An Infectious Disease Spread Simulation to Control Data Bias

Ruochen Kong
Emory University
Atlanta, USA
ruochen.kong@emory.edu

Taylor Anderson
George Mason University
Fairfax, USA
tander6@gmu.edu

David Heslop
University of New South Wales
Sydney, Australia
d.heslop@unsw.edu.au

Andreas Zufle
Emory University
Atlanta, USA
azufle@emory.edu

ABSTRACT

The increased availability of datasets during the COVID-19 pandemic enabled machine-learning approaches for modeling and forecasting infectious diseases. However, such approaches are known to amplify the bias in the data they are trained on. Bias in such input data like clinical case data for COVID-19 is difficult to measure due to disparities in testing availability, reporting standards, and healthcare access among different populations and regions. Furthermore, the way such biases may propagate through the modeling pipeline to decision-making is relatively unknown. Therefore, we present a system that leverages a highly detailed agent-based model (ABM) of infectious disease spread in a city to simulate the collection of biased clinical case data where the bias is known. Our system allows users to load either a pre-selected region or select their own (using OpenStreetMap data for the environment and census data for the population), specify population and infectious disease parameters, and the degree(s) to which different populations will be overrepresented or underrepresented in the case data. In addition to the system, we provide a large number of benchmark datasets that produce case data at different levels of bias for different regions. We hope that infectious disease modelers will use these datasets to investigate how well their models are robust to data bias or whether their model is overfit to biased data.

CCS CONCEPTS

• **Computing methodologies** → **Modeling and simulation**.

KEYWORDS

Data Simulation, Infectious Disease Data, Data Bias, Bias Simulation

ACM Reference Format:

Ruochen Kong, Taylor Anderson, David Heslop, and Andreas Zufle. 2024. An Infectious Disease Spread Simulation to Control Data Bias. In *The 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3678717.3691293

1 INTRODUCTION

Recent epidemics and pandemics caused by infectious diseases such as SARS-CoV-2, monkeypox, and influenza have led to a plethora of research efforts in modeling and forecasting the spread of infectious diseases. The availability of large infectious disease datasets has enabled many data-driven models including sequential [23], graph neural network [6], density estimation [5], ensemble [16], and contrastive predictive coding [19]. A recent survey of data-driven infectious disease forecasting models can be found in [17].

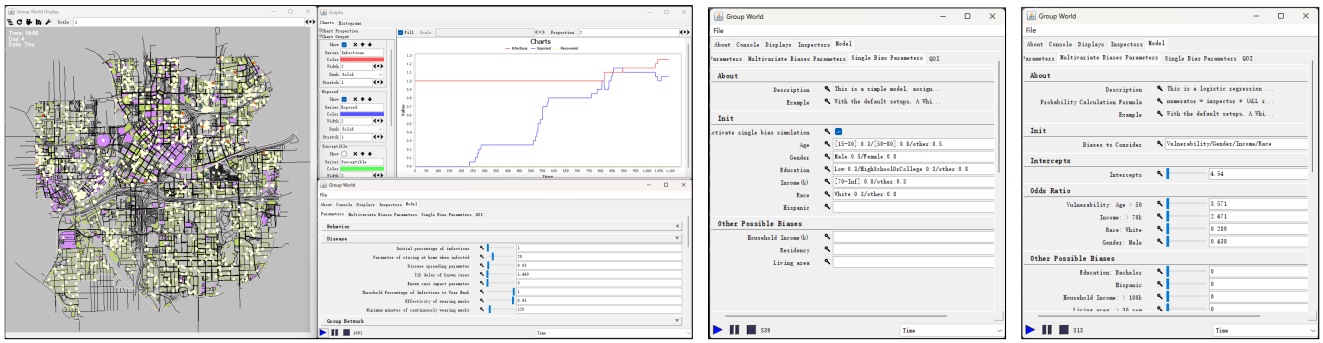
Datasets for these models come from local health organizations collecting clinical test results and voluntary positive at-home test reports. These observed datasets of cases are known to suffer from numerous types of bias across geographic regions and demographic populations [9]. For COVID-19 case data, some of this bias stems from the willingness, access, or ability of certain groups to participate in testing. Participation in testing is influenced by symptom severity [9], symptom recognition [4], occupation [20], ethnicity [7], frailty (susceptibility of more significant adverse effects) [10], place of residence [8], social connectedness [14], internet access [3] and medical/scientific interest [21]. Addressing and correcting the bias in key datasets used as inputs for disease models is essential. However, this is a challenging task. Although we know that these biases exist (as surved in [9]), we don't know the exact degrees of bias to be corrected. This raises an imminent question that has been raised in a recent vision paper [26]: To what degree does biased data yield biased infectious disease predictions?

