road network was constructed from OpenStreetMap data [24], inundation data was collected from TideWatch [29], and the two data sets were spatially join to determine road segment traversability and modified traversal speeds where. Routing was done at household resolution over (A) the baseline road network and (B) the inundated road network where household were matched via their residence to the nearest transportation node.

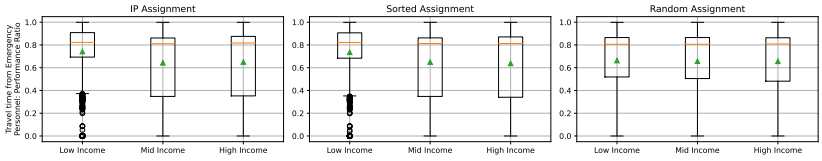

Fig. 2: Distributions of travel time from emergency personnel for affected households under the three assignments.

The following metrics were measured through simulation:

- **Travel time to safety:** We calculate the travel time (τ_1) from household to a safe location (at the periphery of ESVA towards Maryland) for evacuation under baseline conditions an, and the travel time τ_2 for the inundated case. For households that are cut off by inundation and are unable to reach the safety destination, the travel time is set to a very high value $\tau_2 \to \infty$ in case the safe location is not reachable. We then calculate the **performance ratio** $0 < \tau_1/\tau_2 \leq 1$. The closer the metric value to 1, the better.
- **Travel time for emergency personnel:** This measure considers travel time from the nearest fire station to the household in the same manner as for evacuation.
- **Travel time to critical service:** This measure considers travel time to the nearest hospital from the household under two road conditions.

Figure 2 shows the travel times from emergency personnel for the households in each income group that are affected by flooding (i.e., the ones for whom performance ratio < 1). We see that the change for each group is about the same in the random assignment but not for the other two. Table 1 shows the number of households in the three income level groups, affected by road inundation in terms of the three metrics, for different assignment method. For the IP assignment, we

Income Level	Travel time to Safety	Travel time from Emergency Personnel	Travel time to Critical Service
IP assignment			
Low	1927, 17%	2104, 18%	1871, 16%
Mid	1160, 21%	1351, 24%	1461, 26%
High	499, 29%	600, 35%	601, 35%
Sorted assignment			
Low	1960, 17%	2154, 19%	1900, 17%
Mid	1166, 21%	1371, 25%	1499, 27%
High	496, 29%	599, 35%	582, 34%
Random assignment			
Low	2767, 24%	3068, 27%	2981, 26%
Mid	1472, 27%	1351, 24%	1399, 25%
High	430, 25%	480, 28%	469, 27%

Table 1: Number and percentage of households affected by road inundation in each income level, for the IP assignment method.

observe that, by count, low income level households are affected the most. However, if we consider percentage of households, then high income households are affected the most. A similar pattern is observed for Sorted assignment method. However, in the Random assignment method, we observe that the percentage of households impacted in different income levels are close, i.e. $24 - 28\%$.

To understand if the effect of road inundation on different income level households is different, we look at the performance ratio values of the set of households in each income level. We then compare them using the two-sample Kolmogorov-Smirnov (KS) test. The resulting p-values are shown in Table 2. For the IP assignment method, we observe that the p-values are small (i.e. < 0.05) for all three metrics. This implies that the difference in impact of road inundation on different income level households is statistically significant. Similar result is found for the sorted assignment method. However, for the random assignment method, we see that the p-values are large (> 0.05). This implies that the difference in impact of road inundation on different income level households is not statistically significant. This is expected as the households were assigned to residences uniformly at random.

5 Discussion

The IP implementation was done using Gurobi. It is more computationally expensive than doing a random matching or a maximally correlated matching

Income Level	Travel time to Safety	Travel time from Emergency Personnel	Travel time to Critical Service
IP assignment			
Low vs Mid	3.03e-07	1.62e-15	1.42e-40
Mid vs High	2.1e-07	2.46e-12	1.69e-08
Low vs High	5.167e-19	1e-34	1.11e-46
Sorted assignment			
Low vs Mid	1.08e-06	1.92e-14	2.7e-43
Mid vs High	5.77e-07	2.25e-11	4.65e-07
Low vs High	1.74e-17	1.73e-32	6.28e-40
Random assignment			
Low vs Mid	0.65	1	0.7
Mid vs High	0.79	0.98	0.55
Low vs High	1	1	0.76

Table 2: p-values from two-sample Kolmogorov-Smirnov test. For IP and Sorted assignment, the p-values are small (< 0.05), implying statistically significant difference between the income level groups. For Random assignment, p-values are large (> 0.05), implying no statistically significant difference between the groups.

