The Human Exposure Model (HEM)

Residential Population Generator (RPGen) Module

Technical Manual

September 2018

U.S. Environmental Protection Agency, Office of Research and Development

Prepared by

Kathie Dionisio[1]

Graham Glen[2]

Heidi Hubbard[2]

Jessica Levasseur[2]

Contributions from

Kristin Isaacs[1]

Paul Price[1]

Dan Vallero[1]

[1]U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory

[2]ICF

Table of Contents

ACKNOWLEDGMENTS AND DISCLAIMER

The United States Environmental Protection Agency through its Office of Research and Development funded and collaborated in the research and development of this software. This model and its default data are currently under development; this material has been distributed for evaluation purposes only. The model has not been cleared by the United States Environmental Protection Agency for general distribution. While example input data have been provided as an example, it is up to the user to verify appropriate input data are being used for a given application. This manual is draft documentation and has not been cleared for publication.

1. Introduction

1.1 Overview

The Residential Population Generator (RPGen) module generates a simulated population of individuals along with their corresponding individual and household characteristics and a description of their residence which is representative of the U.S. population. RPGen takes as input large, nationally administered databases representing U.S. demographic, household, and housing patterns.

1.2 Purpose of this Technical Manual

This Technical Manual is intended for use by scientists to understand the logic and scientific rationale implemented in RPGen.

2. Overview of Residential Population Generator

2.1 Inputs

The RPGen module takes as input national surveys of individual, household, and housing characteristics for the U.S. population. Though the various national surveys were conducted independently, the survey data is linked on key characteristics within the RPGen module such that variables from all surveys can be assigned in the output data set.

In the RPGen module, three national databases are linked: the Public Use Microdata Sample (PUMS), American Housing Survey (AHS), and Residential Energy Consumption Survey (RECS). PUMS is produced by the U.S. Census Bureau's American Community Survey. The version provided as a default input file for RPGen is the 5-year sample format, covering the years 2012-2014 inclusive, with nearly 9 million data records. The PUMS data includes data on personal income for household members as well as other population-level descriptors. These data are used to represent the demographic patterns of the U.S., to be replicated in the simulated population output by RPGen. Data from both AHS and RECs are used to provide additional information on the household level for the individuals being simulated. The AHS data provided as default input for RPGen are from the 2013 survey, and cover housing type, housing size, and housing age. The RECS data provided as default input for RPGen is from 2009 and focuses on variables related to heating, cooling, types of appliances, and other energy-consuming objects found in homes.

2.2 Outputs

Each time RPGen is run, a text file "pophouse.csv" is produced, which includes a detailed description of the individual, their household members, and characteristics of their residence, for the desired simulated population. For each primary individual in the simulated population, the pophouse.csv file includes the age and gender of the primary individual plus each other person living in the household. The output file also includes physiology related variables for the individuals.

A detailed data dictionary for pophouse.csv can be found in Appendix B.

3. Implementation

3.1 Generation of variables

In many cases, use of consumer products by the primary product user, in addition to bystander exposure in a household when products are used by other household members, will vary by life stage, and by characteristics of the individual's housing situation. As such, we aim to define bins of household

composition and housing characteristics which are most likely to capture differences in an individual's use of consumer products, as well as the background or bystander exposure due to product use in the household. Additionally, we bin households to assist with linking the three input datasets based on the common variables present in all input datasets. The PUMS, AHS, and RECS input datasets all include variables pertaining to the household composition, specifically the total number of occupants of a household, and additional variables which allow for the logical assignment of the number of adults and children in the household.

The PUMS dataset is initially used to generate the simulated population. PUMS includes linked population and housing data. Separate apartments in the same building are considered separate housing units. Statistical sampling weights provided by PUMS were used to balance the selection probability at the individual level.

The PUMS survey includes one record per person, with the potential for multiple records per household.

RECS and AHS contain one record per household. The housing portion of the AHS dataset is a random survey of potential dwellings, and thus includes empty houses. Due to the original survey design, the dataset is no more likely to include a house with many occupants over a house with few (or no) occupants. By assigning a compatible house to each person selected from PUMS, the RPGen sample becomes representative of the overall population.

Please note that though variables generated in this Module and described below may relate to estimation of a home's air exchange rate, the air exchange rate is not calculated or estimated in the RPGen Module. For assignment of air exchange rate for each household, please see the Source-to-Dose Module Technical Manual.