For example, assume an outbreak of a novel infectious disease that spreads equally across all populations. For this disease, kits to test, detect, and report the disease are available but expensive. We can expect that low-income populations will be under-sampled in the observed data, but we don't know to what degree. Now assume a predictive AI model (such as those in [17]) trained on the observed data predicts a steady number of cases in affluent neighborhoods, but a low number of cases in all low-income neighborhoods. Later, as actual case numbers are observed, this model may indeed yield a smaller prediction error than classic infectious disease prediction models that aim to capture underlying processes driving the disease outcomes. But due to the data bias, these predictions, despite being close to the (biased) observed cases, may be far from the (unobservable) ground truth. In contrast, a traditional compartmental Susceptible-Infectious-Recovered (SIR) model [11] may be able to leverage epidemiological information to predict equal spread across all populations. Such predictions would be closer to the ground truth (as the equal spread assumed in the example) but would have a high error compared to the observed but bias data.

This example assumes that AI models may overfit to observable biased data and fail to model unobservable true cases. Whether this

SIGSPATIAL '24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1107-7/24/10.
https://doi.org/10.1145/3678717.3691293



(a) Simulation GUI: Spatial Network with Agents (Left), Epidemiological Curves (Top-Right), Dynamic Disease Spread Parameters (Bottom-Right)

(b) GUI to Generate Single-Parameter Bias (Section 2.3)

(c) GUI to Generate Multi-Parameter Bias (Section 2.4)

Figure 1: Screenshot of Modules of our Demonstration

overfitting really happens is hardly proven in the real-world, where we cannot observe unobservable data as we don't know what we don't know. This demonstration aims to bridge this gap using a simulated world, in which we can control the bias in the collected data and empirically evaluate the robustness of different infectious disease models to data bias. Our goal is not to provide realistic unbiased disease data but to understand the bias in silico [25] in the simulated world. This simulation builds upon the Patterns of Life simulation [12, 28], a scalable agent-based simulation of human behavior that was recently used to generate large-scale and socially plausible location-based social network data [27] and trajectory data [2]. To simulate the spread of disease and the process of generating observable biased datasets, we extended the Pattern of Life Simulation with the following features:

- An infectious disease model as detailed in Section 2.1.
- Functionality to use real-world population census data described in Section 2.2 to simulate any region in the world.
- To generate biased observations, we offer a feature to adjust the proportion of different population groups (e.g., age, gender, income) reporting their cases as described in Section 2.3.
- To understand multivariate bias that is confounded by multiple population attributes, Section 2.4 provides a logistic regression to define the probability of an agent reporting their cases based on all their population characteristic.
- We provide a demonstration of simulation and data generation for showcase at SIGSPATIAL'24 described in Section 3. The implementation details can be found on our Github [18].

2 FRAMEWORK

The Patterns of Life simulator is an open-source software written in Java that mimics human behavior by simulating Maslowian [15] needs of individual agents. Agents need to go home to satisfy their Shelter Need, agents need to go home or to a restaurant to eat and satisfy their Food Need, agents need to go to work to satisfy their Financial Needs, an agents need to meet and interact with other agents to satisfy their Love Need. Based on their needs, agents trigger actions and plan their activities following the Theory of Planned Behavior [1]. A detailed simulation description is found in [28]. This section describes how we extend the Patterns of Life simulation to 1) simulate the spread of an infectious disease among the agents, 2) use real-world census data to sample realistic agent attributes, and 3) generate biased infectious disease case data by controlling the rates at which different populations report cases. For all following simulation runs, we used a population of 5,000 agents for 30 simulation days. Each simulation took approximately 5 hours running single-threaded on an Intel NUC using an Intel i5-1135G7 CPU with 2.40GHz.

