

CONF-9803120--

LA-UR-98: 969

Approved for public release;  
distribution is unlimited.

Title:

SYNTHETIC POPULATION SYSTEM USER GUIDE

RECEIVED  
SEP 02 1998  
OSTI

Author(s):

DOUGLAS J. ROBERTS

Submitted to:

DALLAS COG  
DALLAS, TX  
MARCH 11, 1998

MASTER

Los Alamos  
NATIONAL LABORATORY

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. The Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

## DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

## **DISCLAIMER**

**Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.**

**TRANSPORTATION ANALYSIS SIMULATION SYSTEM  
(TRANSIMS)  
RELEASE 1.0**

**Synthetic Population System  
User Guide**

**March 1998**

**LA-UR - 98-xxxx**

Notice: The Government is granted for itself and others acting on its behalf a paid-up, non-exclusive, irrevocable worldwide license in this data to reproduce, prepare derivative works, and perform publicly and display publicly. Beginning five (5) years after 1995, subject to two possible five year renewals, the Government is granted for itself and others acting on its behalf a paid-up, non-exclusive, irrevocable worldwide license in this data to reproduce, prepare derivative works, distribute copies to the public, perform publicly and display publicly, and to permit others to do so. NEITHER THE UNITED STATES NOR THE UNITED STATES DEPARTMENT OF ENERGY, NOR ANY OF THEIR EMPLOYEES, MAKES ANY WARRANTY, EXPRESS OR IMPLIED, OR ASSUMES ANY LEGAL LIABILITY OR RESPONSIBILITY FOR THE ACCURACY, COMPLETENESS, OR USEFULNESS OF ANY INFORMATION, APPARATUS, PRODUCT, OR PROCESS, OR PROCESS DISCLOSED, OR REPRESENTS THAT ITS USE WOULD NOT INFRINGE PRIVATELY OWNED RIGHTS.

# CONTENTS

1. BACKGROUND .....	3
2. INSTALLING THE SOFTWARE.....	5
3. RUNNING THE SYNTHETIC POPULATION SYSTEM.....	6
3.1 INPUT DATA .....	6
3.2 RUNNING SYN .....	6
3.3 NOTES ON RUNNING SYN .....	9
APPENDIX: FIGURES FOR SYN.....	10

# 1. BACKGROUND

The Los Alamos National Laboratory (LANL) TRansportation Analysis SIMulation System (TRANSIMS) synthetic population system (SYN) is designed to produce populations (family households, non-family households, and group quarters) that are statistically equivalent to actual populations when compared at the level of block group or higher. The methodology used by this system is described in a report entitled *Creating Synthetic Baseline Populations*<sup>1</sup>. The inputs to the system are U.S. Census Bureau data (STF3A and PUMS) and MABLE/GEOCORR data. Census Bureau STF3A and PUMS data formats are commonly used and are available on CD-ROM from the Census Bureau. These data inputs will not be described in any detail in this guide. MABLE/GEOCORR data is relatively new, and an article describing it can be found at

<http://www.oseda.missouri.edu/plue/geocorr/doc/article.html>

The primary function of MABLE/GEOCORR data is to cross-reference STF3 block group data to PUMS areas.

The outputs of the system are files that contain family household, non-family household, and group quarters data in the form of household and person records. Each record will contain the following information:

- PUMS area id
- tract id
- block group id
- a list of user-specified household/person demographics

SYN will run on a variety of Unix platforms. It has been tested on Sun workstations running Solaris2.5, and on Pentium PC workstations running Linux 2.0. In addition, SYN should run on the following platforms, although it has not been tested on all of them.

- Linux/i386 (a.out and ELF, XF86 2.1 and XF86 3.1)
- Linux/alpha (a.out and ELF)
- Linux/m68k
- Linux/sparc (ELF)
- Linux/PowerMac (ELF)
- MkLinux/PowerMac X11R6.1
- SunOS SunOS 4.1 (sun4/sparc and sun3) and Solaris 2.5(SunOS 5.5)
- SGI (5.2, 5.3, 6.0 and 6.2)

---

<sup>1</sup> *Creating Synthetic Baseline Populations*, Richard J. Beckman, Keith A. Baggerley, and Michael D. McKay, Statistics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A., Pergamon PII: S0965-8564(96)00004-3.

