MARGene

Minorities at Risk Data Generation and Management Program

v1.03
Documentation

Documentation updated January 27, 2003

Jonathan Wilkenfeld
Interim Director, Minorities at Risk
Director, Center for International Development and Conflict Management
University of Maryland
145 Tydings Hall
College Park, Maryland 20742-7231
jwilkenfeld@gvpt.umd.edu

MARGene can be accessed through the MAR website:
http://www.cidcm.umd.edu/inscr/mar

MARGene was developed under support provided by a grant from the Carnegie Foundation, International Peace and Security Division. Primary developers of the *MARGene* project were D. Scott Bennett and Christian Davenport. Significant development work and data conversion was performed by Mitchell Brown, University of Maryland.

Table of Contents

Overview

The Minorities at Risk Data Generation and Management Program (*MARGene*) is designed to allow easy access to data in the Minorities at Risk data set developed by Ted Gurr and associates (Gurr et al. 1993). *MARGene* makes it easy for scholars to access variables from this data set, creating subsets of the data, interpolating across unobserved data points, and setting up the data for merging with other state-level data sets. Running in a standard Microsoft Windows environment, the program allows users to select variables from the Minorities at Risk data set, specify subsets of the data based on time and space, and specify various options concerning data creation (such as whether to interpolate missing data or not). The program then creates a new, customized data that can be loaded into statistical analysis software such as State, SPSS, or Limdep, for further manipulation and/or analysis.

The Minorities at Risk (MAR) Project is "an independent, university-based research project that monitors and analyzes the status and conflicts of politically-active communal groups in all countries in the world with a current population of at least 500,000. The project is designed to provide information in a standardized format that will aid comparative research and contribute to the understanding and peaceful accommodation of conflicts involving communal groups. Selected project materials on over 285 groups are available through this site for the information of researchers, students, public officials, journalists, activists, and other interested individuals, including access to the regularly updated MAR database and codebook" (MAR website, http://www.cidcm.umd.edu/inscr/mar, October 2002).

Rationale

The *MARGene* project has focused on improving the accessibility of the Minorities at Risk data set in several ways. While the Minorities at Risk data set has been rich in information, the format of the quantitative data set has made conducting statistical analysis of the data more difficult than necessary.

1) The MAR data has been organized in such a way as to make use in some standard research design setups quite difficult. The unit of analysis of the pre-2003 data set is the group. Individual variables (of the almost 1000 variables) then combine elements of substance and time. For instance, "comcon91" represents the annual communal conflict index in 1991. Rather than a set of substantive variables paired with variables giving with identifying features of the case (state, group, and time), identifying features like time become part of the variable. This makes merging with typical pooled cross-national or time-series data sets difficult, as there is no "time" variable. Users must undertake significant data manipulation in order to analyze trends over time. A more standard setup would have a variable such as "comcon" with the unit of analysis being the group-year. That is, the variable would be communal conflict index, and there would be observations on this variable in separate years, e.g. 1990, 1991, 1992, etc. Key other variables in the data set would identify the group identity and the year of the observation. The data set would then take the form of a standard pooled time-series of groups observed over time. The *MARGene* project has included conversion of the data set to a more standard pooled, cross-national, time-series data set.

2) The MAR data has been inconsistent in terms of the frequency of data gathering on different variables and over time. Some data has been collected at 5 year intervals, other data annually. Some data was collected infrequently at first (e.g. every 5 years) but then collection was modified to seek new data annually or biennially. In most cases, the data was collected at "wider" intervals (rather than annually) because the variable values changed quite slowly. Changing sampling patterns lead to difficulties in analysis, however, because observations are not uniform. A five year interval in which a group faced some conditions in some ways should count as five-times more important than a one year interval in some other condition. But the variables in the data set represent these unequal time periods as the same (one variable). It was incumbent on users of the data set to recognize and adopt these differences in sampling to their own needs. *MARGene* allows users to create interpolations of the data within these larger intervals automatically, leading to a uniform and standard annualized data set appropriate for statistical analysis.
3) Users frequently have different needs in their data analysis. Some wish to analyze only one region of the world or selected groups; others wish to analyze all regions and all groups. Some wish to analyze annualized data on group characteristics; others may wish to examine characteristics summarized at a biennial, quinquennial, or decadal frequency. Some wish to analyze groups individually; others wish to analyze states, some of which may contain several at-risk minorities while others contain none. *MARGene* allows users to select different temporal and physical "aggregation units" (e.g. 1, 2, 5, or 10 year temporal periods, and group, state, regional, or world physical groupings). The program creates an output data set fitting the analyst's needs which can then be analyzed in other statistical analysis software packages.

