CONCEP 2011

Sept 2011

*401 B Street, Suite 800 • San Diego, CA 92101-4231 • (619) 595-5300*



ABSTRACT

|  |  |
| --- | --- |
| TITLE: | CONCEP 2011 |
| AUTHOR: | San Diego Association of Governments |
| DATE: | September 2011 |
| SOURCE OF COPIES: | San Diego Association of Governments  401 B Street, Suite 800  San Diego, CA 92101  (619) 699-1900 |
| NUMBER OF PAGES: | 27 |
| ABSTRACT: | This report presents a basic description of the Consolidated Estimates Characteristics Program, including a general flow of information and some of the key inputs, assumptions, and computations for each of the components. |

TABLE OF CONTENTS

Overview 1

Technical Description 2

POPEST 2

Housing Unit Method 2

MGRA Population Estimates 2

CONTROLGQ 4

PASEE 4

Cohort-Component Method 5

Special Population Estimates 5

Regional Population Age, Sex and Ethnicity Estimates 6

Subregional Area Population by Sex, Ethnicity and Age Estimates 8

Adjusted and Survived Non-Special Population 8

Completing the Non-Special Population Estimates 9

Census Tract Estimates 11

Adjusted and Survived Population 11

Controlling the Survived Population to SRA Estimates 12

Sex and Ethnicity Estimates 12

Estimates of Age within Sex and Ethnic Group 12

Final Adjustment 13

ControlVitals 13

ESTINC 14

Income Model 14

Household Income Distribution Estimates 16

Initial Estimates 16

Controlling 17

IncomeCalib 17

Index of Dissimilarity 18

Algorithm 18

PDHH - POPEST 19

Households by household size (1, 2, 3, 4+) 20

Households by number of children under age 18 in household (0, 1, 2+): 20

Households by workers in household (0, 1, 2, 3+): 21

Allocation to MGRAs 21

Group Quarters by Type 21

Data Definition 22

Data Sources and Storage 22

Base Year Population and Housing 22

Population and Housing Control Totals 22

Housing Stock Changes 23

Occupancy Rates and Household Size 23

Group Quarters Population 23

Vital Events 23

Net Migration 24

Income 24

Detailed Household Characteristics 24

Programmer’s Reference 25

Environment 25

Programming 25

Nomenclature 25

Components 26

VERSION description 27

LIST OF TABLES

Table 2-1 Matrix Organization to Convert Survived Population by Age to SRA Non-Special Population Estimate Hispanic Males 12

Table 2-2 Matrix Organization to Convert Survived Population by Age to Census Tract Non-Special Population Estimate Hispanic Males 14

Table 4-1 CONCEP Software and Data Organization 24

Table 4-2 CONCEP Code Components 25

Table 4-3 CONCEP Utilities 26

Table 4-4 CONCEP Database Tables 26

**Section 1**

Overview

The Consolidated Characteristics Estimates Program (CONCEP) is a modeling system. CONCEP was developed to integrate a collection of interrelated but independently operated estimation models. Conceptually the model performs several different functions which fall into the broad categories of estimates of total population and housing units, detailed demographic characteristics estimates (population by age, sex and ethnicity), detailed household characteristics including households by size categories, households by presence of children, households with workers and household income distribution estimates. Bringing the models under one umbrella streamlines the modeling process and ensures connectivity among the components.

CONCEP includes seven separate modules.

* POPEST – housing and population estimates for the SANDAG Master Geographic Reference Areas (MGRAs), the atomic geographic area adopted for use with the 2011 Current Estimates and the Series 12 Forecast
* PASEE – detailed population estimates by ethnicity, sex and age for the region, subregional areas, census tracts and MGRAs
* ESTINC – household income distribution estimates for the region, subregional areas, census tracts and MGRAs.
* PDHH – POPEST detailed household characteristics for CT and MGRA.
* ControlVitals – utility that controls the detailed birth and death records to regional controls
* ControlGQ – utility that controls the detailed parcel-level group quarters records by type to regional control totals
* IncomeCalib – utility to calibrate ESTINC income equations

The major modules and utilities of CONCEP are described in detail below.

**Section 2**

Technical Description

# POPEST

Population and housing estimates are derived by POPEST. (The actual program is POPEST – POPEST to MGRA). Total housing units are estimated for single and multiple family units and mobile homes and other units. Estimates of occupied units, also known as households, are produced for each type of housing unit. Population estimates are then derived from the number of occupied housing units and the average number persons per household, plus persons not living in households (such as military personnel living in barracks, college students living in dormitories, and prisoners living in correctional facilities).

The model uses a bottom-up approach. Population estimates are first completed for each MGRA, and then are controlled to city and regional totals derived from data supplied by the California Department of Finance.

## Housing Unit Method

The population of any given area can be considered to be the sum of household population (persons living in a housing unit) and non-household population (persons living in group quarters, such as military barracks, college dorms, or convalescent homes). Thus, total population can be represented by the following:

POPULATION = HOUSEHOLD POPULATION + NON-HOUSEHOLD POPULATION

*where the household population is represented by:*

HOUSEHOLD POPULATION = HOUSING UNITS x OCCUPANCY RATE x AVG. HOUSEHOLD SIZE

In theory, if the number of housing units, occupancy rate, average household size, and non-household populations were known precisely, then the true total population would be known. However, in practice these inputs are rarely known with complete accuracy. Therefore, using assessor’s parcel data for housing units, occupancy rates and household size parameters derived from the 2000 Census, and trends in occupancy rates and household size, the individual components are estimated and are used to estimate the population.

## MGRA Population Estimates

For each MGRA, the following procedure is performed:

1. Update total housing units

Total housing units are derived from SANDAG’s Landcore geodatabase and an annual update from the San Diego County Assessor’s Master Property Record. Essentially, SANDAG staff annually updates the parcel-based polygons in landcore and assigns dwelling unit counts according to updated Master Property Records. Structure type is assigned according the SANDAG land use designation. Housing units on “non-residential” land are assigned a structure type based upon the unit count on the parcel.

Theoretically, if housing unit changes represented by the assessor’s annual parcel data are reported by each jurisdiction reports both SANDAG and the California Department of Finance (DOF), the results from step 1 should match DOF’s official housing estimates for each jurisdiction. This is seldom the case, however, as DOF may not receive all inputs in time for the year’s estimates release, and DOF can adjust the jurisdiction-level data sent to them. Because of these issues in the official reporting process, beginning with the January 1, 2003 population and housing estimates, SANDAG decided to no longer control housing stock estimates to DOF estimates. We determine housing stock from official San Diego County Assessor’s records and assign these changes to MGRA based on a master cross-reference table in our Geographic Information System.

1. Estimate occupied housing units

Occupied housing units, also referred to as households, are estimated from the housing stock by applying occupancy rate factors. The occupancy rate is equivalent to one minus the vacancy rate.

Estimates of total housing stock are more detailed than those for occupied households because housing unit permits and annexation information allow an accurate update by structure type. Vacancy rates vary by structure type and single family vacancy rates are generally lower than multi-family rates. While using aggregate rates could produce large errors in areas where the housing mix has significantly changed, research has shown that vacancy rate estimates for small areas are not reliable. In fact, using separate rates for single and multiple family units may introduce more error than using an aggregate rate. Separate estimates of single family and multiple family vacancy rates are not usually attempted for areas with less than 10,000 people.

HH = TOTALHS \* OCC

*where*

TOTALHS is the total number of housing units

HH is occupied housing units, and

OCC is occupancy rate (the occupancy rate = 1 – vacancy rate).

1. Estimate population in households

Household population is estimated from households by applying persons per household factors (household size). Estimates of household size for small areas have the same problems as vacancy rates. Therefore, household size is applied to the households for single and multiple family units combined and to mobile homes households.

