Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo and M. Loper, eds.

GENERATING SYNTHETIC POPULATIONS BASED ON GERMAN CENSUS DATA

Johannes Ponge
Malte Enbergs
Michael Schüngel
Bernd Hellingrath

Department of Information Systems
University of Münster
Leonardo-Campus 3
48149 Münster, GERMANY

André Karch

Institute Epidemiology & Social Medicine
University of Münster
Domagkstraße 3
48149 Münster, GERMANY

Stephan Ludwig

Institute of Virology
University of Münster
Von-Esmarch-Straße 56
48149 Münster, GERMANY

ABSTRACT

Spatial agent-based simulations of infectious disease epidemics require a high-resolution regional population model. However, only aggregated demographic data is available for most geographic regions. Furthermore, the infectious disease application case can require the fusion of multiple data sources (e.g. census and public health statistics), inducing demand for a modular and extensible modeling approach. In this work we provide a novel sequential sample-free approach to generate synthetic baseline populations for agent-based simulations, combining synthetic reconstruction and combinatorial optimization. We applied the approach to generate a population model for the German state of North Rhine-Westphalia (17.5 million inhabitants) which yielded an average accuracy of around 98% per attribute. The resulting population model is publicly available and has been utilized in multiple simulation-based infectious disease case studies. We suggest that our research can pave the way for more geographically granular synthetic populations to be used in model-driven infectious disease epidemics prediction and prevention.

1 INTRODUCTION AND RELATED WORK

The COVID-19 pandemic has demonstrated the profound contribution of spatial agent-based simulations (ABS) to public health decision-making (Bicher et al. 2021). While some models are used to anticipate the disease propagation (Mahmood et al. 2020), others focus on the evaluation of effective intervention strategies (Silva et al. 2020). It is imperative for these models to capture regional demographic structures on individual level (i.e. age distribution or occupation) as well as on household level (i.e. average household sizes). While age might be a proxy for individual risks of hospitalization upon infection, occupation may inform the model about the probability of encounters with infectious individuals. Average household-sizes may greatly impact the effect of stay-at-home orders as useful interventions. This could be seen during the second wave lockdown of the COVID-19 epidemic in Germany, where the city of Münster consistently showed lower incidence and mortality rates as compared to the German average (Robert Koch-Institut 2021). As a so-called student city, Münster combines a low average age and a large amount of single-person

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

households. Seeing that a large portion of infections occur at home (Grijalva et al. 2020), these households seem to be self-evidently beneficial for breaking chains of infections. However, this setup is not representative for the whole country. This implies, that simulation-based propagation and intervention models should not be designed and tested with a secluded population, but require a high-resolution regionally stratified population model. The procedure of generating such virtual representations of real-life populations is referred to as Population Synthesis (Beckman et al. 1996).

Most of the available population synthesis approaches make use of local census data which is usually provided in the form of aggregated statistics on demographic features (such as age distribution) or a representative sample of households and individuals. Thus, these methods are either focusing on disaggregating statistics (sample-free), scaling-up samples (sample-based) or both. The two most popular approaches are the so-called synthetic reconstruction (SR) and combinatorial optimization (CO) (Ye et al. 2017). The SR approach proposed by Beckman et al. (1996) is still one of the most commonly used methods to generate baseline populations. It is a sample-based approach which relies on microdata records containing detailed information about a subset of individuals of the target population (sample). The authors derive joint distributions of features (e.g. age and sex) based on this sample and upscale it by means of iterative proportional fitting (Deming and Stephan 1940) to match the target population size and satisfy overall demographic features (such as the number of people in a certain age group). Although being remarkably applicable, the approach lacks the option of controlling household and individual attributes simultaneously, the integration of multiple data sources, and the zero-element problem which limits the synthetic population to the individuals contained in the sample. Subsequent publications (i.e. Pritchard and Miller 2012) since improved the initial algorithm to address these caveats. Gargiulo et al. (2010) propose a contrasting sample-free approach which generates the target number of individuals solely by their age-groups and combines them to households based on census-provided probabilities of people of certain ages sharing a home. While this approach suggests wide applicability as it does not require a sample, it is limited to a single individual attribute (age) and hence of limited usability in the infectious disease context. Williamson et al. (1998) were the first to suggest a fundamentally different approach to population synthesis by means of combinatorial optimization. It has since been adapted and expanded in many subsequent publications (e.g. Huynh et al. 2016). The general approach is to draw the target number of individuals from a sample and replaces individuals with subsequent drafts from the sample if they improve the overall fitness. In a comparing study, Ryan et al. (2009) found that CO approaches outperform their SR counterparts in terms of accuracy, however, due to their computationally intensive calculations of fitness, they usually require to split the population in mutually exclusive subpopulations to control for the problem complexity (Ye et al. 2017).

With the particular focus on infectious disease simulation, we argue that sample-free approaches provide an advantage over sample-based approaches due to the fact that the model might require in inclusion of multiple data sources. Such data can embody statistics on preexisting health conditions or social contact studies which are usually not provided by the general census. Hence, a census-based sample would not contain the required data. The absence of a suitable sample of households naturally comes with the requirement of introducing descriptive (qualitative) assumptions about grouping individuals to households – what constitutes a family? Such definitions (also referred to as household restrictions throughout this document) can be obtained from the census' collateral and have to be respected during the household generation process.