- Dec-Alpha Digital Unix/OSF1 (V2.1 & V3.x)
- HP/HP-UX
- HP/m68k /HP-UX
- IBM-RS6000 & PowerPC /AIX
- FreeBSD
- NetBSD/i386 NetBSD 1.0
- NetBSD/sparc NetBSD/sparc 1.1
- NetBSD/m68k NetBSD/m68k(Amiga) 1.1
- OpenBSD OpenBSD 2.0 (i386, sparc, pmax(mips), alpha and m68k(Mac))
- BSDi/i386 BSDi 2.0 and 3.0
- unixware Unixware 2.02
- Solaris/i386 Solaris 2.4
- SCO/i386 SCO 3.2V4.2
- DecStation /(mips)Ultrix
- LynxOS 2.2 /m68k
- OpenVMS /Alpha
- OS2
- Cray Unicos(C90 & YMP)
- Convex ConvexOS(C2 & C3)
- DG/UX DG/UX 5.4R3.10 (MC88100)
- PowerMac Machten

An ANSI C++ compiler is required to build the application.

## 2. INSTALLING THE SOFTWARE

The Synthetic Population system requires an ANSI C++ compiler and either X11R6 or OpenWindows. The Gnu C++ compiler is recommended. Perform the following steps to complete the software installation.

- 1) Procure the correct source distribution for your platform: the Sun Solaris and Linux/i386 ELF versions may be obtained from the TRANSIMS WWW Home Page,  
*<http://www-transims.tsasa.lanl.gov/synthetic/>*  
as gzipped tar files.
- 2) Unzip the file and untar it into a selected directory. This directory will be referred to as \$SYNTH\_HOME.
- 3) Change directory (cd) to \$SYNTH\_HOME/Synth
- 4) Edit the Makefile as needed. The make variables that might need to be modified are X11LIBS, X11INLL, INSTALL\_DIR, and CC.
- 5) Change directory (cd) to \$SYNTH\_HOME and do a *make depend*, followed by *make*, and then *make install*.
- 6) Change directory (cd) to \$SYNTH\_HOME/STF3A.
- 7) Run (as root) the script named *link\_cd*. This script takes one argument, the name of the directory on which the CD-ROM is mounted. Example:  
*sh ./link\_cd /cdrom*

## 3. RUNNING THE SYNTHETIC POPULATION SYSTEM

### 3.1 Input Data

The synthetic population system requires three input data sets:

- 1) U.S. Census Bureau Public Use Microdata Samples (PUMS) data,
- 2) U.S. Census Bureau STF3A data, and
- 3) MABLE/GEOCORR data.

MABLE/GEOCORR data is obtained from either of the following two WWW sites:

- 1) <http://plue.sedac.ciesin.org/plue/geocorr/>
- 2) <http://www.oseda.missouri.edu/plue/geocorr/>

The SYN system expects STF3A and PUMS data to be read via CD-ROM. MABLE/GEOCORR data must be obtained via a WWW browser prior to running SYN. Perform the following steps to generate a MABLE data file.

- 1) Go to either of the above web sites and select the state in which the PUMS areas of interest are.
- 2) Select the source and target geocode data as shown in Figure 1 and Figure 2 (see Appendix).
- 3) Select **Comma Separated Value File** → **Codes and Names** (see Figure 3 in the Appendix)
- 4) Click **Run Request**. The web server will require several minutes to generate the file. When it is finished, a page with a *gecorr.csv* link will be displayed.
- 5) Click on the link to view the MABLE data.
- 6) Use your browser's *Save File* feature to save the data to disk. It is advisable to save the file with a meaningful name, such as "Oregon\_MABLE.csv". The file name extent **must** be *.csv*.

### 3.2 Running SYN

Before SYN can be run, three user environment variables must be set. These variables, with sample values, are shown here (Bourne shell).

- 1) `export SYNTH_HOME=/usr/local/src/synthetic`
- 2) `export STF_INFO_DIR=$SYNTH_HOME/Parep2/stf`
- 3) `export LBLCD=$SYNTH_HOME/STF3A`

If you have a C-shell environment, the variables would be set with `setenv`:

```
setenv SYNTH_HOME /usr/local/src/synthetic
setenv STF_INFO_DIR $SYNTH_HOME/Parep2/stf
setenv LBLCD $SYNTH_HOME/STF3A
```

Perform the following steps to run SYN.

