METHODS REPORT

Attribute Assignment to a Synthetic Population in Support of Agent-Based Disease Modeling

James C. Cajka, Philip C. Cooley, and
William D. Wheaton

September 2010

RTI Press

RTI INTERNATIONAL

About the Authors

James C. Cajka, MA, is a senior GIS analyst in RTI International's Geospatial Science and Technology unit.

Philip C. Cooley, MS, RTI Fellow in bioinformatics and high-performance computing, is a principal scientist and assistant director of bioinformatics, with a focus on computational biology. He has more than 35 years of experience developing computer models for the study of environmental health and disease transmission scenarios.

William D. Wheaton, MA, is a senior research geographer and director of RTI International's Geospatial Science and Technology program.

RTI Press publication MR-0019-1009

RTI International is an independent, nonprofit research organization dedicated to improving the human condition by turning knowledge into practice. The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

Suggested Citation

Cajka, J. C., Cooley, P. C., and Wheaton, W. D. (2010). Attribute Assignment to a Synthetic Population in Support of Agent-Based Disease Modeling. RTI Press publication No. MR-0019-1009. Research Triangle Park, NC: RTI International. Retrieved [date] from http://www.rti.org/rtipress.

This publication is part of the RTI Press Methods Report series.

RTI International
3040 Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
Fax: +1.919.541.5985
E-mail: rtipress@rti.org
Web site: www.rti.org

doi:10.3768/rtipress.2010.mr.0019.1009

www.rti.org/rtipress

Attribute Assignment to a Synthetic Population in Support of Agent-Based Disease Modeling

James C. Cajka, Philip C. Cooley, and
William D. Wheaton

Contents

Abstract

Communicable-disease transmission models are useful for the testing of prevention and intervention strategies. Agent-based models (ABMs) represent a new and important class of the many types of disease transmission models in use. Agent-based disease models benefit from their ability to assign disease transmission probabilities based on characteristics shared by individual agents. These shared characteristics allow ABMs to apply transmission probabilities when agents come together in geographic space. Modeling these types of social interactions requires data, and the results of the model largely depend on the quality of these input data. We initially generated a synthetic population for the United States, in support of the Models of Infectious Disease Agent Study. Subsequently, we created shared characteristics to use in ABMs. The specific goals for this task were to assign the appropriately aged populations to schools, workplaces, and public transit. Each goal presented its own challenges and problems; therefore, we used different techniques to create each type of shared characteristic. These shared characteristics have allowed disease models to more realistically predict the spread of disease, both spatially and temporally.

RTI INTERNATIONAL

Introduction

Agent-based models (ABMs) are a new and important paradigm for studying epidemics.[1] Although these models can simulate the realistic propagation of epidemics, they require input data about the social networks that are part of the agents' day-to-day activities. With funding from the Models of Infectious Disease Agent Study (MIDAS), we initially created a national synthetic population database, whereby a record was created for each person living in the United States. The primary attributes of each person included age, gender, marital status, household location, employment status, and commute mode. These attributes enabled us subsequently to create school, workplace, and public transit interactions and then code this information into the data. This report documents the data sources, assumptions, and methods used to develop three additional social contacts: schools, workplaces, and public transit. Each social contact had its own specific requirements, data sources, and problems, which led to the development of three distinct methods.

Agent-Based Models

Throughout this report, *agents* are defined as autonomous individuals—not organisms that can cause infection or disease, as traditionally defined in medical epidemiology. ABMs are computational methods for simulating the actions and interactions of agents for the purpose of assessing their effects on the system of which they are part. The rationale behind ABMs is that the simultaneous operations of multiple interacting agents describe and predict the actions of complex, interacting processes. Individual agents are characterized as rational entities that act by using heuristics, or simple decision-making rules. A typical ABM consists of the following components:

- agents (e.g., people susceptible to disease) that are associated with attributes that influence their actions,
- decision-making heuristics (e.g., go to work or ride the bus, or both),
- learning rules (e.g., go to the doctor if sick),
- an interaction landscape (contacts in different locations), and
- a non-agent environment (e.g., schools, workplaces, malls).

ABMs have been used since the mid-1990s to solve a variety of business and technology problems. Because of the versatility of the ABM method, the use of ABMs is expanding. The heterogeneous property of agents enables more sophisticated and complex environments to be described by the ABMs method than by alternative approaches.[2-4]

In the case of epidemiological ABMs like the kind used in this study, the simulated process represents the interaction patterns of people who are parts of multiple social networks in the modeled area. This simulation requires that social network data and population movement patterns from multiple sources be fused. It begins with the allocation of agents to the households in a given area, consistent with available demographic information, including age, gender, and socioeconomic status. Algorithms also assign these agents to the schools, workplaces, and other elements in the greater community to which they belong (shopping malls, neighborhood organizations, etc.) in ways that are consistent with available source data.

A disease pathogen is then introduced into one or more of the agents to assess how the agents' and their social networks' properties propagate the disease. Pathogens themselves also have properties that influence disease transmission (virulence, generation time, etc.). Notably, the distributed collection of interacting persons susceptible to infection constitutes a disease system and functions without a "leader." That is, the persons interact locally according to simple rules of behavior, responding in appropriate ways to environmental cues and not necessarily striving to achieve an overall goal.

The Models of Infectious Disease Agent Study

Funded by the National Institutes of Health, MIDAS is a collaborative effort of investigators from various research and informatics groups that develop novel computational models, including ABMs, to study the interactions between infectious diseases and their hosts, disease spread, prediction systems, and response strategies. As part of the MIDAS network,

RTI has developed data sets to support disease transmission modeling using these ABMs.[5] One of the most important data sets built by RTI is the synthetic persons data set. It consists of five linked tables, which include the US Census Public Use Microdata Sample (PUMS) household (PUMS_H) and person (PUMS_P) tables.