The goal of the *MARGene* project, specifically the software program and the data format conversion that has accompanied software development, is that users do not have to face these problems and deal with them on an ad hoc basis. Rather, accessing MAR data through the software automates corrections to the problems, ensures replicability, and facilitates easy use of the data. The program is modeled after the *EUGene* software package used in the creation and management of international relations data (as explained in Bennett and Stam 2000).

Contacting MAR

MAR is interested in receiving bug reports, suggestions, and any other feedback about *MARGene*. We plan to make program updates available as we make additions and improvements to the software. Please use email to contact the MAR Coordinator at minpro@cidcm.umd.edu.

Citation

If you use *MARGene* to generate data subsequently used in a published analysis, we ask that you cite *MARGene* as follows (until *MARGene*'s publication of record is in print):

CIDCM/MAR 2003. *MARGene* v1.0. Software. Website: http://www.cidcm.umd.edu/inscr/mar

MARGene makes use of raw data originally collected under the supervision of Ted Gurr. In addition to citing *MARGene*, we ask that you cite the original data source for the minorities at risk data set:

Gurr, Ted Robert (with chapters by Barbara Harff, Monty G. Marshall, and James R Scarritt). 1993. *Minorities at Risk: A Global View of Ethnopolitical Conflict*. Washington, DC: United States Institute of Peace Press.

Program Specifications

MARGene was written using the Borland Delphi language (v 6.0). *MARGene* runs under Microsoft Windows (Windows 95, 98, 2000, NT [version 4.0 or higher], or XP). Once installation of *MARGene* and the MAR data is completed, the final program with all data files will occupy about 30 MB on disk. To perform a full installation of the program and MAR data, you will need approximately 60MB of free disk space.

Installation and Uninstallation

MARGene can only be installed on Windows 95 (or higher) and Windows NT 4.0 (or higher) systems; this includes Windows 98, Windows 2000, and Windows XP.

Installation from download

To install *MARGene*, you must download a setup file to your PC, and then run an installation routine that will unpack all necessary files, including the main program executable file, source code, and input data. The initial set of installation files that must be downloaded is about 10 Megabytes. When unpacked, the files take up about 30 Megabytes of disk space. Most of this space is the original MAR data file.

1. Create or identify a directory (such as "c:\temp") on your machine where *MARGene*'s installation files can be kept. This can be any directory you want. Once installation is complete, you can delete the initial *MARGene* setup file that you download to this directory.
2. Access the *MARGene* web site at http://www.cidcm.umd.edu/*MARGene*.
3. Download the main setup file "SETUP.EXE" by clicking on the appropriate link to download the software. Download the file to the temporary directory you identified in step 1.
4. In the Windows Explorer, double click on the "SETUP.EXE" file in your temporary directory, OR use the "Run" command under the "Start" button to run "SETUP.EXE" from that directory. You will be prompted for installation options, but should normally just accept the defaults. You may install *MARGene* to any directory of your choice; if necessary this directory will be created automatically. Running setup will extract the program and data files, and create a new group in Windows under "***Start*** – ***Programs***".

5. Read the conditions of use noted in the file "LEGAL.RTF" in the installation directory. Note that none of the program files (source code, the full data set, or executable files) may be redistributed. However, you may distribute data extracted from *MARGene* as part of datasets that you create related to publications and as part of replication materials.

6. Program documentation is included as an MS-Word document titled "*MARGene*Documentation.doc", and a text file titled "*MARGene*Documentation.txt".