- 1) Create a working directory and copy the MABLE/GEOCORR .csv file into that directory.
- 2) Change directory (cd) into the working directory.
- 3) Start SYN by typing the executable name, *Syn*.
- 4) Synthetic populations are generated by the following sequence of operations (see Figure 4 in the Appendix):
  - a) Type a two-letter state abbreviation.
  - b) Enter random number seeds, if desired. Up to three integer values may be used as random seeds. If no data is entered, zeros will be used as the random number seeds.
  - c) Type the PUMS ids. These are the ids of the PUMS areas for which synthetic populations are to be generated. Enter one PUMS id per line, and be sure to press **[Enter]** after the last PUMS id.
  - d) Type the base file name. This should be a meaningful string, and it will be pre-pended to the output synthetic population file name.
  - e) If only household id numbers without demographics and without person records are required, click the **Household Only** radio button.

Extract the PUMS data for the selected PUMS ids. This is done by clicking **[Press For Options]** in the upper left corner of the SYN window and selecting the first menu item, *Extract PUMS Data from CD-ROM*.

A warning pop-up window with the message "*Make sure that the CD-ROM containing PUMA data is mounted.*" will be displayed.

Click **[OK]** if the CD is mounted; otherwise mount the CD.

An input pop-up window with the message "*CD-ROM directory name?*" and a default value of */cdrom* will be displayed.

If your CD-ROM drive is mounted as a different directory name, enter that name and click **[OK]**. Otherwise, simply click **[OK]**.

After a few seconds, a pop-up window with the message "*Now make sure that the CD-ROM containing the STF3A data is mounted.*" will be displayed. Do so at this time, and then click **[OK]**.
  - f) At this time, household and person demographics may be specified.

Click **[Press For Options]** and select either "*Specify Household Demographics*" or "*Select Person Demographics*".

A window containing the names of demographic data will be displayed. These are the demographics as they are defined in the "pumsusdd.txt" data dictionary file that is on the PUMS CD-ROM.

Click on each button for which you wish to have demographic data saved in the output synthetic population files. Be sure to click **[Save]** on each page of demographic data names specification. The demographic data is saved to disk, so this data need only be entered once, or until a different set of demographics is required.

- g) Click [**Press For Options**] and select the "*Select MABLE CSV File, Collect PUMS and STF3 Census Data*" option.

A file selector window will appear in which the names of all MABLE/GEOCORR .csv files are listed.

Double click on the appropriate file name. SYN will now read input data (PUMS and MABLE/GEOCORR data from the current working directory, as well as STF3A data from CD-ROM); and it will generate several intermediate files that will be used as inputs to the iterative proportional fitting (rake) section of SYN. The status indicators (the text sub-window above the "State" input box, and the "Percent Done" dial indicator) provide the user with information on processing progress. When processing large PUMAs it may take as long as 5 - 10 minutes per PUMA to complete this phase of the task on a 266 MHz Pentium machine running Linux 2.0.

- h) When the data has been read and processed, the text status sub-window will contain the message "*Ready to run the rake code.*".

Click [**Press For Options**] and select the "*Run Rake Code*" option. This phase of the task can take an hour or more for large PUMAs, and the "Percent Done" dial does not update during this process.

When this phase is complete, the message "*Done Running Rake Code.*" will be displayed in the text status sub-window.

- i) Click [**Press For Options**] and select the "*Generate Populations*" option. This phase can take up to 15 minutes per PUMA.

Upon completion of this task, three files will have been created in the working directory: "*Family\_Synthetic\_HHRecs.out*", "*Non\_Family\_Synthetic\_HHRecs.out*", and "*Group\_Synthetic\_HHRecs.out*". The format of these files is two lines containing the names of the household and person demographic data that was collected, followed by the synthetic household data.

The first line of data will consist of the

- PUMA id,
- tract id,
- block group id,
- an "H" to indicate that it is a household record,
- household id, and
- demographic data in the order listed in line one of the file.

This household record is followed by person records (one per person in the household files; group quarter records have one person per household).