HHP = HH \* HHS

*where*

HHP is household population, and

HHS is household size or average persons per household.

Household population estimates are controlled to DOF annual estimates for each city and the unincorporated area using a factoring algorithm and a +1/-1 adjustment to control for rounding. Changes in household population on military bases are an external input and are not adjusted during controlling.

1. Estimate total population

As noted above, total population is the sum of household population plus non-household population. Non-household population (also referred to as “group quarters”) is further defined as including military non-household population (military group quarters, such as persons living in barracks) and civilian non-household population (civilian group quarters, such as persons living in dormitories or convalescent homes). The group quarters population is maintained at the parcel level and aggregated to MGRAs. The parcel-level group quarters estimates are controlled to city-level totals published by the California Department of Finance.

By adding the estimates of household population, military group quarters population, and civilian group quarters population, the total population estimate is obtained.

TPOP = HHP + MILGQ + CIVGQ

*where*

TPOP is total population;

MILGQ is military group quarters population (uniformed military personnel living in barracks or on ships), and

CIVGQ is civilian group quarters population (civilian population living in college dorms, prisons, convalescent homes and other non-household living quarters).

1. Completing the estimates

At this point, POPEST has derived final estimates of housing stock, households, and household population. POPEST re-computes the vacancy rates and household sizes from these normalized data for use as starting values for the following year’s estimates. Group quarters population estimates are added to household population to determine the total population for MGRA. POPEST estimates both civilian and military group quarters population. These estimates are based on changes added to base year values. If necessary, POPEST controls the civilian group quarters population to DOF estimates for each city and the unincorporated area. Changes in military group quarters population are an external input and are not controlled.

## CONTROLGQ

CONTROLGQ is a POPEST utility that controls parcel-level data for Group Quarters to regional totals. SANDAG annually gets estimates of group quarters population for each parcel in the county. These are typically classified by land use type into military, civilian college and civilian other. Because the data sources are independent, the sum of group quarters population may not equal the total group quarters population as reported by the California Department of Finance. Therefore, this utility does a proportional distribution of the differences to each parcel to ensure that the totals match.

The group quarters controlling uses the specified parcel land use code to determine the character of the residents. Military group quarters are defined as group quarters populations on land use codes 1403, 6700-6709. Civilian college group quarters populations are defined as group quarters populations on land use codes 1402, 6800-6809. Any other group quarters is classified as civilian other.

# PASEE

PASEE produces annual population estimates by single year of age, cross-classified by sex and race/ethnicity for eight groups. Estimates are produced for special and non-special population groups. Special populations include persons living on military bases, in large off-base military housing, and around major universities.

For non-special population groups, the model uses the cohort-component method, described below, to develop the population estimates. For reasons described below, different procedures are applied to accurately estimate the special population characteristics.

The model uses a top-down approach. Estimates are completed first for the region, then for each subregional area (SRA), and finally for census tracts (CT). All estimates are controlled to the total population estimates from POPEST.

## Cohort-Component Method

PASEE uses a cohort-component method to estimate population characteristics annually from base year census data. The cohort-component method models components of population change (natural increase and net migration) for age cohorts grouped by sex and ethnicity. Natural increase is the number of births minus the number of deaths occurring to a population. Net migration is the difference between the number of people moving into an area and the number who leave. The relationship of population change to its components can be expressed as:

Pt– P0 = (B – D) + NM

*where*

Pt is the population at the end of the time period;

P0 is the population at the beginning of the time period;

B is births;

D is deaths, and

NM is net migration, which may be positive or negative.

## Special Population Estimates

While most populations in an area experience the typical life cycle of births, aging and deaths, some populations experience a continual in and out rotation of persons sharing similar demographic characteristics. Consider a typical neighborhood – most residents will remain in the same home, each resident aging as time passes. There may be a few births in the area. There may be a few deaths over the course of a year. A few residents may move to other neighborhoods or leave the region. Most areas in the region experience this trend over time. However, certain special populations do not. Consider a college campus and nearby student housing. Each year, the oldest students graduate and leave the area, and a new group of younger students move in. Over time, the population in this area retains a fairly consistent age range in the late teens and early twenties. We do not expect that a university campus will be filled with thirty year olds ten years from now. Thus, a special estimation procedure must be used for these “special population” areas.

In the San Diego region, special population census tracts include the following:

Military: 38.00, 63.00, 64.00, 94.00, 99.01, 99.02, 106.02, 113.00, 114.00, 187.00

Off-base military: 66.00, 87.01, 92.01, 95.08

Universities and other: 28.01, 29.04, 55.00, 62.00, 83.05, 83.41, 83.42, 83.43, 91.06

A bottom up approach is used to estimate the special population characteristics. Estimates are done first for each special population census tract. Special population estimates for the region and SRAs are created by summing the census tract estimates. For most of these tracts (sct), a raking procedure is used to adjust the base year age, sex and ethnic distribution to the total population estimate from POPEST:

POPEs,e,a,sct = POPB s,e,a,sct \* (POPEsct / POPBsct)

*where*

POPE s,e,a,sct is the POPEST control, and

POPB s,e,a,sct \* is the POPEST base population.

Three military bases, the NavyAmphibious base, Miramar and Camp Pendleton, use changes in group quarters and household populations to update the base year characteristic estimates. These areas have large group quarters and household populations that have unique demographic characteristics. Estimates are generated by:

POPE s,e,a,sct = POPB s,e,a,sct \* + GQCHG s,e,a,sct + HPOPCHG s,e,a,sct

GQCHG s,e,a,sct = (GQEsct - GQB sct) \* UMILDISTs,e,a

HPOPCHG s,e,a,sct = (HPOPE sct - HPOPB s,e,a,sct \*) \* (.75 \* MILDEPDISTs,e,a+ .25 \* UMILDISTs,e,a)

*where*

GQCHG and HPOPCHG are the change in military group quarters and on-base household populations, respectively;

GQE is the military group quarters population estimate from POPEST;

GQB is the base year military group quarters population from POPEST;

UMILDIST is the 2000 regional proportionate distribution of uniformed military in each age, sex and ethnic group from PUMS; (see Data Sources and Storage for details on PUMS)

HPOPE is the on-base household population estimate from POPEST;

HPOPB is the base year on-base household population from POPEST; and

MILDEPDIST is the 2000 regional proportionate distribution of military dependents in each age, sex and ethnic group from PUMS. This method assumes that an average of four persons live in on-base housing and that one person is in the military; therefore, .75 and .25 are used in the weighted average of the two regional distributions.

Because of rounding error, these estimates for special population census tracts do not exactly match the total population estimates from POPEST. A routine (called vector rounding) adds or subtracts one from randomly selected age, sex and ethnicity categories to correct the rounding problem. This routine starts by calculating the difference between the POPEST census tract control and the sum of the estimates over the sex, ethnicity and age categories:

DIFFPOP sct = POPE sct - ΣPOPE s,e,a,sct

*where*

Σ is the summation over the sex, ethnic and age categories; and

DIFFPOP is the difference between the control total and the POPEST estimate.

A random number from 1 to 1,616 is selected (there are 1,616 age, sex and ethnic categories) and one is added to or subtracted from the corresponding POPE s,e,a,sct. One is added if DIFFPOPsct is positive and subtracted if DIFFPOPsct is negative. Categories with zero values are not adjusted and no estimate can have a negative value. The DIFFPOPsct value also is adjusted by plus or minus one. This operation is repeated until DIFFPOPsct reaches zero.

## Regional Population Age, Sex and Ethnicity Estimates

1. Components of change

First determine total population change from POPEST. Then determine which portion of the change is attributable to migration, rather than natural increase. Migration is then separated into special and non-special components.