Determination of household size and distribution of adults/children

Each PUMS record has a variable indicating the total number of persons living in that housing unit. Separate apartments in the same building are considered separate housing units. The total number of persons living in a housing unit ranges from 0-20, with lower numbers being far more common. For the purposes of the RPGen module, houses with no occupants were removed from consideration. The PUMS survey includes variables indicating number of children in the household who are related to the head of household, and a categorical variable with four options: no children, at least 1 child under 6 years, at least 1 child between 6-17 years, or at least 1 child under 6 years and at least 1 child from 6-17 years. For the purposes of the RPGen module, the number of children in the household was assumed to be either the number of children related to the head of household, or the smallest number consistent with the categorical variable. A minimum of one adult is required in each household. Adults were assumed to be individuals 18 years of age or older.

The AHS and RECS variables representing household composition were more straightforward, with RECS including variables providing the total number of individuals in the household, and the binned age of each member of the household. AHS included variables providing the total number of residents in the household, the number of children (<18 years), number of adults (≥18 years), and number of elders (≥65 years).

Binning of households by age distribution and number of household members

Due to available information within each survey on total number of household members and age and gender distribution, and corresponding potential for differentiated product use by age and gender, the determination was made to match survey records by grouping households into 4 bins based on household composition.

Table 1. Household composition bins

	Adults	Children
Bin 1	1	0
Bin 2	1	1+
Bin 3	2+	0
Bin 4	2+	1+

Statistical weighting variables and study design were taken into account to implement statistical sampling on a per-person basis, though the population and housing data will be linked on a per-household basis.

Location

Due to the inability, given current data, to determine differences in product use by fine-scale geographic regions (e.g., city or state), the RPGen module identifies location of an individual's home as being in one of 4 geographic regions (Northeast, Midwest, South, and West), and as being in a rural or urban setting. The PUMS data set from which individual demographic variables are sampled identifies the Public Use Microdata Area (PUMA) in which each individual resides. Using the U.S. Census Topologically Integrated Geographic Encoding and Referencing (TIGER) dataset, the population density of each PUMA was determined. The RPGen module then classifies an individual's location of residence as urban if the population density is >129.8 people/km^2, and rural for a lower population density. The four geographic regions were coded as 1, 2, 3, or 4 (defined as Northeast, Midwest, South, and West, respectively) and are defined in Appendix A.

Both the AHS and the RECS surveys include variables providing data on region of the country and rural or urban designation, to be used for binning and survey matching purposes.

Household Income

Dependent on product category, use of consumer products is sometimes correlated with wealth. In the RPGen module, household income was utilized as the indicator of wealth which may correspond to product use. However, because purchasing power associated with income is relative to cost of living, income was first ranked by region and urban/rural status, then assigned to bins corresponding to the households with the top, middle, and bottom third of income within that region and urban or rural designation. The number and size of bins related to purchasing power were chosen arbitrarily in the absence of data indicating how consumer product use varies with income.

House Type

A variety of housing types are identified in the AHS and RECS datasets. Within the RPGen module, housing types have been simplified and condensed as: single-unit (stand-alone) structures, multi-unit structures, and other (mobile homes, boats, etc.). It is believed that the largest influence of house type

on exposure will be in the determination of air exchange rates which influence indoor air concentrations. Additional impacts of housing type on product use relate to the presence or absence of required yard or outdoor maintenance, which are typically not required when one does not live in an owned, stand-alone unit. Additional descriptive variables such as if the household owns or rents can impact this distinction as well.

3.2 Linking of Survey Data/Simulation of the Population

Using the above defined key demographic and household related variables which can be identified in all datasets, the PUMS dataset providing demographic and household composition data was linked with the AHS and RECS datasets which each provided various housing characteristics. Records in each of the 3 surveys were binned on the variables defined previously. To link survey data, records from the same bins were matched from each of the surveys.

Surveys were linked based on 4 variables: location, household income, house type, and household composition. Bins utilized for each of the 4 key variables are identified in Table 2.

Table 2. Linking variables and bins

Linking variable	**Bins**
Location	Northeast, Midwest, South, West Urban vs. rural
Household income	Top, middle, and bottom third of household income by geographic region and urban/rural designation
House type	Stand-alone structure, multi-unit structure, other
Household composition	Household composition as defined by number of adults and children in the household, by age group (see Table 1)

3.3 Generation of physiology data

For the primary person in each household (randomly chosen, so it may be a child or infant), the httk R package is used (modified slightly for random number reproducibility) to generate a set of physiologic variables. These modifications include adding upper and lower bounds on variables such as height and weight to ensure that physiologically realistic descriptions are generated for the simulated population. Variables output in pophouse.csv include properties such as height, weight, skin area, organ masses, and blood flows to each organ. The data are available for later use, for example, by a model which tracks the chemical after it has entered the body, such as a physiologically-based pharmacokinetic (PBPK) model.