2.1 Infectious Disease Model

We implemented the spread of an arbitrary infectious disease in the ABM following an SEIR model [11]. In the SEIR model, an agent is initially Susceptible (S) and can be infected by Infectious agents. Then after being infected, the agent immediately becomes Exposed (E) and is unable to spread the disease or be infected again by another source, but will become Infectious after d_E simulation days. Once Infectious (I), the agent may spread the disease to other agents at the same physical location with an infection probability p_I. An Infectious agent will stay at home for a number of d_{home} days. The infectious stage lasts d_I days after which the agent becomes Recovered (R) who will be immune to the disease for d_R days. The simulator will collect when, where, and by which agent, an agent is infected, and write into the output file in chronic order.

The functionalities of the disease model are presented with the Graphical User Interface (GUI) shown in Figure 1a. The left part shows the environment (roads and buildings). Figure 1a shows the environment for Downtown Atlanta, GA, USA, but users can obtain the environment data from OpenStreetMap as described in [13]. Agents are color-coded by their disease status. The percentages of agents in each disease status (S,E,I,R) over time are plotted on the top right window of Figure 1a. The parameters of the infectious disease, such as d_E, p_I, d_{home}, d_I, and d_R can be changed at simulation runtime using the GUI elements shown at the bottom right window.

2.2 Agent Generation with Census data

Instead of uniformly assigning attributes of agents and locating them over the map, agent generation is informed by real census data. We first split the entire map used in the simulation into census regions such as census tracts for the United States or Townships in China. Based on the census population of each census region, we create a stratified population sample to ensure that regions having a large population in the real world have a large population in the simulation. For each generated agent, we calculate the distribution of agent attributes based on census distributions. For example, Figures 2a)-d) depict the map of Atlanta downtown, divided into census tracts. Figures 2a and 2d show that the age and gender distributions in this area are quite uniform but Figures 2b and 2c show a trend of increasing income and White population from East to