POPCHNG = ΣPOPEct - ΣPOPB ct

NMIG = POPCHNG – BIRTHS + DEATHS

SPECMIG = ΣPOPEsct - ΣPOPBsct

NSPECMIGreg = NMIG – SPECMIG

*where*

Σ is the summation over all census tracts (or over special population census tracts when denoted by “sct”);

POPE is POPEST estimate;

POPB is POPEST base population;

NMIG is total net migration;

BIRTHS is total region births;

DEATHS is total region deaths;

SPECMIG is net migration of special populations, and

NPECMIG is net migration of non-special populations.

1. Adjusted and survived non-special population

To determine the non-special population by sex, ethnicity and age, start with the base population and subtract the special population deaths. The remaining population is “aged” into the next year of age group. Births become the new 0-year group.

APOPs,e,a = POPB s,e,a – SPECPOPB s,e,a – DEATHSs,e,a

SPOPs,e,0 = BIRTHSs,e

SPOP s,e,a = APOPs,e,a-1 (for populations aged 1 through 99)

SPOPs,e,a,100 = APOPs,e,a,99 + APOPs,e,a,100

*where*

APOP is adjusted base year population;

s is sex;

e is ethnicity/race;

a is age;

SPOP is survived population one year from the base year, and

SPECPOPB is base year special population aggregated from census tracts.

1. Non-special net migration

NSPECMIG s,e,a = SPOP s,e,a \* (CIVMIGRATE s,e,a / 1,000)

NSPECMIGSUMs,e = ΣNSPECMIG s,e,a

NSPECMIGFAC s,e = NSPECMIG reg/ NSPECMIGSUM s,e

NSPECMIG s,e,a = NSPECMIG s,e,a \* NSPECMIGFAC s,e

*where*

CIVMIGRATE is the civilian net migration rate, and

NSPECMIGSUM is the sum of migration of non-special population over all age groups,

NSPECMIGFAC is the sex and ethnic specific adjustment factors used to control the initial age-specific migration estimates.

1. Non-special and total population estimates

To get the region total population by sex, ethnicity and age, complete the non-special estimates by adding the survived and migration populations. Then add the special and non-special estimates by sex, ethnicity and age.

NSPECPOPE s,e,a = SPOP s,e,a + NSPECMIG s,e,a

POPE s,e,a = SPECPOPE s,e,a + NSPECPOP s,e,a

*where*

NSPECPOPE is the non-special population estimate, and

SPECPOPE is the special population estimate aggregated from census tracts.

## Subregional Area Population by Sex, Ethnicity and Age Estimates

The main objective is to develop non-special population estimates for each SRA. Once this is done, they are added to the special population estimates to get the total population estimates by age, sex and ethnic group.

### Adjusted and Survived Non-Special Population

The same procedure described for the regional estimates is used to adjust and survive the base year population in each SRA. Births, deaths and base year special population are aggregated from census tracts. There are two modifications of the adjusted population before calculating the survival equations. It is possible in age, sex and ethnic groups with small population sizes that the subtraction of deaths could cause a negative adjusted population. If this occurs, the adjusted population is set to zero. Also, two SRAs (16- Miramar and 43- Pendleton) are entirely military so their non-special adjusted populations are set to zero. There are a few births and deaths in these SRAs, but they are ignored.

If these adjustments were not made, the sum over the SRAs of the survived non-special population would match the regional survived non-special population in each age, sex and ethnic group within rounding error. Because of this rounding error and the modifications to the adjusted population, a routine called Pachinko is used to control the SRA survived population to the regional totals. This controlling allows for the proper calculation of SRA overall net migration level without the use of births and deaths (discussed below in the section on Net Migration Estimates by Age, Sex and Ethnicity). The first step is to create a regional survived population total by adding across the SRAs for each sex, age and ethnic group and then calculating the difference between the sum and the regionwide control:

SPOPSUM s,e,a = ΣSPOP s,e,a,sra

DIFFSPOP s,e,a = SPOP s,e,a - SPOPSUM s,e,a

*where*

Σ is the summation over the SRAs;

SPOPSUM is the resulting survived population regional sum, and

DIFFSPOP is the difference between the regional control and the summation over the SRAs.

The second step is to create probability and cumulative probability distributions across the SRAs for each sex, age and ethnic group:

PRSPOP s,e,a,sra = SPOP s,e,a,sra / SPOPSUM s,e,a \* 100

CPRSPOP s,e,a,1 = PRSPOP s,e,a,1

CPRSPOP s,e,a,2 = CPRSPOP s,e,a,1 + PRSPOP s,e,a,2, etc.

CPRSPOP s,e,a,41 = CPRSPOP s,e,a,40 + PRSPOP s,e,a,41

*where*

PRSPOP is the probability distribution across SRAs (i.e., the percent ranging from 0 to 100 of the region’s survived population in each SRA);

1, 2, 40 and 41 represent a sequential count of the SRAs, and

CPRSPOP is the cumulative probability distribution across SRAs.

Starting with the first age, sex and ethnic group, a random number from 1 to 100 is selected. This number identifies the SRA to be adjusted based on its cumulative probability value. For example, a random number of 20 is selected and the third SRA has a cumulative probability of 15 percent and the fourth SRA has a cumulative probability of 25 percent, the fourth SRA would be adjusted. One is subtracted from or added to SPOP s,e,a,sra depending on the value of DIFFSPOP s,e,a. One is added if DIFFSPOP s,e,a is positive and subtracted if DIFFSPOP s,e,a is negative. No adjustment is made if there are fewer than five survived persons. DIFFSPOP s,e,a is then adjusted by plus or minus one. This process is repeated until DIFFSPOP s,e,a reaches zero. Pachinko is then applied to the next age, sex and ethnic group and continues until all groups are processed. This method causes the size of the controlling adjustment to be directly related to the size of the survived population in an SRA.

## Completing the Non-Special Population Estimates

The survived population is controlled directly to the regional non-special population estimates by age, sex and ethnic group. Controlling the survived population to regional non-special population estimates is done in three steps.

##### Sex and Ethnicity Estimates

Estimates are first developed by sex and ethnicity. The initial step is to aggregate the survived population by age, sex and ethnicity into sex and ethnic categories for each SRA. The SPOPs,e,sra matrix is adjusted to match the SRA non-special population estimate (NSPECPOPEsra) as well as the regional estimate of non-special population by sex and ethnic group (NSPECPOPEs,e). The SRA control total is computed by:

NSPECPOPEsra = POPEsra – SPECPOPEsra

*where*

POPEsra is the POPEST total population estimate based on the aggregation of all census tracts, and

SPECPOPE is the POPEST special population estimate based on the aggregation of the special population tracts.

An iterative proportionate adjustment method is used to factor the SPOP s,e,sra matrix to both row totals (SRA non-special population estimate) and column totals (regional sex and ethnicity non-special population estimate). It is important that the grand totals of the NSPECPOPE sra and NSPECPOPE s,e match; that is:

TROW = ΣNSPECPOPsra

TCOL = ΣNSPECPOPs,e

TROW = TCOL

*where*

TROW is the sum over the SRA non-special population estimates, and

TCOL is the sum over the regional sex and ethnicity non-special population estimates.

The adjustment occurs by an iterative sequential raking of the SPOP matrix, alternating with rows and columns. Starting with a proportional adjustment to the SRA non-special population estimate (rows), the matrix is changed by:

SUMROWsra = ΣSPOPs,e,sra

NSPECPOPEs,e,sra(1) = SPOPs,e,sra \* (NSPECPOPEsra / SUMROWsra)

*where*

SUMROW is the summation across sex and ethnicity categories for each SRA, and

(1) represents the first adjustment to the matrix.