Using the PUMS_H table as a guide, RTI randomly placed the correct number of households within each Census block group, with the help of a geographic information system (GIS). Using the PUMS_P table, RTI created a database record for each person in the United States and then linked it to one of the synthetic households. This linking created a set of individual households and persons that, when aggregated in a geospatial area (such as Census tract, county, or state), would be consistent with the Census data for that area.[5] Each person was linked by means of a unique identifier to the PUMS_P table, which contains a variety of attributes, including age and school enrollment. This unique identifier provided a linkage between people living in the same household, which is an essential point of social contact. This database structure provided a starting place for the creation of additional social contact points, one of the MIDAS research tasks and the focus of this report.

Methods

The shared-characteristic assignment methods we created had to use nationally available input data so that the assignments could be performed anywhere in the United States. In addition to assigning shared characteristics at the individual level, the methods maintained the geographic integrity of the output data so that the counts would match the original data when aggregated. Aside from these overarching commonalities, each of the school, work, and public transit assignment methods was a unique solution, using specific data sets and presenting its own set of challenges.

Schools

Infectious diseases are known to be transmitted at a higher rate in schools because of the close contact that students have with one another.[6] To create this type of social contact, we collected the information necessary to assign elementary, middle, and high school–aged students to public schools and non-residential private schools. We did not create school assignments for college students, because they do not necessarily attend a school in close proximity to their residences; nor did we create them for preschoolers, because publicly available enrollment data were insufficient to support the assignment process. However, if schools reported preschool enrollments, then those assignments were made.

The synthetic population data reported students' ages, whereas the school enrollment data reported the number of students by grade. To equate grade to age, we used Table 1 to assign students (e.g., 4-year-olds were assigned to preschool slots; 5-year-olds, to kindergarten slots).

Table 1. The assignments based on grade-age equivalents

Grade	Age
Preschool	4
Kindergarten	5
1	6
2	7
3	8
4	9
5	10
6	11
7	12
8	13
9	14
10	15
11	16
12	17

Besides age, another variable stored in the PUMS_P table was the "enroll" field,[7] which contained values 0–3, which corresponded to the educational enrollment status for the 2000 Census (Table 2).

Table 2. The code and corresponding description for the "enroll" field

Code	Description
0	N/A (less than 3 years old)
1	No, has not attended in the last 3 months
2	Yes, public school or college
3	Yes, private school or college

Source: US Census Bureau. 2000 Census of Population and Housing, public use microdata sample, United States: technical documentation, 2003. 2003 [September 2, 2010]; Available from: http://www.census.gov/prod/cen2000/doc/pums.pdf.

For the MIDAS assignments, only individuals in the synthetic population table coded as 2 were selected for assignment to public schools, and only those coded as 3 were selected for assignment to private schools.

Public School

In reality, determining which person attends what public school is a complicated process with tremendous variability both within and between school systems across the country. This determination depends on the individual school district's specific goals, such as minimizing busing, mixing ethnicities, or maintaining neighborhood schools,[8] and includes many factors, such as geographic proximity, socioeconomic characteristics, physical barriers, availability of busing, and politics. There is no one formula or set of criteria that would apply to school assignments nationwide. To create a process by which students were assigned to public schools in a systematic way and one that could be repeated anywhere in the country, we created our own process using the following assumptions:

- Geographic proximity would be the best objective criterion for making assignments.
- Students would be assigned to a school on the basis of distance along a network (roads) rather than distance along a straight line.
- Students would attend school only in their county of residence.
- Students would be assigned to a school according to the school's capacity for their grade.
- No special allowances would be made to assign siblings to the same school, other than the fact that they shared the same geographic location and therefore should be assigned to the closest school that had capacity for their grade levels.

To assign students to public schools, the following data sets were required:

- The location and enrollment by grade of all public schools in the United States,
- A spatial data layer of public roads, and
- The location and age of eligible persons.

School locations and enrollment. The National Center for Education Statistics (NCES)[9] maintains downloadable files, available at http://nces.ed.gov/ccd/bat, that contain information about all known public schools in the United States. Using this Web site, we were able to retrieve enrollment data by grade for each US public school, as well as additional information, including the school's name, address, and NCES School ID. These school data were converted into a text file and sent to Tele Atlas, North America, to be geocoded according to the school's address. With the Environmental Systems Research Institute's (ESRI's) ArcGIS software product, we used the latitude and longitude pairs in the geocoded data set to convert the text file into a spatial data layer. We interactively processed this spatial data layer to resolve any ambiguous geocoding results (e.g., address not found, Post Office box address), using a variety of Internet mapping resources and aerial photography. We also checked to ensure that, at a minimum, the school was located in the correct county. Spot checking revealed that, overall, schools were very well located. After the geographic locations were checked, the data were loaded into a SQL Server database running ESRI's ArcSDE middleware.