In this work, we propose a sequential sample-free approach for generating synthetic baseline populations. Synthetic reconstruction is utilized to generate a pool of individuals and households with all required features based on aggregated German census data (Destatis 2011c). We then assign these individuals to households with respect to the qualitative definitions of households and apply combinatorial optimization to increase accuracy, both, on individual- and household level. The results are then validated for synthetic populations of all 396 municipalities of the German state North Rhine-Westphalia (NRW) with three individual- and three household attributes using a weighted mean absolute percentage error measure (WMAPE). We found that this approach generates fairly precise population models with average accuracies among the various attributes of 98%. Meanwhile, we are maintaining the flexibility to adapt the

generation process to the requirements of a particular simulation scenario. The approach has been utilized in multiple case studies, notably in a simulation retracing the initial COVID-19 outbreak in Germany in February 2020. Our work contributes to the ongoing efforts of infectious disease simulation in two ways. First, while our population model was developed for Germany, we suggest that our approach is adaptable to other geographic regions and their particular data availability. Second, we made our synthetic German population model publicly available to be reused in further simulations.

The following section provides a brief overview of the data we utilize for generating populations. We then present our sequential approach to population synthesis and provide validation of our work in sections 3 and 4. Lastly, we discuss our approach and its application in an ABS project in section 5.

2 DATA

Our population generation solely relies on aggregated data, both regarding the frequency of individual characteristics (e.g. a certain age) and the frequency of household characteristics (e.g. a certain household size). For definition purposes, we regard characteristics as feasible values for attributes (e.g. *Male* and F*emale* are characteristics of the attribute *Sex*). Whenever we refer to census attributes and characteristics specifically, they are being put in cursive letters throughout this document. Although we use the German census data (Destatis 2011a), it is possible to incorporate data from arbitrary sources as long as the data provides the same geographical resolution (e.g. stratified by municipalities) and the same subdivisions of joint characteristics (e.g. both sources stratifying age-groups in five-year compartments). Thus, when combining data from multiple sources, additional preprocessing may be required.

The census data provides tables with marginal distributions of up to 18 individual and up to seven household attributes (Destatis 2011b) stratified by 11,340 geographical regions (municipalities) (Destatis 2011a) as well as several joint attribute tables. Moreover, it contains qualitative definitions of household compositions, i.e. regarding their family forms or senior status (Destatis 2011c). With the particular application focus on infectious disease propagation, we selected the three most relevant household attributes *Size* (HSI), *Family Status* (HLA), and *Senior Status* (HSC) and three individual attributes *Age* (AG2), *Sex* (SEX) and *Employment* (EMP) for our synthetic population (census identifiers in parentheses). While these attributes (except *Employment*) are mandatory for our synthesis, the generic approach can be extended to incorporate more of the available attributes by adding the respective tables. In addition to the mentioned single-attribute tables we utilize joint attribute tables for *Age-Sex* and *Sex-Employment*.

Due to data privacy reasons, especially in scarcely populated areas, the census may not provide values for all characteristics and combinations, as it would allow to conclude assumptions about living persons. Thus, it contains missing values or noisy data (Amt für Statistik Berlin-Brandenburg 2013). These occurrences induce the requirement of additional data preprocessing scaling the sum of single characteristics per attribute and hence matching the actual total of individuals or households.

A fortunate feature of the German census is that it provides grid data, informing us about the number of households in 100x100m cells across the whole country (Destatis 2011b). There are also worldwide population density maps (e.g. Oak Ridge National Laboratory 2019) with a resolution of 1x1km which could be utilized for other countries.

3 POPULATION SYNTHESIS

The generation of population for each municipality consists of five steps (Figure 1). First, the data is preprocessed to harmonize table totals. Then, separate pools of individuals and households instances are created using a Synthetic Reconstruction approach. In a third step, individuals get assigned to households based on a predefined ruleset. However, it is virtually impossible to obtain a perfect match of individuals and households when working with synthesized populations based on aggregated data. For this reason, we relaxed the restriction to assign every individual to a household only once in order to achieve the best fitting composition of households. Naturally, this procedure impedes the accuracy of the individuals' demographic. For this reason, we apply combinatorial optimization to increase accuracy, both, on individual

and household level in a fourth step. Finally, the households are assigned to a grid equipping them with a geographical location.

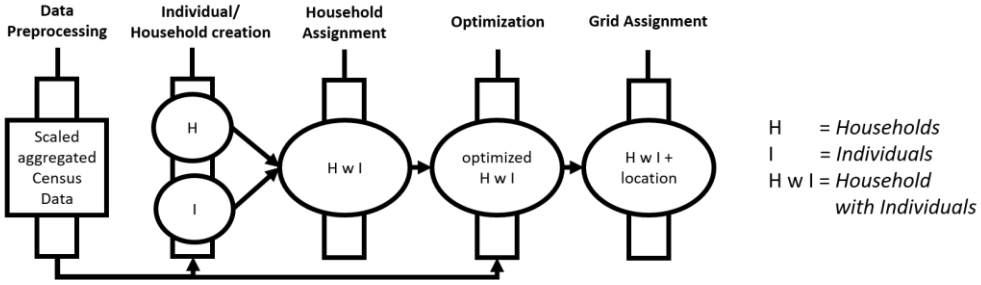

Figure 1: Synthetic Population Generation Procedure.

Most design decisions were being made in order to balance the conflicting goals of flexibility and run-time. Especially the combinatorial optimization phase is computationally intensive. This applies to large municipalities in particular as the runtime grows linearly with the number of households. While the runtimes for the vast majority of municipalities are a matter 1-5 minutes, very few very large municipalities (e.g. Cologne with ~1 million inhabitants) can take several hours. As we opted for a Java-based implementation we are able to parallelize the generation of municipalities. A synthetic population of NRW with roughly 17.5 million inhabitants in 396 municipalities takes around three and a half days to generate on a XEON quad-core server with 16GB memory. However, we expect the runtime to drop by about 70% when running our code on a high-performance cluster with as many threads as target municipalities as it is then limited by the generation time of the largest municipality. In further versions of the software, we also plan to parallelize the optimization procedure itself.

