

Data Sources

Introduction

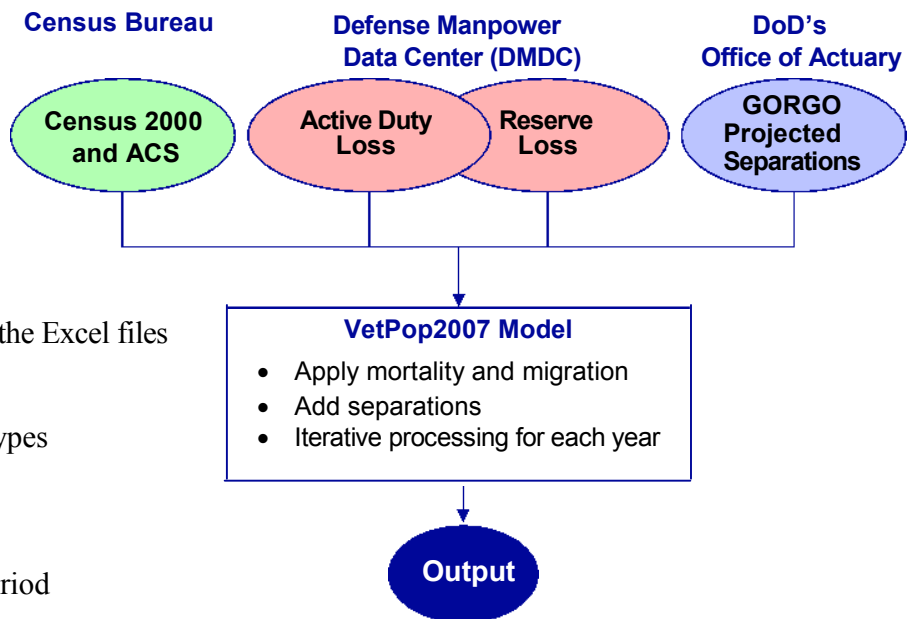
This section (part of the VetPop2007 Technical Documentation) describes in general terms the model output and the sources of veteran data used in the VetPop2007 model. The diagram below provides a simple representation.

Model Output

The final product for estimating and projecting the veteran population from April 1, 2000 to September 20, 2036 is prepared in two formats – statistical analysis software (SAS) and Microsoft Excel. The description of main variables in the SAS files is included in the Appendix; the structure and use of the Excel files are found in the Tutorial document.

The output files contain the following types of information:

- Cumulative living at a given date
- Deceased veterans within a time period
- New separations within a time period



The Excel files are intended to ease user interaction for obtaining veteran distribution by specific characteristics: Age, Gender, Race, Branch, Rank (Officer/Enlisted), and State. Due to the large size of the SAS files and limitations in Excel, the SAS files are converted into a number of Microsoft Access files as an intermediate step before transporting the data into Excel.

Data Sources

These are the main data sources¹ for VetPop2007:

- Census Bureau – Census 2000
- Defense Manpower Data Center (DMDC) – Active Duty
- Defense Manpower Data Center (DMDC) – Reserve
- Office of the Actuary of the Department of Defense – GORGO projection
- Census Bureau – American Community Survey (ACS)

1. The data sources for county are not included in this document.

2 VETERAN POPULATION MODEL

VetPop 2007

A combination of Census 2000, DMDC losses, and GORGO projections determines the veteran counts by the specific characteristics listed under Model Output. ACS also holds an important role by providing the state distributions in recent years. Some data highlights are described below, while the general analysis approach is described in the Analysis section.

Data on actual separations from April 2, 2000 through September 30, 2006, as well as actual deaths of those veterans who separated in this time frame, came from DMDC. Though DMDC data on separations prior to April 2, 2000, are available, the Census data are assumed to be more reliable. For each of the regular military and reserve components, DMDC provides two files, a master file and a duplicate file. The master file contains demographic information along with the most recent separation data for the veteran; the duplicate file contains information on all previous separations from active duty since 1970.

You will find additional information on these and other data sources in the Parameters section and Appendix.

Census 2000 Data

As part of the Census 2000, the Census Bureau published veteran data in the Summary Files, SF3 and SF4. Because the veteran distribution information in these files was not sufficient to meet VA's needs, the Census Bureau produced more detailed tabulations, based on the long form data, as requested by VA.