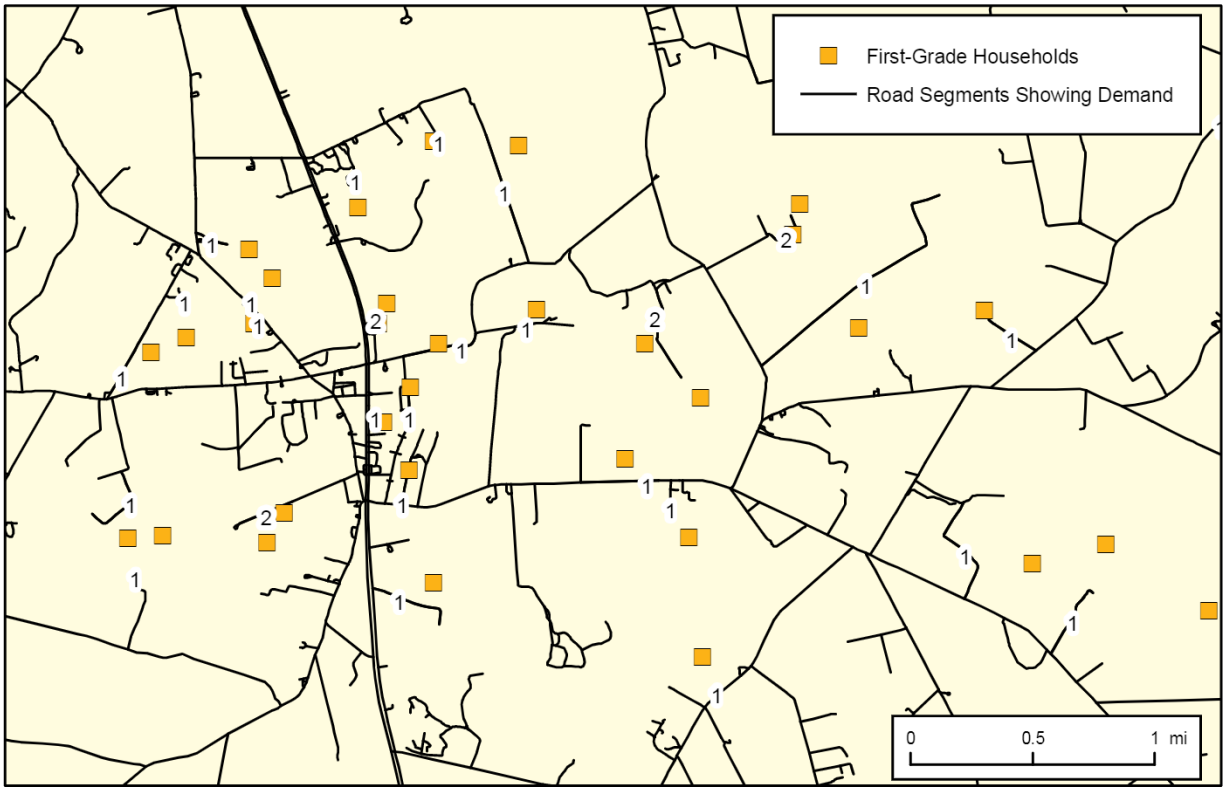
Roads. We selected the LocateAllocate command from within the ArcPlot module of the workstation version of ArcGIS to assign students to schools. One of the required parameters was specifying a network along which the assignments could occur. In this case, the roads network provided the means by which potential students were connected to their

neighboring schools. The US Census Bureau's 2000 TIGER/Line files were the source of the roads data.

Location and age of eligible persons. The three parameters that guided the extraction from the MIDAS synthetic population data set were enrollment, age, and location. All public school students who were 4 to 17 years old and who lived in a given county were extracted to a text file. Because each person was associated with a household, and each household had a latitude and longitude, the household latitude and longitude were transferred to each person. A geographic information system was then used to create a spatial data layer of eligible public school–aged students.

Technical approach. A fundamental assumption was that students could attend schools only in their county of residence; therefore, school assignments were performed one county at a time. Persons aged 4 to 17 years whose enrollment code was 2 (public) were extracted from the synthetic population database; public schools with total enrollment greater than zero were extracted from the schools layer; and all roads were extracted from a nationwide spatial data layer of TIGER/Line roads. Because the allocation occurred along a network, both the demand (enrollment) and the supply (students) values were transferred to the spatial data. The enrollment for each school was known by grade, so this information was transferred to the closest node in the network. Similarly, the number of students eligible to be assigned to a specific grade was transferred to each road segment. This assignment entailed finding the road segment nearest to each household and then generating a count of students by internal road segment ID. The count of eligible students was then transferred back to the segment so that each segment possessed the number of students available for a given grade (Figure 1).

Figure 1. Count of grade-eligible students by road segment



Notes: Squares represent the individual households where first-grade students lived. The total number of available students is labeled for each segment. More than one grade-eligible student could live in a household.

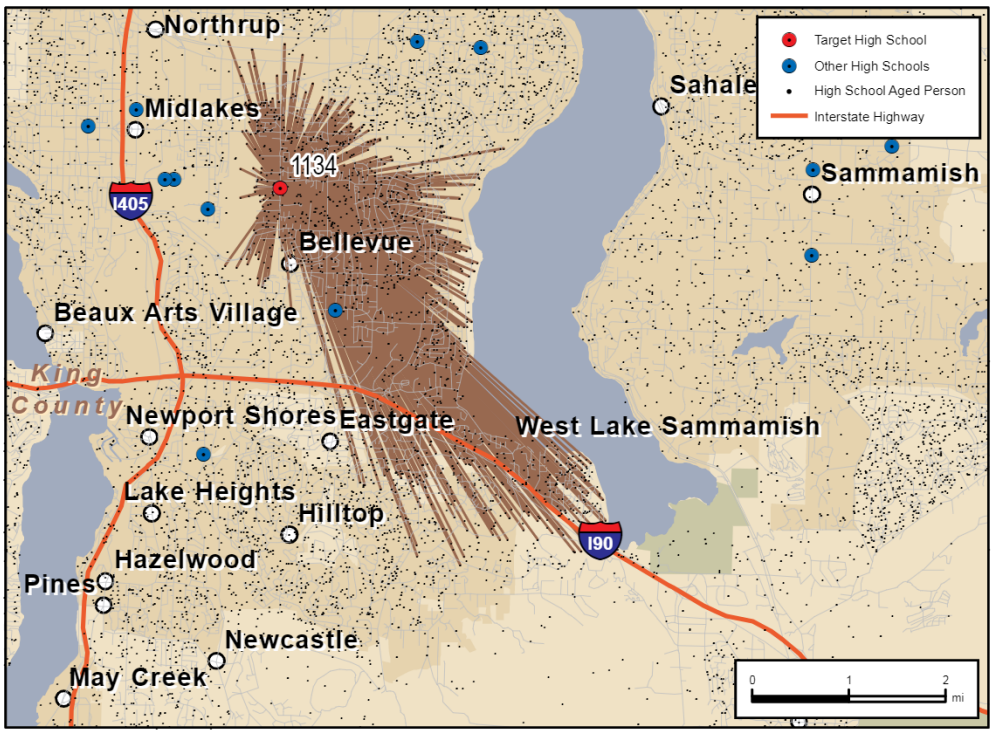
After these values were populated, we used the ArcPlot LocateAllocate command, which simultaneously assigned road segments to schools according to their proximity, until the capacity of the grade for a given school was reached. This command stopped when all grades reached their capacity (or when the network ran out of students). The resulting route system contained a network of road segments, with each segment assigned to a school. Through a series of data table relates, the NCES School ID was transferred from the schools layer to the road segment and, finally, to the individual student.

