LABORATOIRE DE SIMULATION DÉMOGRAPHIQUE

THE UNITED STATES MICROSIMULATION MODEL (LSD-USA)

Methodological Document

Institut national de la recherche scientifique
Centre Urbanisation Culture Société
385, rue Sherbrooke Est
Montréal, Québec, Canada

Population Research Institute
The Pennsylvania State University
University Park, PA 16802

February 2019

THE UNITED STATES MICROSIMULATION MODEL (LSD-USA)

Principal Investigator

Alain Bélanger
Institut national de la recherche scientifique Urbanisation Culture Société
385, rue Sherbrooke Est
Montréal, Québec, Canada
H2X 1E3
514 499-4023
http://www.inrs.ca/alain-belanger

Collaborators and Research Assistants

Patrick Sabourin
Samuel Vézina
Guillaume Marois
Kevin D'Ovidio
Olivier Lafontaine
Jennifer Van Hook
Anne Morse

Model Version History

Changes in the first number indicate a complete restructuring of the model. Changes in the second number indicate the addition or the removal of a module, as well as major bug fixes. Changes in the third number indicate minor modifications to a module, as well as minor bug fixes.

Version 1.0.0

February 2019

Abstract

This document provides an overview of how the *LSD-USA* projection model works. *LSD-USA* is a microsimulation demographic projection model designed to project the United States population according to various characteristics. *LSD-USA* stands for *Laboratoire de Simulations Démographiques-United States of America*. The model has been developed by the authors at the *Institut national de la recherche scientifique* (INRS) in Montreal, Canada, in collaboration with researchers at the Population Research Institute at the Pennsylvania State University. More precisely, the *LSD-USA* model simultaneously projects demographic (age, sex, place of residence, place of birth, generation, immigrant status, age at immigration), ethnocultural (language spoken at home, English proficiency, race/ethnicity, religion) and socioeconomic (education, literacy, labour force participation) characteristics of the United States population. It allows for changes in individual characteristics over the life course as well as for intergenerational transfers of some characteristics from the mother to her child. This document describes the base population, data sources and methods for every projection modules contained in the model. The conceptualization of the model events and the derivation of the corresponding parameters are also described. This document provides a technical supplement to analytical reports and papers published in peer-reviewed journals presenting results generated by *LSD-USA*.

Keywords

Microsimulation; Population Projection model; Methodology; United States

Résumé

Le présent document donne un aperçu du fonctionnement du modèle de projections *LSD-USA*. Il s'agit d'un modèle de projections démographiques par microsimulation conçu pour projeter la population des États-Unis selon plusieurs caractéristiques. LSD-USA signifie *Laboratoire de Simulations Démographiques-United States of America*. Le modèle a été développé par les chercheurs de l'Institut national de la recherche scientifique (INRS) à Montréal, Canada, en collaboration avec les chercheurs du Population Research Institute de la Pennsylvania State University. Le modèle *LSD-USA* projette la population des États-Unis selon certaines variables démographiques (âge, sexe, lieu de résidence, lieu de naissance, statut de génération, statut d'immigration, âge à l'immigration), ethnoculturelles (langue parlée à la maison, connaissance de l'anglais, race/ethnicité, religion) et socio-économiques (éducation, littératie, statut d'activité sur le marché du travail). Le modèle permet de modifier dynamiquement les états des cas simulés et permet les transferts intergénérationnels des caractéristiques de la mère à ses enfants. Ce document décrit la population de base du modèle, les sources de données et la méthodologie utilisée pour chacun des modules de projection du modèle. La conceptualisation des événements, des états dérivés et les paramètres utilisés pour la modélisation sont également décrits. Ce document technique vient complémenter les rapports d'analyses et les articles scientifiques décrivant les résultats de projections générés par *LSD-USA*.

Mots-clés
Microsimulation, Modèle de projections démographiques; Méthodologie; États-Unis

TABLE OF CONTENT

General Description of the Model

Context

Most developed countries are facing similar demographic challenges: population aging, possible labor shortages and reduced population growth, if not population decline. Consequences of population aging and labour force population decline are posing a serious challenge to Western countries' policy makers regarding long-term sustainability of social security programs, healthcare and retirement plans. In response to these challenges, most developed countries have increased immigration levels such that current immigration has reached historical levels. International net migration is actually accounting for two thirds of the total population growth of developed nations. In consequences, the ethno-cultural makeup, especially in large urban areas, is changing rapidly, forcing political decision-makers to deal with a number of challenges in the areas of urban development, labor market integration, health and social services, and public institutions. In addition, with the replacement of older less educated cohorts by younger more educated cohorts, developed countries are also facing rapid changes in the educational composition of their labor force, generating potential mismatch between labor force demand and supply in terms of skills.

The presented U.S. microsimulation projection model is theoretically grounded on the idea that forecasting socioeconomic changes can be realized through population forecasts following the *Demographic Metabolism* theory developed by Ryder and more recently by Lutz. It puts forward the importance of cohort succession to explain social changes, building on the fact that the process of population replacement provides opportunity for social change through the constant flow of new people entering the social process and continuous withdrawals of older individuals through death.

The first objective of the project is to analyse the demographic and socio-economic behaviours differentials of ethnocultural groups in the United States using several data sources (censuses, vital statistics, social and labor force surveys). Estimated parameters from this analysis are then used to develop LSD-USA, a microsimulation projection model that simulates the future education attainment and labour force participation of the projected populations by ethnocultural characteristics. This is important because immigration alters the population demographically, socio-economically, spatially, and culturally. According to some authors, this process, described as a "Third Demographic Transition" can be a source of growing political conflict, cultural disunity and loss of community or cohesion. It poses challenges that should not be ignored or dismissed, as *laisser-faire* may lead to more fractionalisation and ultimately to more inequalities. The ultimate objective is thus to project populations beyond age and sex so that generated results may guide policy makers in their decisions regarding immigration policy, social cohesion, labor market needs and changes, poverty and inequalities, and education and language skill formation needs.

The Model and Its General Structure

LSD-USA is a dynamic microsimulation projection model of the U.S. population coded using the Modgen programming language. Its point of departure is 2015 and its starting population is based on the 2015 American Community Survey 1% sample, available from the University of Minnesota's Integrated Public Use Microdata Sample archive (ACS-IPUMS). The ACS-IPUMS is supplemented by three other data sets. The 2012/2014 US survey of the Program for the International Assessment of Adult Competencies (PIAAC) provides data on literacy, the Pew 2014 Religious Landscape Survey (RLS) provides data on religion, and the 1972-2012 General Social Survey (GSS) provides data on mother's education and generational status for ACS respondents under age 30 who do not reside with their mother. The reliance on publicly available databases distinguishes the LSD system of models from other models produced by Statistics Canada (such as Demosim) which are based on confidential microdata and whose code is not available to the general public. With LSD-USA, the goal is to create an accessible and transparent model, so that other similar models may be built for other countries, thus making international comparisons possible.

The LSD-USA model simultaneously projects demographic (age, sex, place of residence, place of birth, generation, immigrant status, age at immigration), ethnocultural (language spoken at home, English proficiency, race/ethnicity, religion) and socioeconomic (education, literacy, labour force participation) characteristics of the U.S. population. It allows for changes in individual characteristics over the life course as well as for intergenerational transfers of some characteristics from the mother to her child. Individuals from the base population are simulated one by one and their characteristics are modified through scheduled events whose timing is determined by the values of their specific input parameters at any given time during the projection period. The parameters are themselves derived from a variety of sources and methods: census, surveys, vital statistics.