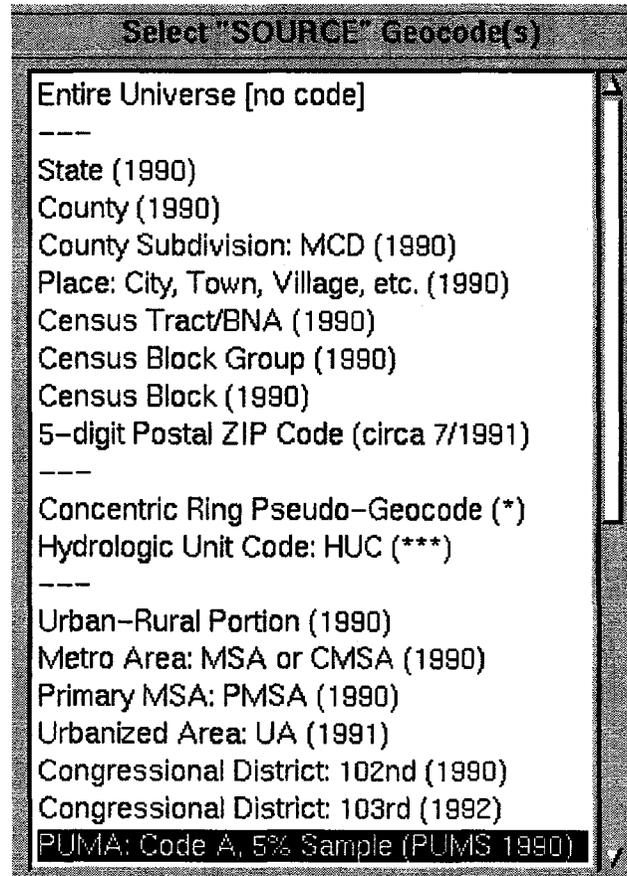
### 3.3 Notes on Running SYN

The iterative proportional fitting computational process requires a fairly large amount of memory. SYN was run on PUMAS 01000, 01200, 01300, 01400, and 01500 for the urban area 6442 (Portland--Vancouver, OR--WA (pt.): FIPS.STATE=41, URBAREA=6442), and the running process required 110.6 MB of memory at its peak usage. For this reason, it is advisable to run one PUMA at a time for areas of interest.

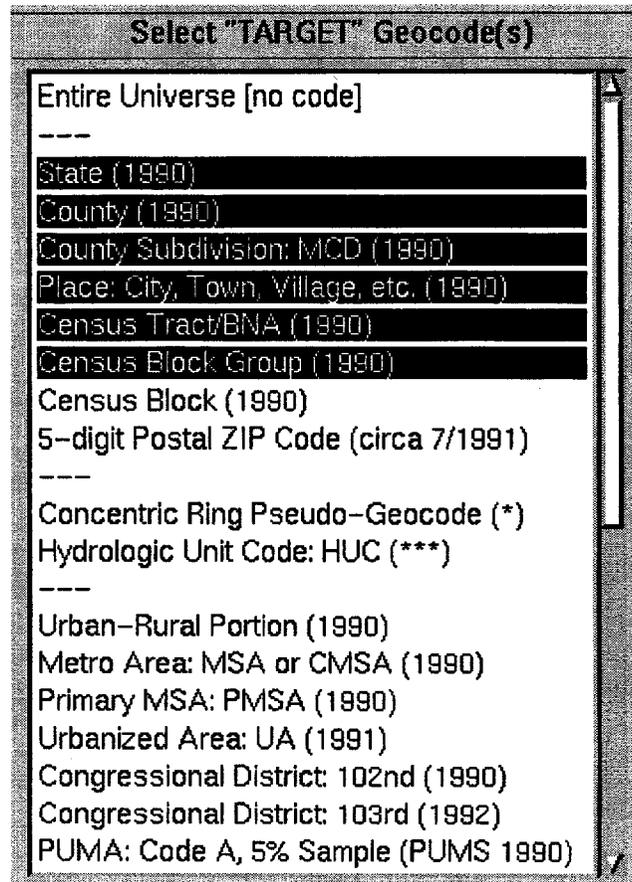
The synthetic population output files are large as well. The family synthetic population file for the above-mention case was 67.4 MB, the non-family file was 13.2 MB, and the group quarters file was 0.86 MB. The processing time for this set of PUMAs was approximately four hours from start to finish, again on a 266 MHz Pentium machine running Linux.

A final note: SYN always generates the synthetic population files using the same names: Family\_Synthetic\_HHRecs.out, Non\_Family\_Synthetic\_HHRecs.out, and Group\_Family\_Synthetic\_HHRecs.out. Remember to rename these files if cases are being run for an area comprising more than one PUMA, because SYN will overwrite them otherwise.