Highlights

Of the many tabulation files requested, a single file of veteran data — aggregated by State, Age, Gender, Period of Service, and Race² — is incorporated into the model to form the veteran count as of 4/1/2000.

- State includes DC, Puerto Rico, and US Island Areas³ (Virgin Islands, American Samoa, Guam, and Northern Marianas)
- Age 17+, where 90+ is lumped into a single category
- Race categories, as defined by OACT, are Hispanic of any race, White Non-Hispanic, Black Non-Hispanic, American Indian Non-Hispanic, Asian Non-Hispanic, Pacific Islander Non-Hispanic, and Other or Multiple Race Non-Hispanic
- Period of Service is categorized by the 13 periods specified by VA

-
2. Additional variables (income and POSTVNE—post Vietnam era indicator) were requested but ultimately not used in VetPop2007.
 3. US Island Area data from SF4 were also used in the model.

Because of disclosure concerns, the Census Bureau rounded each cell⁴ in the files received by VA as follows:

- 0 rounds to 0
- 1–7 rounds to 4
- 8–12 rounds to 10
- above that, round to the nearest multiple of 5

Active Duty Data

Data on separations from regular military components — Army, Navy, Air Force, Marines — reside in the Active Duty loss file.

Highlights

The DMDC person level file used within VetPop2007 contains the following information: Date of Birth, Gender, Branch, Rank, Race, Activation and Separation dates, Date of Death (if applicable), and Active flag.

- Age at Separation is calculated from the Date of Birth and the Separation period. (Separation period indicates the respective fiscal year based on the separation date.)
- Period of Service is calculated from the Activation and Separation dates.
- Date of Death comes from a match to the Social Security Administration (SSA) Death file, which may miss some deaths that were not reported to SSA.
- Active flag indicates that the person was on active duty as of September 30, 2006.

Reserve Data

The Reserve Loss file contains information on Reserve and National Guard members. It includes those who were serving in the Reserve Forces as of September 30, 2006, as well as those who served in the Reserves in the past and been discharged. As such, the file includes persons who completed their tour of duty, retired from service, or died before being discharged. In addition, all types of reserves - Selected, Active/Guard Reserve (AGR), Military Technician, Individual Ready Reserve/Inactive National Guard, Standby Reserve, and Retired Reserve - are considered, but only those with presidential activation are counted as veterans in VetPop.

Highlights

A person level file obtained from DMDC provides Age, Gender, Branch, Rank, Race, Activation and Separation dates (from contingency data), Date of Death (if applicable), and Active flag.

4. Each cell is determined by State, Age, Gender, Race, Period of Service, POSTVNE, and Income.

4 VETERAN POPULATION MODEL

VetPop 2007

- Age is calculated from the Date of Birth.
- Period of Service is calculated from the Activation and Separation dates.
- Date of Death comes from a match to the SSA Death file.
- Active flag (or Active/Guard Reserve flag) is as of September 30, 2006.
- Veteran Status is determined using the Contingency data, which indicates a presidential activation such as Desert Storm, and thus a “veteran” by VA’s definition.

Although the reserve master file should contain only one record per person, it actually contains duplicates. This is because the reserve data are matched by SSN *and* branch, not by SSN alone. That is, DMDC treats “John Smith,” who started in the Navy Reserve, as a separate record from “John Smith,” who later joined the Army Reserve. The duplicate file contains some additional information for persons with more than one separation from active duty since 1980. The most recent separation occurs in the master file, and other separations fall into the duplicate file.

GORGO Data

VA obtains GORGO data annually from DoD’s Office of the Actuary. The data file contains DoD projections for the number of separations from active duty for each of the next 100 fiscal years. The purpose of the GORGO model is to value retirement liabilities annually.

For VetPop2007, OACT uses GORGO2006, the active military projections as of September 30, 2006, as the projected separations data.

Highlights

DoD’s OACT provides VA with celled data, including Fiscal Year, Age, Type of Separation, Number of Years of Service, and Rank.

- Age 18+ (age nearest birthday at the end of the Fiscal Year)
- Separation types: retirement, withdrawal, permanent disability, temporary disability, and death
- Average exact years of service at (mid-year) termination
- Active duty military is defined as active duty personnel plus full-time reservists; it excludes reservists who are not full time as well as uniformed Non-defense personnel

The projection uses an open-group basis, which assumes that new service members will enter the military in sufficient number to maintain a specified force level.