Each event depends on a selection of characteristics. Interregional mobility, for instance, varies with age, sex, language spoken at home, race/ethnicity, generational status, and education. All events are thus highly interdependent.

The conceptualisation of the model events and the derivation of the corresponding parameters are further described below in the *Modules* section.

Modgen

Modgen is a meta language of C++ developed and maintained by Statistics Canada. Modgen and its documentation may be downloaded for free on the Statistics Canada website[1].

Modgen models are coded and implemented in the Microsoft Visual Studio software suite. Once compiled, a Modgen model takes the form of a stand-alone executable file (.exe) that allows manipulation of the model through a graphical user interface. From this interface the user is able modify simulation parameters and create customized scenarios.

[1] http://www.statcan.gc.ca/eng/microsimulation/modgen/modgen

LSD – USA is a microsimulation model programmed in Modgen. It is case-based, meaning that each individual is simulated separately from other individuals and that no interactions between individuals are allowed (with the exception of interactions between mother and children). The model is also dynamic and in continuous time, meaning that characteristics of individuals are modified continuously in "real time", in contrast to discrete-time models where characteristics are changed within predefined time units (typically one year).

For more details on Modgen, please visit Statistics Canada's website. For a full didactical presentation of Modgen, you may consult the book "Microsimulation and population dynamics: an introduction to Modgen 12" published by Springer.

Content of This Report

This methodological document presents a high-level technical overview of the LSD-USA model. It does not provide specific code or algorithms used to derive parameters or build modules.

The next section will describe the variables extracted from the ACS-IPUMS, PIAAC, and RLS and included in the microsimulation model. It will also provide simple descriptive statistics for these variables.

The following *Modules* section will describe in as much detail as possible the modules of the model in terms of their events, their parameters and their outputs. Interdependence between the different modules will also be described.

The Base Population

A base population is the database that is used as the starting point of a simulation: it is the synthetic representation of a population at a given time.

The base population in LSD-USA is constructed by recoding and extracting relevant variables from the 2015 American Community Survey 1% Integrated Public Use Microdata file (ACS-IPUMS). Because the ACS-IPUMS does not include information about literacy and religion, we added this information using cross-survey imputation. This technique borrows information about literacy and religion from supplemental data sources while preserving the relationships between these two variables and all of the demographic, sociocultural, and socioeconomic data in the ACS-IPUMS. Specifically, we used data from three supplemental data sources that contain data on literacy, religion, and mother's education and generational status, as well as all of the demographic variables included in the ACS-IPUMS. The 2012/2014 US survey of the Program for the International Assessment of Adult Competencies (PIAAC) provided data on literacy; the Pew 2014 Religious Landscape Survey (RLS) provided data on religion; and the 1972-2012 General Social Survey (GSS) provided data on mother's education and generational status for ACS respondents not residing with their mother. We harmonized the categories of the demographic variables in the PIAAC, RLS, and GSS to match those in the ACS-IPUMS. Next, we pooled the PIAAC, RLS, and GSS with the ACS-IPUMS and imputed literacy, religion, and mother's education and generational status for the individuals in the ACS-IPUMS. Finally, we dropped the observations from the PIAAC, RLS, and GSS leaving only the individuals in the ACS-IPUMS with all of their observed demographic, sociocultural, and socioeconomic characteristics and their imputed values for literacy, religion, and mother's education and generation. The extracted and imputed information is saved in a flat file (CSV) readable by Modgen.

The ACS-IPUMS contains 3,147,005 records, which corresponds to .98% of the total U.S. population in 2015. When weighted, these records amount to a population of 321,418,821. In the ACS-IPUMS, weight values range from 1 to 3,596, for an average of 102.1.

Each of the 3,147,005 records includes values for 28 variables, all of which are further described in the next section. Those variables includes age, sex, language spoken at home, English proficiency, immigrant status, education, age at migration, state, region, place of residence, race/ethnicity, place of birth, sampling weight, generational status, religion, literacy, and household and person identifiers; and resident mother's language spoken at home, English proficiency, immigrant status, education, age at immigration, race/ethnicity, mother's education, generational status, and religion.

Variables

Each record in the base population contains the information on 28 variables, all of which are presented below in the order in which they appear in the CSV file (this order is further specified in parenthesis as the position number).

Age: Single year of age (position 0)

Age is provided in single years ranging from 0 to 97.

Code	Description	n	%
[0,97]	Continuous age	3,147,005	100

Male: Sex (position 1)

M*ale* is directly recoded from the *SEX* variable in the ACS-IPUMS.

Code	Description	n	%
0	Female	1,610,169	51.2
1	Male	1,536,836	48.8

Language: Language Spoken Most Often at Home (position 2)

Language **most often** spoken at home is derived from the variable *LANGUAGE* in the ACS-IPUMS. English and Spanish are the first two categories of this variable. Languages other than the two languages are grouped in a third category *Others*. Values are imputed for the population aged 0 to 4.

Code	Description	n	%
0	English	2,554,131	81.7
1	Spanish	334,847	10.7
2	Others	238,709	7.6

Mom_Language: Mother's language Spoken Most Often at Home (position 3)

Language **most often** spoken at home by the person's mother is derived from the variable *LANGUAGE* in the ACS-IPUMS. English and Spanish are the first two categories of this variable. Languages other than English or Spanish are grouped in a third category *Others*. Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. This variable has values for the 834,051 persons who reside with their mother.

Code	Description	n	%
0	English	612,214	73.4
1	Spanish	134,049	16.1
2	Others	87,788	10.5

Speakeng: English Proficiency (position 4)

English proficiency is directly recoded from the variable *SPEAKENG* in the ACS-IPUMS. Those who report speaking English "well" or "very well" are coded as "English proficient." Those who report speaking English "not very well" or "not at all" are coded as "Not English proficient." Values are imputed for the population aged 0 to 4. The frequencies in the table below include only non-imputed values.

Code	Description	n	%
0	English only	2,553,834	81.7
1	English proficient	464,988	14.9
2	Not English proficient	108,865	3.5

Mom_speakeng: Mother's English Proficiency (position 5)

Mother's English proficiency is directly recoded from the variable *SPEAKENG* in the ACS-IPUMS. Mothers who report speaking English "well" or "very well" are coded as "English proficient." Mothers who report speaking English "not very well" or "not at all" are coded as "Not English proficient." Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. This variable has values for the 834,051 persons who reside with their mother.

Code	Description	n	%
0	English only	612,100	73.4
1	English proficient	155,920	18.7
2	Not English proficient	66,031	7.9

Immstat: Immigrant Status (position 9)

The immigrant status variable is recoded from the *BPL* variable in the ACS-IPUMS. Immigrants are defined as those who were born outside the United States or U.S. outlying areas.

Code	Description	n	%
0	U.S.-born	2,751,965	87.4
1	Foreign-born	395,040	12.6

Mom_immstat: Mother's immigrant Status (position 10)

Mother's immigrant status is recoded from the *BPL* variable in the ACS-IPUMS. Immigrants are defined as those who were born outside the United States or U.S. outlying areas. Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. This variable has values for the 834,051 persons who reside with their mother.