This is followed by a proportionate adjustment to the regional sex and ethnicity estimates (columns) by:

SUMCOLs,e = ΣNSPECPOPEs,e,sra(1)

NSPECPOPEs,e,sra(2) = NSPECPOPEs,e,sra(1) \* (NSPECPOPEs,e / SUMCOLs,e)

*where*

SUMCOL is the summation across SRAs for each sex and ethnic category, and

(2) represents the second adjustment to the matrix.

The second adjustment to the regional estimate throws the figures out of line with the SRA non-special population estimate; that is, the sum across sex and ethnicity categories no longer matches the SRA population figure. The next step is to readjust the figures with respect to the row totals and obtain NSPECPOPEs,e,sra(3). This readjustment to the row totals throws the figures out of line with the regional non-special population estimate by sex and ethnicity; that is, the sum across SRAs no longer matches the regional estimates, and so the figures are again adjusted to conform to the column totals to obtain NSPECPOPEs,e,sra(4). This adjustment sequence is continues until convergence with a final unchanging matrix, which adds to the required SRA non-special population estimate and regional non-special population estimate by sex and ethnicity. The program allows for 15 iterations that reaches convergence within rounding error. The rounding error can be significant if many cells of the matrix have fewer than 10 persons.

Because rounding error still remains after the application of the iterative proportionate adjustment method, a routine (called matrix rounding) was developed. This routine adds or subtracts one from each cell of the matrix until the rounding error for both the rows and columns is eliminated.

The first step is to compute the difference between the SRA control and the sum across sex and ethnic categories and the difference between the regional sex and ethnicity controls and the sum across SRAs. Because of the mathematics of the iterative proportionate adjustment routine, the totals of the row and column differences match. These differences along with their totals are computed by:

DIFFROWsra = NSPECPOPEsra - SUMROWsra ;

DIFFCOLs,e = NSPECPOPEs,e - SUMCOLs,e;

TDIFFROW = ΣDIFFROWsra;

TDIFFCOL = ΣDIFFCOLs,e; and

TDIFFROW = TDIFFCOL

*where*

DIFFROW is the rounding error of the rows;

DIFFCOL is the rounding error of the columns;

TDIFFROW is the sum over the row (SRA) differences, and

TDIFFCOL is the sum over the regional sex and ethnicity differences.

The matrix rounding procedure starts with the first row with a non-zero difference. A random number for 1 to 16 is selected that identifies the sex and ethnicity category to be adjusted. A one is added to or subtracted from NSPECPOPEs,e,sra, if that category has at least five persons. One is added if DIFFROWsra is positive and subtracted if DIFFROWsra is negative. DIFFROWsra is then adjusted by plus or minus one and DIFFCOLs,e for the appropriate sex and ethnic group is also adjusted like DIFFROWsra. This process is repeated until DIFFROWsra reaches zero for a particular SRA. Matrix rounding is applied to the next SRA with a non-zero DIFFROWsra and continues until all SRAs are processed.

##### 2. Age within Sex and Ethnic Group Estimates

The last step is to estimate the non-special population by age within each sex and ethnic group. This is done by re-adjusting the SPOPs,e,a,sra matrix to match the SRA non-special population estimate by sex and ethnic group described in the preceding section (NSPECPOPEs,e,sra) as well as the regional estimate of non-special population by age within each sex and ethnic group (NSPECPOPEs,e,a). For these estimates, the SPOP array is essentially split into sixteen matrices representing age by SRA within each sex and ethnic category. The iterative proportionate and matrix rounding methods convert the survived population into the non-special population by age and SRA. They are applied sixteen times to complete the non-special population estimate by age, sex and ethnicity (NSPECPOPEs,e,a,sra). Table 2-1 illustrates the matrix organization using Hispanic males as an example.

|  |
| --- |
| Table 2-1 |
| Matrix Organization to Convert Survived Population by Age to SRA Non-Special Population Estimate |
| Hispanic Males |



##### Total population estimates

The total population estimate is the sum of the special and non-special population estimates:

POPEs,e,a,sra = SPECPOPEs,e,a,sra + NSPECPOPEs,e,a,sra

*where*

SPECPOPE is the special population estimate aggregated from census tracts (see Special Population Estimates section).

## Census Tract Estimates

There are two groups of census tracts, special population (i.e., military and college areas) and non-special population. Estimates for the special population census tract are described in the SPECIAL POPULATION ESTIMATES section. A modified cohort-component method that does not estimate migration is applied to the non-special population census tracts. The survived population estimates for these census tracts are controlled to the SRA non-special population estimates.

### Adjusted and Survived Population

A procedure similar to that for the region and SRAs is used to adjust and survive the base year population in each non-special population census tract. The difference is that the base year population is only adjusted for deaths, because by definition there are no special populations in these census tracts. If the subtraction of deaths causes a negative adjusted population, the population in the tract is set to zero. Also, the adjusted population in all special population census tracts is set to zero. Setting the adjusted population to zero lets the matrices, which contain all census tracts, represent only non-special population tracts for computational purposes.

### Controlling the Survived Population to SRA Estimates

The survived population by census tract is controlled to the non-special population estimates by age, sex and ethnic group for its SRA and the total population estimate for the census tract. A two step estimation process is used. Non-special population census tract estimates are developed by sex and ethnicity. These estimates are then disaggregated in age categories within each sex and ethnic group.

### Sex and Ethnicity Estimates

The initial step is to aggregate the survived population by age, sex and ethnicity to sex and ethnicity categories for each non-special population census tract (nsct). The SPOPs,e,nsct matrix is adjusted to match the census tract population estimate (POPEnsct) and the SRA estimate of non-special population by sex and ethnic group (NSPECPOPEs,e,sra). The POPE nsct control represents the POPEST estimate. The iterative proportionate adjustment method matches the SPOPs,e,nsct matrix to both row totals (non-special population census tract estimate) and column totals (SRA sex and ethnicity non-special population estimates). The matrix rounding routine completes the sex and ethnic group estimates in the non-special population tracts (POPE s,e,nsct).

### Estimates of Age within Sex and Ethnic Group

The last step is to estimate population by age within each sex and ethnic group for non-special population census tracts. This is done by adjusting the SPOP s,e,nsct matrix to match the sex and ethnic group estimates described in the preceding section (POPE s,e,nsct) and the SRA estimate of non-special population by age within each sex and ethnic group (NSPECPOPE s,e,a,sra). For these estimates, the SPOP array is essentially split into sixteen matrices representing age by non-special population census tract within each sex and ethnic category. The iterative proportionate and matrix rounding methods convert the survived population estimate into the population by age and non-special population census tracts. These methods applied sixteen times yield estimates by age, sex and ethnicity for non-special population census tracts (POPE s,e,a,nsct). Table 2-2 illustrates the matrix organization using Hispanic males.



### Final Adjustment

A top down approach is used to prepare the census tract estimates by age, sex and ethnicity. SRA estimates are controlled to regional estimates and census tract estimates are controlled to their parent SRA. Several routines are used to match both the POPEST totals for each geographic area and the age, sex and ethnicity estimates for the larger geographic area (e.g., region for the SRAs). Of primary concern is that the sum over all age, sex and ethnicity categories for any geographic area matches the POPEST total population estimates. A final adjustment is made to the age, sex and ethnicity estimates for the non-special population census tracts. This is the same method used to adjust the special population census tract detail to the POPEST control.

The vector rounding routine adds or subtracts one from randomly selected age, sex and ethnicity categories to correct the rounding problem. This routine starts by calculating the difference between the POPEST census tract control and the sum over the sex, ethnicity and age estimates:

DIFFPOPnsct = POPEnsct - ∑POPEs,e,a,nsct , where ∑is the summation over the sex, ethnicity and age categories; POPEnsct is the population estimate from POPEST; and DIFFPOP is the difference between the control total and the estimate total.