For each grade (kindergarten through 12) subjected to this process, students usually remained who were not assigned to a public school. Typically, students were unassigned because the schools closest to them were filled to capacity, but also, possibly, the road network did not connect to an eligible school (sometimes roads in adjacent counties were required to make a network connection), or the public school inventory was incomplete and therefore the overall capacity was too low. To compensate for these shortcomings, and in an effort to assign every eligible student to a school, we created a public school post-processing step that assigned leftover students to schools that were both close in proximity and had the highest capacity.

This post-processing step started with the creation of a distance matrix of each unassigned student to all public schools that taught the grade that he or she would attend. The distance was then multiplied by the school's enrollment for that grade. In this way, a measure of accessibility was created. For example, if two schools had the same enrollment for the same grade but one was half as far, then the student was assigned to the closer school. However, if one school had 10 times the enrollment and was twice as far as the second school, then the student was assigned to the school that was larger and farther away. This post-process step allowed larger schools to be proportionately overfilled, while still allowing for the student to be assigned to a school that was in close proximity.

Although the overall distribution of students assigned to schools produced a pattern of concentration, there were no definitive geographic boundaries. The processing and post-processing steps allowed students of the same age that lived near each other to be assigned to different schools. This situation might have occurred when one school was filled to capacity but another one was not. Figure 2 presents a distribution of public high school assignments in King County, Washington (Seattle). This type of

Figure 2. Students assigned to a single high school in King County, Washington



Notes: The map shows the distribution of high school–aged students assigned to the target high school. Other high school and eligible student locations are shown for reference.

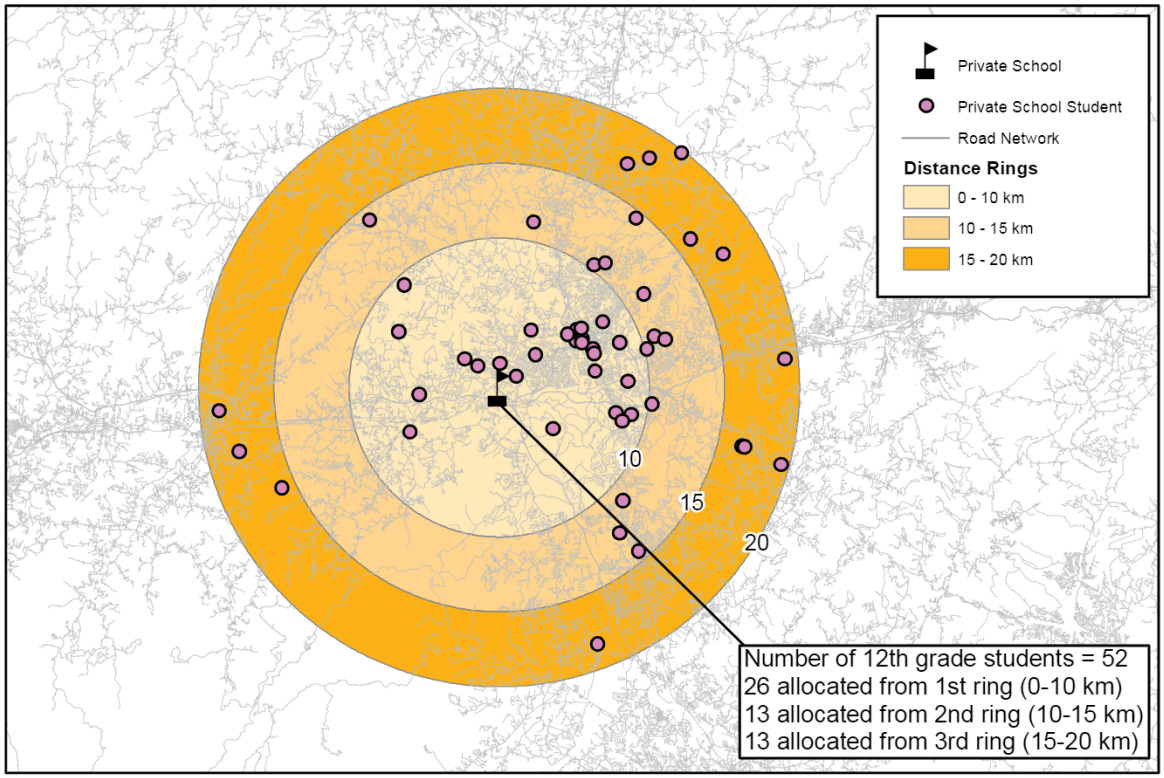
distribution typifies school assignments throughout the United States.