Code	Description	n	%
0	U.S.-born	647,312	77.6
1	Foreign-born	186,739	22.4

Ed: Educational attainment of persons age 25+ (position 11)

The education variable describes the highest level of educational attainment and is derived from the variable *EDUCD* in the ACS-IPUMS for all persons ages 25 and older. It contains four categories:

1. *No diploma*: includes individual without at least a high school diploma as well as individuals aged 0 to 14.
2. *High school*: includes individual with a high school degree or equivalency certificate.
3. *Associates*: includes individual with a two-year associates degree or occupational certificate.
4. *College*: includes individuals with a bachelor's degree.
5. *Professional degree*: Includes all individuals with degree beyond a bachelor's degree.

Code	Description	n	%
0	No diploma	271,562	12.3
1	High school	1,069,965	48.4
2	Associates	182,640	8.3
3	College	417,289	18.9
4	Professional degree	270,682	12.2

Momed: Mother's educational attainment (position 12)

The education variable describes the highest level of the respondent's mother's educational attainment and is derived from the variable *EDUCD* in the ACS-IPUMS for all persons living with a mother. It is imputed[2] for those under age 30 who do not live with a mother. Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. It contains four categories:

1. *No diploma*: includes individual without at least a high school diploma as well as individuals aged 0 to 14.
2. *High school*: includes individual with a high school degree or equivalency certificate.
3. *Associates*: includes individual with a two-year associates degree or occupational certificate.
4. *College*: includes individuals with a bachelor's degree.
5. *Professional degree*: Includes all individuals with degree beyond a bachelor's degree.

Code	Description	n	%
0	No diploma	223,002	14.6
1	High school	700,012	45.9
2	Associates	156,983	10.3
3	College	287,007	18.8
4	Professional degree	157,523	10.33

Ageimm: Age at immigration (position 15)

Ageimm provides the age at immigration of foreign-born persons, coded in single years ranging from 0 to 94. It is recoded from the *BPL* and *YRIMMIG* variables in the ACS-IPUMS. It is coded to zero for U.S.-born persons.

Code	Description	n	%
[0,94]	Age at immigration	3,147,005	100

[2] For ACS-IPUMS respondents under age 30 who do not live with their mother, mother's education and generational status is imputed simultaneously as a function of sex, region, age, place of birth, and race/ethnicity based on pooled ACS and GSS data.

Mom_ageimm: Mother's age at immigration (position 16)

Mom_ageimm provides the mother's age at immigration of foreign-born mothers, coded in single years ranging from 0 to 94. It is recoded from the *BPL* and *YRIMMIG* variables in the ACS-IPUMS. It is coded to zero for U.S.-born mothers. Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. This variable has values for the 834,051 persons who reside with their mother.

Code	Description	n	%
[0,94]	Mother's age at immigration	834,051	100

Statefip: State of residence (position 17)

This variable indicates state of residence. It is identical to the STATEFIP variable in the ACS-IPUMS.

Code	Description	n	%
0	Alabama	47,476	1.5
1	Alaska	6,619	0.2
2	Arizona	67,014	2.1
3	Arkansas	29,605	0.9
4	California	374,943	11.9
5	Colorado	53,570	1.7
6	Connecticut	35,787	1.1
7	Delaware	9,017	0.3
8	District of Columbia	6,610	0.2
9	Florida	194,548	6.2
10	Georgia	97,854	3.1
11	Hawaii	14,124	0.5
12	Idaho	15,725	0.5
13	Illinois	126,642	4.0
14	Indiana	66,045	2.1
15	Iowa	31,900	1.0
16	Kansas	28,774	0.9
17	Kentucky	44,749	1.4
18	Louisiana	43,892	1.4

19	Maine	13,059	0.4
20	Maryland	59,332	1.9
21	Massachusetts	68,785	2.2
22	Michigan	98,008	3.1
23	Minnesota	54,811	1.7
24	Mississippi	29,600	0.9
25	Missouri	61,586	2.0
26	Montana	9,841	0.3
27	Nebraska	19,089	0.6
28	Nevada	26,988	0.9
29	New Hampshire	13,378	0.4
30	New Jersey	87,815	2.8
31	New Mexico	19,072	0.6
32	New York	195,742	6.2
33	North Carolina	98,184	3.1
34	North Dakota	7,869	0.3
35	Ohio	118,123	3.8
36	Oklahoma	37,251	1.2
37	Oregon	39,992	1.3
38	Pennsylvania	128,145	4.1
39	Rhode Island	10,563	0.3
40	South Carolina	48,023	1.5
41	South Dakota	8,742	0.3
42	Tennessee	65,549	2.1
43	Texas	259,224	8.2
44	Utah	29,290	0.9
45	Vermont	6,326	0.2
46	Virginia	83,472	2.7
47	Washington	71,804	2.3
48	West Virginia	18,051	0.6

49	Wisconsin	58,578	1.9
50	Wyoming	5,819	0.2

Region_US: Region of residence (position 18)

Region describes region of residence. It is recoded from the *STATEFIPS* in the ACS-IPUMS as indicated below.

Code	Description	n	%
1	Northeast (Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont)	559,600	17.8
2	Midwest (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin)	680,167	21.6
3	South (Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia)	1,172,437	37.3
4	West (Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming)	734,801	23.3

Res: Detailed region of residence (position 19)

This variable provides a more detailed description of place of residence, based on metropolitan area of residence, state, or Census division. It is recoded from the MET2013, STATEFIP, and REGION variables in the ACS-IPUMS.

Code	Description	n	%
0	San Francisco Bay Area	49,112	1.6
1	Southern California	201,176	6.4
2	Rest of California	124,655	4.0
3	Rest of Pacific Division	132,539	4.2
4	Houston	50,765	1.6
5	Dallas/Ft. Worth	67,029	2.1
6	Rest of West/South/Central Division	252,178	8.0
7	Miami	56,718	1.8
8	Orlando	20,727	0.7

9	Tampa	32,433	1.0
10	Rest of Florida	84,670	2.7
11	Washington, DC	59,562	1.9
12	Atlanta	49,379	1.6
13	Raleigh/Durham	13,002	0.4
14	Rest of South Atlantic Division	292,679	9.3
15	New York City	182,058	5.8
16	Philadelphia	53,478	1.7
17	Rest of Middle Atlantic Division	182,087	5.8
18	Chicago	80,271	2.6
19	Detroit	33,685	1.1
20	Rest of East/North/Central Division	345,154	11.0
21	Boston	66,285	2.1
22	Rest of New England Division	81,613	2.6
23	Las Vegas	19,343	0.6
24	Phoenix	44,203	1.4
25	Denver	34,614	1.1
26	Rest of Mountain Division	129,159	4.1
27	East/South/Central Division	187,374	6.0
28	St. Louis	25,777	0.8
29	Minneapolis	24,476	0.8
30	Rest of West/North/Central Division	170,804	5.4

Raceth: Racial-ethnic identity (position 20)

The racial/ethnic identity variable is recoded from the *RACE* and *HISPAN* variables in the ACS-IPUMS.