   To run *MARGene*, select the icon labeled "*MARGENE*" in the *MARGENE* program group under the start menu (Start | Programs | *MARGene*), or double click the *MARGENE*.EXE icon in the Windows Explorer in the "C:\Program Files*MARGENE*" directory.

Creating Shortcuts

To run *MARGene*, you will usually select the icon labeled "*MARGENE*" in the *MARGENE* program group under the start menu (Start | Programs | *MARGene*), or double click the *MARGENE*.EXE icon in the Windows Explorer in the installation directory. However, you may also want to create a shortcut to *MARGene* on your desktop or in another program group. If you do this, you should check the "properties" of the shortcut to ensure that the location in the "start in" directory (under the "shortcut" tab) is set to the directory where you installed *MARGene* (typically C:\Program Files*MARGENE*).

Uninstalling MARGene

To uninstall *MARGene*, either

1. Navigate through the Start menu to the program group for *MARGene* and select "Uninstall *MARGene*;" usually this will simply be ***Start – Programs – MARGene – Uninstall MARGene***.

2. Run "***Add/Remove Programs***" in the Windows Control Panel [select ***Start*** - ***Settings*** - ***Control Panel***, or ***Start*** - ***Control Panel*** in Windows XP]. You will see an entry for *MARGene*. Select it, and click "*Remove*."

MARGene's files will be removed from your system. If you have created data sets using *MARGene*, or modified the input files, you will have to delete those files (and the "C:\Program Files\MARGENE" directory) manually.

Running *MARGene*

The following is a summary of what the user sees and the user selections that can be made when the program runs. When *MARGene* runs, a main window will appear onscreen, with a set of menus at the top and a set of tabs to select from in the main window.

Menus

File Menu

Exit: This will exit *MARGene*.

Save Settings: Selecting "Save Settings" will save all of the current specifications you have entered in your *MARGene* run in terms of population of cases selected, output files and format, variables selected, and so on.

Load Settings: Selecting "Load Settings" will load your previously saved settings.

Create Data Set Menu

Create Now: Select *Create Now* under this menu to generate a data set according to the specifications entered in the main window. Functions identically to the "*Go"* button at the bottom of the main window.

Show Output Window: This will bring up the program output window, in which data on groups/states/etc. will be displayed if the output location is set to "To Screen".

Trace Menu

On: If set to on, a window will appear while *MARGene* is running that will trace through the program's internal routines.

Off: No trace window will appear during *MARGene* execution.

Note: Trace is set to "off" by default.

Help Menu

Minorities at Risk Documentation: Opens a window containing documentation for the Minorities at Risk dataset, including a discussion of individual variables and coding rules.

MARGene Quickstart Guide: Opens a window containing a short introduction to using *MARGene*, including a walkthrough of how to obtain basic output. For complete documentation, see the next menu option.

MARGene Documentation: Opens a window containing the full documentation for the *MARGene* software program. Note that this documentation is about the software program; documentation about the data itself is obtained under the menu item **"***Minorities at Risk Documentation*".

About: Displays basic information about *MARGene*, including the version number and program copyright.

Main Program Selections

Main Settings Tab

Physical Aggregation

The "Physical Aggregation" options allow users to choose the physical unit of analysis for the output data set. This unit defines the grouping of the MAR variables within combined units, as follows. Note that if the user selects a physical aggregation other than the group, *MARGene* creates a set of new variables to mark key information about the aggregated cases. Those variables are described below in the section "Variable aggregation across multiple countries and time points".

Group: If selected, then output data will have separate cases for each group.

State: If selected, then output data will have separate cases for each state. Multiple groups within each state will be aggregated to the level of the state. For each MAR variable selected by the user, three variables will be included in the output data set representing the minimum, maximum, and mean aggregated values of the variable in question within the state. The variables will have new, unique names to reflect the varied aggregations. For instance, if the user chooses the variable "CulDiff" with the state as the level of physical aggregation, the output data set will contain the variables "CulDiff_min" "CulDiff_max" and "CulDiff_mean". Procedures for computing such aggregated variables are discussed further below in section "Variable aggregation across multiple countries and time points". Note that while new variables names default to the convention above, if the user has requested command files to read the data into SPSS or LIMDEP (which only allow 8-character names), the names will be altered.