**APPENDIX**  
**Figures for SYN**



**Figure 1: Select "SOURCE" Geocode(s) Window**



**Figure 2: Select "TARGET" Geocode(s) Window**

**Comma Separated Value File**

---

Generate a CSV file

---

Just Codes (No Names)  
 Codes and Names  
 Just Names (No Codes)

---

Use tabs (not commas) as delimiter

---

● Process time for large areas may be several minutes.

**Figure 3: Comma-Separated Value File Window**

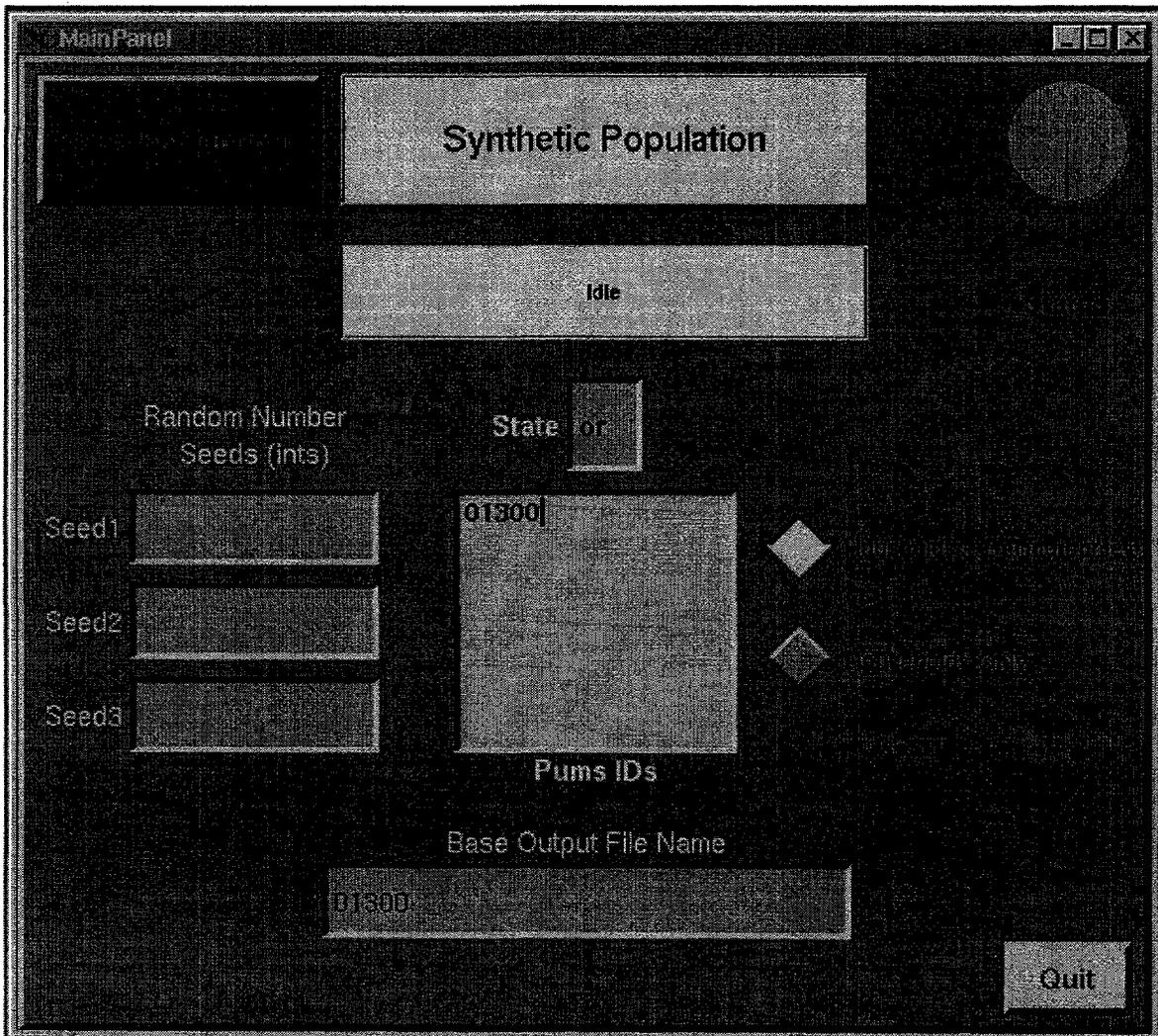


Figure 4: SYN Main Panel