Code	Description	n	%
0	Non-Hispanic White	2,114,159	67.2
1	Black	318,215	10.1
2	Asian	161,549	5.1

3	Hispanic	449,024	14.3
4	Other or multirace	104,058	3.3

Mom_raceth: Mother's racial-ethnic identity (position 21)

Mother's racial/ethnic identity is recoded from the *RACE* and *HISPAN* variables in the ACS-IPUMS. Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. This variable has values for the 834,051 persons who reside with their mother.

Code	Description	n	%
0	Non-Hispanic White	507,383	60.8
1	Black	90,487	10.8
2	Asian	55,230	6.6
3	Hispanic	157,102	18.8
4	Other or multirace	23,849	2.9

POB: Place of Birth (position 22)

Place of Birth is directly recoded from the *BPL* variable in the ACS-IPUMS.

Code	Description	n	%
0	United States	2,751,965	87.4
1	Mexico	93,030	3.0
2	Central America and West Indies	60,174	1.9
3	South America	25,526	0.8
4	Canada	10,644	0.3
5	Europe	61,413	2.0
6	East Asia	42,187	1.3
7	Southeast Asia	40,479	1.3
8	Rest of Asia	42,299	1.3
9	Africa	4,143	0.5
10	Other	10,053	1.1

Perwt: Sampling weight (position 23)

Sampling weights are taken from the *PERWT* variable in the ACS-IPUMS.

Range	Description	n	%
[1, 3596]	Sampling weight	3,147,005	100

Gen: Generational Status (position 24)

The generational status variable is recoded from immigration status and age at immigration. Among the foreign-born, those who arrived before age 15 are classified as the 1.5 generation, and those who arrived age 15 or older are classified as the 1.0 generation.

Code	Description	n	%
0	U.S.-born	2,751,965	87.4
1	Foreign-born, 1.5 generation	112,221	3.6
2	Foreign-born, 1.0 generation	282,819	9.0

Mom_gen: Mother's generational Status (position 25)

Mother's generational status variable is recoded from mother's immigration status and mother's age at immigration. Among foreign-born mothers, those who arrived before age 15 are classified as the 1.5 generation, and those who arrived age 15 or older are classified as the 1.0 generation. Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. This variable has values for the 834,051 persons who reside with their mother.

Code	Description	n	%
0	U.S.-born	647,312	77.6
1	Foreign-born, 1.5 generation	37,445	4.5
2	Foreign-born, 1.0 generation	149,294	17.9

Gen2: Generational Status, including 3rd+ generation (position 26)

This generational status variable is based on immigration status and mother's generational status. It is imputed[2] for those under age 30 not living with a resident mother.

Code	Description	n	%
0	3rd+ generation	1,100,584	72.2
1	1st generation (foreign-born)	160,630	10.5
2	2nd generation (U.S. born with a foreign-born mother)	263,313	17.3

Religion: Religion (position 27)

The religion variable is recoded from the variable *qe1* in the Pew Religious Landscape Study and imputed for the respondents in the ACS-IPUMS using cross-sample multiple imputation[3].

Code	Description	n	%
0	No religion	737,818	23.4
1	Christian	2,170,363	69.0
3	Muslim	30,239	1.0
4	Jewish	70,192	2.2
6	Other	138,393	4.4

Mom_religion: Mother's religion (position 28)

Mother's religion is recoded from the variable *qe1* in the Pew Religious Landscape Study and imputed for respondents in the ACS-IPUMS using cross-sample multiple imputation. Mothers are identified using the MOMLOC identifier in the ACS-IPUMS. This variable has values for the 834,051 persons who reside with their mother.

Code	Description	n	%
0	No religion	176,322	21.1
1	Christian	589,942	70.7
3	Muslim	8,015	1.0
4	Jewish	16,745	2.0
6	Other	43,027	5.2

Predlit: Literacy (position 29)

Literacy is measured in the 2012/2014 PIAAC and imputed[4] for the 1,628,755 adults age 25-64 in the ACS-IPUMS. According to PIAAC, "Literacy is understanding, evaluating, using and engaging with written text to participate in the society, to achieve one's goals and to develop one's knowledge and potential." PIAAC measures literacy of adults aged 25-64 for OECD countries, including the United States.

Range	Description	n	%

[3] Religion and mother's religion are imputed simultaneously as a function of age, sex, race/ethnicity, and place of birth.

[4] For the U.S.-born respondents of the ACS-IPUMS, literacy is imputed as a function of age, sex, educational attainment, region, and labor force participation. For the foreign-born respondents of the ACS-IPUMS, literacy is imputed as a function of age, sex, educational attainment, region, labor force participation, English language proficiency, age at immigration, and place of birth.

[143, 355]	Literacy Level	1,628,755	100

Serial: Unique household identifier (position 30)

The unique household identifier is identical to the variable *SERIAL* in the ACS-IPUMS.

Range	Description	n	%
[1, 1375481]	Household identifier	3,147,005	100

Pernum: Person identifier (position 31)

The person identifier is identical to the variable *PERNUM* in the ACS-IPUMS. It is unique within households (SERIAL).

Range	Description	n	%
[1, 20]	Person identifier	3,147,005	100

Example of a Record in the Base Population

Below is an example of a typical record found in the base population:

33, 0, 0, , 0, , 0, , 3, 1, , 3, 1, 1, 1, 24, , 1, 3, 70, 0, , 5, 270, 1, , 1, 0, 0, 282.95, 263, 3

The record is from a 33 year-old English-speaking woman residing in Alabama. She was born in Europe, has a B.A. and immigrated to the United States at the age of 24 (she is therefore Generation 1.0). She does not live with her mother. She identifies as non-Hispanic white. She is imputed to be Muslim, to have a literacy score of 282.95, and to have a mother with a high school education. Her sampling weight is *270*. Her household id is 263 and her person id is 3.

Modules

Typical Modgen models are based on two main elements. The first element is the main model file, which is the driver of the simulation. This file includes control flow structures looping through all cases in the simulation, creating and starting the simulation of each actor sequentially (states for a case are retrieved from a record in the base population file). The second main element of the model is the *Person* class (essentially a C++ class) which contains all of the elements required for the simulation of a case: event functions (rules for changing state variables and creating new actors), state variables (age, sex, etc.) and output tables. The description of the *Person* class is usually divided in several files or "modules" according to functional criteria (mortality, fertility, migration, etc.).

The main module is usually called the *PersonCore* module, as it is the only one automatically created by the Modgen model Wizard. Technically, all the code pertaining to the description of the actor (the *Person* class) could be contained in this single file, but this would prove impractical for big models. The extension of Modgen files (main file or module files) is .mpp.

Each of the modules also contains a declaration of Modgen-specific data structures (range, classification, etc) and parameters. Parameters declared in the modules are containers whose values are externally defined in a separate data file. Parameter values are defined in a separate file so that users may build alternative scenarios through the user interface without having to alter the source code.

LSD-USA is mainly comprised of the following files:

LSD-USA.mpp: This is the main file. It has been modified to include the simulation of future immigration (this will be discussed later on).

PersonCore.mpp: This is the main module. In LSD-USA, this file is mostly used for basic simulation parameters and time keeping (age and calendar time).

ModgenInputMicroDataModule.mpp: This module serves the only purpose of reading data in the base population file.

Tables.mpp: This module contains all of the output tables generated by the model.

