GENERATING SYNTHETIC INDIVIDUAL-HUMAN POPULATION AND ACTIVITY MODELS

Emily Schmidt
CSE Department, Miami University
Oxford, OH 45056, USA.
email: schmidee@miamiOH.edu

Dhananjai M. Rao
CSE Department, Miami University
Oxford, OH 45056, USA.
email: raodm@miamiOH.edu

KEYWORDS

Demographics, Contact Networks, Individual-based Model, Model Generation, Epidemiology

ABSTRACT

Simulation-based analysis is rapidly gaining importance for developing public policy for dealing with a broad spectrum of issues ranging from natural catastrophes to seasonal communicable diseases. Simulation-based analysis requires the use of high fidelity models for conducting in-depth studies of different scenarios. However, generating valid, large-scale human models from raw demographic, geographic, and other statistical data is challenging. This paper proposes a "first principles" approach to generating models using summary statistics from authoritative sources. Our generated model is called Synthetic, Individual-human Population and Activity Model (SIPAM). SIPAM characterizes population demographics, schools, businesses, and typical daily activities of individuals at a given level of detail (currently 1 km, based on data availability). This paper discusses: ① our method for generating a SIPAM, ② the exhaustive set of verification experiments (10 million replications with ~1.04 billion activities), and ③ case studies of seasonal influenza epidemic for comprehensive validation. The experiments establish that the proposed method yields valid synthetic models that are amenable to a variety of analysis.

INTRODUCTION

Population growth and increasing urban densities pose many challenges for dealing with a broad spectrum of issues ranging from natural catastrophes, seasonal communicable diseases, to routine city planning. An example of such a issue was witnessed around the Gulf Coast of United States during the 2005 hurricane Katrina – it was tragically discovered the evacuation plans were not adequate to evacuate the impacted area (Daniels, 2007). The root issue is that responses to particular catastrophes are usually based on historical examples of similar events and not on current scenarios. Conventional statistical analyses are not conducive for "what-if" type analyses that is required for policy assessments. Furthermore, they do not yield sufficiently detailed and intuitive information about individuals in the population and their stochastic behaviors. Consequently, innovative approaches are needed to *proactively* address these growing challenges.

Simulation-based analyses are rapidly growing in importance to meet the aforementioned needs. Catalyzed by rapid advancement in computational infrastructures, simulations enable systematic and multifaceted analysis required for policy development (Giridharan and Rao, 2016). Simulations fundamentally rely availability of a valid, comprehensive, and robust model. Important model characteristics include: ❶ Realism: The model must be realistic and mirror geographic, demographic, and behavioral characteristics; ❷ Reusability & accuracy: Investments into model development and validation are effectively amortized only when models can be reused or easily adapted for different types of analyses; and ❸ Computational costs: Time and resources required for model generation, validation, simulation, and analysis need to be balanced with realism, accuracy, and effective use (Chen and Zhan, 2008).

Generating realistic, reusable, and cost effective models for comprehensive, multifaceted analysis of human populations and day-to-day activities is a daunting task (Barrett et al., 2009) due to the following diverse challenges: ① identifying data necessary for model generation, ② finding suitable data sources, ③ preprocessing and streamlining data for modeling, ④ developing and validating methods for model generation, and ⑤ verification of generated models.

Overview & advantages of proposed approach

In this paper we propose a novel method for generating realistic synthetic models of individuals in a population along with their associated daily activities. Our model is called Synthetic Individual-human Population and Activity Model (SIPAM). A SIPAM is generated from anonymous summary statistics that can be readily obtained from authoritative data sources, such as: USCB (2011), USBLS (2015), and demographic databases (CIESIN, 2016). SIPAM preserves demographic, socioeconomic, and geospatial characteristics consistent with the resolution of available data – that is, aggregate characteristics of a SIPAM are statistically indistinguishable from the census data used to create it.

The generated model includes temporospatial activities (for a full week) for each individuals based on their age and employment status (working, unemployed, retired). The activities are structured around a variety of buildings, such as: homes, schools, businesses, etc. The buildings are generated as part of the model using census data. These are key aspects of the SIPAM that ensure it is realistic. The data is supplied via a comprehensive input configuration file to enable effective reusability and extensibility. The activities are scheduled in configurable time intervals or time-blocks (currently set to 10 minute blocks) thereby striking a balance between realism versus computational costs for model generation and simulation. The models can be used for a variety of simulation-based analyses using different approaches, including: cycle-based simulations, agent-based simulations, or discrete event simulations. Our exhaustive suite of verification and validation experiments (see Section EXPERIMENTS) establish that the proposed method generates valid, realistic SIPAMs that are amenable to a variety of analyses.

BACKGROUND & RELATED WORKS

Realistic and comprehensive modeling and simulation-based analysis of specific aspects of human populations has been an active area of research in diverse fields, including: computational epidemiology, transportation, and economics. Humans are either modeled as collection of interacting individuals or groups. Group models represent a collection of collocated individuals modeled as an indivisible entity. Different types of group models have been proposed by several investigators including Balcan et al. (2010), Rao et al. (2009), and Keeling (2005). In their models, the groups are organized based on their geographic locations resulting in a logical structure similar to a Voronoi tessellation. Interactions between groups is modeled implicitly based on their adjacency or via explicit mobility networks. The benefit of using a group model is it reduces the computational cost because the model consists of fewer number of entities which significantly reduces computations preformed at each time step. Furthermore, such models do not require detailed, voluminous data about the population, which can be hard to obtain.