Region: If selected, then output data will have separate cases for each region, but not for separate states or groups. Multiple groups and states within each world region will be aggregated to the level of the region. For each MAR variable selected by the user, three variables will be included in the output data set representing the minimum, maximum, and mean values of the variable in question within the region.

Diaspora: <**NOTE: This option is not yet available in *MARGene*.** We expect to add a variable identifying the diaspora of each MAR group in a future version.> If selected, then output data will have separate cases for each minority diaspora, but not for separate states, groups, or regions. Multiple groups within each world diaspora will be aggregated

to the level of the diaspora. For each MAR variable selected by the user, three variables will be included in the output data set representing the minimum, maximum, and mean values of the variable in question within the diaspora.

Globe: If selected, then output data will have no separate cases for groups or states, but instead will have one case (per temporal unit, e.g. one case per year, decade, etc) for the world as a whole. All groups and states within the international system will be aggregated to this level. For each MAR variable selected by the user, three variables will be included in the output data set representing the minimum, maximum, and mean values of the variable in question across all groups.

All vs. Subset of Groups, States, or Regions

When the group, state, or region is selected as the physical aggregation unit, the user may choose to output either all groups in the MAR data set (or states, or regions), or only a selected subset. To select a subset of cases, check the "Subset" box and click the corresponding "Specify Subset" button to open a new window that will allow you to identify the particular groups (or countries or regions) to be included in the output data. To select all of the cases in a class (groups, states, or regions), just click the "All Groups" "All States" or "All Regions" box.

Specifying State/Group Subsets

When the specify subset button is clicked, a window appears allowing the user to define the subset of cases that are to be included in the output. Initially, a list of groups (or states, or regions) will appear in the left pane of the window, while the right pane will be blank. From this window the user may select a subset of countries to be included in the output by highlighting a name and pressing the ">" button to move the highlighted entry to the selected list. The "<" button can be pressed to move an entry out of the selected list. The ">>" and "<<" buttons will move all groups, states, or regions.

Non-MAR Control Groups

Data in the Minorities at Risk data set is collected only on groups that are at-risk (per definitions in Gurr 1993). Normally, an output data set created using *MARGene* will include only data points for which there is actual data (years without data on a country, and countries without MAR groups, are omitted). Having data only on the "events" of MAR groups does not allow easy comparison to groups that are not at risk, or states that do not contain such groups. The "Include States and Years without MARs" option instructs *MARGene* to output "dummy"

placeholder data for countries that do not contain any at-risk minority groups. The resulting data set can then be merged easily with other cross-national data bases to allow comparison between groups in states with MARs, and states without.

Specifically, if non-MAR control groups are added, the data set will be modified to include a newly created artificial group in each country that does not contain a MAR. For each of these artificial cases, the group name is "Not a MAR group". The Numcode variable for the new group (which provides a unique case id) will be a new number equal to ((ccode*100) + 0), where ccode is the Correlates of War country code number for the state. For actual at risk minority groups within states, the Numcode for each group is computed as ((ccode*100) + 1) for the first group, ((ccode*100) + 2) for the second group, and generally ((ccode*100)+n) for each of n groups in a state. So for instance, if ccode 365 has 2 at risk groups in it, these groups are numbered 36501 and 36502. For an artificial, dummy, group, the group number is set as 0 within the state, and the final number is ((ccode*100) + 0). So if ccode 365 had no at risk groups, a dummy group would be created with code 36500.

To more easily distinguish between groups/states with MARs and groups/states without, you may check the variable "is_mar"; this variable is a dummy variable with 0 indicating that a group is not a MAR group (i.e. it is an artificially created dummy group), and 1 indicating that the group is a MAR group.

To include "dummy" groups from states without at-risk minorities, check the box labeled "Include States and Years without MARs." You may then check one of two further sub-options:

All Non-MAR States: If checked, a dummy group will be included for every state in the international system that does not have a MAR in each year of the output data set.