Modules (12): Education.mpp, Emigration.mpp, EnglishProficiency.mpp, Fertility.mpp, HomeLanguage.mpp, Immigration.mpp, Labour.mpp, Literacy.mpp, Mobility.mpp, Mortality.mpp, RaceEthnicity.mpp, Religion.mpp

Each of these modules will now be described in more detail.

The Main Simulation File (LSD-USA.mpp)

The Main Simulation file is not a module *per se*, as it does not contain any information on the *Person* class. Rather, it is the main file and driver of the model: it creates the

actors and simulates each one of them in succession, one at a time. The main file contains two functions: *Simulation* and *CaseSimulation*.

The *Simulation* function contains a simulation loop whose number of iterations is specified in the scenario. Each iteration of this loop reads a record from the base population file and creates an actor (in C++ parlance, an actor object is instantiated from the *Person* class).

The simulation is then launched by calling the *CaseSimulation* function, which in turn calls the *start*() function which will initialize all of the actors state variables. After an actor dies or leaves the simulation (through emigration or death), the loop iterates again and moves on to the next record, and so on until the end of the base population file.

The Main Module (PersonCore.mpp)

The PersonCore module contains basic characteristics (state variables) of the actor such as age and sex. It also keeps track of calendar time.

The module also contains the definition of a very important function, the *Start* function, which is called at the creation of a new actor. This function initializes all the states of the actor.

Ranges and classifications	
AGE	Age range. Can take integer values between 0 and 100.
YEAR	Simulation year range. Can take values between 2015 and 2065.
SEX	Sex of actor. Male or female.

In addition to these data structures, other classifications and partitions are defined for table production or parameter definition. These can be found in the program code. Parameters are listed below.

Parameters	
yearACS	Year of the ACS (2015)
SimEnd	Last year of simulation
nPopBase	Number of cases to read in the population file
sexratio	Sex ratio at birth (105/205)
CohortOnly	Allows the simulation of a specific cohort (birth cohort or immigration cohort)

The preceding Modgen elements are declared outside the *Person* class and therefore are not members of this class. State variables and event function declaration are made inside the *Person* Class.

Main state variables	
alive	Logical. It is turned to FALSE after a death event.
sex	SEX. Sex of the actor.
age_int	AGE. Age of the actor in completed years.
BasePop	Logical. Indicates if an actor was taken from the base population or was created during the simulation through birth or immigration.
year_int	Integer. Calendar year during the simulation.
cohort	Integer. Year of arrival for immigrants and year of birth for native.

The *Person* class also includes declaration of event functions. Event functions are used to change the value of state variables.

Events and functions	
Birthday	Increments the age of the actor by one year
New year	Increments the calendar year by one unit
Start	This function is very important. It is called to create and start a new actor in the simulation. It initializes all of the actor's state variables according to the type of actor that is being simulated (an actor taken from the base population, an actor born during the simulation or an immigrant actor arriving during the simulation).
Finish	This function is called to terminate the simulation of an actor, typically after a mortality or an emigration event.

The Education Module (Education.mpp)

Highest educational attainment is taken directly from the base population file for individuals 25 years and older. Individuals less than 25 years old are assumed to have an incomplete education and are thus reassigned a new educational pathway based on the results of a multinomial logit regression model and distributions of age at graduation.

The multinomial model predicts educational attainment. The model is stratified by generational status (1st, 2nd, and 3rd+) and includes mother's education, age, sex, race/ethnicity, and region (4 categories) as independent variables. Analyses were performed on the 2015 ACS-IPUMS.

Newborns and actors with incomplete educational history are reassigned a highest educational attainment upon the start of their simulation. Educational history is reconstructed from distributions of graduation ages derived from the National Longitudinal Survey of Youth (1997 cohort). Specific graduation ages are assigned for all

diplomas (high school, college and university). Additional rules are included so that graduation ages stay coherent: graduation from high school can only occur at least three years before graduation from college, and graduation from college must occur at least two years before obtaining a professional degree. Associate degrees are obtained at least one year after graduation from high school.

Ranges and classifications	
EDUCATION	There are five possible levels of educational attainment: no diploma, high school, associate's, bachelor's, or professional degree (see the description of the education variable in the base population for more detail)

Parameters included are regression coefficients from the multinomial logit as well as age at graduation distributions.

Parameters	
Coefficients	Coefficients by mother's education, age, sex, race/ethnicity, and region are provided for all strata (generational status) and education levels.
Age distributions	Distribution for age at diplomation are included for every level of education (high school, associate's, bachelor's and professional)

The education module has a single relevant state variable.

Main state variables	
education	EDUCATION. May take any of the five values listed in the EDUCATION classification.

Events and functions	
Highest degree event	Determines the highest education attainment of the actor and the age at graduation for all degrees. This function is run at birth for actors born in the model, and upon the start of the simulation for individuals under 25 years old in the base population.
Secondary	Changes the value of the education state variable to high school
Associates	Changes the value of the education state variable to associate
College	Changes the value of the education state variable to college
Professional	Changes the value of the education state variable to professional

The Emigration Module (Emigration.mpp)

Emigration is modeled as an annual rate of out-migration. Rates of out-migration for the foreign-born are estimated by the U.S. Census Bureau (Leach 2017)[5] and vary by country or region of birth, sex, and duration of residence as shown below. Rates of out-migration for the U.S. born are very low (less than .02%)[6] and are assumed to be zero in the LSD-USA model.

Group	Annual Emigration Rate
Born in Mexico	
10 years of U.S. residence or less	
Male	0.045
Female	0.012
More than 10 years of U.S. residence	0.006
Born in Canada or Europe	
10 years of U.S. residence or less	0.031
More than 10 years of U.S. residence	0.009
Born in Asia	
5 years of U.S. residence or less	0.023
More than 5 years of U.S. residence	0.000
Born in other Latin American Country	0.000
Born in other country outside the U.S.	0.005

Source: Leach 2017

An emigration event terminates the simulation of the out-migrating actor.

Ranges and classifications
The emigration module does not contain any classification of its own.

Parameters	
Emigration hazards	Emigration hazards vary according to country of birth, duration of U.S. residence, and sex.

[5] Leach, Mark. 2017. "Recent Innovations in the U.S. Census Bureau's Method of Estimating Foreign-born Emigration." Poster presented at the 2017 annual meetings of the Population Association of America, Chicago, IL.

[6] Fernandez 1995. Fernandez, E.W. 1995. "Estimates of the Annual Emigration of U.S. Born Persons by Using Foreign Censuses and Selected Administrative Data: Circa 1980." Technical Working Paper No.1 O. Population Division, U.S. Census Bureau, Washington, DC. Available online at http://www.census.gov/population/www/techpap.html; Gibbs, J., Harper, G., Rubin, M., & Shin, H. (2003). Evaluating components of international migration: Native-born emigrants. *U. S. Census Bureau, Population Division Working Papers*, (63).

Main state variables	
emigration	LOGICAL. A flag that identifies emigrants. Mainly used for tabulation.

Events and functions	
emigration	Calculates time before emigration. If emigration occurs, turns emigration state variable to TRUE and removes actor from simulation.

The Fertility Module (Fertility.mpp)

Fertility in LSD-USA is modeled in three steps.

In the first step, age-specific fertility rates by state of residence are taken from vital statistics and standardized so that the TFR is equal to 1 in any given state. These curves provide the tempo information of fertility.