A random number from 1 to 144 is selected (there are 144 age, sex and ethnic categories) and one is added to or subtracted from POPEs,e,a,nsct. One is added if DIFFPOPnsct is positive and subtracted if DIFFPOPnsct is negative. Categories with zero values are not adjusted and no estimate can have negative value. The DIFFPOPnsct value also is adjusted by plus or minus one. This operation is repeated until DIFFPOPnsct reaches zero.

## ControlVitals

ControlVitals is a utility used to control detailed annual birth and death data to regional totals.

We obtain detailed birth and death information at the zip code level from the California Department of Health and/or San Diego County Health and Human Services. The zip code data is allocated to MGRA based on the prevent women of childbearing age b yMGRA (within the zip code). The babies are aggregated to get 0-year-old population by census tract and controlled to regional totals.

# ESTINC

ESTINC uses a mathematical model that represents the shape of the household income distribution curve. This model has undergone extensive evaluation and accurately replicates known income distributions for the region and its geographic areas. Three factors are needed to specify this model. The first, median household income represents the central point of the household income distribution. The second factor represents the distribution’s variability. The third factor is a non-linear adjustment that provides a more accurate estimate of the income distribution.

Estimates of households are produced for 10 income ranges: less than $15,000, $15,000 to $29,999, $30,000 to $44,999, $45,000 to $59,999, $60,000 to $74,999, $75,000 to $99,999, $100,000 to $124,999, $125,000 to $149,999, $150,000 to $199,999 and $200,000 and above. These income distribution estimates are expressed in constant (real) 1999 dollars. This follows the convention used in the 2000 census, which reports income for 1999. Median household income is calculated in both constant 1999 dollars and current dollars.

The estimates are produced using a top-down approach. Estimates are done first for the region, then for each SRA, and finally for census tracts.

## Income Model

Income distributions are characterized by a single mode and are positively skewed. ESTINC uses a mathematical model, known as the modified lognormal curve, to estimate this distribution. The classical lognormal curve fits the middle of the income distribution well, but it underestimates the proportion of low-income households and overestimates the proportion of high-income households. A modification was made to the lognormal curve to remedy this problem. To better understand the development of the modified lognormal curve, the lognormal model is first presented.

A variate X, int this case, income, is log normally distributed if the natural logarithm of X follows a normal curve. The lognormal distribution, Y, is minimally specified by its mean and standard deviation:

MED(Y) = ln[MED(X)]

SD(Y) = [2 \* (ln[MEAN(X)] - ln[MED(X)])]1/2

*where*

ln is the natural logarithm;

MED is the median;

SD is the standard deviation; and

MEAN is the average.

Cumulative probabilities for units with incomes less than X are computed by:

Yi = ln(Xi);

Zi = (1 / [SD(Y) \*√(2)]) \* [Yi - MED(Y)]; and

CUMPi = (1 + erf[Zi /√(2)]) / 2

*where*

i is income class;

Y is the upper limit of the income class;

Z is the lognormal standard score;

CUMP is the cumulative probability for income less than class Xi, and

erf is the error function of the normal curve.

The modification of the lognormal curve is made to the Zi equation by:

PARM = 1 / [SD(Y) \*√2]

Zi = PARM \* YiEXP \* [Yi - MED(Y)]

*where*

i is the income class;

PARM reflects the constant terms, and

YiEXP is a non-linear adjustment parameter.

The addition of YiEXP to the lognormal curve helps to correct its estimation problems. PARM and EXP are determined empirically by iteration to minimize the index of dissimilarity between the estimated curve and the observed income distribution (INCOMECALIB).

As the Zi equation shows, three parameters are needed to estimate the household income distribution: median income, PARM and EXP. There are six parameter types allowed for use in ESTINC, however generally only types 1, 2, and 3 are used. Types 4, 5, or 6 are assumed to be null (i.e. not used) unless overrides (i.e. setting the parameter value for type 4, 5, or 6 equal to 1) are specified by the model user.

* Type 1 = "asd" dispersion parameter, generally assumed to remain constant except during a recalibration period
* Type 2 = "nla" non-linear adjustment parameter, generally assumed to remain constant except during a recalibration period
* Type 3 = median income
* Type 4 = use base year distribution applied to current hh control
* Type 5 = use sra distribution applied to hh control
* Type 6 = use sra % change applied to base year distribution and current hh control

The derivation for these paramesters for the post-censal time period is described in Volume III, Data Definition.

The cumulative probabilities (CUMP) are translated into a probability distribution, P, of a household being in an individual income class by:

P1 = CUMP1, where 1 is the lowest income category

Pi = CUMPj+1 - CUMPj, where i ranges from 2 to n-1; j ranges from 1 to n; and n is the number of income categories

Pn = 1 – CUMPn-1

∑Pi = 1

*where*

∑is the summation over the income groups.

To provide estimates that are consistent with the 2000 census, this probability distribution is adjusted by the difference between the 2000 census proportionate income distribution and the 2000 proportionate distribution estimated from the modified lognormal. An additive adjustment insures that the resulting probability distribution will sum to one. This process is represented by:

ADJ2000i = CEN2000i – EST2000i

∑ADJ2000i = 0

*where*

∑is the summation over the income groups;

ADJPi = Pi + ADJ2000i

*where*

i is the income class;

ADJ2000 is the census calibration factor;

CEN2000 is the proportionate distribution from the census;

EST2000 is the 2000 estimated proportionate distribution from the modified lognormal curve; and

ADJP is the adjusted proportionate distribution for the estimate year, and

∑ADJPi = 1

*where*

∑is the summation over the income groups.

Households in each income class are determined by multiplying the adjusted probability distribution by the total households. This method insures that the sum over the households in each income class will equal the total households for a geographic area, within rounding error. The calculation of the household income distribution is done by:

ESTINCi = HH \* ADJPi

*where*

i is the income class;

ESTINC is the estimated household income distribution; and

HH is the estimated households from POPEST.

## Household Income Distribution Estimates

A household income distribution is first developed for the region based on the modified lognormal curve. Next distributions are developed for SRAs, which are controlled to the regional distribution. Finally, census tract household income distributions are controlled to their parent SRA. This multi-stage controlling process insures consistent information across all geographic levels.

### Initial Estimates

The initial or uncontrolled estimates of the household income distributions apply to SRAs and census tracts but not to the region, since the region’s income distribution estimate is not affected by controlling.

Also, the modified lognormal curve is not used for every SRA or census tract to generate the uncontrolled distribution, because the calibration showed too large a discrepancy between the observed and estimated 2000 income distributions. The initial income distributions for these areas are generated in one of two ways. For the SRAs and for census tracts with under 500 households, the base year (e.g., 2001 for the 2002 estimate) income distribution is used. The initial distribution for the other census tracts is the base year distribution adjusted by the change in the SRA income distribution between the base and estimate years. This is done as follows:

INCHGi,sra = ESTINC i,sra / BYINC i,sra

ESTINC i,ct = BYINC i,ct \* INCHG i,sra

*where*

i is the income class;

INCHG is the change in SRA income;

ESTINC is the estimated income distribution, and

BYINC is the base year income distribution.

### Controlling

An iterative proportionate adjustment method is used to factor the initial SRA income distribution matrix (ESTINC i,sra) to both row totals (SRA household estimate from POPEST (HHsra)) and column totals (regional income distribution estimate (ESTINC i)). It is important that the grand totals of HHsra and ESTINC i match; that is:

TROW = ∑HH sra;

TCOL = ∑ESTINC i, and

TROW = TCOL

where

TROW is the sum over the SRA household estimates; and

TCOL is the sum across the regional household income distribution estimates.