The primary disadvantage of aggregate or group models is that information about each individual is not preserved. The models do not preserve heterogeneity that may be present within the group. However, such information maybe vital for certain types of analyses and design of public policies. Consequently, several researchers have proposed the use of individual-based models, where each individual is independently modeled. Such models essentially embody contact-networks which define temporal interactions that occur between individuals. Longini et al. (2005) discuss the generation and use of synthetic, individual-based models for containing influenza epidemics. They use a variety of data sources to generate individuals and their temporospatial activities. Recently Bhatele et al. (2017) discuss enhancements to modeling and simulation of individual-based model of the United States generated as part of prior work by Barrett et al. (2009). Their model is fine-grained and has been generated from a variety of public and proprietary data sources.

The advantages of individual-based models are: ① they yield temporospatial characteristics for epidemics because they explicitly model the location and contacts between individuals in the populations, ② since they are typically less prescriptive, they can be used for analyzing epidemics whose parameters are not well established, and ③ they enable vivid visualization. The drawbacks of individual-based models are: ① they are computationally demanding for model generation and simulation (Bhatele et al., 2017), ② it is hard to fit such models to surveillance data due to the significant differences in resolution, and ③ the volume of data generated by the models can pose bottlenecks for analysis.

Similarities & differences to our approach

The proposed method for generating a Synthetic, Individual-human Population and Activity Model (SIPAM) is a novel combination of group models and individual-based models. The design rationale is to preserve advantages of the two types of models while minimizing their drawbacks. SIPAM is similar to individual-based models in that daily activities for each individual is explicitly generated. This enables tracking interactions between individuals for various analysis. However, spatial resolution for buildings and people is limited (currently to 1 km^2) similar to aggregate or group models. Another objective is to enable generation of SIPAMs using freely available summary data sets. This enables their ready use by the community for different applications and geographic regions.

MODEL GENERATION METHOD

The proposed method for generating a Synthetic Individual-human Population and Activity Model (SIPAM) is summarized in Figure 1. The method consists of four main phases, namely: ❶ Population/family generation, ❷ Building generation, ❸ Daily schedule generation, and ❹ Output generation. The statistical data required for SIPAM generation is supplied via a text-based configuration file. The generated SIPAM can be used for a different types of simulation-based analyses by combining it with a suitable *policy* model implemented during simulation. Policies essentially modify attributes of individuals and their daily schedules. For example, in this study an epidemic policy is used to simulate progression of seasonal influenza in the generated synthetic population. Simulation of epidemics using a synthetic model has been enabled via a cycle-based simulator called HAPLOS, developed as part of this investigation. The details of each of the four phases are described in the following subsections.

Phase 1: Population/Family Generation

The first phase in generation of a SIPAM involves creation of households of different "sizes" (*i.e.,* number of individuals), with each individual meeting the specified age and sex distributions observed in a census. In our example models, we have obtained household size, individual age, and sex distribution summaries from the USCB (2011). Census data is summarized as probabilities in the input configuration file as shown in Figure 2. The population generation process beings with creating families of different sizes

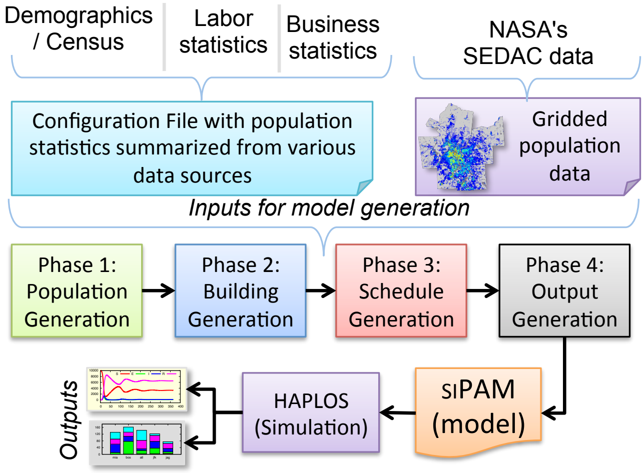
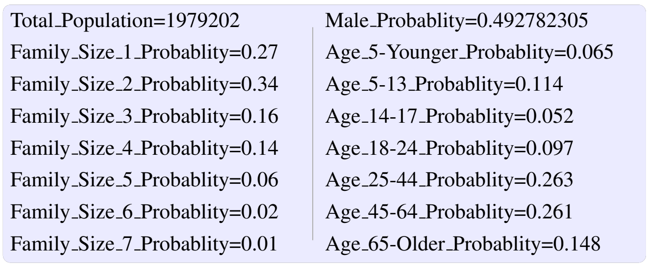