Private School

The assumption for private-school assignments was that students might attend a school anywhere within their state of residence but, all other things being equal, would more likely attend a school closer to their residence. To start this process, persons aged 4 to 17 years whose enrollment code was 3 (private) were extracted from the synthetic population database, and private schools with total enrollment greater than zero were extracted from the schools layer. The private-school assignment process divided the area around each private school into three concentric rings (Figure 3). The first ring extended out from the school location to a distance of 10 km. The second ring encompassed an area starting at 10 km and extending out to 15 km. The third ring extended from 15 km to 20 km. The assignment process used the ArcPlot Reselect command and selected 50 percent of the students from the first ring, 25 percent from the second ring, and the remaining 25 percent from the third ring. We established these proportions after running the command several times in rural, suburban, and urban areas and after reviewing the results for assignment completion rates.

Despite our having established these proportions, at the end of the assignment process not all privately enrolled students were assigned to a private school. The unassigned portion was a function of private schools' reaching their capacity or privately enrolled students' living more than 20 km from a private school that taught their grade. To assign these unassigned students, we created a private-school post-processing step that selected unassigned students and assigned them to the nearest private school within 80 km that taught their grade. Although this step made the allocation much more complete, it did overfill some schools and still left students beyond 80 km unassigned.

Figure 3. The private-school allocation process



Notes: The map shows the distribution of students selected to attend 12th grade at one particular private school. Half the students came from within a 10 km radius of the school, while a quarter came from each of the next two 5 km rings. Post-processing allowed students to come from up to 80 km away if no private schools existed within 20 km of the student's location.

Workplace

We could not assign workers to workplaces the same way we assigned students to schools. Not only did we lack a realistic way to create accurate locations for workplaces, but also the underlying assumption that people lived close to their places of work was not necessarily true. For these reasons, we decided to simply create appropriately sized virtual workplaces and then randomly select and assign workers to them. An advantage to making the assignments this way was that the process could be conducted entirely within a database, using SQL commands and Python scripting. To write a program to perform this type of workplace assignment, we required two pieces of information: (1) a count of the number of persons who lived in one Census tract but worked in another and (2) a count of firms by firm size, by the same Census geography.

The US Census Bureau has published a data file entitled Census 2000 Special Tabulation Product 64 (STP64),[10] which summarizes the number of persons by Census tract of work and Census tract of residence, combined. The commercial company InfoUSA has compiled the number of US firms by firm size category and Census block group. These two data sets fulfilled our data input requirements.

The first task was to account for persons not assigned a specific work tract in the STP64 data. In most cases the STP64 data file contained the number of persons who lived in one Census tract and worked in another. However, some records listed the number of persons living in one Census tract and working somewhere in that same county. To assign these workers to workplaces, we apportioned them to tracts within the county in accordance with the number of other persons working in those tracts. In this way, those tracts that already employed the most people received most of these additional workers. We also calculated the percentage of persons who were 15 to 54 years old and who were 55 to 74 years old for each Census tract, using basic US Census counts. This determination allowed the program to assign workers proportionately in areas with a larger proportion of older people (e.g., retirement areas) and in areas with a larger proportion of younger people (e.g., college towns).

The InfoUSA data contained counts of firms by the number of employees. The data summed the number of firms by using the following size categories:

- 1 to 4
- 5 to 24
- 25 to 99
- 100 to 499
- 500 to 4,999
- 5,000 or greater

In some cases the same total number of employees was listed for a given tract, but with no breakdown for the number of firms in each category. In these cases an exponential distribution was assumed, where each smaller category had twice as many firms as the next largest.

Each firm size category was assigned a mean number of employees as a starting point (Table 3). We calculated these mean numbers at a national level by dividing the total number of persons working in a given size category by the number of firms in that same category. Because the national level data were incomplete, we made an ad hoc decision to use the mean rather than another measure, such as median or mode.

The primary challenge of assigning workers to workplaces was that the STP64 data file count of persons working in a tract did not match the number of job slots available in the same tract according to the InfoUSA data (i.e., if we multiplied the number of firms in each size category by the mean number of employees in each category). For example, if a single firm existed in each size category for a given Census tract, then the total job slots in that tract would be 9,231 (2 + 9 + 50 + 170 + 1,000 + 8,000). If the two data sets matched, then the STP64 data file would have 9,231 persons available for these job slots; however, the number of people available for jobs rarely, if ever, matched the number of job slots available.

Table 3. The size category and national mean number of employees

Size Category	Mean
1 to 4	2
5 to 24	9
25 to 99	50
100 to 499	170
500 to 4,999	1,000
5,000 or greater	8,000