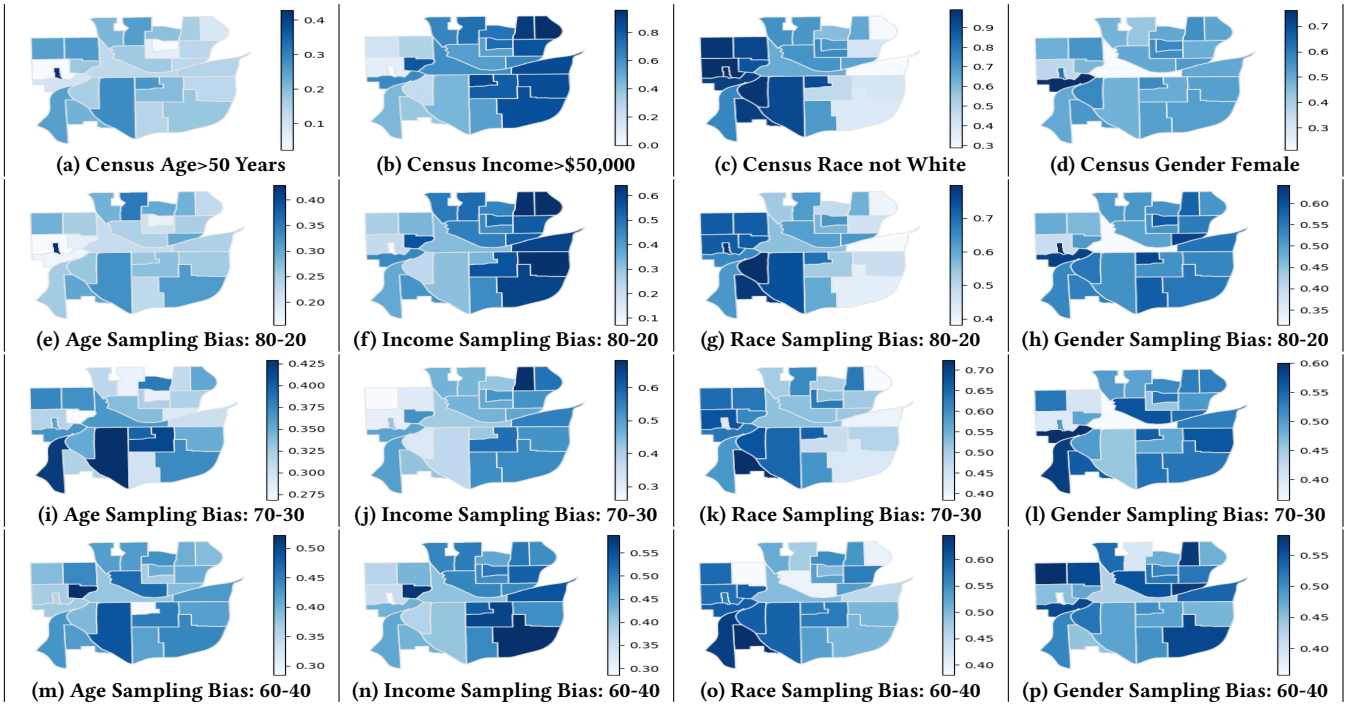


Figure 2: Visualization of sampling rates in the Atlanta Downtown study region. Figures a)-d) show census data. Figures e)-t) show biased sampling rates of a simulated infectious disease using respective sampling rates. For example, Figure e) shows resulting sampling rates for agents aged 50 or older having a probability of 80% and a probability of 20% for other agents.

West. We also observe a small census tract to the West that appears to have an unusual high aged population with extremely low income. This is explained by the low population (and thus, high variance of the mean) of this census as shown in Figure 3a which shows the census population for each region. In addition to the attributes depicted in Figure 2 our simulation also supports any other attributes provided by census data, such as education level, annual household income, or residency status as used in Tianhe, Guangzhou, China. Technically, any attributes with census data could be considered by slightly modifying the source files. Census data and corresponding shapefiles [22] across the United States is available online and instructions on how to obtain the data for a new study region (for any place in the United States) is described in our Github repository [18]. We note that, as census data provides attribute distributions of agents independently for each attribute, it is possible that the simulation may generate a 16-year-old agent having a graduate degree earning $300k a year.

2.3 Data Generation with Single-Parameter Bias

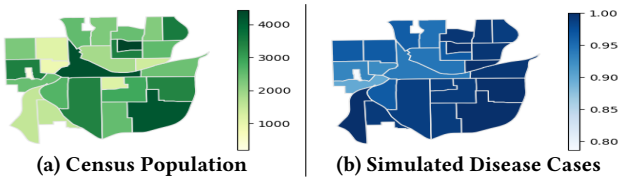
The simulation GUI, as in Figure 1b, provides the capability to inject different biases into the case dataset, simulating reporting bias based on population demographics. Users may select an attribute (e.g. age) and inject a bias within it, where some groups (e.g. ages 15-30) are underrepresented or overrepresented in the case data. In the GUI, for each attribute, multiple key-value pairs can be assigned. The key of a pair can be a consecutive range of value (e.g., [15-30] for age), a specific characteristic (e.g., "White" for race), or "other". The value of the pair should be a positive real value in $[0, 1]$ representing the corresponding reporting rate. Following the previous example of "Age", the line `[15-30]:0.3/[50-80]:0.8/other:0.5` implies that agents between the ages of 15 and 30 will have a 30% chance of reporting, agents between the ages of 50 and 80 will have an 80% chance, and all others will have 50% chance.

The system allows to generate different types and degrees of biased data as we exemplary show for the Atlanta region in Figures 2e-2p. In this example, each attribute is binary: age over 50 years, annual income exceeding $70,000, belonging to a non-white racial group, and being female. We generated biased observational datasets using three scenarios having a reporting rate of 0.8/0.7/0.6 for the corresponding group and 0.2/0.3/0.4 for all other groups. We observe that, depending on the data collection bias, the corresponding infectious disease cases become similar to the corresponding census distributions. To understand the bias in these datasets, we can compare these datasets to the corresponding simulated "Ground Truth"[1] of case rates shown in Figure 3b. As we see, the simulated disease was oblivious of population characteristics and affected all populations equally (subject to random variance due to agents chance of coming into contact with an infected agent). We can see that the Ground Truth of case rates differs drastically from the bases observed in the biased datasets shown in Figures 2e-2p. We share all datasets used to create these figures in our GitHub repository [18].