Demographics / Census
Labor statistics
Business statistics
NASA's SEDAC data
Configuration File with population statistics summarized from various data sources
Gridded population data
Inputs for model generation
Phase 1: Population Generation
Phase 2: Building Generation
Phase 3: Schedule Generation
Phase 4: Output Generation
Outputs
HAPLOS (Simulation)
SIPAM (model)

Figure 1: Overview of proposed method for generating a Synthetic Individual-human Population and Activity Model (SIPAM)



```
Total_Population=1979202              Male_Probablity=0.492782305
Family_Size_1_Probablity=0.27         Age_5-Younger_Probablity=0.065
Family_Size_2_Probablity=0.34         Age_5-13_Probablity=0.114
Family_Size_3_Probablity=0.16         Age_14-17_Probablity=0.052
Family_Size_4_Probablity=0.14         Age_18-24_Probablity=0.097
Family_Size_5_Probablity=0.06         Age_25-44_Probablity=0.263
Family_Size_6_Probablity=0.02         Age_45-64_Probablity=0.261
Family_Size_7_Probablity=0.01         Age_65-Older_Probablity=0.148
```

Figure 2: Fragment of input configuration file with demographic data of desired synthetic population

based on the probability values specified in the input configuration. A standard uniform, discrete random number generator (`std::discrete_distribution`) is used to determine family sizes. This random number generator produces integers on the interval $[0, n)$, where the probability of each individual integer i is defined as $w_i \div S$, where $S = \Sigma w_i \quad (0 \leq i \leq n)$, that is the probability of the ith integer divided by the sum of all probabilities.

Next, the given number of individuals of different ages are generated for each family. The age and gender of each person in a family is determined based on the demographic probabilities specified in the input configuration file (see Figure 2), with an added restriction that the first member of a family must be an adult. A uniform, discrete random number generator is used to determine age and sex for each person. For persons older than 64 years, the fraction of unemployed adults is used to assign a "retired" status. The family generation process is repeated until the number of persons in the model exceed the specified population.

It must be noted that currently our method does not track other demographic properties such as race, ethnicity, etc., for each person. Consequently, the aforementioned method yields households that are statistically indistinguishable from the census data. However, if additional demographic properties are desired in the generated SIPAM, then Iterative Proportional Fitting (IPF) statistical method along with Public Use Microdata Sample (PUMS) can be used (Beckman et al., 1996). PUMS essentially provides an n-dimensional table of typical family configurations and marginals. The IPF statistical procedure iteratively adjusts cell values to estimate family size and configurations such that marginal totals remain fixed (Beckman et al., 1996). The resulting fitted tables can be used for generating synthetic households.

Phase 2: Building Generation

The second phase of model generation involves creation of variety of buildings associated with different activities. Our method currently generates the following types of buildings – ① *business*: general buildings where people may visit or work, ② *medical*: hospitals where people work, visit, or are interned, ③ *school*: further subdivided into elementary, middle and high schools where people may visit, study, or work, ④ *daycare*: children are interned while adults may visit or work, ⑤ *transport hub*: used as temporary holding area for commuters using public transport, and ⑥ *building*: used a general purpose placeholder for homes, apartments, etc.

Building generation commences with assigning homes to each family generated in Phase 1. Location of homes is determined based on population counts at a given spatial resolution. Our building generation method uses gridded population counts to appropriately distribute homes and families to reflect spatial population characteristics. We use gridded ASCII data format from NASA's Socioeconomic Data and Applications Center (SEDAC) (CIESIN, 2016). SEADC provides gridded population data at a resolution of 30 arc-seconds or approximately 1 km^2 as shown in Figure 3. Furthermore, a transport hub is assigned to each grid.

In the next step, sufficient number of day care centers as well as elementary, middle, and high schools are generated to accommodate children and school-age persons generated in Phase 1. The school capacities and occurrence probabilities are supplied via configuration data as shown in Figure 4. The sizes and occurrence probabilities have been determined from summary tables published by USCB (2012). The locations of the schools are determined based on population densities in the gridded population, with higher population areas having more buildings.

Similarly businesses are generated to accommodate all working adults. The size and occurrence for businesses in the generated SIPAM is determined based on probabilities (computed from data published by USCB (2012)) specified in the input configuration data as shown in Figure 4. The businesses are also spatially distributed to various grids based on relative population densities in the grids. Moreover, as buildings are generated, references to them are stored in different datastructures to enable rapid access during schedule generation. In this context, it must be noted that unlike homes that have persons in a family assigned to them, people are not yet assigned or "mapped" businesses and schools. Instead, this mapping occurs during schedule generation in Phase 3.