Appendices

A. Geographic region definitions

Region code	State name
3	Alabama
4	Alaska
4	Arizona
3	Arkansas
4	California
4	Colorado
1	Connecticut
3	Delaware
3	District of Columbia
3	Florida
3	Georgia
4	Hawaii
4	Idaho
2	Illinois
2	Indiana
2	Iowa
2	Kansas
3	Kentucky
3	Louisiana
1	Maine
3	Maryland
1	Massachusetts
2	Michigan
2	Minnesota
3	Mississippi
2	Missouri
4	Montana
2	Nebraska
4	Nevada
1	New Hampshire
1	New Jersey
4	New Mexico
1	New York
3	North Carolina
2	North Dakota
2	Ohio
3	Oklahoma
4	Oregon
1	Pennsylvania

1	Rhode Island
3	South Carolina
2	South Dakota
3	Tennessee
3	Texas
4	Utah
1	Vermont
3	Virginia
4	Washington
3	West Virginia
2	Wisconsin
4	Wyoming

B. Output files

Data dictionary for pophouse.csv output file

Please note, not all variables listed are used in subsequent modules of HEM. Variables are maintained however for potential future use in HEM, or if found useful when using RPGen to generate simulated populations for use in other modeling efforts beyond HEM. A '[-]' indicates the variable is unitless.

Variable Name	Description
Population Variables from PUMS	
gender	gender of selected person (primary individual); Male or Female
reth	ethnic group (httkpop categories)
compid	7-digit code; first two digits = state FIPS, last 5 digits = 2010 PUMA
recno	PUMS record number
race	W=White, B=Black, N=Native American, A=Asian, P=Pacific Islander, O=Other, M=Multiple
ethnicity	M=Mexican hispanic, O=other hispanic, N=not hispanic
age_years	age in full years, rounded down (range= 0 to 96)
pwgtp	statistical sampling weight
pool	combination of database matching variables family type, house type, income, census region, and urban/rural (range 1-288)
income	annual household income
ages	40-character string, each pair is age of one household member (range 00-96)
genders	20-character string, each is M or F
state	2-digit FIPS code for one of the 50 states or DC (range 01=Alabama to 56=Wyoming)
Housing Variables from RECS	
afuel	fuel used for air conditioning ; 1-electricity, 2=gas or propane, 3=other, -6=NA
baths	number of full bathrooms; 0-10 (capped at ten), -6=NA
bedrms	number of bedrooms; 0-10 (capped at ten), -6=NA
built	year house was built; each year for 1990+, rounded down to 5x for 1970-1989, rounded down to 10x for 1920-1969, earlier=1919
cars	number of cars; 0-5 (capped at 5), -6=NA
cellar	type of basement; 1=full basement, 2=partial basement, 3=crawl space, 4=slab, 5=other, -6=NA

Variable Name	Description
hequip	main heating equipment; 1=forced air furnace, 2=steam radiators, 3=heat pump, 4=electric baseboard, 5-14=others
lot	square footage of lot; range is 200 - 999,997 square feet (almost 22 acres)
pwt	statistical weight within AHS; used for random selection
rooms	number of rooms; 1-21 (capped at 21)
sewdis	type of sewage disposal; 1=septic tank, 2=chemical toilet, 3=outhouse, 4=other, 5=none, -6=municipal system
unitsf	square footage of house (excl. garage, unfinished areas); 99-99998 (minimum allowed=99, capped at 99,998)
water	source of water (for washing and bathing); 1=water system, 2=well, 3=spring, 4=cistern, 5=stream or lake, 6=bottled, 7=other
waterd	source of drinking water; 1=water system, 2=well, 3=spring, 4=cistern, 5=stream or lake, 6=bottled, 7=other
control	record number from full 2013 AHS database
Housing Variables from AHS	
doeid	record number from 2009 RECS database
nweight	sampling weight
hdd30yr	average annual heating degree days; range 0 – 13346
cdd30yr	average annual cooling degree days; range 0 – 5357
kownrent	own or rent house; 1=owned, 2=rented, 3 = stay without rent
condcoop	part of condo or coop; 1=condominium, 2=cooperative, -2=NA
naptflrs	number of floors in apartment; range 1-4, -2 = not an apartment
stories	number of stories in single-family home; 10=one, 20=two, 31=three, 32=4+, 40=split-level, 50=other, -2=not a single family home
stoven	number of oven-cooktop combinations; range 0-10
stovenfuel	fuel used for stove; 1=gas, 2=propane, 5=electric, 21=other
stove	number of cooktops (not combined with ovens); range 1-10
stovefuel	fuel used for cooktop; 1=gas, 2=propane, 5=electric, 21=other
oven	number of ovens (not combined with cooktops); range 0-10
ovenfuel	fuel used for oven; 1=gas, 2=propane, 5=electric, 21=other
ovenuse	frequency of oven use; 0=not used, 1=3+ per day, 2=twice per day, 3=once per day, 4 = few times per week, 5=1/week, 6=less
outgrill	outdoor grill used; 0=no, 1=yes
dishwash	dishwasher used in home; 0=no, 1=yes
cwasher	clothes washer used in home; 0=no, 1=yes
washload	frequency clothes washer used; 1=1/week or less, 2=2-4 per week, 3=5-9 per week, 4=10-15 per week, 5= 16+ per week
dryer	clothes dryer used in home; 0=no, 1=yes
dryruse	frequency clothes dryer used; 1=every time clothes washed, 2=sometimes when clothes washed, 3=rarely, -2=NA
tvcolor	number of televisions used in home; range 0-15
computer	computer used at home; 0=no, 1=yes
numpc	number of computers; range 0-15
pcprint	number of printers used at home; range 0-9, -2=NA
moisture	humidifier used at home; 0=no, 1=yes