(which can be done by sorting and matching). However, as this method is run independently for each blockgroup, it can be parallelized easily. The method is also agnostic to the attributes, so it can be used wherever we have beliefs about the correlation between attributes, but lack data.

Empirical studies have consistently shown disparities by socioeconomic class, race, and ethnicity, in risks due to flooding and other environmental hazards. This is an important area of research in multiple domains. The use of digital similars in these contexts is helpful in the evaluation of detailed and geographically contingent policies and procedures for mitigating these risks. Our methodology in this work brings additional veridicality to these efforts.

Data availability

The digital similar that formed the basis for this work, albeit without the IP implementation, is available as "Virginia: DP-VA-2.4.0" at `https://doi.org/10.18130/V3/5LSDCY`.

Acknowledgments. This work was funded in part by CoPe: NSF Award 2053013: Focused CoPe: Building Capacity for Adaptation in Rural Coastal Communities, by

NASA Applied Sciences Program Grant #80NSSC22K1048, by the AI Research Institutes program supported by NSF and USDA-NIFA under the AI Institute: Agricultural AI for Transforming Workforce and Decision Support (AgAID) award No. 2021-67021-35344.

References

1. Adiga, A., Agashe, A., Arifuzzaman, S., Barrett, C.L., Beckman, R.J., Bisset, K.R., Chen, J., Chungbaek, Y., Eubank, S.G., Gupta, S., Khan, M., Kuhlman, C.J., Lofgren, E., Lewis, B.L., Marathe, A., Marathe, M.V., Mortveit, H.S., Nordberg, E., Rivers, C., Stretz, P., Swarup, S., Wilson, A., Xie, D.: Generating a synthetic population of the United States. Tech. Rep. NDSSL 15-009, Network Dynamics and Simulation Science Laboratory (2015), https://drive.google.com/file/d/1S8Z3sqCMxBGBB7WbNoPHy7ff7NtJo6GR/view?usp=drive_link
2. Aleta, A., Martín-Corral, D., Pastore y Piontti, A., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N.E., Halloran, M.E., Longini Jr, I.M., Merler, S., Pentland, A., Vespignani, A., Moro, E., Moreno, Y.: Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. Nature Human Behaviour **4**(9), 964–971 (Sep 2020). https://doi.org/10.1038/s41562-020-0931-9
3. Bhattacharya, P., Chen, J., Hoops, S., Dustin, M., Lewis, B., Venkatramanan, S., Wilson, M.L., Klahn, B., Adiga, A., Hurt, B., Outten, J., Adiga, A., Warren, A., Baek, H., Porebski, P., Marathe, A., Xie, D., Swarup, S., Vullikanti, A., Mortveit, H., Eubank, S., Barrett, C.L., Marathe, M.: Data-driven scalable pipeline using national agent-based models for real-time pandemic response and decision support. International Journal of High Performance Comput. Appl. **37**(1), 4–27 (2023)
4. BuildingFootprintUSA: (2019), https://www.buildingfootprintusa.com/, last accessed 15 September 2019
5. van Dam, K.H., Bustos-Turu, G., Shah, N.: A methodology for simulating synthetic populations for the analysis of socio-technical infrastructures. In: Jager, W., Verbrugge, R., Flache, A., de Roo, G., Hoogduin, L., Hemelrijk, C. (eds.) Advances in Social Simulation 2015. pp. 429–434. Springer, Cham (2017)
6. Duan, R., Pettie, S.: Linear-time approximation for maximum weight matching. J. ACM **61**(1), 1:1–1:23 (2014)
7. Edmonds, J.: Paths, trees and flowers. Canadian Journal of Mathematics **17**, 449–467 (1965)
8. Gabow, H.N.: A scaling algorithm for weighted matching on general graphs. In: Proc. 26th IEEE Symposium on Foundations of Computer Science (FOCS). pp. 90–100 (1985)
9. Gallagher, S., Richardson, L.F., Ventura, S.L., Eddy, W.F.: SPEW: Synthetic populations and ecosystems of the world. Journal of Computational and Graphical Statistics **27**(4), 773–784 (2018). https://doi.org/10.1080/10618600.2018.1442342
10. Geographic Information Science & Technology, Oak Ridge National Laboratory: Landscan, https://landscan.ornl.gov/
11. Gridded Population of the World (GPW), v4, https://sedac.ciesin.columbia.edu/data/collection/gpw-v4
12. He, B.Y., Zhou, J., Ma, Z., Chow, J.Y., Ozbay, K.: Evaluation of city-scale built environment policies in New York City with an emerging-mobility-accessible synthetic population. Transportation Research Part A: Policy and Practice **141**, 444 – 467 (2020)