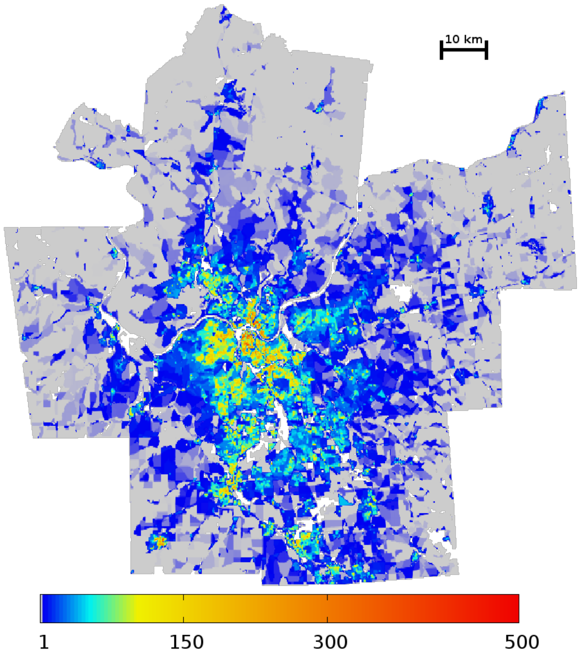


Figure 3: Example of gridded population data (persons per km^2) for city of Cincinnati, Ohio, USA (total population: ~2 million). Data freely available from CIESIN (2016)

```
Business_Size_0-4_Pr=0.478        School_Size_0-4_Pr=0.0008
Business_Size_5-9_Pr=0.135        School_Size_5-9_Pr=0.002
Business_Size_10-19_Pr=0.085      School_Size_10-19_Pr=0.005
Business_Size_20-99_Pr=0.092      School_Size_20-99_Pr=0.013
Business_Size_100-499_Pr=0.049    School_Size_100-499_Pr=0.009
Business_Size_500_Pr=0.161        School_Size_500_Pr=0.001
```

Figure 4: Fragment of input configuration file with building probabilities from USCB (2012)

Phase 3: Schedule Generation

The third phase deals with generation of weekly schedules for each individual depending on their age and employment status. A weekly schedule includes two distinct sub-schedules, namely: 5 weekday (working days) schedules and 2 weekend (non-working day) schedules. The daily schedules are rounded to 10 minute activity blocks, *i.e.,* each day is represented by 60 minutes×24 hours ÷10 = 144 blocks. The primary motivation for organizing schedules into blocks is to strike a tradeoff between model complexity, accuracy, memory, and runtime of simulations.

The schedule associated with a person is determined primarily on their age and employment status. The employment probabilities based on age have been determined based on total statistics reported by USBLS (2015). The travel distances to work and time taken to travel have been estimated based on type of transportation to work reported by USCB (2014). The probabilities and associated travel distances are specified as part of the input configuration file. Currently, we have included the following restrictions for travel: ① distance traveled by walking is limited to 2 miles, ② maximum distance traveled by public transport is limited to 10 miles, ③ school attending children are limited to travel to school in a 10 mile radius.

The schedules generated for each person is generated from one of the following templates:

- Young child template (age < 5 years): Young persons are assigned the same schedule as an adult in their family. If the adult works then the adult will be assigned to take the individual to a daycare prior to leaving to work location. The child will be assigned to the closet daycare (generated in Phase 2) to their home location that is not filled to capacity. When the adult in the family is no longer at work, they are scheduled to retrieve the child from the daycare. The child will then follow the adults schedule till the next day.
- School Age Child Template (5–17 years of age): Each person is assigned to attend a school (generated in Phase 2) during weekdays only. School is assigned based on the person's age and the nearest school to their home that provides the necessary grade level and is not at capacity. The schedule of young school children (*i.e.,* 5–13 years of age) is generated based off a designated care-giving adult in their family. However, school children older than 13 years are assigned an independent schedule with limited travel radius.
- Working adult template (>18 years): Weekday schedules are anchored around working at a business location generated in Phase 2. The distribution of working hours model the data from USBLS (2015). If the weekly work hours exceed 40, then the maximum number of hours/day is set to 10 hours. Otherwise the maximum number of works hours/day is assumed to be 8 hours. During periods when an adult is not at work, they are scheduled to visit other buildings or return home. The capacity for visitors for businesses will range between 1–500 visitors/hour depending on the number of employees. Furthermore, the schedules are further modified if the person is also designated as a child-care adult. In this scenario, the schedules are updated to include travel to-and-from daycare location.
- Non-working adult template (>18 years): The non-working adult template is meant for unemployed or retired adults (designated in Phase 1). This differs from the working adult template by removing the need of having to be at a job. Thus their schedules will be much more random in terms of locations visited during the day. The time spent at each of these locations will then be distributed through out the week randomly.

Similar to the above weekday travel patterns, weekend travel patterns are also specified via configuration file parameters. Weekend travel patterns are determined based on age of a person, with younger children spending more time at or near home. Adults are set to spend 50% of daytime traveling with a fixed radius. Currently, the schedules do not include special scenarios such as vacations, holidays, etc. Such schedules need to be suitably incorporated into the model during simulation depending on analysis needs.

Phase 4: Output generation

The previous three phases operate using in-memory datastructures to enable rapid model generation. The final phase of model generation stores the resulting model to disk. A custom textual file format has been used to store the resulting model. The format has been chosen to ease the use of the model for conducting simulations and performing various analysis. Furthermore, summary statistics on the synthetic population demographics and buildings are also generated to aid verification and validation.