3.1 Data Preprocessing

To preprocess the data every attribute table is checked for the same column totals. The comparison is done using a base-table containing individual- and household sums for each municipality provided by the census (Destatis 2011b). If an attribute table sum differs from the base-table for the respective municipality the values are scaled to match the total while retaining proportions of characteristics. These deviations are generally below 1% of the target totals with a few outliers of around 5.5%. The scaling enables an exact match with the created population.

3.2 Individual & Household Generation

The goal of this phase is generating pools of individuals and households so that their overall demographic features match with the aggregated census data. While there are approaches to generate individuals and households simultaneously (e.g. Moreno and Moeckel 2018) we opted for a sequential approach as we see the risk of having the number of potential attribute combinations (of combined households and individuals) exceed the actual target quantity of individuals, especially for smaller municipalities. This would strongly impact any form of stochastic reliability of the resulting population model.

We rely on a well-established Synthetic Reconstruction approach utilizing Iterative Proportional Fitting (Beckman et al. 1996) to determine the frequency of co-occurring characteristics enabling the generation of the aforementioned individuals and households. Figure 2 depicts the following workflow for the pool of individuals, however, the household instances are generated in exactly the same way:

1. Transforming absolutes to probabilities. In a first step, the absolute values from the attribute tables are normalized by dividing every cell by the column total to obtain percentage values characteristic frequencies (e.g. 40 males out of 100 individuals is translated to a fraction value of 0.4 males).

2. Combining probabilities. The attributes' characteristic probabilities are now joined to a probability matrix by means of multiplication. The very first iteration is done with two of the input tables. Subsequent iterations join the Combined-table (output of step 6) with the next input table. As mentioned in section 2,

the census does also provide joint attribute tables which we utilize in our synthesis. For these tables, we check for the existence of what we call a pivotal attribute first, referring to an attribute contained in the Combined-table and available in the joint input table. In absence of a pivotal attribute, the process is identical with the case of single-attribute tables. In presence of a pivotal attribute, we generate one probability matrix for each characteristic of the pivotal attribute. Considering a joint input table of *Age x Sex* and the Combined-table consisting of *Age x Employment*, the pivotal attribute would be *Age* and thus trigger the generation of one *Sex x Employment* table per age group. While the marginal values of the Combined-table remain in the newly generated tables to retain the initial distribution of characteristics (i.e. *Age* & *Employment*), the added attribute (i.e. *Sex*) is scaled to achieve that marginal values sum up to 1 (within the respected age group) and thus not impede the initial distribution of characteristics.

3. Enforcing restrictions. As some combinations of characteristics should be considered infeasible (e.g. an employed five-year old), we maintain a list of such "impossible" occurrences. In each iteration, we apply this list to the newly joint table and set all probabilities to 0 for matching combinations.

4. Optimizing the table. After the previous step, the row- and column-wise sums of cells do not match the marginal distributions anymore (as some of the cells were set to 0). To converge the remaining values, we apply Iterative Proportional Fitting according to Beckman et al. (1996). This procedure scales the cells of rows and columns to match their marginal values alternately. The fitting procedure is terminated either at a threshold value of 0,1% deviation from the anticipated marginal values or after 200 iterations which were found to be an appropriate limit (Huang and Williamson 2001) to prevent infinite loops.

5. Transforming into vector. The resulting matrix is then transformed into a vector by combining the attribute characteristics of the two attribute tables. This way we can use the resulting table for further iterations.

6. Generating instances. Once all attributes were combined in one table the individual- and household-instances are created. This is done by scaling the resulting probability vector to the initial population size and generating the respective quantity of instances.

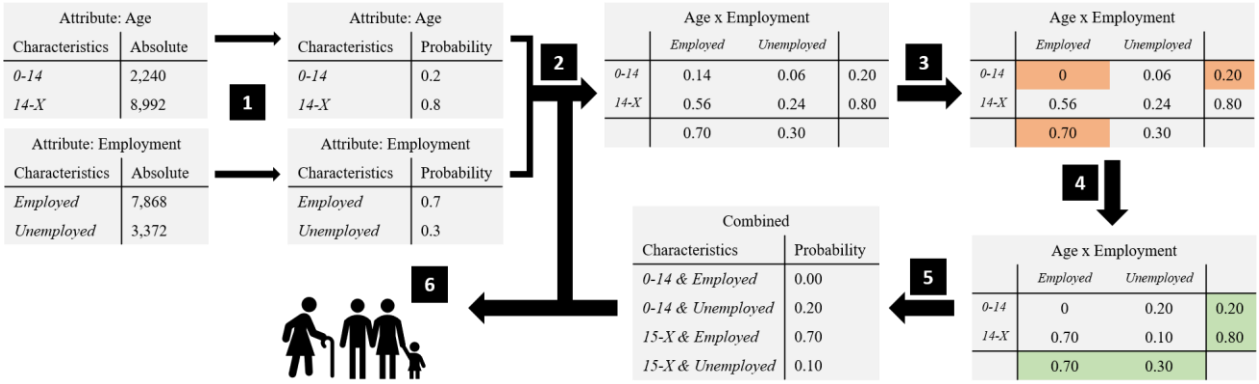

Figure 2: Individual Pool Generation.