Variable Name	Description
prkgplc1	have an attached garage; 0=no, 1=yes
prkgplc2	have a detached garage or carport; 0=no, 1=yes
cooltype	type of air conditioning system; 1=central, 2=window/wall, 3=both, -2=none
tempniteac	temperature setting at night (in warm weather); range 45-96, -2=no AC
numberac	number of window/wall AC units; range 1-15, -2=NA
numcfan	number of ceiling fans used; range 0-15
notmoist	dehumidifier used at home; 0=no, 1=yes
highceil	high ceilings in home; 0=no, 1=yes
windows	number of windows in heated areas of home; 0=none, 10=1-2, 20=3-5, 30=6-9, 41=10-15, 42=16-19, 50=20-29, 60=30+
adqinsul	level of insulation; 1=well insulated, 2=adequate, 3=poor, 4=none
drafty	home drafty in winter; 1=always, 2=mostly, 3=sometimes, 4=never
swim	swimming pool or hot tub; 0=none, 1=hot tub only, 2=pool only, 3=both
Physiological variables from httk	
mean_logh	mean of log(height) for this age-gender group
mean_logbw	mean of log(body weight) for this age-gender group
weight	body weight in kilograms; calculated from mean_logbw and logbw_resid
height	height in centimeters; calculated from mean_logh and logh_resid
blood_mass	mass of blood (kg)
brain_mass	mass of brain (kg)
gonads_mass	mass of gonads (kg)
heart_mass	mass of heart (kg)
kidneys_mass	mass of kidneys (kg)
large_intestine_mass	mass of large intestines (kg)
liver_mass	mass of liver (kg)
lung_mass	mass of lungs (kg)
muscle_mass	mass of muscular tissue (kg)
pancreas_mass	mass of pancreas (kg)
skeleton_mass	bone mass (kg)
skin_mass	mass of skin tissue (kg)
small_intestine_mass	mass of small intestines (kg)
spleen_mass	mass of spleen (kg)
stomach_mass	mass of stomach (kg)
adipose_flow	blood flow to adipose tissue [-]
brain_flow	blood flow to brain [-]
CO	Cardiac output (L/h)
gonads_flow	blood flow to gonads [-]
heart_flow	blood flow to heart muscle (not into heart) [-]
kidneys_flow	blood flow to kidneys [-]
large_intestine_flow	blood flow to large intestines [-]
liver_flow	blood flow to liver [-]
lung_flow	blood flow to lung tissue [-]

Variable Name	Description
muscle_flow	blood flow to muscles [-]
pancreas_flow	blood flow to pancreas [-]
skeleton_flow	blood flow to bone tissue [-]
skin_flow	blood flow to skin tissue [-]
small_intestine_flow	blood flow to small intestines [-]
spleen_flow	blood flow to spleen [-]
stomach_flow	blood flow to stomach [-]
other_mass	mass of other tissues (kg)
adipose_mass	mass of adipose tissue (kg)
org_flow_check	relevant to httkpop r package
weight_adj	adjusted body weight (sum of organ masses) (kg)
BSA_adj	adjusted body surface area (cm^2)
million.cells.per.gliver	Hepatocellularity, million cells/g liver
hematocrit	Percent volume of red blood cells in the blood
serum_creat	Serum creatinine, mg/dL
gfr_est	Estimated glomerular filtration rate, mL/min/1.73m^2 BSA
bmi_adj	adjusted body mass index
bmi	body mass index
BSA	body surface area (cm^2)