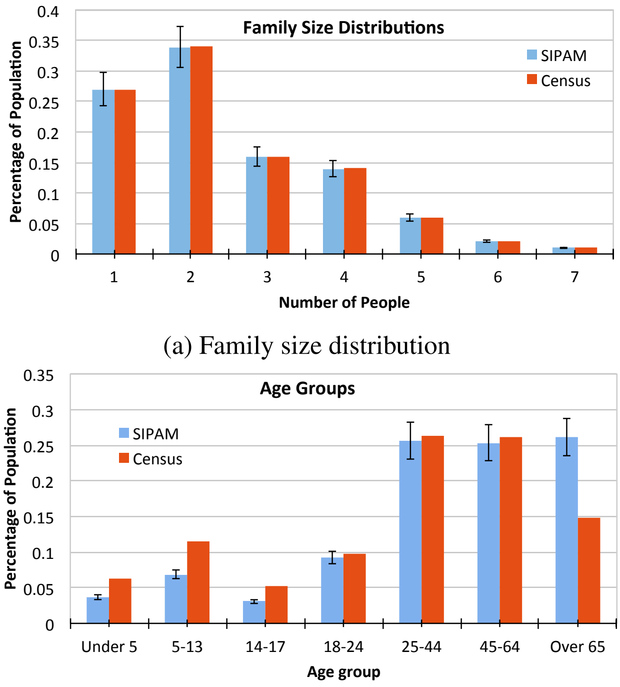
EXPERIMENTS

The proposed method for generating a Synthetic Individual-human Population and Activity Model (SIPAM) has been validated using both real-world data and a number of test data sets. The real-world dataset for validation was based on the greater metropolitan area of Cincinnati, Ohio, a typical city in mid-west United States. Figure 3 shows the gridded population data from NASA's SEDAC data set that is freely available from CIESIN (2016). The metropolitan area has a population of ~2 million with a significantly varying population spread over 1440 × 960 grid from CIESIN (2016). In addition to the large real-world data set a smaller test city with 22,000 residents spread on a 500 × 500 grid using a triangular distribution has also been used. The primary motivation of using a smaller test city was to enable extensive validation of schedule generation, which is a time consuming for large models.

Validation of the model generation method and the generated SIPAM has been conducted in two steps. First, the key characteristics of the generated model has been validated using real-world data set as discussed in the following subsection. Next, the schedule generation has been verified to ensure that valid schedules are generated in a broad range of scenarios. Importantly, the SIPAM as also been validated using an influenza epidemic scenario as discussed in subsection Full SIPAM validation

Validation of population & building generation

The core method for generating synthetic populations and buildings has been validated using Cincinnati, Ohio as the reference. Since the model generation method is probabilistic, each run will yield a slightly different SIPAM. Consequently, 10 different models were generated (from exactly the same input configuration) and collectively analyzed for validation. The chart in Figure 5(a) shows a comparison of family size distributions. The chart in Figure 5(b) shows a comparison of individuals in different age groups. The error bars in the chart show the 95% Confidence Intervals (CIs) computed from variance observed in the 10 replications. As illustrated by the charts the family size, the primary parameter, is statistically indistinguishable between the SIPAM and input census data. The age groups also show consistent distribution in most cases except for age $>$ 65 years. Our analysis suggests that the source of this discrepancy (for age $>$ 65 years) is with using a discrete distribution which rounds values more disparately.



(a) Family size distribution

(b) Population age distribution

Figure 5: Comparison of key demographics of SIPAM versus census data

The chart in Figure 6 shows a comparison of size of business (*i.e.,* number of employees) between the synthetic model and the supplied census data. The error bars in the chart show the 95% Confidence Intervals (CIs) computed from variance observed in the 10 runs. As illustrated by the charts the business sizes is statistically indistinguishable between the SIPAM and input census data. We did observe that for small models the variance in business sizes was larger. However, as the size of the model increases, the synthetic model produces statistically identical distributions. The experimental analysis establishes that the proposed method produces valid synthetic models.

Verification of schedule generation

Verification of generated synthetic schedules was conducted in two steps. First, the number of generated schedules for different categories of individuals in the SIPAM was compared with the summary census data. The chart in Figure 7 shows a comparison of different schedule types. The error bars in the chart show the 95% Confidence Intervals (CIs) computed from variance observed in 10 replications of model. Note that

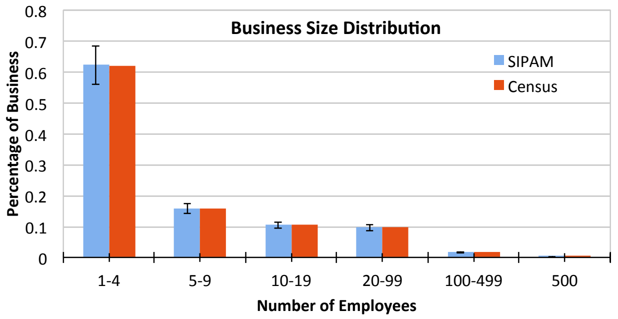


Figure 6: Comparison of business size distribution of SIPAM to census data

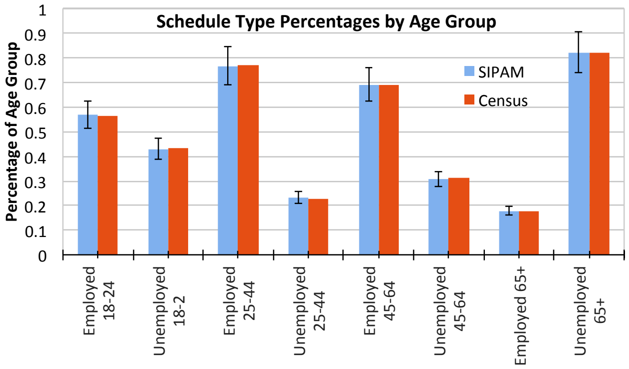


Figure 7: Comparison of different schedule types for different age groups in SIPAM to census data