The adjustment occurs by an iterative sequential raking of the ESTINCi,sra matrix, alternating with rows and columns. Starting with a proportional adjustment to the SRA household estimate (rows), the matrix is changed by:

SUMROWsra = ∑ESTINC i,sra

ESTINC i,sra(1) = ESTINC i,sra \* (HH sra / SUMROWsra)

where

SUMROW is the summation across income categories for each SRA; and

(1) represents the first adjustment to the matrix.

This is followed by a proportionate adjustment to the regional income distribution estimate (columns) by:

SUMCOL i = ∑ESTINC i,sra(1)

ESTINC i,sra(2) = ESTINC i,sra(1) \* (ESTINC i / SUMCOL i)

*where*

SUMCOL is the summation across SRAs for each income category; and

(2) represents the second adjustment to the matrix.

The second adjustment to the regional estimate throws the figures out of line with the SRA household estimate; that is, the sum across income categories no longer matches the SRA household figure. The next step is to readjust the figures with respect to the row totals and obtain ESTINC i,sra(3). This readjustment to the row totals throws the figures out of line with respect to the regional income distribution estimate; that is, the sum across SRAs no longer matches the regional estimates, and so the figures are again adjusted to conform to the column totals to obtain ESTINC i,sra(4). This adjustment sequence is continued until convergence with a final unchanging matrix, which adds to the SRA household estimate and regional income distribution estimate. The program allows for 15 iterations that reaches convergence within rounding error. The iterative proportionate adjustment routine is then used to control the initial census tract income distribution matrix to both row totals (census tract household estimate from POPEST) and column totals (SRA income distribution estimate).

## IncomeCalib

The income modified lognormal curve is calibrated using the CONCEP utility program INCOMECALIB. It calibrates to known income distributions such as the census. Beginning in 2011, the model inputs were re-benchmarked using data from the 2005-09 American Community Survey for the region and census tracts. Using the median income and income distribution as inputs, this program estimates the other two parameters needed to fully specify the modified lognormal income distribution model (PARM and EXP). An iterative method is used to find the parameters that minimize the error between the observed and the estimated income distributions. This error is measured by the index of dissimilarity.

### Index of Dissimilarity

The index of dissimilarity (IOD) is often used to measure the inequality of two distributions. It is computed by comparing two percentage distributions as follows:

IOD = Σ⏐Acti - Esti⏐ \* .5

*where*

Σ is the summation across income groups;

⏐ ⏐ is the absolute value;

Act is the observed percent of households in an income group; and

Est is the estimated percent based on the modified lognormal curve

IOD is interpreted as the percentage that the estimated distribution would have to change in order to match the actual distribution. Given that the objective is to provide reliable estimates, and based on the analysis of the calibration to 2000 census data for the region, SRAs, census tracts, the following guidelines were used to evaluate the IOD. An IOD less than 7 percent indicated that the modified lognormal curve generated an accurate income distribution estimate, while an IOD greater than 9 indicated that the curve’s estimate was not precise enough. Areas with IODs between 7 and 9 percent were examined on a case-by-case basis. The modified lognormal curve produced an acceptable estimate for most of them. For those areas where the modified lognormal curve could not be used to estimate the income distribution, other procedures were developed (see Initial Estimates section in the previous chapter).

### Algorithm

For each geographic area, INCOMECALIB produces an estimated income distribution based on the modified lognormal curve for 9,250 combinations of PARM and EXP in conjunction with the observed median and income distribution. The PARM and EXP values are controlled by user defined calibration limits, which are 125 to 310 for EXP and 1 to 50 for PARM. It determines which of the 9,250 (185 values for EXP X 50 values for PARM) estimated income distributions are the most accurate using the following logic:

DISSL = 1,000,000,000

FOR EXP 125 to 310

FOR PARM 1 to 50

EXP = EXP / 100 /\* Convert EXP to appropriate scale

PARM = PARM / 1000 /\* Convert PARM to appropriate scale

Solve Curve

Compute IOD

IF (IOD < DISSL)

X1 = EXP

X2 = PARM

DISSL = IOD

CONTINUE LOOPS

ELSE

CONTINUE LOOPS

The first time through the loop (EXP = 125 and PARM = 1) the IOD will be less than the initialized value of DISSL and it will replace the DISSL value, and EXP and PARM will be stored in X1 and X2, respectively. The next time through the loop (EXP = 125 and PARM = 2), the IOD from the newly created estimate is compared to the prior value. If it exceeds the prior value, the loop continues and X1, X2 and DISSL are unchanged. If the IOD is less, X1, X2 and DISSL are replaced before continuing the loop. This process is repeated until all values for EXP and PARM, represented in the loops, are examined; therefore, the final values of X1, X2 and DISSL represent the EXP, PARM and IOD of the income distribution estimate that best matches the observed income distribution. INCOMECALIB then re-generates the income distribution estimate based on the optimal values for EXP and PARM.

The accuracy of the estimated income distribution is evaluated by examining the IOD and the actual and estimated percentages. The other information in the table relates to mechanics of the modified lognormal curve and is not that pertinent to the end user. The last income class contains no information for either Z-score or the adjustment, because its estimated percentage is derived as a residual.

The actual Z-score is computed from the observed (e.g., census) cumulative proportionate income distribution by:

FOR I = 1, 1000

ZOBS = (I - 500) / 100 /\* Convert to Z-score value scale (-4.99 to 5.00)

PROB = (1 + erf((Z / √2)) /2

IF (PROB ≥ CUMP) END LOOP

CONTINUE LOOP

This routine computes the cumulative probability under the normal curve for Z scores that can range from -4.99 to 5.00 (PROB). It checks PROB for a given Z-score against the observed cumulative proportion for an income group. The observed Z-score for a given income class is the one that yields a PROB that matches the observed cumulative portion (CUMP). This routine is repeated for each income class.

The estimated Z and adjustment come from the derivation of the modified lognormal curve as follows:

Zi = PARM \* YiEXP \* [Yi - MED(Y)]; and

ADJi = PARM \* YiEXP

*where*

i is the income class;

Z is the estimated Z-score;

Y is the natural logarithm of the upper limit of the income class; and

MED(Y) is the natural logarithm of the median income.

# PDHH - POPEST

Beginning in 2011, POPEST estimates included detailed household characteristics for households. These are ultimately intended as input to the POPSYN synthetic population estimates program used by the transportation modeling group in conjunction with the activity-based transportation model (AMB). The new detailed household characteristics include:

* Households by household size (1, 2, 3, 4+)
* Households by presence of children (0, 1+)
* Households by number of workers in household (0, 1, 2, 3+)
* Group quarters
  + Military
  + Civilian - college
  + Civilian – other

Estimates are produced at the census tract level, and then allocated to MGRA. From MGRA, data are grouped to TAZ.

### Households by household size (1, 2, 3, 4+)

The distribution of households by size is based on a Poisson distribution as:

hhs(j) =

*where*

m = household population / households

If m = 1, set j(1) = 1, and set j(2,3,4,5,6,7)=0, else

distribution mean (d) = m - 1

distribution of households by size for tract for all tracts (c) and household sizes (j), where 1<=j<=7

The distribution works best with many categories at the high-end of the spectrum. After the initial (seed) value is derived, the estimated distribution is adjusted with an error term. The error term is derived from the relationship of estimated values to observed values from the 2000 and 2010 census. For each Census Tract, an error term is derived for the estimates years 2001 – 2009 by interpolating between the observed 2000 and 2010 data points. After the estimated distributions are adjusted for errors, the distribution is normalized to ensure a sum to 1.

At this point any override values for CT distributions are added. Some special tracts, particularly military tracts with small numbers of households, don’t change over time. We override them with observed data. The distributions are applied to the total number of households to estimate the number of households in each size category.