Subset of Non-MAR States: If checked, dummy groups will be included only for the subset of states designated by the user. The specific groups may be selected by clicking the "Specify Subset" button.

Note that the artificial, non-MAR groups will contain dummy (missing value) data for each time point corresponding to the temporal aggregation selected. That is, if the user selects annual data, an annual observation will be output for each year of the output data set. But note that for MAR groups, the MAR data set frequently has observations at less-than-annual frequency. If the user selects annual output data, and some non-MAR groups, but does NOT select interpolation or 'Output Missing "In Series" Cases', then the non-MAR data will be created and output annually, while observations in states with MAR groups will be at the interval of the MAR data set.

This might result in (for instance) observations for 1950, 1955, 1960 etc. for states with MARs, and 1951, 1952, 1953, etc. for states without MARs. It is incumbent on the user to look carefully at the data frequency and the output data set to ensure that the full data set meets their expectations, and normally to select either interpolation or 'Output Missing "In Series" Cases'.

Temporal Aggregation

This selection specifies the temporal unit of analysis that will be used in the final data set created by *MARGene*. Note that if the user selects a temporal aggregation other than the year (annual), *MARGene* creates a set of new variables to mark key information about the aggregated cases. Those variables are described below in the section "Variable aggregation across multiple countries and time points".

Annual: the final data set will contain an annual observation on the groups/states/regions selected.

2 year: the final data set will contain one observation for each 2 year period on the groups/states/regions selected. 2 year periods are defined so that each period begins in an even-numbered year, so periods cover 1960-1961, 1962-1963, 1964-1965, etc.

5 year: the final data set will contain one observation per 5 year period (quinquennium) on the groups/states/regions selected. Quinquennia are uniformly defined as running from 1960-1964, 1965-1969, 1970-1974, etc.

10 year: the final data set will contain one observation per decade for the groups/states/regions selected. Decades are uniformly defined as running from 1940-1949, 1950-1959, etc.

Note that an observation will appear for a country (or MAR group) if the state in question is recognized as a Correlates of War state for any part of that period. However, a new variable will also be created in the output data that indicates for what sub-years a state was actually recognized as a state. See the section "Variable aggregation across multiple countries and time points" for details.

Interpolation

While fundamentally based on observations made on groups in particular years, the Minorities at Risk data set does not include information on every group in every year. In some cases the data include variables measured at a lower frequency (e.g. variables measured every 5 years rather than annually), and in other cases the values of particular variables are missing in a given year (because data

could not be located, for instance). By default, output data sets created using *MARGene* only include observations that are actually in the Minorities at Risk data. For instance, if the original data contain observations on Brazil in 1960, 1961, and 1964, then only those 3 years would be included in the output data set, as three lines of output. Frequently, users would like an output data set modified in one of two ways. 1) Frequently, we want a data set that shows us the missing observations, rather than deleting them. So, we would like a line in the output data set for Brazil in 1960, 1961, 1962, 1963, and 1964, even if the values on all of our variables of interest are missing. 2) Because missing data in the MAR data set are frequently missing just because the values change infrequently, we may wish to interpolate missing values, that is, replace missing values with values computed from nearby data points. *MARGene* provides a set of options vis-à-vis these scenarios.

Two methods are available for interpolating data, "linear" and "carry forward." Interpolation only computes between existing values – there is no extrapolation beyond the range of the data where there are actual values. In addition, users may choose to create or exclude lines of data in the output file for groups or states in years when the Minorities at Risk data set does not have data in it.