2.4 Data Generation with Multivariate Biases

We also implemented a model considering multivariate biases by calculating the reporting rate with a logistic regression model. Due to the current lack of data about the relationship between human characteristics and chances of reporting, we used a model originally fitted to predict mask usage in the United States [24] and assume that wearing a mask is a proxy for self-reporting of positive tests.

[1] Ground Truth in the simulated world, not the real-world.



(a) Census Population (b) Simulated Disease Cases

Figure 3: Choropleth maps of the census population numbers and the simulated infectious diseases cases.

The original model takes the input of five binary variables, where 'true' indicates the following: an individual's perceived vulnerability is greater than 50%, their annual income is greater than 70k, their race is white, their political leaning is Democratic, and their gender is male. The model from Von Hoene et al. [24], trained on individual-level survey data, learns the odds that individuals with these different attributes will report their positive test. We modified the model slightly, assuming an agent age over the age of 50 has a "perceived vulnerability > 50" (true), and ignoring the political variable, as these features are not available in the simulation. The model is used to predict the reporting rate of each agent, stored in file **bias.properties** in our system. These parameters are also editable for future research. Figure 1c presents the settings of this model that the user can interact with. The considered attributes are specified under the "Init" tag, separated by the symbol "/". The intercepts and the odds ratios can be edited. With this default setup, a 16-year-old white male with a 20k income would have a reporting probability of $\frac{4.54\times(3.571^0\times 2.471^0\times 0.289^1\times 0.438^1)}{1+4.54\times(3.571^0\times 2.471^0\times 0.289^1\times 0.438^1)} \approx 0.365$. Results of this model applied to the Atlanta, GA, USA population are omitted for brevity but can be found on our GitHub [18].

3 DEMONSTRATION

We will present the demonstration for SIGSPATIAL'24 with the following functionalities. The SIGSPATIAL audience may change parameters and run the simulation, single bias, and multivariate bias models on the Atlanta downtown map. By clicking on the run bottom, agents shown on the map start to move, and the epidemiological curves are calculated as shown in Figure 1a. During the agent generation step, the attributes of each agent and the distribution of attributes of each region are shown on the terminal. While the simulation runs, data is continuously collected including: The file `DiseaseReports.tsv` collects all epidemiologic data including information about whether an agent is included in the (biased) data sampling. The file `patterns_of_lifes.log` stores general logging information, and the file `AgentCharacteristics.tsv` contains all agent attributes including age, income, race, gender, and their chance of reporting their infection. During our demonstration, we will show how to switch to other maps such as San Fransisco, and the Guangzhou Tianhe District, and how to control the parameters for reporting rates. We will also present the way of running the simulation without GUI for data generation. Lastly, we will show the generated datasets and reproduce the corresponding bias plots such as shown in Figure 2.

We hope that researchers and health professionals may find these datasets useful to evaluate their infectious disease spread prediction models to understand the robustness of their models to data bias. We also provide instructions to help users to simulate their own study regions using globally available OpenStreetMap data and location-specific census data.

REFERENCES

[1] Icek Ajzen. 1991. The Theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.

[2] Hossein Amiri, Shiyang Ruan, Joon-Seok Kim, Hyunjee Jin, Hamdi Kavak, Andrew Crooks, Dieter Pfoser, Carola Wenk, and Andreas Züfle. 2023. Massive Trajectory Data Based on Patterns of Life. In *ACM SIGSPATIAL*.

[3] Christopher Antoun, Chan Zhang, et al. 2016. Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field methods* 28, 3 (2016), 231–246.

[4] Pierre-Yves Boëlle, Cécile Souty, Titouan Launay, et al. 2020. Excess cases of influenza-like illnesses synchronous with coronavirus disease (COVID-19) epidemic, France, March 2020. *Eurosurveillance* 25, 14 (2020), 2000326.

[5] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. 2018. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology* 14, 6 (2018), e1006134.

[6] Songgaojun Deng, Shusen Wang, et al. 2020. Cola-GNN: Cross-location attention based graph neural networks for long-term ILI prediction. In *CIKM*. 245–254.

[7] Catherine Dodds and Ibidun Fakoya. 2020. Covid-19: ensuring equality of access to testing for ethnic minorities.

[8] Justin Elarde, Joon-Seok Kim, Hamdi Kavak, Andreas Züfle, and Taylor Anderson. 2021. Change of human mobility during COVID-19: A United States case study. *PloS one* 16, 11 (2021), e0259031.