3.3 Household Assignment

This phase aims at assigning the generated individuals to households so that the pool of households provide the same demographic features as the reference data (regarding household composition and size). Therefore we iterate over every household and assign individuals who match the qualitative definition of the respective household characteristic (e.g. *Family Status*, *Senior Status* and *Size*). Table 1 illustrates the assignment ruleset which is based on the census' definition of household types (Destatis 2011c) and will be described in the following:

The *Family Status* (HLA) generally informs the composition of households and the residents' relationships. It is based on the definition of a core family, which consists of a defined index person and at least one further family member living in the household, who is either a child or the partner of the index person. The household can consist of either only the core family or also include additional members, who

do not belong to the core family (Michel 2014). Therefore the characteristics of the *Family Status* are the different types of core families. The characteristics can be grouped in three categories: *No Core Family or Single-Person Households*, *Couples* and *Single Parents*. *Single-Person Households* and households with *No Core Family* have no restrictions on the relations of their residents. Furthermore, there are three census characteristics of *Couples* consisting of a couple and optional children as core families: *Married Couple*, a *Not Formally Registered Couple* or a *Registered Couple*. The vary in that the *Registered Couple* is same-sex, while the other types imply mixed-sex couples. This is built into the assignment, but for the sake of simplicity not visualized in Table 1. Two individuals forming a couple are assigned based on probabilities of age differences and their sex. According to the German Federal Office of Statistics, in 17% of couples, the woman is older than the man, in 73% the man is older and in 10% their age is the same. Furthermore, 52% have an age difference of 1-3 years, 41% a difference of 4-9 years and 7% more than ten years. Married couples have to be at least 18 years of age (Destatis 2018). During our assignment, we sample "suitable counterparts" for the index person based on these probabilities. We apply the same approach for assigning children who have a 31% chance to be over the age of 18, and thus a 69% chance to be minors. Among the latter, 82% are between the ages 0-14 (Destatis 2018). The last group among families consists of *Single Parents*, divided into the census characteristics S*ingle Mothers* and *Single Fathers*. These households have to consist of at least one parent of the defined *Sex* and a child (Destatis 2011c).

Table 1: Household Assignment Ruleset.

Family Status	***Senior Status***	**Rule**	
No Core Family/ Single-Person Household	Only Seniors	Fill with seniors	
	Mix	Assign senior and one non-senior + fill without restriction	
	No Seniors	Fill with non-seniors	
Family	Only Seniors	Assign senior couple + fill with seniors	
	Mix	*If household size = 2*	Assign mixed couple
		Else	assign senior + add non-senior family
	No Seniors	Assign non-senior couple, fill with children	
Single Parent	Only Seniors	Not possible (constraint)	
	Mix	*If household size = 2*	assign senior and fitting child
		Else	assign senior + parent + fill with children
	No Seniors	parent + fill with children	

The S*enior Status* of an household (HSC) has three characteristics: The household can have *No Seniors*, *Only Seniors* or a *Mix*. A senior is defined as a person of age 65 or older (Destatis 2011c).

Finally, the *Household Size* (HSI) is used to determine the number of individuals per household and groups households in *sizes from 1 to 6+*. It is also used to "fill up" a household that already consists of the required individuals to match *Family Status* and *Senior Status* but has not reached capacity.

These definitions restrict the individuals, who are suitable for the respective household. Therefore they have to be filtered by their attributes *Sex* and *Age* (in eleven age groups) (Destatis 2011c). The assignment is then performed using a rule-based approach guided by the census definitions. The two individuals attributes (*Sex* and *Age*) attributes as well as the three household attributes (*Family Status*, *Senior Status* and *Size*) are mandatory input parameters. As mentioned in section 3.2, additional individual- and household attributes can be incorporated. However, if they affect the household assignment (for example the number of members in the core family), it requires an adaption to the assignment ruleset.

Assigning individuals to households "correctly" is a rather complex task as many individuals naturally fit in various households. Furthermore, we cannot be certain that there is a perfect match of individuals and

households as both pools were only generated based on aggregated data. Our initial tests had shown that generating household assignments sequentially, emptying the available pool of individuals, leads to a considerable decline in household accuracy. This particularly applies to the households being assigned towards the end, as the shrinking pools of individuals may not hold any fit anymore. For this reason, we suggest to relax the requirement to assign every individual exactly once. In fact, we saw much better results in sampling individuals from the pool with replacement. While this approach leads to a perfect match on household level and causes a reduction of unique individuals (and thus population heterogeneity) of 0.13%, it decreased the accuracy on individual level to around 88% in our test cases. This necessitates an additional optimization presented in the next section.

3.4 Optimization

Combinatorial Optimization is used to rearrange and replace households to find a set of households which fits the census data the best. In order to guarantee the integrity of households regarding their *Family Status* and *Senior Status*, we do not replace individuals in the respective households. Furthermore, optimizing among all individuals in all households is highly computationally intensive and of limited applicability for large populations as Ponge et al. (2016) have shown. We opted for a recombination approach where we draw households including their assigned individuals from the generated pool and allow for selecting the same household more than once which yields the risk of losing population heterogeneity. However, our tests show that on average, this leads to a reduction of unique households of 3.31% and unique individuals by 0.25%. Thus, at least 96,7% of the population heterogeneity is retained after optimization.

The recombination is performed using a genetic algorithm according to Luke (2013) working with ten candidate solutions. A candidate solution contains the target number of households for a given municipality according to the census. At initialization, each candidate solution is constructed by drawing the respective number of households from the pool at random. We measure the fitness of candidate solutions by means of Z-Scores (RSSZ*) as proposed by P. Williamson (Tanton and Edwards 2013). RSSZ* is an established measure of fit for combinatorial optimization approaches generating synthetic small-area microdata.

In each evolutionary cycle, we build eight crossovers of the two most promising candidate solutions by randomly sampling households from either candidate solution. The new solutions as well as their two parents remain in the set of candidate solutions in order to ensure keeping the best fitting candidate. With a chance of 20% the newly created candidate solutions are then mutated, which means a replacement of 0-100% of the households by randomly selecting substitutes from the pool. The initial solutions not selected for crossovers are disregarded.

The termination criterion is an overall RSSZ* of less than 1, as this guarantees, that every attribute table has a Z-Score of less than one and is therefore defined as a so-called fitting table according to Tanton and Edwards (2013). As there is no definitive guarantee for this termination criterion to be met, our fallback-stopping criterion is triggered after 5,000 generations of this recombination.

