Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

DEVELOPMENT OF LARGE-SCALE SYNTHETIC POPULATION TO SIMULATE COVID-19 TRANSMISSION AND RESPONSE

Chaitanya Kaligotla
Jonathan Ozik
Nicholson Collier
Charles Macal

Argonne National Laboratory
9700 S Cass Avenue
Lemont, IL 60439, USA

Abby Stevens
Bogdan Mucenic
Anna Hotton
Kyoung Whan Choe

University of Chicago
5801 South Ellis Avenue
Chicago, IL 60637, USA

ABSTRACT

This research describes the development of city to multi-county scale synthetic populations for application to an agent-based model (CityCOVID) that simulates the endogenous transmission of COVID-19 and measures the impact of public health interventions.

1 DATA AND METHODS TO CONSTRUCT SYNTHETIC POPULATIONS

CityCOVID is a large-scale (city to multi-county) agent-based model (ABM). Individuals represented as synthetic "agents" have a set of socio-demographic characteristics, behaviors (hourly activity schedules), and places they go to (according to their schedules). Agents also react to disease symptoms and non-pharmaceutical interventions (NPIs). CityCOVID recreates the dynamics of disease spread through the entire population. Each simulation scenario is based on a set of assumptions (informed by data and literature, and updated regularly) concerning NPIs and agent behaviors in response to the interventions.

The synthetic environment in CityCOVID consists of 3 elements - a population of synthetic agents $\{\mathbf{P}\}$, activity schedules $\{\mathbf{A}\}$, and places $\{\mathbf{L}\}$, constructed from various data sources. Together, they represent *in silico* a statistically representative population of the area under study. The development of the synthetic population described here is an expansion of Macal et al. (2018), which has been used to study disease transmission and other social processes (Macal et al. 2014; Kaligotla et al. 2018; Ozik et al. 2018). Our contribution here is in combining new and diverse data sets to extend the scale, fidelity, and granularity of synthetic populations.

Agents: Synthetic agents matching the socio-demographic characteristics of Chicago, Cook County, and a 7-County region in North Eastern Illinois (NE IL) are built from RTI synthetic household population databases (Cajka et al. 2010). Each agent is represented by $p[i_1,\ldots,i_n] \in \mathbf{P}$, where $i_1,\ldots,i_n$ is a vector of characteristics (e.g., age, gender, race, household or group quarter location). These characteristics statistically correspond to *Public Use Microdata* and census aggregated data at a census block group level. Agents are also associated with geolocated households, schools, and workplaces. The three versions of our synthetic population (Chicago, Cook County, and 7-County NE IL) have 2.72M, 5.16M, and 8.54M synthetic agents respectively.

Activity Schedules: Agents' daily activity schedules are constructed from two comprehensive activity surveys of the general US population – the American Time Use Survey (ATUS) and the Panel Study of Income Dynamics (PSID). Specifically, we use ATUS 2018 (https://www.bls.gov/tus/datafiles-2018.htm) for agents aged ≥ 18, and PSID data for < 18 population (https://simba.isr.umich.edu). Each activity schedule $a_j[\texttt{start},\texttt{end},\texttt{loc\_type}] \in \mathbf{A}$ contains a list of activities j for unique schedule a that includes a vector of activity start time, end time, and location type. In all, we construct $12,419$ unique schedules.

Places: Places consist of geolocated points such as households, schools, workplaces, hospitals, and general quarters (e.g., nursing homes, dormitories, jails), where the synthetic agents co-locate based on their activity schedules and associated locations. Places are constructed from the RTI database and Safegraph data (https://www.safegraph.com/). Our synthetic population includes 1.2M (Chicago), 3.49M (Cook County) and 5.15M (7-county) unique places.