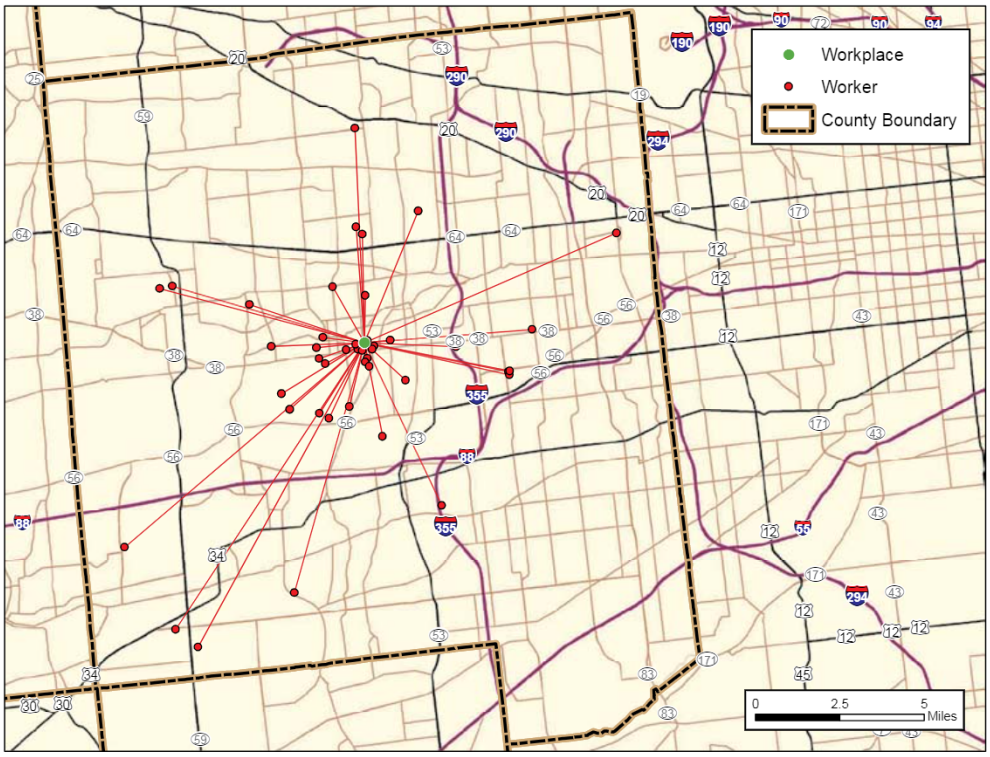
To compensate for this mismatch, we needed to either adjust the values in one table to agree with the other table, or adjust both tables to come to an average of the two. After examining both data sets, we judged that the US Census Bureau's STP64 data file table was more reliable than the InfoUSA's table; therefore, we adjusted the average of each size category by taking the ratio of the total number of STP64 workers to the total number of InfoUSA slots so that, when each was summed, the number of available slots matched the number of persons eligible for work. If, for example, twice as many slots as workers existed, we reduced the average number of slots in each category by a factor of 2. Conversely, if half as many slots as workers existed, then we *increased* the average number of slots in each category by a factor of 2. In this way, the overall distribution of firm sizes (mostly large, mostly small, or some of each) was maintained. Only the average number of workers assigned to each firm was adjusted. The only exception to this rule was to maintain at least two workers per firm in the smallest size category. When these sizes were held to a minimum value, we reduced the firm sizes in the other categories to compensate.

The program we wrote created a serially generated Workplace ID for each workplace. The Workplace ID consisted of the place-of-work Census tract ID and a 6-digit number, starting with the number 000001. The program assigned the Workplace IDs to persons using a two-step process. The first step went through the synthetic population table in two passes: one pass for persons aged 15 to 54 years and the second pass for persons aged 55 to 74. The program randomly selected the correct number of persons who did not already have a Workplace ID and coded them with the Census place-of-work tract ID.

The second step extracted the persons who had the correct residence tract ID and the correct Work Tract ID and created a table. This step also created a table of Workplace IDs, with multiple records with the same Workplace ID being created for each employee working at the target firm. For example, if a firm had 170 employees, the same Workplace ID would be generated and written out 170 times to create 170 new records. At the end of this process, the two files were joined and the Workplace ID was calculated into the table containing the extracted persons. This merged file was then linked back via the Person ID to the master synthetic population table, where the Workplace ID was updated.

An example of a typical workplace assignment distribution, Figure 4 shows the workers that were randomly selected to work at a given virtual firm.

Figure 4. The worker assignments for a given workplace



Notes: The map shows the distribution of workers assigned to a single workplace. When totaled, the number of workers living in one Census track and working in another matched the Census 2000 Special Tabulation Product 64 data tables.

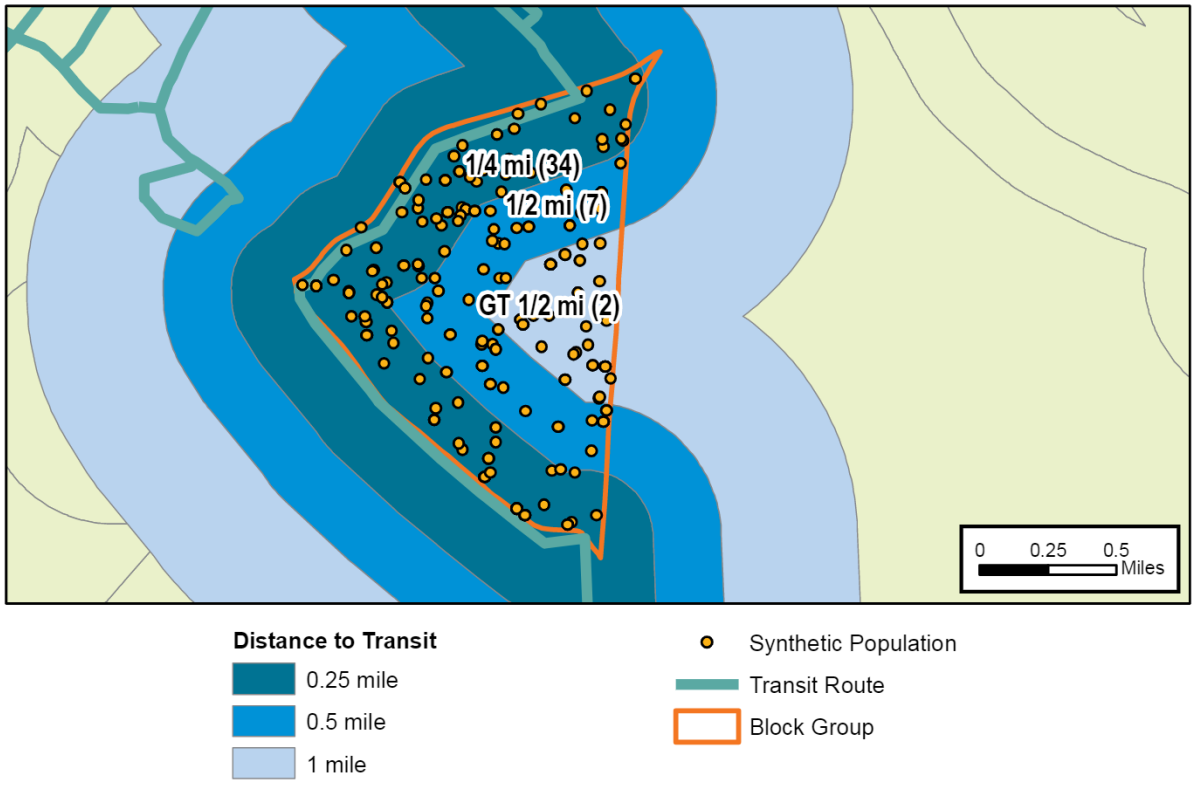
The location of the workers (their Census tract of residence) mirrors the STP64 data for this same parameter.