American Community Survey Data

ACS is a nationwide survey conducted on an ongoing basis to provide the nation with demographic data on a yearly basis instead of every 10 years.

At present, VetPop uses ACS data only for interstate migration. In the future, more extensive use of the survey is anticipated as more detailed geographic data become available.

Highlights

For VetPop2007, VA obtained special tabulations about veterans from ACS for the periods 2003 through 2006. The tabulations consisted of aggregated veteran counts by State, Gender, and Age group.

- State includes DC and Puerto Rico, whenever possible
- 4 age groups: <30, 30-44, 45-64 and 65+

During the analysis for migration, data for the less than 30 age group were found to be too volatile between years; thus, the <30 and 30-44 age groups were combined.

Analysis

Why?

OACT staff must analyze the input data sources in the VetPop model for accuracy, validity, and a general understanding of the data.

How?

When the files are received, conduct the following analyses:

- Check the range of values for each variable to ensure a clear understanding of what each value means. For example, age is expected to be greater than or equal to 17. Finding a negative or otherwise unreasonable age, such as separating at age 10, would obviously be an error.
- Check the distribution of one variable against another to ensure our general understanding of the data. For example, you would not expect to see someone who is 25 years old in 2002 with a history of service from World War II.
- To further ensure the accuracy of the data, compare the current year's data extract with the previous year's data extract.
- Finally, in the case of GORGO data, you can compare predicted values from the previous year's data extract with the actual values from the current year's DMDC data.

Issues

For problems that still exist after you have completed these checks, there are two options: 1) go back to the data source (Census Bureau, DMDC, or DoD's OACT) to see if the data can be corrected at the source; 2) reconcile issues in the model at the earliest possible stage.

6 VETERAN POPULATION MODEL
VetPop 2004 Version 1.0

The next step is to look at certain factors that may require further processing in the program.

- A variable with missing values may require imputation; for example, missing gender in the DMDC data files.
- Processing may be required to deal with values that are outside the expected range or that appear unreasonable; for example, finding a specific date like 19000000 in the DMDC data file, when you are expecting a separation date in the form YYYYMMDD.
- You might assume that the person level file would have one record per person. Because of the complexity of the reserve files, though, this is not the case. In fact, looking at the multiple records per person will give you a much better understanding of the data — and can often provide ideas for future modifications to further improve the quality and usability of the data. This issue is discussed in greater detail in the Programs section. In general terms, however, you should always ask the following questions:
 - Do I want the most recent data?
 - Should one thing take priority over another (such as active over reserves, or federally activated over state or not activated)?
 - Should I be sure to pick the record that does not have a missing value?
- In a few instances, multiple yet differing sources of information may exist. If certain variables obtained from different sources are not the same, then the trick is to understand why they are different and decide which one to use. A perfect example of this is mortality. Sources of information are the Social Security Death File, the Veterans Affairs Compensation & Pension File, and the DMDC data (such as reason for separation).
- Make sure that you understand how the data are delivered. In the case of Census data, this involves understanding the process of rounding the data. In the case of DMDC data, you would need to know exactly which databases are used to create the file that we receive. This step is necessary because DMDC combines a “point-in-time master file” (representing current military personnel as of September 30, 2006 for example) with a “cumulative loss file” (representing previous military personnel at different points in time).
- You will need to make other decisions, either in creating the data or processing the data, to adjust for persons whom we want or do not want in our population. In the case of DMDC data, this may mean deciding how to handle reenlistments. In the case of GORGO data, this would mean understanding that the file does not include persons who enter and leave in the same fiscal year (termed “reentrants”).

An important question to ask is, “How is a given item created in the first place?”

Consider, for example, Period of Service; in the Census data, there is a special Period of Service variable for VA. This data is ultimately based on understanding the long form question that asks respondents which period they served in, given a specific range of dates. In the DMDC data, actual activation and separation dates are provided, allowing us to generate the Period of Service ourselves. However, such data prior to about July 1970 are not available.

(Information on how the VetPop2007 model processes the data is covered in the Programs section.)

OFFICE OF POLICY (008A2)