variance is expected due to the stochastic nature of model and schedule generation. As illustrated by Figure 7, the distribution of different types of schedules is statistically indistinguishable between the SIPAM and input census data. This chart provides baseline verification of the proposed schedule generation method.

Next, a series of independent unit tests were developed to test the accuracy and viability of the generated schedules. The tests included checks to ensure school children are scheduled to attend school on weekdays and are not scheduled to go to work. The schedules for young children were follow the same timelines as the assigned child-care adult in the family. Checks were added to ensure employed adults are scheduled to work at least once and meet their total work hours. In additional several general checks such as being at home for 6 to 8 hours every 24 hours etc. was included in the tests. The tests also checked for consistency of schedules for members in a family.

The resulting tests covered various combinations of 16 different scenarios. The tests were used to validate schedules generated from 10 million replications with different parameter settings. The tests covered ~1.04 billion unique schedules for over 280 million families. The tests were useful to identify unique, conflicting scenarios which required inclusion of necessary logic to resolve the conflicts. With the schedule conflicts resolved, the final model generation passed all of the billion tests, thereby verifying the schedule generation method discussed in Section Phase 3: Schedule Generation

Full SIPAM validation

Having verified the proposed model generation method, we pursed validation of the complete synthetic model through case studies of an epidemic. For this experiment, we chose to model a seasonal influenza (*i.e.,* flu) epidemic and compare our results against validated results from similar models proposed by Nsoesie et al. (2012). Progression of the influenza epidemic in an individual has been characterized using classical compartmental models, with the following 4 compartments: Susceptible (S) → Exposed (E) → Infective (I) → Recovered (R). Transitions between the SEIR states is governed by probabilistic transitions. The incubation period (*i.e.,* E → I) has been set to $\{1, 2, \text{or } 3\}$ days with probabilities of $\{0.3, 0.5, \text{or } 0.2\}$ respectively. Likewise, the infection period (*i.e.,* I → R) has been set to $\{3, 4, 5, \text{or } 6\}$ days with probability $\{0.3, 0.4, 0.2, \text{or } 0.1\}$ respectively.

Infection transmission rate from an infective to a susceptible person (*i.e.,* S → E transition) has been characterized using the following equation proposed by Nsoesie et al. (2012):

$$Pr(w(i,j)) = 1 - (1 - \tau)^{w(i,j)} \qquad (1)$$

where, i denotes infected person, j denotes susceptible person, $w(i, j)$ denotes the contact time between person i and j, and τ is the disease transmission probability per unit of time. Contacts between persons arise when they are collocated in a building. As the time of collocation $w(i, j)$ increases, the probability of infection also increases.

Human Population and Location Simulator (HAPLOS)

Simulation of epidemics using a synthetic model has been enabled via sequential simulator called HAPLOS. It is a conventional cycle-based simulator in which activities of each individual are performed in each time step. Recollect that the time step of a SIPAM has been set to 10 minute increments to strike a tradeoff between model complexity, accuracy, memory, and runtime of simulations. In each time step, a person's location is suitably updated based on their generated weekly schedules. Next, epidemic progression between S → E → I → R compartments is simulated. Finally, contacts between susceptible and infective individuals collocated in each building simulated. Periodically HAPLOS saves the full state of the model to enable visualization, analysis, and validation.

Epidemic scenarios & model validation

Model validation using influenza epidemic as a case study has been conducted using HAPLOS and a SIPAM with a population of 22,000 residents spread (using a triangular distribution) on a 500 × 500 grid. The model was simulated with 5 randomly selected individuals to be exposed to the infection. The model was simulated with 5 different values of τ (see Equation 1). Results from multiple stochastic simulations for each value of τ have been averaged for analysis and

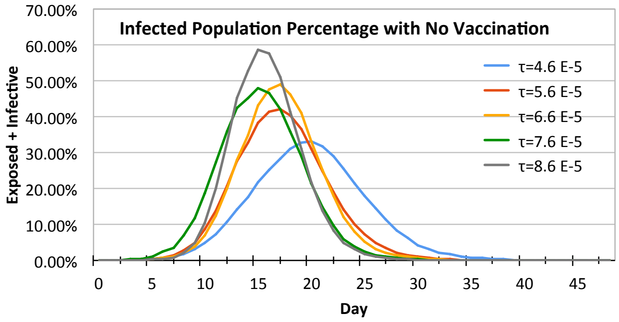


Figure 8: Fraction of exposed + infective population (averaged from multiple simulations) for different values of τ