Public Transit

The way we made public transit assignments in support of disease propagation modeling varied with the availability of input data. In some areas of the United States, public transit is widely used and the route data are freely available. In other areas public transit systems may not exist, or, if they do exist, then transit data are difficult to obtain. We developed two methods of assigning persons as public transit riders: one method was based on geospatial data of bus routes to assign people in close proximity, and a second method was based solely on tabular data. Both methods depended on the US Census Bureau's commuting data.[11] Specifically, persons aged 16 or older who indicated that they worked at some time during the previous full calendar week (Sunday through Saturday) were asked to indicate their main mode of travel or type of conveyance from home to work during the week. Data were tabulated at the block group level and are available to the public in the Summary File 3 (SF3) Census data file.

If bus (or subway) routes were available in geospatial format, we created two buffers along the route. The first buffer was 0.25 miles on either side of the route centerline, and the second buffer was 0.25 miles to 0.5 miles on either side. The size of these buffers reflected the results of many public transportation surveys stating that most people are willing to walk approximately 0.25 miles to a bus stop.[12-14] These same surveys found that approximately 80 percent of ridership came from this 0.25 mile distance. We used the same distance decay factor (that ridership would decline an additional 80 percent over the next 0.25 miles) to determine that 16 percent of the ridership would come from between 0.25 miles and 0.50 miles. The remaining 4 percent came from a distance of greater than 0.5 miles. These buffers were overlaid with the Census block group boundaries to form smaller polygons. The buffers that we created for a sample block group are shown in Figure 5.

Figure 5. The public-transit buffers and synthetic population for a given block group



Notes: The map shows the synthetic population distributed in a single Census block group that has 43 public-transit riders in it, according to the Census data. According to our algorithm, 80 percent of riders (34) would come from a distance of 0.25 mile or less, 15 percent (7) would come from a distance of 0.25–0.50 mile, and the rest (5 percent, or 2 persons) would come from a distance of 0.25 mile or farther.

By following this process, we were able to use a program to randomly select more persons living close to a bus route than persons farther away. We added to the composite layer an attribute field that indicated how many persons to select within each respective polygon. To populate this field, the bus ridership value assigned to the entire block group was multiplied with the previously mentioned buffer weights now associated with the smaller-buffer-distance polygons within each block group. These values were normalized so that the sum of all the small polygons equaled the number of total riders in each block group. We then wrote a GIS program to randomly select the correct number of synthetic people in each polygon and set the public-transit flag to 1. Figure 5 shows the number of persons who were bus riders by buffer distance. This Census block group contained 43 bus riders. The process we created would randomly select 34 bus riders (80 percent) from the 0.25 mile buffer, 7 (16 percent) from the 0.25–0.50 mile buffer, and 2 (4 percent) from the buffer of 0.5 miles or farther.

In the absence of geospatial transit route data, we still were able to designate persons as public-transit riders. The number of people who took a given type of transit was tabulated by Census block group in the STP64 data file; therefore, we created a Python/SQL program, similar to the workplace assignment program, that randomly selected the correct number of people in the appropriate Census block group and set their public-transit flag to 1. Although this process created the correct the number of public-transit riders, the riders were randomly located within their Census block group, not clustered more closely to routes as they would have been if we had used the first method.

Results

We performed school assignments for 52,146,712 US children aged 4 to 17 years. The success rate was 99.13 percent, given the total number of eligible students in the PUMS data: 52,605,801. These assignments encompassed 111,228 schools nationwide. We assigned 90.04 percent of the children to public schools and the remainder (9.96 percent) to private schools. The high overall assignment rate was also quite high among individual states: The lowest assignment rate was 97.05 percent (South Dakota), while several states' assignment rate was 100 percent (Arizona, Hawaii, Illinois, Connecticut, Massachusetts, and the District of Columbia). The primary reason for children's being left unassigned was the absence of schools with the appropriate grade capacity within the post-processing distance threshold (20 km for public schools and 80 km for private schools). This problem was more prevalent in rural areas than in urban ones.

Of the adult working population aged 15 to 74 years, a total of 65,764,162 were employed in 2000 according to the Census. Using our method of assignment, we were able to assign 65,442,190 (99.51 percent) of the synthetic population to workplaces. On a state basis, the assignment percentage ranged from 97.4 percent (Nevada) to 100 percent (Arizona and Utah). The primary assignment problem was the matching of the number of job slots with the number of persons working in a given Census tract. This problem was mainly with the number and size of the firms in the InfoUSA data set, which was incomplete in many parts of the country. Using our adjustment approach, however, we were able to match the distribution of workplace sizes to the InfoUSA data while still assigning workers to match the counts contained within the STP64 data for 2000. Because the assignments were random among working-aged adults, no bias existed in age or gender of assigned worker by workplace size. Consequently, the distribution of worker age and gender in each of the six workplace size categories mirrored these distributions in the local population.