[9] Gareth J Griffith, Tim T Morris, Matthew J Tudball, et al. 2020. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature communications* 11, 1 (2020), 5749.

[10] Melanie Henwood. 2020. Care home deaths: The untold and largely unrecorded tragedy of COVID-19. *British Policy and Politics at LSE* (2020).

[11] William Ogilvy Kermack and Anderson G McKendrick. 1932. Contributions to the mathematical theory of epidemics. II.—The problem of endemicity. *Proceedings of the Royal Society of London. Series A* 138, 834 (1932), 55–83.

[12] J. S. Kim, H. Jin, H. Kavak, O. C. Rouly, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. 2020. Location-Based Social Network Data Generation Based on Patterns of Life. In *MDM*. 158–167. https://doi.org/10.1109/MDM48529.2020.00038

[13] Will Kohn, Hossein Amiri, and Andreas Züfle. 2023. EPIPOL: An Epidemiological Patterns of Life Simulation (Demonstration Paper). In *4th ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology*. 13–16.

[14] T Kuchler, D Russel, and J Stroebel. 2020. *The Geographic Spread of COVID-19 Correlates with Structure of Social Networks as Measured by Facebook (2020)*. Technical Report. CESifo Working Paper.

[15] Abraham H Maslow. 1943. A theory of human motivation. *Psychological review* 50, 4 (1943), 370.

[16] Nicholas G Reich, Logan C Brooks, Spencer J Fox, et al. 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences* 116, 8 (2019), 3146–3154.

[17] Alexander Rodríguez, Harshavardhan Kamarthi, Pulak Agarwal, Javen Ho, Mira Patel, Suchet Sapre, and B Aditya Prakash. 2022. Data-centric epidemic forecasting: A survey. *arXiv preprint arXiv:2207.09370* (2022).

[18] Ruochen Kong. Accessed 04/04/2024. Source Code, Data, and Supplemental for This Submission (https://github.com/RuochenKong/disease-simulator).

[19] Anish Susarla, Austin Liu, Duy Hoang Thai, Minh Tri Le, and Andreas Züfle. 2022. Spatiotemporal Disease Case Prediction Using Contrastive Predictive Coding. In *3rd ACM SIGSPATIAL Workshop on Spatial Computing for Epidemiology*.

[20] Alma Tostmann, John Bradley, et al. 2020. Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. *Eurosurveillance* 25, 16 (2020), 2000508.

[21] Jessica Tyrrell, Jie Zheng, et al. 2021. Genetic predictors of participation in optional components of UK Biobank. *Nature communications* 12, 1 (2021), 886.

[22] United States Census Bureau. Accessed 04/04/2024. United States Census Data Shapefiles (https://www2.census.gov/geo/tiger/TIGER2020PL/STATE/).

[23] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one* 12, 12 (2017), e0188941.

[24] Emma Von Hoene, Amira Roess, Shivani Achuthan, and Taylor Anderson. 2023. A Framework for Simulating Emergent Health Behaviors in Spatial Agent-Based Models of Disease Spread. In *ACM SIGSPATIAL GeoSim Workshop*. 1–9.

[25] Andreas Züfle, Dieter Pfoser, Carola Wenk, et al. 2024. In Silico Human Mobility Data Science: Leveraging Massive Simulated Mobility Data (Vision Paper). *ACM Transactions on Spatial Algorithms and Systems* 10, 2 (2024), 1–27.

[26] Andreas Züfle, Flora Salim, Taylor Anderson, et al. 2024. Leveraging Simulation Data to Understand Bias in Predictive Models of Infectious Disease Spread. *ACM Transactions on Spatial Algorithms and Systems* 10, 2 (2024), 1–22.

[27] Andreas Züfle, Goce Trajcevski, Dieter Pfoser, and Joon-Seok Kim. 2020. Managing uncertainty in evolving geo-spatial data. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 5–8.

[28] Andreas Züfle, Carola Wenk, Dieter Pfoser, Andrew Crooks, Joon-Seok Kim, Hamdi Kavak, Umar Manzoor, and Hyunjee Jin. 2023. Urban life: a model of people and places. *Computational and Mathematical Organization Theory* (2023).