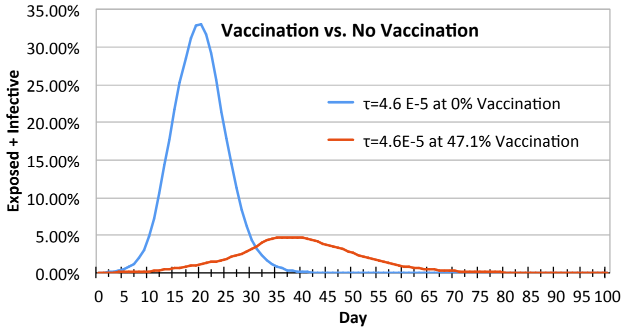


Figure 9: Comparison (Fraction of exposed + infective population) of epidemic progression with and without vaccination

plotted Figure 8. As illustrated by the chart, the epidemic curves follow the expected characteristic trend (see Nsoesie et al. (2012) for analytical details), with higher values of τ causing earlier infection peaks. Importantly, the characteristic curves establish the overall validity of the synthetic population and schedules.

Experiments were also conducted to validate initial settings and model behavior for conducting simulation-based analysis of different scenarios. Specifically, we explored the impact of vaccinating 47.1% of the population, analogous to the current vaccination rates. At this vaccination rate the epidemic is expected to have a significantly reduced peak, but with a slightly extended epidemic period (Nsoesie et al., 2012). Vaccination of population is modeled by randomly initializing 47.1% of the population to the Recovered (R) SEIR state. The chart in Figure 9 compares the fraction of infective population with and without vaccination. Consistent with expectations, vaccination decreases the peak infection while extending the epidemic period.

Summary of experimental results

The proposed method for generating a Synthetic Individual-human Population and Activity Model (SIPAM) and the generated model has been extensively validated using a variety of approaches. The experimental results in Figure 5 and Figure 6 essentially validate Phase 1 and Phase 2 of model generation that deal with generating synthetic populations, homes, schools, and businesses. Generated schedules were verified to have distribution consistent with census data as shown in Figure 7. Moreover, extended verification of schedules was conducted by generating 10 million randomized SIPAMs and verifying ~1.04 billion unique schedules. Having verified the each phase of model generation, the resulting SIPAM has been further validated using an seasonal influenza case study as discussed in Section Full SIPAM validation. The comprehensive set of verification and validation experiments establish that the proposed method generates valid synthetic models.

CONCLUSIONS

The need to analyze, design, and proactively implement sophisticated public health policies has rapidly grown due to population growth, urbanization, and emergent communicable diseases. Simulation-based methods are gaining broad applicability in this area due to their advantages. Simulations requires valid, realistic temporospatial models that effectively characterize human demographics and daily activities. The current state-of-the-art models models are broadly classified into two categories, namely: ① individual-based models in which each person and their activities are explicitly represented; and ② group-based or aggregate models in which a set of collocated humans are modeled as an indivisible entity. These methods have their respective advantages and disadvantages (see BACKGROUND & RELATED WORKS). Immaterial of the type of model being used, generating realistic, reusable, and cost effective models for comprehensive, multifaceted analysis of human populations and day-to-day activities is a challenging task Barrett et al. (2009). The challenges arise due to a myriad of issues, including: availability of data sources, computational costs, verification, and validation of complex models.

This work discussed a novel, "first-principles" method for generating a realistic Synthetic Individual-human Population and Activity Model (SIPAM). The design rationale underlying our modeling approach is to preserve advantages of both individual and group models without succumbing to their drawbacks. SIPAM includes daily activities for individuals but buildings and population densities are grouped into small regions. The size of the region is determined by resolution of input data, which is currently at 1 km^2.

Statistical analysis on generated models for various regions shows that SIPAM preserves demographic, socioeconomic, and geospatial characteristics – that is, aggregate characteristics of a SIPAM are statistically indistinguishable from the census data used to create it. The model and method have been validated using 10 million model replications with different parameter settings. The tests covered ~1.04 billion unique schedules for over 280 million families.

In addition to exhaustive verification, the model has also been validated using influenza epidemics as case studies. The case studies also explored impact of vaccination policies. Simulations for the case studies were conducted using a custom cycle-based simulator called Human Population and Location Simulator (HAPLOS). HAPLOS has been developed as part of this investigation. The experimental data shows that SIPAM faithfully reproduces epidemic characteristics consistent with other validated models. The comprehensive set of verification and validation experiments establish that the proposed method generates valid synthetic models. Our investigations also establish that SIPAM strikes an effective balance between key model characteristics discussed in Section , namely: realism, reusability & accuracy, and computational costs.

A conspicuous advantage of our method is that uses anonymous summary census statistics that can be readily obtained from authoritative data sources. Since our method relies only on small subset of census data it can be readily used for other countries and geographic regions. The resolution of the model can increased or decreased via different gridded population data represented via simple ASCII text files. The activities are scheduled in configurable time intervals or time-blocks (currently set to 10 minute blocks) thereby striking a balance between realism versus computational costs for model generation and simulation. The models can be used for a variety of simulation-based analyses using different approaches, including: cycle-based simulations, agent-based simulations, or discrete event simulations.