2 ALGORITHMS AND WORKFLOW

We use statistical techniques, including iterative proportional fitting (Gange 1995) and multinomial sampling, to construct each synthetic person in the population while maintaining the accuracy of marginal and conditional distributions across census and survey data sources. A selection of the algorithms we use are described below:

We define an activity mapping function $f : \mathbf{A} \rightarrow \mathbf{P}$, where each agent in $\mathbf{P}$ is assigned multiple activity schedules $\mathbf{A}$, by corresponding age, gender, and race, for weekdays and weekends. For every simulated day of a CityCOVID simulation, each person is randomly assigned one of these schedules. A location matching function is defined $f : \mathbf{L} \rightarrow \mathbf{P}$, where each agent in $\mathbf{P}$ is assigned specific locations in $\mathbf{L}$ for each unique place type (e.g., restaurant, hospital, recreation) that are not preassigned in the RTI data (as in case of work and school locations). Our matching algorithm follows a location-weighted assignment (locations are assigned based on an agent's household or work location). CityCOVID is an MPI-distributed ABM (Collier et al. 2015). We apply a *Load Balancing* algorithm $f : \mathbf{L}, \mathbf{A}, \mathbf{P} \rightarrow \mathbf{L}$, where each unique place in $\mathbf{L}$ is assigned to a specific computing rank. Rank assignments are based on both distributing across computing processes and minimizing interprocess communication. During a simulated day in CityCOVID, agents move from place-to-place, hour-by-hour, engaging in interactions with other co-located agents. Simulated COVID-19 transmission occurs through these co-locations generated contact network. We produce summary analytics of our integrated *in silico* population, depicting agent distributions by social characteristics, activity types, and locations, to verify the statistical representations.

ACKNOWLEDGEMENT

Research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. This research was completed with resources provided by the Argonne Leadership Computing Facility and the Laboratory Computing Resource Center at Argonne National Laboratory. We also thank Safegraph for providing data.

REFERENCES

Cajka, J. C., P. C. Cooley, and W. D. Wheaton. 2010. "Attribute Assignment to a Synthetic Population in Support of Agent-Based Disease Modeling". Technical report, RTI Press publication No. MR-0019-1009, Research Triangle Park, North Carolina.

Collier, N., J. Ozik, and C. M. Macal. 2015. "Large-Scale Agent-Based Modeling with Repast HPC: A Case Study in Parallelizing an Agent-Based Model". In *Euro-Par 2015: Parallel Processing Workshops*, Number 9523 in Lecture Notes in Computer Science, 454–465. Vienna, Austria: Springer International Publishing.

Gange, S. J. 1995. "Generating Multivariate Categorical Variates using the Iterative Proportional Fitting Algorithm". *The American Statistician* 49(2):134–138.

Kaligotla, C., J. Ozik, N. Collier, C. M. Macal, S. Lindau, E. Abramsohn, and E. Huang. 2018. "Modeling an Information-Based Community Health Intervention on the South Side of Chicago". In *Proceedings of the 2018 Winter Simulation Conference*, 2600–2611. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Macal, C. M., N. T. Collier, J. Ozik, E. R. Tatara, and J. T. Murphy. 2018. "ChiSIM: An Agent-Based Simulation Model of Social Interactions in a Large Urban Area". In *Proceedings of the 2018 Winter Simulation Conference*, 810–820. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Macal, C. M., M. J. North, N. Collier, V. M. Dukic, D. T. Wegener, M. Z. David, R. S. Daum, P. Schumm, J. A. Evans, J. R. Wilder, S. J. Eells, and D. S. Lauderdale. 2014. "Modeling the Transmission of Community-Associated Methicillin-Resistant Staphylococcus Aureus: A Dynamic Agent-Based Simulation". *Journal of Translational Medicine* 12(1):1–12.

Ozik, J., N. T. Collier, J. M. Wozniak, C. M. Macal, and G. An. 2018. "Extreme-Scale Dynamic Exploration of a Distributed Agent-Based Model With the EMEWS Framework". *IEEE Transactions on Computational Social Systems* 5(3):884–895.