Interpolate Intertemporal Data Points – Linear Interpolation

Linear interpolation computes new values to replace missing values by taking the difference between the previous and next data points with actual values, dividing by the number of missing intervals, and filling in the values stepwise. As 2 examples, the following would be data filled by this procedure:

Year	**Original Data**	**Interpolated Data (linear)**
1960	10	10
1961	missing	20
1962	30	30

Year	**Original Data**	**Interpolated Data (linear)**
1960	10.1	10.1
1961	missing	22.2
1962	missing	34.3
1963	46.4	46.4

Interpolate Intertemporal Data Points – Carry Forward Interpolation

Carry forward interpolation replace missing data values by carrying the last known value in a data series forward over time until a new value is seen. So, if actual values were available for years 1980 and 1990 only, the carry forward method would apply the 1980 value to 1981, 1982, 1983, etc., with the value changing only in 1990. For example:

Year	Original Data	Interpolated Data (carry forward)
1960	10.1	10.1
1961	missing	10.1
1962	missing	10.1
1963	46.4	46.4

Interpolation and Data Types

When the raw data consist of real number values, the interpolated values will represent actual direct mathematical averages. When the raw data are discreet integer values, the interpolation procedure will round the interpolated value to the nearest whole integer. Finally, if the variable is a string or alphanumeric variable (for instance, the name of a group), interpolated values will always take the last known string value and apply it until a new value is seen, essentially applying the carry forward method for the interpolated string values.

No Interpolation

The default in *MARGene* is to output only data for which the Minorities at Risk data set contain actual values. If the default settings are left, then no data will be output for years when the Minorities at Risk data set does not have an observation on the country/group.

No Interpolation – Output Missing "In Series" Cases

Checking this option will include lines of data in the output file during the time period when an at-risk minority existed in a country, even if no data on that particular year is contained in the original Minorities at Risk data. So for instance, if group 36501 had observations in the data set in 1970 and 1973 (only), checking this option would include a line of data in the output data set for 1971 and 1972 as well as 1970 and 1973. With no interpolation selected, all variables will have missing values for those 2 years.

Years to Include in Output

Specifies what years should be reported in the output. Clicking "All Years" will include all possible years for which any data are available in the Minorities at Risk data set. Entering values under "Specified Range" will result in the output containing only those years within that range. Note that if data is not available in some year, a missing value code will be reported in the output file.

Output Options Tab

Output Destination

Specifies where the output should be sent (screen, printer, or file). If the user selects "File" they will be prompted for file name and location. If a selected disk file already exists, the user will be prompted for whether to overwrite it. Files marked as “read-only” in the operating system cannot be selected. If "Screen" is selected a new window will be opened for the output. Data on variables selected under the “Variables” tab will be sent to this output location.

Variable Separator

Specifies whether the numeric variables in the output data will be separated by a tab, space, or comma. *MARGene* output will always be in a flat ASCII text file.

Create Command File(s)

Requests that *MARGene* create command files to make it easier to read data into other software packages. Between 0 and 3 boxes may be checked to create command files for the listed programs.

File Header Information

When checked, a header line will be put in the output file. Including a header line will add a line as the first line of the output file with a label (name) for each variable. If no header line is included, only numbers will be in the file and the user should record separately the variables that were selected for the output file.

Variables Tab

This tab contains a number of sub-tabs, each of which contains a set of MAR data set variables. Each tab contains variables that are logically related. Holding the cursor over one of these variables for approximately 1 second will result in a brief pop-up description of the variable. Checking the check-box next to the variable name will select the variable so that it will be included in the output data set created by *MARGene*.

Available Variables and Temporal Aggregation

Not all variables from the Minorities at Risk data set will be available for selection at all times. Because the original MAR data gathered some variables only at intervals (e.g. every 5 years, or every 2), it is problematic to output data on a variable in the unmeasured interval unless the user chooses to interpolate missing data (see section “Interpolation” below). Similarly, if the user is requesting a data set with a 5-year aggregation unit, it would be difficult to know how to deal with data measured every 2 years within the requested

aggregation period. For this reason, the availability of variables for selection depends on user choices in the main window.

Specifically, any given variable from the Minorities at Risk data set will be available for inclusion in an output data set if and only if:

1) Interpolation is turned "on." When interpolation is turned on (either linear or carry forward), all data are computed annually, and so any variable can be selected regardless of the final temporal aggregation.