Relative risks were also calculated using the 2015 ACS-IPUMS. A logistic regression on the probability of giving birth in the past year was estimated. The independent variables included generation status (1, 1.5 or 2+), race/ethnicity, education, age, state, and combined language and English proficiency. Risks relative to the population as a whole were computed using the *contrast* function in Stata.

Finally, the model was calibrated so that the number of births in the model corresponded to the observed number of births in each state as given by vital statistics. The calibration year was 2015.

After the occurrence of a fertility event, a new actor is created and simulated. This new actor is linked to its mother through a Modgen link, a programming element akin to a C++ pointer. This link allows for the intergenerational transfer of characteristics between mother and child, such as place of residence, language used at home, religion, race/ethnicity, etc.

In Modgen, wait time before next event is recalculated as soon as the event has occurred, so that a female actor could be at risk of giving birth immediately after a fertility event. To avoid abnormally short birth intervals, the model allows female actors to give birth only once per year of age.

Ranges, classifications and links	
AGEFERTILE	This range determines the ages at which a female actor is able to give birth, between 15 and 49.
Links	Modgen links are created between the mother and her children (one link goes from the mother to her child, and a vector of links goes

	from the children to their mother). See the developer's guide for more information on links

Parameters	
Fertility hazards	Fertility hazards vary according age and state of residence.
Relative risks	Fertility risks are modulated according to generation status (1, 1.5 or 2+), race/ethnicity, education, and combined language and English proficiency.

Main state variables	
fertileyear	LOGICAL. Flag that prevents actor from giving birth twice in any given year of age.

Events and functions	
fertility	Calculates time before a fertility event. If a birth occurs, a new actor is created and simulated (its Start function is called). This new actor is linked to its mother actor through a Modgen link. This link allows the transmission of characteristics between mother and child (this is done in the Start function of the child).

The English Proficiency Module (EnglishProficiency.mpp)

This module simulates the acquisition of English proficiency.

Acquisition rates are based on survival curves derived from birth- and year-of-arrival cohorts followed from the 2011 ACS to the 2012, 2013, 2014, and 2015 ACSs[7]. Survivors are defined as actors that are not yet English proficient (i.e., they speak English "not well" or "not at all"). The survival curves are based on age for the U.S.-born or on time since migration for immigrants. They vary according to language spoken at home, region of residence and, additionally in the case of immigrants, age at migration. We used a collapsed version of detailed region of residence with 11 categories[8]. Rates of acquisition are calculated each year based on the simple survivorship formula 1-($Survivors_{t+1}$/$Survivors_t$).

[7] For more details on the derivation of these survival curves, see Sabourin, P., Bélanger, A. (2015). The Dynamics of Language Shift in Canada. Population, 70(4), 727-757.

[8] California, Florida, and remainders of the following census divisions: Pacific, West South Central, South Atlantic, Middle Atlantic, East North Central, New England, Mountain, East North Central, and West North Central.

Ranges, classifications and links	
LANGUAGEPROF	English proficiency classification. It has two possible values: English proficient and not English proficient. Those who speak only English at home are assumed to be English proficient. See section on base population for further details.

Parameters	
Acquisition survival curves: U.S.-born	Survivors are actors who are not English proficient at age *t*. Survival curves vary according to acquired English proficiency, language spoken at home, and region of residence (11 categories).
Acquisition survival curves: Immigrants	Survivors are actors who are not English proficient at a duration *t* following migration. Survival curves vary according to acquired English proficiency, language spoken at home, region of residence (11 categories), and age at migration.

Main state variables	
languageprof	LANGUAGEPROF. Indicates if an actor is English proficient.
languageprofmother	LANGUAGEPROF. Indicates if an actor's mother is English proficient.

Events and functions	
LanguageProfEvent	This event simulates the acquisition of knowledge of English during the lifecourse.

The Language Most Often Spoken at Home Module (HomeLanguage.mpp)

The module for the language most often spoken at home (thereafter the *home language* module) simulates changes in the language used at home occurring during the lifecourse. Changes in the language used at home are particularly common for children of immigrants and linguistic minorities, as many of them gradually stop using their mother tongue at home, for instance through linguistic assimilation in school or through exogamy after leaving the family home. When an actor starts predominantly using at home a language that is different from its mother tongue, it is said that the actor has performed a language shift. Only language shifts from Spanish or other languages to English are considered, as shifts from English to other languages are rare. Since language spoken at home and language proficiency are modeled separately, it is possible that an actor be scheduled for a language shift whereas it is still not English proficient. This

situation should occur rarely as English language acquisition is faster and more frequent than language shift. Therefore, to maintain consistency, English proficiency is set to Proficient whenever a language shift occurs.

Language shift rates are based on survival curves derived from birth- and year-of-arrival cohorts followed from the 2011 ACS to the 2012, 2013, 2014, and 2015 ACSs[9]. Survivors are defined as actors that do not speak English at home. The survival curves are based on age for the U.S.-born or on time since migration for immigrants. They vary according to race/ethnicity, region of residence and, additionally in the case of immigrants, age at migration. We used a collapsed version of detailed region of residence with 11 categories[10]. Rates of acquisition are calculated each year based on the simple survivorship formula 1-($Survivors_{t+1}$/$Survivors_t$).

Ranges, classifications and links	
LANGUAGEHOME	Language most often spoken at home classification. Can take three values English, Spanish or Other. See section on base population for further details.

Parameters	
Language shift survival curves: U.S.-born (LanguageIntraShiftNative)	Survivors are actors who still do not speak English at home at age *t*. Survival curves vary according to race/ethnicity and region of residence (11 categories).
Language shift survival curves: Immigrants (LanguageIntraShiftImmig)	Survivors are actors who do not speak English at home at a duration *t* following migration. Survival curves vary according to race/ethnicity, age at migration and region of residence (11 categories).
Intergenerational language shift: origin-destination matrices (LanguageInterShiftODM)	Origin (home language of mother) – Destination (home language) matrices varying according to region of residence, immigrant status of mother and home language. Some language shifts occur "at birth" when the declared home language for the child differs from the home language of the mother.

[9] For more details on the derivation of these survival curves, see Sabourin, P., Bélanger, A. (2015). The Dynamics of Language Shift in Canada. Population, 70(4), 727-757.
[10] California, Florida, and remainders of the following census divisions: Pacific, West South Central, South Atlantic, Middle Atlantic, East North Central, New England, Mountain, East North Central, and West North Central.

Main state variables	
languagehome	LANGUAGEHOME. Indicates the actor's home language.
languagehomemother	LANGUAGEHOME. Indicates the home language of the actor's mother. Used in the derivation of intergenerational language shifts.

Events and functions	
LanguageHomeInterShiftEvent	This event simulates the occurrence of a language shift at birth (i.e. home language of mother differs from home language of child).
LanguageHomeIntraShiftEvent	This event simulates the occurrence of a language shift during the lifecourse.

The Derived Language Variables Module (Language.mpp)

The Derived Language Variables module is not a module *per se*, as it doesn't drive any event by itself. Its sole purpose is to generate states derived from the other two language variables (home language and English proficiency). The only event in this module is called externally, that is from other event functions generating a change of state in one of the core language variables.

The derived variables are used either analytically (as variables of interest for the analysis) or for modeling (as dimensions of parameters). They are further described in the table below.