The next step controls the estimated households by size to regional totals. Overridden Census Tracts are not included in the controlling and the regional totals are adjusted accordingly. The program then compares the resulting implied household population with the POPEST household population and makes minor adjustments in the distribution to ensure that the implied hhp is consistent with the POPEST total; that is HHP >= 1\*HH1+2\*HH2+3\*HH3+…+7\*HH7.

### Households by number of children under age 18 in household (0, 1, 2+):

This methodology is based on the limited observed data for households with and without children. Starting with the number of households with children from 2000 Census, and the total number of children (pop 0 – 17) from the current estimates data, we derive a factor: #children/household for households with children. Then for any given estimates year, get total children (pop 0 – 17) from estimates and divide total kids by factor to get an estimate of households with children (hhwc). Compute the households without children (hhwoc) as a residual: hhwoc = hh – hhwc. Then, control to regional control totals.

Controlling has constraints. For 2000, the estimates are tautological. For other years, we likely overestimate hhwc because we aren’t “modeling” the kids/household factor from the 2000 Census. In any case, controlling can significantly modify the estimated hhwoc and hhwc values. At this point we established a threshold of about 10% as minimum hhwc estimate to be eligible for controlling. Based on initial runs for 2010, most of the CTs with fewer than 500 estimated hhwc were excluded from controlling. With the established threshold, for all CTs not conforming, hhwc and hhwoc were summed and excluded from the regional controls in the 2-way controlling algorithm. These are treated the same as overrides. The presence of children should be consistent with local age data:

* Pop age 18 and younger must be >= 1\*HH\_child1+2\*HH\_child2
* Pop age 18 and younger must not be so large as to imply excessively large number of children per household (relative to observed data).

### Households by workers in household (0, 1, 2, 3+):

Estimating the number of workers in households starts with household by household size distribution aggregated to 4 categories multiplied by the percent distribution of presence of workers to households by size, based on observed distribution. This distribution is a user-defined input based on Census 2000 data.

**Example (numbers are for illustrative purposes only)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Workers \ HH size | 1 person HH | 2 person HH | 3 person HH | 4+ person HH |
| 0 workers | 50% | 30% | 10% | 20% |
| 1 worker | 50% | 40% | 30% | 30% |
| 2 workers | 0% | 30% | 50% | 30% |
| 3+ workers | 0% | 0% | 10% | 20% |
| Sum | 100% | 100% | 100% | 100% |

The estimates are controlled to regional totals. The presence of workers should be consistent with local age data:

* Pop age 16-64 must be >= 1\*HH\_worker1+2\*HH\_worker2+3\*HH\_worker3
* Pop age 16-64 must not be so large as to imply excessively large number of workers per household in 3+ category

### Allocation to MGRAs

The estimates of detailed household characters from the three categories are then allocated to Series 12 MGRAs using a straight cumulative probability distribution. This assumes that each MGRA will essentially have a distribution that mirrors its containing Census Tract. There is, however, no round off error.

### Group Quarters by Type

See POPEST section on CONTROLGQ for details.

**Section 3**

Data Definition

# Data Sources and Storage

Data for the CONCEP models are gathered from a wide variety of sources. Data are then stored in tvarious places, depending on data type. MS Excel files are stored in various subdirectories in the either the directory **M:\RES\estimates & forecast\concep** or in the directory **M:\RES\estimates & forecast\popsyn.** Database tables are stored in the **CONCEP\_SRXX** database on the SQL server, where XX represents the current forecasting series (12, 13, etc.). As of this writing the active CONCEP database is concept\_sr12 reflecting the Series12 forecast time period.

## Base Year Population and Housing

2000 population counts from the U.S. Census Bureau comprise the base data for the population by age, sex, and ethnicity. The Census Bureau data file that contains 100% population counts for the region and its census tracts and sufficient detail about age, sex and ethnicity is Summary File 1 (SF1). The Summary File 3 (SF3) contains detailed information on housing stock by structure type, as well as household income distributions.

The U.S. Census Bureau did not control its Summary File 3 (sample survey) variables to the 100% count Summary File 1. To ensure proper model function, SANDAG performed controlling procedures to ensure, for example, that housing stock by structure type counts in Summary File 3 summed to total housing stock counts from Summary File 1. In addition, some minor adjustments were made (such as correcting the location of housing units that were erroneously allocated by the Census Bureau to roads and open space) to more accurately reflect the region’s true population and housing distribution.

The SQL server table **census\_MGRA\_sf1\_sr10** contains basic population characteristics, and **census\_MGRA\_sf1\_5ya\_sr10** contains detailed demographic (five year age group) characteristics from the Census 2000 Summary File 1. Summary File 3 data from Census 2000 is found in the SQL server tables **census\_MGRA\_sf3\_1\_sr10**; **census\_MGRA\_sf3\_2\_sr10**, and

**census\_MGRA\_sf3\_3\_sr10**. Starting with Version 3.0, these tables were used to build compatible tables for MGRAs.

## Population and Housing Control Totals

Each year the California Department of Finance (DOF) produces estimates of population and housing stock for each county and city in California. Known as the “E-5 City/County Population and Housing Estimates,” these county and jurisdiction estimates serve as the control totals for the POPEST model for the San Diego region.

## Housing Stock Changes

## Occupancy Rates and Household Size

Occupancy rates are estimated from the 2000 Census. Average household size is based on trends between the 1990 and 2000 Censuses and on data including school enrollment and vital statistics counts.

## Group Quarters Population

The California Department of Finance collects military population information and SANDAG is authorized to receive that information from DOF. The Department of Finance provides an annual tabulation of on-base military population for each military installation in the region. In addition, DOF provides a report of group quarters populations for local universities and correctional facilities. These estimates form the basis of the annual non-household population estimate used in the POPEST model. Allocations of group quarters population to MGRAs are completed in a spreadsheet that is stored in **M:\RES\estimates & forecast\concep\inputs\GroupQuarters**.

Detailed demographic information about the characteristics of the uniformed military population and military dependent population is not available at the census tract level, but the Public Use Microdata Sample (PUMS) from the Census 2000 provides characteristics at the PUMA level. Census data from 1990 were used to develop some of the parameters of the estimates model.

## Vital Events

For all vital events (births and deaths) a certificate is issued. Historically a summary table of these certificates was used to determine the region’s vital events. These records provide the most detail regarding vital events. However, reporting restrictions and data errors have reduced their effectiveness in tracking vital events for modeling purposes. Therefore SANDAG has shifted to a two-pronged method for tracking deaths, and a three-pronged method for tracking births. In both cases, vital records are used to determine region controls for births by ethnicity/race and sex, and for deaths by age, ethnicity and sex. The controlling of births and deaths is done with the CONCEP utility program CONTROLVITALS.

For deaths, SANDAG uses sex, age, and race/ethnicity-specific survival rate data to produce initial census tract level estimates of deaths. These estimates are then controlled to the state control totals for deaths by sex, age and race/ethnicity.

For births, SANDAG uses both region control totals as well as zip code control totals. The zip code birth information from the Department of Health Services has a much higher degree of accuracy than the census tract records do. The use of zip code control totals reduces errors in small area birth estimates that would result from simply applying fertility rates to the population. In the case of small areas surrounding universities, for example, there are large proportions of women of childbearing age, but few births occur in these areas. Thus, rather than applying fertility rates to the population (as survival rates are applied to calculate deaths), the zip code birth records are distributed directly to MGRAs. The MGRA birth estimates are controlled to the region totals. The MGRA estimates are then aggregated back up to census tracts, which are the primary modeling unit in CONCEP.