3.5 Grid Assignment

In the last step, we assign locations to the household selection that was generated in the previous step. Therefore the grid data of the census is used. By determining all inhabited grid-cells for the municipality, the percentage of the households, which are in the respective grid-cell can be calculated. Every household is then assigned to a grid-cell, following the household density distribution given by the census grid data. This approach will always assign all households of the final set. The assignment of households is performed at random, as the census does provide no information on what particular households shall be located in a particular grid-cell for privacy reasons. In fact, the approach of reverse-engineering information about individual households (e.g. through cells with single households) might be a violation of census regulations. However, as the German census contains more than eleven thousand municipalities, assigning households per municipality does still provide a substantial level of region heterogeneity. Figure 3 shows the frequencies and locations of households in the synthetic population for the municipality of Gangelt in the Heinsberg county (Western Germany) using ArcGIS Pro (Esri 2019).

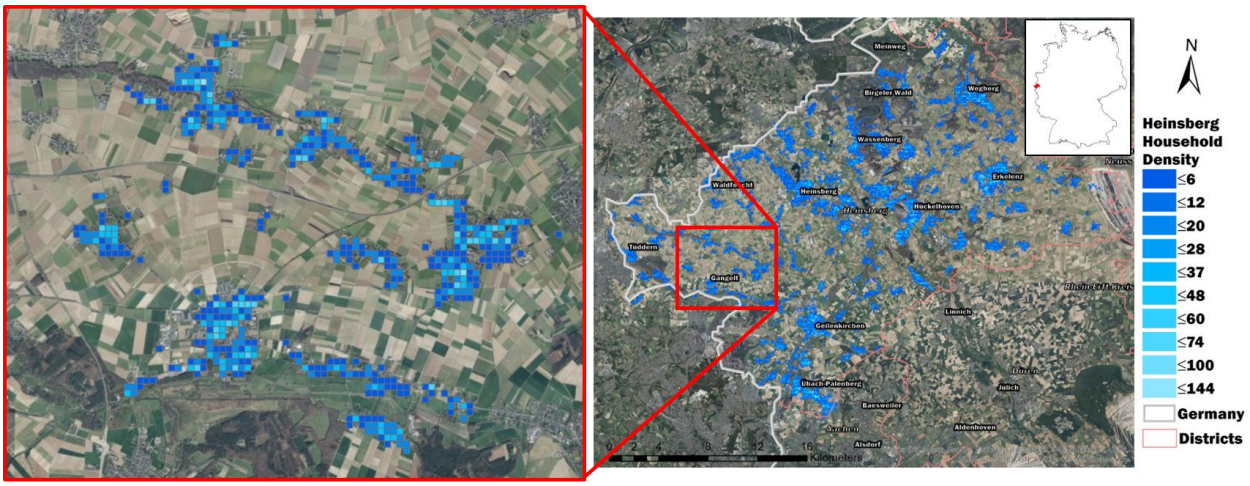

Figure 3: Household Locations for the Gangelt Municipality (Left) in the Heinsberg County (Right)**.**

4 VALIDATION

We validate the resemblance of our synthetic population to its real-life equivalent in two steps: An in-depth examination of a single municipality's demographic features and a comprehensive validation comparing separate population models for all 396 municipalities of the state of NRW. For the in-depth validation, we chose the Gangelt municipality in the western part of Germany (s. Figure 3). Gangelt was the epicenter of Germany's COVID-19 outbreak in February 2020 (Streeck et al. 2020) and thus the first municipality we generated a synthetic population for using the approach presented in this paper.

We apply an artificial census to our synthesized population and contrast the absolute values of the results with the original census data. The results of the used attributes (census identifiers in parentheses) *Age* (AG2), *Sex* (SEX), *Employment Status* (EMP), *Household Size* (HSI), *Family Status* (HLA) *and Senior Status* (HSC) can be seen in Figure 4. Furthermore we calculate the Weighted Mean Absolute Percentage Error (WMAPE) for each attribute:

$$WMAPE_a = \frac{\sum_{c=1}^{n} |E_c - S_c|}{\sum_{c=1}^{n} |E_c|}$$

E_t, S_t = Empirical and Synthetic frequencies for characteristic c of attribute a

This measure weights the proportion of characteristics in comparison to the total amount of instances. It prevents the overemphasis of percentage errors in characteristics with comparably few individuals in the respective category (e.g. age group [0-3] vs. age group [40-49]) and thus creating a robust measure for comparison. In the following, the terms error and WMAPE are used interchangeably. We generally observe a very good fit of our synthetic population and the Gangelt census data with errors between 0.2% (*Senior Status*) and 6.28% (*Household Size*). By comparison to other municipalities in NRW (Figure 5), the 6.28% error in household sizes constitutes an outlier. The target population size for the synthetic population is informed by the overall sum of household sizes. However, this target sum (10,722) does not match with the reported number of individuals (11,405) (Destatis 2011c). A fraction of this deviation may be traced back to our limit of six individuals per household, since the census does not report on the actual number of individuals in large households. Yet, to account for the full deviation of 683 people, the average size of the 90 larger-than-six person households in Gangelt (Destatis 2011c) would have to be 13.6 which we assume to be rather unlikely. Another more probable explanation is the fact that a considerable number of people in Gangelt (617) live in communal accommodation (Destatis 2011a). While this discrepancy also causes the synthetic population to be slightly smaller than the real population (11,180 individuals vs. 11,405), it does not seem to be a reoccurring effect throughout all municipalities. In fact, among the 396 municipalities of NRW, we see 59 entities with a *larger* target sum of household sizes as opposed to the reported number of individuals. The peculiarity of household sizing requires further examination in the future.