Language Derived Variable	
Combination of home language and English proficiency (languagehomeprof)	This variable is used as a determinant of educational attainment, literacy, fertility, and mobility. It can be set to one of five values: Speaks only English at home, speaks Spanish at home and is English proficient, speaks Spanish at home and is not English proficient, speaks some other non-English language at home and is English proficient, and speaks some

	other non-English language at home and is not English proficient.

The Labour Force Participation Module (Labour.mpp)

Labour force participation is a dichotomous state variable: an actor may be considered *active* (employed or unemployed looking for work) or *inactive* (unemployed and not looking for work). The value of the labour force participation state variable is derived from the actor's characteristics: there are no specific transition probabilities between the *active* and the *inactive* states.

The probability of being active at any point in time is calculated using a logistic regression that predicts labour force participation as a function of age (five-year age groups), detailed region of residence (31 categories), race/ethnicity and generational status as regressors. The regression was further stratified by sex and level of education. Whenever a state variable affecting labour force participation changes, the probability of being active is recalculated and the labour force status reassigned following a Monte-Carlo trial.

This method of deriving participation rates provides adequate aggregate cross-sectional descriptions, but yields incoherent individual lifecourses, as individual actors are susceptible to change their participation status every time they move from one region to another or reach another age group. Further development of the model may introduce transition probabilities between states according to characteristics of the actor as well as according to duration in a given state.

Ranges, classifications and links	
LFP	Labour force participation status. Can take two possible values: active (at work or unemployed and looking for work) or inactive (unemployed and not looking for work).

Parameters	
Logistic regression coefficients	Logistic regression coefficients for age (*LaborAge*, five-year age groups), detailed region of residence (*LaborPOR,* 31 categories), race/ethnicity (*LaborRace)* and generation status (*LaborGen)*. Coefficients are further stratified by sex and education. Probability of being active is derived from these coefficients and is recalculated every time that there is a change in characteristics. Only actors aged 15 or more can be active.

Main state variables	
lfp	LFSTAT. Indicates if actor is active or inactive.

LFPcount	INT. A derived variable who is incremented every time a factor affecting labour force participation is changed. A change in the value of this variable triggers the labour force participation event.

Events and functions	
LFPEvent	This event computes the probability of being active and randomly assigns a new labour force status based on a Monte-Carlo trial.

The Mortality Module (Mortality.mpp)

The modeling of mortality in LSD-USA is straightforward. Each actor is subjected to an annual risk of death according to age, sex, race/ethnicity and year of simulation. Rates were generated from projected numbers of deaths and population in U.S. Census Bureau's 2014 population projection[11].

Additionally, immigrants of varying lengths of U.S. residence are subjected to a decreased risk of death to account for the 'healthy immigrant' effect. The value of this relative risk was estimated using the linked 1986-2009 National Health Interview Survey-National Death Index file. The risk of death was estimated using a Cox proportional hazards model for the foreign-born of 0-4, 5-9, 10-14, 15-19, and 20 or more years of U.S. residence versus the U.S. born. Risks relative to the population as a whole were then computed using the *contrast* function in Stata. Finally, we used these relative risks to adjust the Census Bureau's death rates to account for immigration status and duration of residence.

There are no specific classification or state variable associated with the mortality module.

Parameters	
Mortality rates	Mortality rates vary according to year of age, calendar year, sex and race/ethnicity. Projection of mortality rates were done by the U.S. Census Bureau.

Events and functions	
MortalityEvent	This event computes the waiting time before a death event. It also takes into account the healthy immigrant effect and the maximum allowable age. After a death event, the *alive* state variable is set to *FALSE* and the actor is removed from the simulation.

[11] U.S. Census Bureau, Population Division. 2014. "NP2014_D1: Projected Population by Single Year of Age, Sex, Race, and Hispanic Origin for the United States: 2014 to 2060" and "NP2014_D3: Projected Deaths by Single Year of Age, Sex, Race, and Hispanic Origin for the United States: 2014 to 2060", [Machine-readable files], released December 2014.

The Race/Ethnicity Module (Race.mpp)

The race/ethnicity module takes care of intergenerational transfers of race/ethnicity identity from mother to child. Transfers are established according to origin – destination matrices (from mother's race/ethnicity to child's race/ethnicity) extracted from the ACS-IPUMS file. The origin -destination matrices are further derived according to the mother's immigration status.

At birth, an actor is automatically given the race/ethnic identity of its mother and is then immediately and instantaneously subjected to perform an intergenerational transfer of race/ethnic identity. Race/ethnicity is then fixed for life.

Ranges, classifications and links	
RACETH	Race/ethnic identity. Can take 5 possible values (see relevant section above in the description of the base population).

Parameters	
Race/ethnic identity origin-destination matrix (RaceInterTransferODM)	Origin (mother's race/ethnicity) – Destination (actor's race/ethnicity) matrices for intergenerational race/ethnic identity transfers. Vary according to the mother's immigrant status.

Main state variables	
raceth	RACETH. Race/ethnic identity.
racethmother	RACETH. Race/ethnic identity of the mother.

Events and functions	
Intergenerational transfer of race/ethnicity (RaceInterTransferEvent)	This event assigns a race/ethnic identity based on an origin – destination matrix. This event only occurs once at birth.

The Religion Module (Religion.mpp)

The religion module takes care of intergenerational transfers of religious affiliation from mother to child. Intergenerational transfer probabilities were established according to origin – destination matrices (from mother's religious affiliation to child's religious affiliation) extracted from the 2014 Pew Religious Landscapes Study. The origin - destination matrices were further derived according to the mother's immigrant status.

At birth, an actor is automatically given the religious affiliation of its mother and is then immediately and instantaneously subjected to perform an intergenerational transfer of religious affiliation. Religious affiliation is then fixed for life.

Ranges, classifications and links	
RELIGION	Religious affiliation. Can take 5 possible values (see relevant section above in the description of the base population).

Parameters	
Religious affiliation origin-destination matrix (ReligionShiftODM)	Origin (mother's religious affiliation) – Destination (actor's religious affiliation) matrices for intergenerational religious affiliation transfers. Vary according to the mother's immigrant status.

Main state variables	
religion	RELIGION. Religious affiliation.
religionmother	RELIGION. Religious affiliation of the mother.

Events and functions	
Intergenerational transfer of religious affiliation (ReligionShiftEvent)	This event assigns a religious affiliation to a newborn actor based on an origin-destination matrix. This event only occurs once at birth.

The Literacy Module (Literacy.mpp)

The literacy score in LSD-USA is a continuous derived variable taking a value between 0 and 500. It is derived from parameters of linear regressions on log scores with the following regressors: sex, age, region (4 categories), education, language used at home, English proficiency, labor force participation status, and for immigrants, age at migration and place of birth. Regressions were further stratified by immigrant status.

Only the population aged 25 to 64 may be assigned a literacy score. Literacy scores are updated whenever relevant actor's characteristics undergo changes.

Since literacy and labor force participation share many determinants and are both derived variables, labor force participation status takes precedence when both states are scheduled to be updated at the same time. This is necessary as labor force participation is a determinant of literacy.

The source of data used for the analysis was the Program for the International Assessment of Adult Competencies (PIAAC) database of 2012 & 2014. The analysis was based on more comprehensive work done by Samuel Vézina as part of his PhD thesis.