We conducted transit assignments for bus and subways in two specific areas of the United States (central North Carolina and New York City). We made assignments by using buffers when geospatial transit data were available and by using Census ridership data at the block group level when these data were not. In both cases, 100 percent of the ridership was assigned to the synthetic population. These assignments were also random within the working-aged population and therefore were not biased towards age or gender.

Conclusions

The school assignment process was challenging because the public school system assignment methods differ substantially between districts, as well as differing from private school methods. The bases of variability include the degree of emphasis on neighborhood schools, busing, and socioeconomic mixing. Therefore, we implemented a consistent, repeatable method for public school assignment based on proximity. This method created a school assignment pattern that clustered students around their schools. Consequently, although our method does not capture the heterogeneity that likely exists within the school systems, it does capture the variety of daily group interactions of children between households, on the basis of the common school assignments that are required to successfully model the spread of infectious disease in the community.

The most common daily adult group interactions occur in workplaces and transit systems. The characteristics of these interactions are quite different from a disease modeling perspective, and the nature of the assignment and assignment process reflected this difference. In the workplace, interactions are likely to occur regularly between the same groups of people, whereas in transit systems the interactions are much more random. Using the InfoUSA data set, we created a representative distribution of firm sizes in each Census tract and then adjusted the number of people assigned to each firm to match the STP64 Census data. This process allowed us to retain the relative distribution of small, medium, and large firms in a given area, while still assigning all workers to firms. This approach worked well, except in cases in which the InfoUSA firm data were severely lacking or absent. Using the Census counts to drive the process, rather than the number and size of firms, was consistent with our goal to prioritize the creation of person-to-person contacts and not the economic impact of altering the number and size of firms.

While both the school and workplace assignments benefitted from the existence of standardized, national data sets, the transit assignment process was impeded by the lack of consistent data from system to system. We presented two methods: one that used available geospatial data to designate individuals as transit riders who lived close to transit routes, and a second that randomly designated transit riders within block groups according to the number tallied by the Census. Both methods yielded 100 percent assignment rates and are equally useful for disease models.

The methods used for these three types of assignments differ because the underlying available data differed. Although they each presented their own challenges, they were each successful. This suggests that future researchers may be able to adapt these techniques to create more types of social contact assignments, such as places of worship or other points of congregation.

References

1. Axelrod R. The complexity of cooperation: agent-based models of competition and collaboration. Princeton, NJ: Princeton University Press; 1997.
2. Ferguson NM, Cummings DA, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. Nature. 2006 Jul 27;442(7101):448-52.
3. Halloran ME, Longini IM, Jr., Nizam A, Yang Y. Containing bioterrorist smallpox. Science. 2002 Nov 15;298(5597):1428-32.
4. Longini IM, Jr., Halloran ME, Nizam A, Yang Y, Xu S, Burke DS, et al. Containing a large bioterrorist smallpox attack: a computer simulation approach. Int J Infect Dis. 2007 Mar;11(2):98-108.
5. Wheaton WD, Cajka JC, Chasteen BM, Wagener D, Cooley PC, Ganapathi L, et al. Synthesized population databases: a US geospatial database for agent-based models. Research Triangle Park, NC: RTI Press; 2009.

6. Longini IM Jr, Halloran ME. Strategy for distribution of influenza vaccine to high-risk groups and children. Am J Epidemiol. 2005 Feb 15;161(4):303-6.
7. US Census Bureau. 2000 census of population and housing. Public use microdata sample [Internet]. Washington, DC: US Dept. of Commerce, Economics and Statistics Administration; 2003 [cited 2010 Sep 3]. 723 p. Report No.: PUMS/16-US (RV). Available from: http://www.census.gov/prod/cen2000/doc/pums.pdf
8. Saporito S, Sohoni D. Coloring outside the lines: racial segregation in public schools and their attendance boundaries. Sociol Educ. 2006 April;79(2):81-105.
9. National Center for Education Statistics. Common Core of Data Build a Table [Internet]. Washington, DC: US Dept. of Education [date unknown] [cited 2010 Sep 3]. Available from http://nces.ed.gov/ccd/bat
10. US Census Bureau. Census 2000 special tabulation: Census tract of work by Census tract of residence (STP 64) [CD-ROM]. Washington, DC: US Dept. of Commerce, Economics and Statistics Administration; 2004 [cited 2010 Sep 3]; Available from: http://www.census.gov/mp/www/cat/decennial_census_2000/census_2000_special_tabulation_census_tract_of_work_by_census_tract_of_residence_stp_64.html
11. US Census Bureau. Census 2000 summary file 3 (SF 3)—sample data [Internet]. Washington, DC: US Dept. of Commerce, Economics and Statistics Administration; [date unknown] [cited 2010 Sep 3]. Available from: http://factfinder.census.gov/servlet/DatasetMainPageServlet
12. Seneviratne PN. Acceptable walking distances in central areas. J Transp Eng-Asce. 1985;111(4):365-76.
13. Dittmar H, Ohland G, editors. The new transit town: best practices in transit-oriented development. Washington, DC: Island Press; 2004.
14. Ewing R. Pedestrian- and transit-friendly design: a primer for smart growth. US Environmental Protection Agency, Urban and Economic Development Division. Washington, DC: Smart Growth Network; 1999.

Acknowledgments

The project described was supported by grant number U01GM070698 (Models of Infectious Disease Agent Study—MIDAS) from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institute of General Medical Sciences or the National Institutes of Health.

www.rti.org/rtipress **RTI Press publication MR-0019-1009**