

Putting data on the map - data science seminar highlights synthetic populations

By Gabriel Knudsen

Bill Wheaton, director of the Geospatial Science and Technology Program at [RTI International](http://www.rti.org/), presented a July 30 NIEHS Data Science Seminar Series lecture, discussing "Synthetic Populations: Concepts, Implementation, and Applications in Environmental Health Sciences." The talk was hosted by Allen Dearry, Ph.D., director of the NIEHS Office of Scientific Information Management.

Wheaton uses computerized representations of households or individuals, known as synthetic populations, to build realistic models portraying population statistics, geography, and behavior. These models can support public health decisions, such as those necessary to control outbreaks of infectious diseases or mount response to natural disasters.

At RTI, Wheaton oversees the [Synthetic Microdata Household Viewer](http://synthpopviewer.rti.org/),

(<http://synthpopviewer.rti.org/>)

developed in conjunction with the Models of Infectious Disease Agent Study at the National Institute of General Medical Sciences.

"In order to protect privacy, the interactive map doesn't show actual households in their exact locations like Google Earth. Nonetheless, the data represent real households in reasonably accurate detail, enabling the map to show complex population distributions," he said.

Computer models employing synthetic populations can be used to provide realistic information for public health decision-makers, by forecasting conditions, such as the aging of a population, in combination with an element of interest, such as obesity or heart disease.

Building models is difficult work even when laboratory data are available. Building models based on field data is doubly so. Disease modeling has often applied a measured value, such as infections in prior years, to current population data. Wheaton explained how using synthetic populations allows for more precise health forecasts. For example, information such as age, household size, school location, and other statistics were combined with geographical data to test the efficacy of closing schools for various numbers of days, in an exercise to model the spread of a virus.

Linked Video

[Watch "Stats in Action," a U.S. Census Bureau video in which Wheaton discusses the use of American Community Survey data in synthetic population generation. \(04:39\)](#)

Finding the right data

Wheaton works with microdata, data that are not aggregated, but are read on an individual household level with identifying information removed. In contrast, most users of census data analyze aggregated values, typically in the form of a given number of households with a given characteristic, such as income and household size, within a given geographic area.

These microdata can then be used to model an entity or behavior. Most household and personal microdata that RTI uses come from the American Community Survey and the 2010 U.S. census.

Mapping data on food insecurity

Another method for generating the microdata uses probabilities to interpolate individual data points from marginal data. National Health and Nutrition Examination Survey (NHANES) data are being used in this way to map food insecurity, by combining data on gender, race and ethnicity, income, and education levels, and mapping probabilities of food insecurity for different members of the synthetic population. Food insecurity means that people do not always know where they will get their next meal.

(Gabriel Knudsen, Ph.D., is an Intramural Research Training Award fellow in the National Cancer Institute Center for Cancer Research Laboratory of Toxicology and Toxicokinetics.)



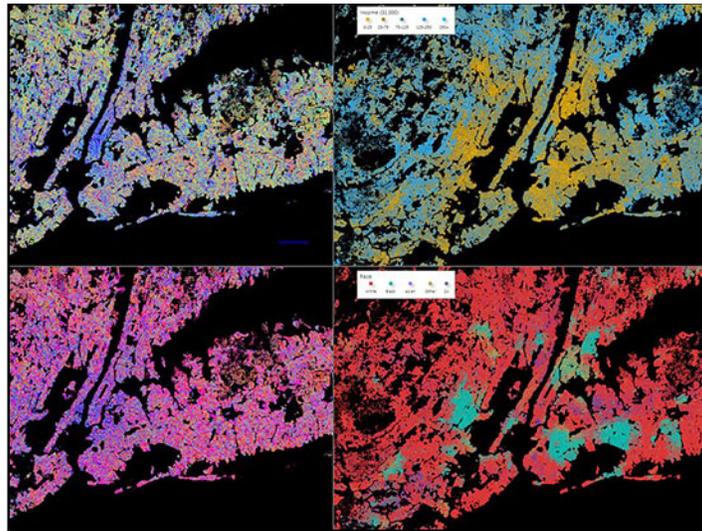
As Dearry explained, "This is the first in a new series of data science seminars, that will introduce researchers to the challenge and advantages of sharing and reanalyzing datasets and moving along the continuum of data, to information, to knowledge." (Photo courtesy of Michael Garske)



According to Wheaton, synthetic populations can help emergency managers planning evacuation routes, such as from the coast during a hurricane, or from the vicinity of a nuclear power plant, because they can map traffic loads, as well as the likely number of households without personal transportation. (Photo courtesy of Michael Garske)



Wheaton demonstrated the synthetic population viewer modeling the hypothetical spread of influenza. (Photo courtesy of Michael Garske)



"By representing each and every household as a point on the map, a wealth of complex patterns becomes apparent," Wheaton said, describing the synthetic population viewer. (Photo courtesy of RTI International, Inc.)

The Environmental Factor is produced monthly by the [National Institute of Environmental Health Sciences \(NIEHS\)](http://www.niehs.nih.gov/)

(<http://www.niehs.nih.gov/>)

, Office of Communications and Public Liaison. The content is not copyrighted, and it can be reprinted without permission. If you use parts of Environmental Factor in your publication, we ask that you provide us with a copy for our records. We welcome your [comments and suggestions](#).

(bruskec@niehs.nih.gov)

This page URL: NIEHS website: <http://www.niehs.nih.gov/>

Email the Web Manager at webmanager@niehs.nih.gov