Additionally, as a result of reporting differences (birth records show a high degree of divergence from Census reporting in terms of race/ethnicity) births control totals for small populations (American Indian, Asian, Hawaiian & Pacific Islanders, Other and 2 or More Races) are smoothed to more closely reflect the Census distribution of 0-4 year olds.

Vital events data are stored in **M:\RES\estimates & forecast\concep\inputs\VitalEvents**.

## Net Migration

Census Bureau data are used to calculate net migration. The files used in the age, sex and ethnicity estimates model include the 1990 Modified Age, Race, Sex (MARS) file, and the 1990 and 2000 Public Use Microdata Sample (PUMS) files. The survival rates used in the migration calculations were based upon vital events records. (See above.)

Migration rates by sex, age and ethnicity were calculated between the 1990 and 2000 Census. These migration rates are stored in **M:\RES\estimates & forecast\concep\inputs\Migration**.

Net migration estimates from 1990 to 2000 are calculated for the civilian population. Civilian population excludes the uniformed military. For the region, this is done in two stages. Civilian net migration is first calculated by sex and ethnic group using the residual method. The life table survival rate procedure is then used to calculate net migration by age within each sex and ethnic group. The regional net migration levels are converted to annual rates. Civilian net migrants by sex and ethnic group also are estimated for each SRA using the residual method.

## Income

Base year household income distribution information is derived from the decennial census, conducted by the U.S. Census Bureau (see above).

Trends in median income are determined from a range of data sources, including the U.S. Census Bureau’s Current Population Survey (CPS) which reports data for the state of California and American Community Survey (ACS) which tracks data for the San Diego Region on an annual basis, and for smaller geographies (e.g. cities and census tracts) on a rolling basis of 5-year aggregated survey data. The first of these 5-year releases occurred in 2010 for the period 2005-09, and income estimates are re-benchmarked to that point. Other sources used to estimate the regional trend for median income include the California Department of Finance, and the California Franchise Tax Board. During model testing, the Census Bureau growth rates (from the CPS and ACS) tracked most closely with historical trends, and thus were used for estimation purposes.

Income trends are calculated and stored in **M:\RES\estimates & forecast\concep\inputs\Income**. Income trends are estimated for the region and cities from data reported in the U.S. Census Bureau’s American Community Survey. Tracts are estimated based upon the regional and city trends.

## Detailed Household Characteristics

The raw data used to estimate households by household size, households by the presence of children and households with workers were obtained primarily from the 2000 Census, the 2010 Census and the American Community Survey. Description TBD HERE.

**Section 4**

Programmer’s Reference

This document provides general programming information for CONCEP, including a description of source files, data, and utilities. The discussion covers the current program organization and operation, and provides general detail to assist in making changes to source code to prepare executable programs. Additional technical information describing the various relationships and equations is available in Section 2 of this document. Detailed program documentation is provided in each source file as described below. Furthermore, some of what is discussed here may be described in more detail in Section 3 – Data Design and Database Documentation, which contains the detailed database formats and organization for the input and output data discussed in each section.

CONCEP encompasses several computer programs and a variety of data sources. In the following discussions, the environment is described in terms of the primary programs and ancillary programs, utilities, and procedures.

# Environment

The public version of CONCEP is stored on SANDAG’s shared PC network, in the common directory **m:\res\estimates & forecast\concep***.* Most of the CONCEP data is maintained in the SQL Server 2010 relational database management system on the SANDAGNET server **pila**. The primary database for CONCEP is **concep\_sr12***.* (In practice, the CONCEP project is copied to a local workstation for execution. Table 4-1 summarizes the organization of the principal CONCEP development and operational software.

# Programming

CONCEP is written in Microsoft’s C# using the Visual Studio.Net development environment. Data originating in or being sent to database tables are managed in a variety of ways. Programmatically, access to database tables is provided through Active-X Data Objects (ADO).NET. Some utilities for constructing or analyzing data tables are written as SQL scripts that are executed with SQL Server query tools.

# Nomenclature

In general, a file or table name is descriptive and its extension, if applicable, identifies its purpose. The Visual Studio.Net environment is case sensitive. Most program names are in mixed (upper and lower) case, may be up to 64 characters, and may use the underscore character (“\_”) or the dot character (“.”) in the name or extension. Database tables have no extension, but may make use of the underscore character. The following nomenclature is used to help identify the assorted programs, tables, files and scripts used in CONCEP (the <> symbols represent a name or parameter and are not typed). Only the principal source and data files are listed here; some source files used for cosmetic effects are not listed. Consult the documentation for the Visual Studio.Net for a complete description of the available file types.

* **<file\_name>.csproj**: identifies C# project files.
* **<file\_name>.cs*:*** identifies C# source files.

Unless stated otherwise, files are assumed to reside in the **code** directory. Path and file names are listed in **bold face**. Files with various year designators are listed as “\_YYYY” for YYYY = 2000, 2001, and so on. Files with year designators lagged one year are listed as “\_L” for \_L = YYYY -1.

# Components

Table 4-2 lists the major source code components of CONCEP. There are other source files in the CONCEP project that are used to display help windows; they are not described here.

**Table 4-2**

**CONCEP Code Components**

|  |  |
| --- | --- |
| **Component** | **Description** |
| **concep.csproj** | CONCEP project file |
| **cmain.cs**  **concepUtils.cs**  **controlVitals,cs** | CONCEP C# main menu source file  Collection of CONCEP Utilities  ControlVitals program source file |
| **estinc.cs**  **IncomeCalib.cs** | ESTINC C# source file  Income calibration program source file |
| **pasee.cs** | PASEE C# main source file |
| **pasee\_compute.cs** | PASEE C# utilities source file |
| **pasee\_ct.cs** | PASEE C# census tract classes and methods source file |
| **pasee\_mgra.cs** | PASEE C# MGRA classes and methods source file |
| **pasee\_sra.cs**  **pdhh.cs** | PASEE C# SRA classes and methods source file  Detailed HH characteristics program source file |
| **POPEST.cs** | POPEST C# source file. |

**Section 5**

VERSION description

CONCEP Version 3.3:

* Major recoding to add estimates of detailed household characteristics for use with PopSyn.
* Recoding to consolidate utilities in a separate source file.
* Changes to interface to include new functionality
* Reverting to Series 12 MGRAS as the atomic geographic unit for estimates
* Introduced the software version control environment TortoiseSVN

CONCEP Version 3.2:

* Minor changes for 2006 Current Estimates. Added processing to control the Group Quarters populations. Added “estimates\_year” column to migration rates table to permit changing them over time.

CONCEP Version 3.1:

* Replaced change files for housing stock and group quarters with total count files based on landcore parcels aggregated to MGRAs

CONCEP Version 3.0:

* Replaced MGRAs with SGRAs.
* Changed methodology to compute HS changes to parcels and Landcore assignments
* Removed suffix “\_newgroups” from CONCEP income tables
* Changed database to concept\_sr11 to distinguish MGRA-based code and tables from those based on MGRAs.

CONCEP Version 2.5:

* Eliminated POPEST split tract processing in favor of POPEST MGRA-level computations. POPEST was recoded to do controlling for cities rather than split tracts, adding controlling for households by structure type.
* Changed ESTINC to do income estimates for 10 income groups. Recalibrated the income curve base on 2000 Census data for the 10 income groups.
* Added the suffix “\_newgroups” to all income table names. Moved old ESTINC tables to **kahuna.old\_concep**.
* Changed controlling for births to control to ZIP CODE estimates and allocate to MGRAs, then rebuild **births\_ct\_YYYY** for input to PASEE.
* Modified CONTROLVITALS program to handle the new controlling for births to zip codes.

CONCEP Version 2: (skipped)

CONCEP Version 1:

* Coded in Visual Studio C# combining POPEST, POPEST, PASEE and ESTINC with one interface.