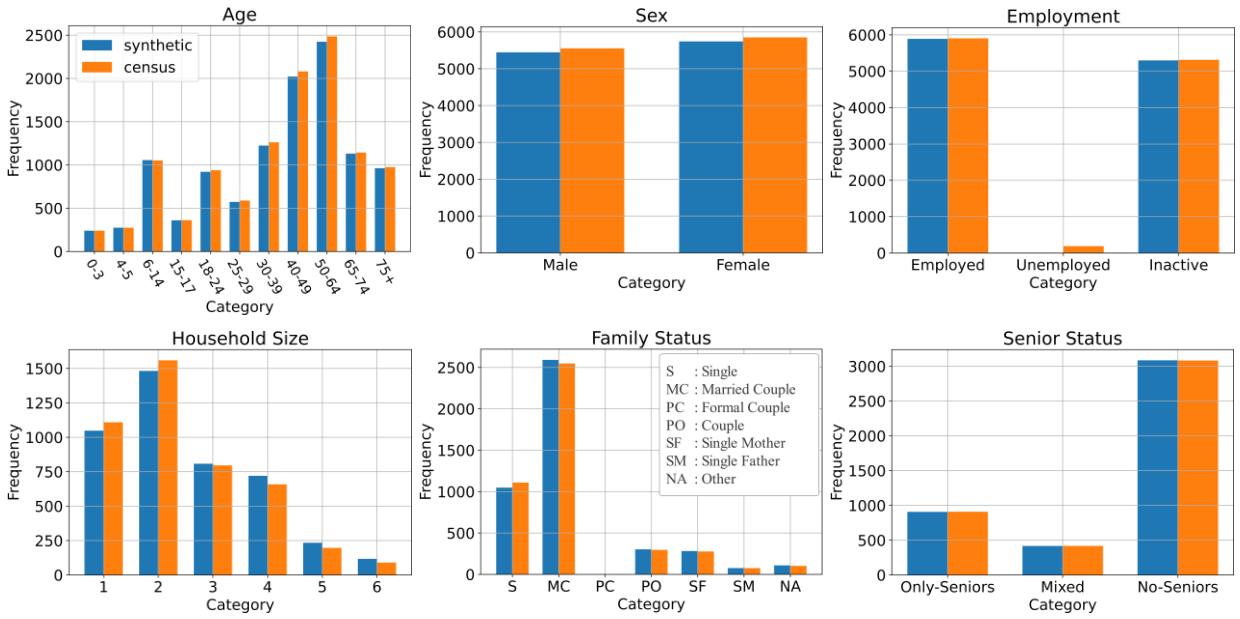

Figure 4: Absolute Frequencies of Individual- & Household Characteristics for Gangelt considering *Age*, *Sex* and *Employment* (Top Row) and *Household Size*, *Family Status* and *Senior Status* (Bottom Row)**.**

In order to obtain statistically reliable validation results, we generated 100 synthetic populations for Gangelt and calculated the WMAPE for every attribute. The results are given in Table 2. We observe, that the errors for the various attributes are rather consistent with maximum standard deviation of 0.14% and a maximum range of 0.67%.

Table 2: WMAPE in Percent per Attribute Across 100 Synthetic Populations for Gangelt (Heinsberg).

Attribute	**Mean Error**	**St.-Deviation**	**Min Error**	**Max Error**
Age	2.10 %	0.14 %	1.86 %	2.53 %
Sex	1.89 %	0.03 %	1.81 %	2.00 %
Employment	1.89 %	0.03 %	1.81 %	2.00 %
Household Size	6.28 %	0.08 %	6.08 %	6.53 %
Household Family Status	2.88 %	0.07 %	2.68 %	3.13 %
Household Senior Status	0.20 %	0.08 %	0.05 %	0.41 %

However, as we have seen for Gangelt, validation results can be highly dependent on the particularities of the municipality in focus. To collect a more comprehensive assessment on the synthesis performance, we generated synthetic populations for all 396 municipalities of the state of NRW in a second validation step. We selected NRW as it is the largest German state and it shows significant demographic differences among the municipalities regarding the number of citizens (4,197 to 1,005,775), share of minors (13% to 25%), and the share of senior citizens (14% to 31%) as well as the composition of households regarding the share of single-person households (17% to 51%), share of married-couple households (32% to 68%), and no-senior households (55% to 77%). We consider this variety to be a reasonable stress test for our algorithm.

For every municipality we calculated the WMAPE for each attribute. Figure 5 contains the distribution of errors visualized in a boxplot with the errors for *Household Family Status* (0.91%), *Household Senior Status* (0.48%), *Household Size* (1.53%), *Sex* (0.46%), *Age* (1.62%), and *Employment* (0.88%).

As all average error margins are contained below a 2% threshold, we conclude that our approach yields an average accuracy of around 98% per attribute. We can observe a general trend, that attributes with more characteristics (*Age*, *Family Status and Household Size*) tend to yield higher error margins. Nonetheless, the number of characteristics shall not be an universal estimator for an attribute's error margin since the

error can also be induced by mismatches in the input data as it was previously shown for *Household Size*. From this observation we conclude the requirement of thorough input data analysis, both, for the attributes at hand and especially for attributes which might be added in future works.