FUTURE WORK

This paper presented the current checkpoint in our ongoing investigations on generating realistic models. We are continuing to refine and extend our methods to incorporate more detailed real world data when available. Currently, we are investigating the use of freely available street maps to further refine locations of homes, schools, businesses, and other buildings. The use of additional demographic data from NASA's SEDAC data sets is also being explored. Furthermore, we are working on performance improvements to our model generation routines to reduce time and memory footprint. We are optimistic that with these ongoing enhancements will yield improved, realistic, cost optimal synthetic models that can be used for a broad range of analysis in a variety of fields.

ACKNOWLEDGMENTS

This work was funded in part by Miami University's Center for Analytics & Data Science (CADS).

REFERENCES

Balcan D.; Gonalves B.; Hu H.; Ramasco J.J.; Colizza V.; and Vespignani A., 2010. *Modeling the spatial spread of infectious diseases: the GLobal Epidemic and Mobility computational model*. *Journal of computational science*, 1, no. 3, 132–145. ISSN 1877-7503. doi:10.1016/j.jocs.2010.07.002.

Barrett C.L.; Beckman R.J.; Khan M.; Kumar V.A.; Marathe M.V.; Stretz P.E.; Dutta T.; and Lewis B., 2009. *Generation and analysis of large synthetic social contact networks*. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*. ISSN 0891-7736, 1003–1014. doi:10.1109/WSC.2009.5429425.

Beckman R.J.; Baggerly K.A.; and McKay M.D., 1996. *Creating synthetic baseline populations*. *Transportation Research Part A: Policy and Practice*, 30, no. 6, 415–429. ISSN 0965-8564. doi:10.1016/0965-8564(96)00004-3.

Bhatele A.; Yeom J.S.; Jain N.; Kuhlman C.J.; Livna Y.; Bisset K.R.; Kale L.V.; and Marathe M.V., 2017. *Massively Parallel Simulations of Spread of Infectious Diseases over Realistic Social Networks*. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. 689–694. doi:10.1109/CCGRID.2017.141.

Chen X. and Zhan F.B., 2008. *Agent-based modelling and simulation of urban evacuation: relative effectiveness of simultaneous and staged evacuation strategies*. *Journal of the Operational Research Society*, 59, no. 1, 25–33. ISSN 1476-9360. doi:10.1057/palgrave.jors.2602321.

CIESIN, 2016. *Gridded Population of the World, Version 4 (GPWv4): Population Count*. NASA Socioeconomic Data and Applications Center (SEDAC). doi:10.7927/H4X63JVC. Center for International Earth Science Information Network, Columbia University.

Daniels R.S., 2007. *Revitalizing emergency management after Katrina*. *Public Manager*, 36, 16–20. ISSN 2381-4160.

Giridharan N. and Rao D.M., 2016. *Eliciting Characteristics of H5N1 in High-Risk Regions Using Phylogeography and Phylodynamic Simulations*. *Computing in Science Engineering*, 18, no. 4, 11–24. ISSN 1521-9615. doi:10.1109/MCSE.2016.77.

Keeling M.J., 2005. *Models of foot-and-mouth disease*. In *Proceedings of the Royal Society B: Biological Sciences*. 1569. ISSN 0962-8452, 1195–1202. doi:10.1098/rspb.2004.3046.

Longini I.M.; Nizam A.; Xu S.; Ungchusak K.; Hanshaoworakul W.; Cummunings D.A.T.; and Halloran M.E., 2005. *Containing pandemic influenza at the source*. *Sience*, 309, no. 5737, 1083–1087. doi:10.1126/science.1115717.

Nsoesie E.O.; Beckman R.J.; and Marathe M.V., 2012. *Sensitivity Analysis of an Individual-Based Model for Simulation of Influenza Epidemics*. *PLOS ONE*, 7, no. 10, 1–16. doi:10.1371/journal.pone.0045414.

Rao D.M.; Chernyakhovsky A.; and Rao V., 2009. *Modeling and analysis of global epidemiology of avian influenza*. *Environmental Modelling & Software*, 24, no. 1, 124–134. ISSN 1364-8152. doi:10.1016/j.envsoft.2008.06.011.

USBLS, 2015. *U. S. Bureau of Labor Statistics: Employment status of the civilian noninstitutional population by age, sex, and race*. URL https://www.bls.gov/cps/cpsaat03.htm.

USCB, 2011. *U. S. Census Bureau: Statistical Abstract of the United States*. URL https://www2.census.gov/library/publications/2011/compendia/statab/131ed/tables/pop.pdf.

USCB, 2012. *U. S. Census Bureau: Number of Firms, Number of Establishments, Employment, Annual Payroll, and Estimated Receipts by Enterprise Employment Size for the United States, All Industries, Establishments the United States*. URL http://www2.census.gov/econ/susb/data/2012/us_6digitnaics_2012.xls.

USCB, 2014. *U. S. Census Bureau: Means of transportation to work*. URL http://factfinder.census.gov/bkmk/table/1.0/en/ACS/14_1YR/B08301.