13. HERE Premium Streets Data set for the U.S. (2020), https://www.here.com/
14. IDM: Synthpops (2025), https://docs.idmod.org/projects/synthpops/en/latest/
15. Islam, K.A., Chen, D.Q., Marathe, M., Mortveit, H., Swarup, S., Vullikanti, A.: Incorporating fairness in large-scale evacuation planning. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. p. 3192–3201. New York, NY, USA (2022)
16. Islam, K.A., Marathe, M., Mortveit, H., Swarup, S., Vullikanti, A.: A simulation-based approach for large-scale evacuation planning. In: IEEE International Conference on Big Data, 2020. pp. 1338–1345 (2020)
17. Lum, K., Chungbaek, Y., Eubank, S.G., Marathe, M.V.: A two-stage, fitted values approach to activity matching. International Journal of Transportation **4**(1), 41–56 (2016)
18. Maalouly, N.E.: Exact matching: Algorithms and related problems. In: Proc. Symposium on Theoretical Computer Science (STACS). pp. 36:1–36:24 (2023)
19. Marathe, M., Mortveit, H., Parikh, N., Swarup, S.: Prescriptive analytics using synthetic information. In: Hsu, W.H. (ed.) Emerging Trends in Predictive Analytics: Risk Management and Decision Making, pp. 1–19. IGI Global, Hershey, PA (2014)
20. Microsoft: U.S. building footprints (2018), https://github.com/Microsoft/USBuildingFootprints
21. Mistry, D., Litvinova, M., Pastore y Piontti, A., Chinazzi, M., Fumanelli, L., Gomes, M.F.C., Haque, S.A., Liu, Q.H., Mu, K., Xiong, X., Halloran, M.E., Longini, I.M., Merler, S., Ajelli, M., Vespignani, A.: Inferring high-resolution human mixing patterns for disease modeling. Nature Communications **12**(1), 323 (2021)
22. National Center for Education Statistics (NCES), T.: http://nces.ed.gov, last accessed: February 2020
23. Nejad, M.M., Erdogan, S., Cirillo, C.: A statistical approach to small area synthetic population generation as a basis for carless evacuation planning. Journal of Transport Geography **90**, 102902 (2021)
24. OpenStreetMap points of interest, https://www.openstreetmap.org/, last accessed: 7 Feb 2021
25. Renardy, M., Eisenberg, M., Kirschner, D.: Predicting the second wave of COVID-19 in Washtenaw County, MI. Journal of Theoretical Biology **507**, 110461 (2020)
26. Tatem, A.: WorldPop, open data for spatial demography. Scientific Data **4** (2017)
27. Thorve, S., Baek, Y.Y., Swarup, S., Mortveit, H., Marathe, A., Vullikanti, A., Marathe, M.: High resolution synthetic residential energy use profiles for the United States. Scientific Data **10**, Article number 76 (2023)
28. US Census: Public use microdata sample (pums), https://www.census.gov/programs-surveys/acs/microdata.html, last accessed: 24 May 2021
29. Virginia Institute of Marine Sciences: Tidewatch (2025), https://cmap2.vims.edu/SCHISM/TidewatchViewer.html
30. West, D.B.: Introduction to Graph Theory. Prentice-Hall, Englewood Cliffs, NJ (2001)
31. Wheaton, W.D., Cajka, J.C., Chasteen, B.M., Wagener, D.K., Cooley, P.C., Ganapathi, L., Roberts, D.J., Allpress, J.L.: Synthesized population databases: A US geospatial database for agent-based models. Tech. Rep. MR-0010-0905, RTI International, Research Triangle Park, NC, USA (2009)