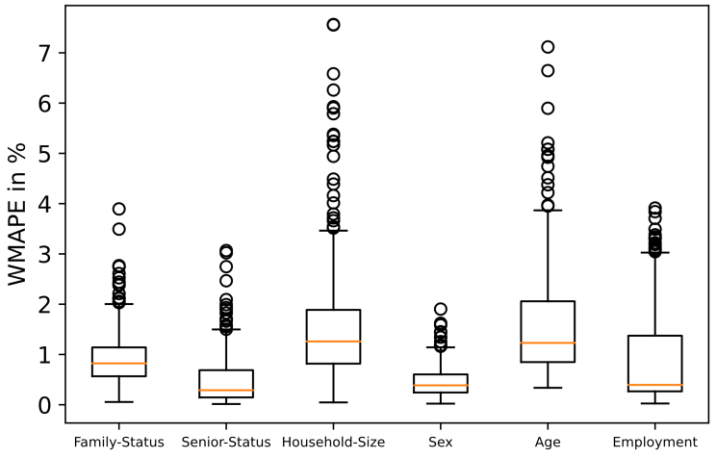

Figure 5: WMAPE in Percent of household- & individual attributes across all 396 municipalities of the state of North Rhine-Westphalia (Largest State in the West of Germany with roughly 17.5 million inhabitants). Each observation corresponds to one municipality.

5 DISCUSSION AND APPLICATION

In the previous chapters we demonstrated how a two-step hybrid approach for population synthesis can benefit from both, the input data flexibility of synthetic reconstruction and the accuracy improvement of combinatorial optimization approaches. While our methodology is based on the initial work of Beckman et al. (1996), we can omit the zero element problem by calculating all possible attribute combinations based on the census tables. Moreover, this approach eliminates the need for a microsample, but introduces the risk of generating non-feasible attribute combinations (employed five-year-olds) which have to be controlled for manually. However, especially in the infectious disease simulation context, we see the increased input data flexibility as a significant advantage. In contrast to the sample-free approach by Gargiulo et al. (2010), our approach is able to base household assignments and compositions not only on individuals' age but on an arbitrary number of attributes by means of the rule-based assignment approach complemented with combinatorial optimization. Through separating the generation of individuals and household composition we benefit from the advantages of combinatorial optimization (Huynh et al. 2016; Ryan et al. 2009; Williamson et al. 1998) without being dependent on the availability of a microsample.

The model presented in this paper serves as the baseline population layer to an in-progress agent-based simulation project investigating the effectiveness of non-pharmaceutical interventions in wake of the German COVID-19 epidemic. On top of the population layer we synthesize a contact network layer including workplace and school connections as well as random encounters in supermarkets or during leisure activities. Our high-resolution population model enables us to evaluate interventions with respect to regional demographical features. We retraced the initial outbreak in the municipality of Gangelt which occurred after an infected couple attended a carnival event in February 2020 (Streeck et al. 2020). While containment efforts (i.e. closure of schools and non-essential businesses) were imposed promptly, empirical studies have found that roughly 15% of inhabitants were infected within the next six weeks (Streeck et al. 2020). A scenario which we could also demonstrate in our simulation. Then, we set out to explore, whether such an event and the subsequent interventions would have caused the same disease progression in a municipality of similar size but with considerably different demographic features (in terms of age structure and household sizes). Preliminary results indicate that the same interventions would have prevented up to 25% of the overall infections in more "demographically advantageous" regions (i.e. regions with more single-person households and a lower average age). The conclusions we draw from our observations are twofold: first, demographical features do have an influence on the course of an epidemic and second, these

features should be considered when trying to design minimal-invasive interventions. We suggest that agent-based simulations in accord with our regional high-resolution population models can support the design of effective and efficient containment measures.

6 CONCLUSION

In this work we presented a novel sequential sample-free approach to generate synthetic baseline populations for agent-based simulations. The five-step approach is modular and extensible, both, regarding the input data as well as the ruleset guiding the composition of households. We applied the approach to create synthetic populations for the German state of NRW and could show that we achieve consistent average accuracies across all attributes of at least 98%. The resulting dataset has been made publicly available at https://github.com/JohannesPonge/SyntheticPopulations and we suggest that the spatial ABS research community will benefit from our efforts. It has been used in agent-based simulation case studies to evaluate the effectiveness of immediate interventions in wake of the initial COVID-19 outbreak.

Still, there is a lot potential for further research. So far, only census data has been used to generate and parametrize synthetic populations. However, the infectious disease simulation context can require the combination of multiple data sources (such as data on preexisting conditions). We anticipate the extension our model and thus provide more comprehensive synthetic population datasets in the next years.

ACKNOWLEDGMENTS

This work is part of the EpiPredict project, funded by the German Federal Ministry of Education and Research, project 01KI1913.

REFERENCES

Amt für Statistik Berlin-Brandenburg. 2013. "SAFE – Verfahren zur sicheren Anonymisierung für Einzeldaten zur Wahrung des Statistikgeheimnisses beim Zensus 2011".

Beckman, R. J., K. A. Baggerly, and M. D. McKay. 1996. "Creating synthetic baseline populations". *Transportation Research Part A: Policy and Practice* 30(6):415–429.

Bicher, M., C. Rippinger, C. Urach, D. Brunmeir, U. Siebert, and N. Popper. 2021. "Evaluation of Contact-Tracing Policies Against the Spread of SARS-CoV-2 in Austria - An Agent-Based Simulation". *medRxiv* :2020 - 05.

Deming, W. E., and F. F. Stephan. 1940. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known". *The Annals of Mathematical Statistics* 11(4):427–444.

Destatis. 2018. "Familie, Lebensformen und Kinder: Auszug aus dem Datenreport 2018".

Destatis. 2011a. Ergebnisse des Zensus 2011. https://ergebnisse2011.zensus2022.de/datenbank/online, accessed 28th March 2021.

Destatis. 2011b. Ergebnisse des Zensus 2011 zum Download - erweitert. https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html, accessed 2nd April 2021.

Destatis. 2011c. "Zensus 2011 - Bevölkerung & Haushalte: Übersicht über Merkmale und Merkmalsausprägungen, Definitionen".

Esri. 2019. *Imagery [basemap]*: *World Imagery.* https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9, accessed 2nd April 2021.