Ranges, classifications and links
No specific range or classification for litteracy

Parameters	
Age (LiteracyAge), sex (LiteracySex), region of residence (LiteracyRegion), education (LiteracyEduc), language used at home (LiteracyLangHome), English proficiency (LiteracyLangProf), labor force participation status (LiteracyLFP), age at immigration (LiteracyAgeImm), place of birth (LiteracyPOB)	Regression parameters to estimate literacy scores. All parameters are available for immigrants and non-immigrants.

Main state variables	
literacy	DOUBLE. A derived variable giving the literacy score of the actor based on the result of a linear regression (see above).
LiteracyCount	INT. A derived variable who is incremented every time a factor affecting literacy is changed. A change in the value of this variable triggers the literacy event.

Events and functions	
LiteracyEvent	This event computes the literacy score based on the result of a linear regression.

The Mobility Module (Mobility.mpp)

Interregional mobility in the model unfolds in two separate steps. First, waiting time before an outmigration event is calculated from the parameters of a logistic regression. Second, once an outmigration event has been scheduled, a destination region is chosen according to an origin – destination matrix, whose values vary according to the actor's characteristics.

Parameters used to calculate outmigration rates are derived from the results of logistic regressions performed on the 2011-2015 ACS-IPUMS. Migration is defined as having lived in a different detailed region (31 categories) one year before than the current region of residence. For individuals aged 1 year or older, the probability of migration was regressed against sex, age, language (combination of language spoken at home and English proficiency), race/ethnicity, generational status, duration since migration, and education. Regressions were estimated separately for children younger than 15 and individuals aged 15 or older. For individuals aged less than 15 years old, sex, labor force participation status and education were removed from the equation. Regressions were further stratified by region of residence (all 31 regions were used).

Origin-destination matrices between all 31 regions were derived separately for young adults age 18-24 (to capture home-leaving mobility) and all other individuals, and by sex, education, race/ethnicity, and generational status.

Since the probability of outmigration is modeled over a one-year time span, i.e. migrants are counted and not migrations, only one migration event is allowed in a single year of age.

Classifications related to the geography of the model are included in the mobility module.

Ranges, classifications and links	
POR	Place of residence. All 31 regions included in the model. See relevant section above.
REGION	Aggregation of POR into 11 larger regions (see footnote number 8)
REGION4	Aggregation of REGION into 4 larger regions (Northeast, Midwest, South, West).

Parameters	
Outmigration parameters: MobilityAge, MobilityLanguage, MobilityRace, MobilityGen,	Includes all regression coefficients necessary to calculate probability of outmigration. These parameters include the intercept, two sex categories (male, female), age (12 categories: 0-4, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+), home language (3 categories: English only, speaks Spanish at home, speaks other

MobilityDuration, MobilityEducation	language at home), race/ethnicity (5 categories: see relevant section above), generational status (3 categories: U.S.-born, immigrant arrived before age 15, immigrant arrived age 15 or older), duration of U.S. residence (continuous), and for those aged 15+ only, education (5 categories, see relevant section above). The parameters are stratified according to age: one stratum for 15 years old and more and one stratum for less than 15 years old.
Origin-destination matrices: MobilityODYouth MobilityODNonYouth	Matrices containing the distribution of movers in destination regions (31 categories) according to region of origin one year ago (31 categories), age (2 categories to distinguish student migrations: between 18 and 24 years old, all others), sex (2 categories), education (5 categories, see relevant section above), race/ethnicity (5 categories: see relevant section above), and generational status (3 categories: U.S.-born, immigrant arrived before age 15, immigrant arrived age 15 or older).

Main state variables	
por	POR. A state variable giving the place of residence of the actor (31 categories).
region	REGION. A state variable giving the region of residence of the actor (11 categories). The regions are aggregated from the por variable.
region4	REGION4. A state variable giving the large region of residence of the actor (4 categories). It is aggregated from the region variable.

Events and functions	
MobilityEvent	Calculates waiting time before outmigration and picks a new region of residence based on an origin-destination matrix. Actors may only move once per year of age.

The Immigration Module (Immigration.mpp)

The immigration module includes all classifications, state variables and parameters relevant to immigrants and immigration: immigration level and composition, immigrant status, age at immigration, duration since immigration, generation status and place of birth.

The module includes a single event whose purpose is to increment the number of years since migration. Immigration itself (i.e. the creation and simulation of a new immigrant

actor during the projection) is not included in this module. Why is this so? Unlike actors born during the projection, immigrant actors are not related to existing actors in the simulation. Whereas a newborn actor has a link to one of the female actors, an immigrant is not linked to any preexisting actor in the simulation[12]. Hence, the creation of immigrants must be «external», that is defined outside the description of the *Person* class. Immigrant actors must be created in the same way as are the actors from the base population: in the main simulation file (see section on the main simulation file above).

Whereas the starting time for the actors of the base population is the year of the ACS-IPUMS (2015), starting times for new immigrants must be spread all along the course of the simulation.

The characteristics of recent immigrants in the ACS-IPUMS are used as a basis to determine the characteristics of future immigrants in the simulation. Recent immigrants are extracted from the base population and saved in a separate file. Each case from this immigration fileis simulated once in every year of the projection (2015 to the end of the projection) as a new immigrant.

To achieve this in Modgen, a second "immigration" loop is inserted after the main simulation loop (in the *Simulation* function of the Main file). For each iteration of this second loop, the *CaseSimulation* function is called to create an immigrant actor arriving in the simulation at year t + a random number between 0 and 1, where t is the iteration number corresponding to the year of the simulation.

This method for generating immigrants implies that all immigrant cohorts have the same characteristics as recent immigrants from the base population. The immigration file may be modified to change immigrant characteristics or general immigration volume (by including new cases or by changing weights).

For more information on the integration of immigration in a Modgen microsimulation model, please consult chapter 5 of the book *Microsimulation and population dynamics* published by Springer.

Ranges, classifications and links	
IMMSTAT	Immigrant status. Includes two categories: U.S.-born, immigrants.
GEN	Generation Status. Includes three categories: generation 2+ (U.S.-born), generation 1.5 (foreign-born from foreign-born parents, arrived in U.S. before the age of 15), generation 1 (foreign-born from foreign-born parents, arrived in U.S. at the age of 15 or older).

[12] This is a specificity of this model and is not an inherent characteristic of microsimulation models. One could equally produce a microsimulation model simulating chain migration, in which actors could generate new immigrants.

POB	Place of birth. See relevant section above for categories.
AGEIMM	Age at immigration. An integer number between 0 and 100.
DUR5	Years since migration in three categories: 0-4 years, 5-9 years, 10+ years..

Parameters	
ImmigFile	Path and name of immigration file. The content of the file determines immigration composition and volume.

Main state variables	
immstat	IMMSTAT. Immigration status.
ageimm	AGEIMM. Age at immigration.
ageimmmother	AGEIMM. Age at immigration of the mother.
gen	GEN. Generation status.
genmother	GEN. Generation status of the mother.
dur_res	INTEGER. Number of years since migration.
dur5	DUR5. Duration since migration, in three categories.
pob	POB. Place of birth.

Events and functions	
DurImmEvent	Increments the number of years since migration for an immigrant actor.