Gargiulo, F., S. Ternes, S. Huet, and G. Deffuant. 2010. "An iterative approach for generating statistically realistic populations of households". *PloS one* 5(1):e8828.

Grijalva, C. G., M. A. Rolfes, Y. Zhu, H. Q. McLean, K. E. Hanson, E. A. Belongia, N. B. Halasa, A. Kim, C. Reed, A. M. Fry, and others. 2020. "Transmission of SARS-COV-2 infections in households—Tennessee and Wisconsin, April - September 2020". *Morbidity and Mortality Weekly Report* 69(44):1631.

Huang, and Williamson. 2001. "A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata". Working Paper 2001/2, Department of Geography, University of Liverpool. http://pcwww.liv.ac.uk/~william/microdata/pop91/methodology/workingpapers/hw_wp_2001_2.pdf, accessed 2nd April 2021.

Huynh, N. N., J. Barthelemy, and P. Perez. 2016. "A heuristic combinatorial optimisation approach to synthesising a population for agent-based modelling purposes".

Luke, S. 2013. *Essentials of Metaheuristics*: *A Set of Undergraduate Lecture Notes.* 2nd ed. Morrisville, N.C.

Mahmood, I., H. Arabnejad, D. Suleimenova, I. Sassoon, A. Marshan, A. Serrano-Rico, P. Louvieris, A. Anagnostou, S. J. E. Taylor, D. Bell, and others. 2020. "FACS: a geospatial agent-based simulator for analysing COVID-19 spread and public health measures on local regions". *Journal of Simulation* :1–19.

Michel, N. 2014. "Zensus 2011: Was uns der Zensus über Haushalte und Familien verrät: Teil 1: Haushalts- und Familientypen". *Statistisches Monatsheft Baden-Württemberg* 8.

Moreno, A. T., and R. Moeckel. 2018. "Population synthesis handling three geographical resolutions". *ISPRS International Journal of Geo-Information* 7(5):174.

Oak Ridge National Laboratory. 2019. LandScan Global 2019. https://landscan.ornl.gov/, accessed 28th June 2021.

Ponge, J., D. de Siqueira Braga, D. Horstkemper, B. Hellingrath, S. Ludwig, and F. B. de Lima Neto. 2016. "Automated scalable modeling for population microsimulations". In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7.

Pritchard, D. R., and E. J. Miller. 2012. "Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously". *Transportation* 39(3):685–704.

Robert Koch-Institut. 2021. Robert Koch-Institut: COVID-19-Dashboard. http://corona.rki.de/, accessed 2nd April 2021.

Ryan, J., H. Maoh, and P. Kanaroglou. 2009. "Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms". *Geographical Analysis* 41(2):181–203.

Silva, P. C. L., P. V. C. Batista, H. S. Lima, M. A. Alves, F. G. Guimarães, and R. C. P. Silva. 2020. "COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions". *Chaos, Solitons & Fractals* 139:110088.

Streeck, H., B. Schulte, B. M. Kümmerer, E. Richter, T. Höller, C. Fuhrmann, E. Bartok, R. Dolscheid-Pommerich, M. Berger, L. Wessendorf, and others. 2020. "Infection fatality rate of SARS-CoV2 in a super-spreading event in Germany". *Nature communications* 11(1):1–12.

Tanton, R., and K. L. Edwards, eds. 2013. *Spatial Microsimulation*: *A Reference Guide for Users*. Dordrecht: Springer.

Williamson, P., M. Birkin, and P. H. Rees. 1998. "The estimation of population microdata by using data from small area statistics and samples of anonymised records". *Environment and Planning A* 30(5):785–816.

Ye, P., X. Hu, Y. Yuan, and F.-Y. Wang. 2017. "Population synthesis based on joint distribution inference without disaggregate samples". *Journal of Artificial Societies and Social Simulation* 20(4).

AUTHOR BIOGRAPHIES

JOHANNES PONGE is a Research Assistant at the Chair for Information Systems and Supply Chain Management at the University of Münster, Germany. He holds an M.Sc. in Information Systems. As part of the chair's crisis-management and humanitarian logistics research team, his work focuses on the development of simulation-based decision support systems for infectious disease mitigation and intervention. His email address is johannes.ponge@ercis.uni-muenster.de

MALTE ENBERGS is a Graduate Assistant at the Chair for Virology at the University of Münster, Germany. He holds an B.Sc. in Information Systems. He is now doing his M.Sc. in Information Systems and specializing in Information Systems Development and Buisness Networks. His email address is malte.enbergs@uni-muenster.de

MICHAEL SCHÜNGEL is an Undergraduate Assistant at the Chair for Information Systems and Supply Chain Management at the University of Münster, Germany. He is about to graduate and continue his study in the M.Sc in Information Systems. His email address is michael.schuengel@uni-muenster.de

BERND HELLINGRATH is a Professor and Head of the Chair for Information Systems and Supply Chain Management at the University of Münster, Germany. His research interests deals with the broader area of modeling and simulating with a distinct focus on the context of crisis-management and humanitarian logistics. His email address is bernd.hellingrath@ercis.uni-muenster.de

ANDRÉ KARCH is a Professor of Clinical Epidemiology, and the Chair of the Department of Clinical Epidemiology at the University of Münster, Germany. He holds an MD and an MSc in Epidemiology. His research in the field of infectious disease epidemiology focuses on the use of dynamic transmission models for public health decision-making. His email address is akarch@uni-muenster.de

STEPHAN LUDWIG is Full Professor and Director of the Insitute of Virology (IVM) at the Medical Faculty, University of Münster, Germany. Research at the IVM is centered around interactions of respiratory viral pathogens with the host and the immune system. He is furthermore interested in species transmission and epidemiology of zoonotic viral pathogens. His email address is ludwigs@uni-muenster.de