2) The variable in question has a "compatible" time unit to the temporal aggregation selected. This can happen in two cases:

    a. If the variable was measured in the original Minorities at Risk data set at a time interval equal to what the user has specified as the temporal aggregation. For example, if a variable was measured at 2 year intervals and the user asks for a 2 year aggregation, the variable is compatible.

    b. If a variable was computed at an interval that is evenly divisible into the selected aggregation unit. Variables measured annually are always available in any other temporal aggregation, as other time intervals (2, 5, 10) are always divisible evenly by 1. A variable measured at decade intervals will only be available if the selected temporal aggregation period is 10 years. A variable measured at 5 year intervals will be available if either a 5 or 10 year temporal aggregation has been selected. Finally, a variable measured at 2 year intervals will be available if the temporal aggregation period is either 2 or 10 years. A 2 year interval variable cannot be selected if the user chooses to output in 5 year periods without interpolation.

    Some variables in the minorities at risk data set change their availability over the course of the data set, for instance being measured every decade initially, and then measured every 2 years beginning at some point in time. If the user chooses to interpolate missing data, then these variables are available regardless of the user choice on temporal aggregation. If interpolation is *not* turned on, though, the availability of these variables will be based on the smallest time unit on which it is measured, but missing values may be output for specific time points before the measurement moves to the higher frequency.

For example, consider a variable measured every decade in the 1960s, and every 2 years in the 1970s and beyond. This variable is available for selection (inclusion in the output data set) if the user chooses 2 years or 10 years as their temporal aggregation unit (and does not choose to interpolate). If selected, however, the value for 1960 reported in the original data set would be reported for 1960 and 1970, but missing values would be output for 1962, 1964, 1966, and 1968. From 1970 forward, actual values from the data set would be output, for instance in 1972, 1974, and so on.

In a few other cases, variables were measured only once, for instance in 1980, 1998, or 1995. If the measurement occurred at a decade marker (e.g. 1980, 1990) then the variable is treated and made available as if it were measured regularly at decade intervals. If the measurement occurred at a 5 year marker (e.g. 1985, 1995) then the variable is treated and made available as if it were measured regularly at regular 5 year intervals. If the measurement occurred at any other year (e.g. 1996, 1997, 1998) then it is treated and made available as if it were measured annually. However, note that even though the variable is available for selection in these cases, all data points will be missing except for the one year where the data was actually measured.

"Go! Create Data Set" Button

Clicking the *Go* button generates a data set according to the specifications entered in the main window. Functions identically to the "Create Now" selection under the "Create Data Set" Menu.

"Exit MARGene" Button

Clicking the *Exit MARGene* button will exit *MARGene*.

Interrupting MARGene

While *MARGene* is running, a progress bar will appear on screen to show *MARGene*'s progress. The "STOP" button provided on that bar will abort the program run, interrupting the current run of the program, leaving the user in *MARGene* where another data set may be created. The "PAUSE" button will pause the program, as in instances when the user may need to perform some other CPU intensive task without fully stopping *MARGene*.

Variable aggregation across multiple countries and time points

MARGene gives users the ability to create data sets that can be analyzed over either space or time, and merged with data sets that may have many different units of analysis. The most common merging of MAR data is likely to be with other state-year data sets such as the Polity data, but some users may wish to portray global or regional trends, count groups over time, or perform other analyses that we cannot anticipate. As a result, *MARGene* allows users to select any combination of physical and temporal "aggregation units" that will define the structure of the output data set.

Definitions and Procedures

A physical aggregation unit consists of essentially the physical level of analysis of the output unit, be it the group, state, region, or globe. A temporal aggregation unit is defined by the time period that will characterize each output unit, or, if you will, each line of data in the output data set. At its base, the Minorities at Risk consists of observations on particular *groups* in particular *years*, and so the group-year is the lowest unit of analysis that we can analyze.

Analyzing states, regions, or decades (or any of the other aggregation units) requires us to *aggregate* information on multiple group-years in some way. That is, if we request an observation on a state in a give year, then information about all of the groups in the state must be combined in some way to give a single value for the state.

Aggregated MAR Variables

For each variable from the MAR data set selected by the user, three variables will be included in the output data set representing the minimum, maximum, and mean values of the variable across the groups encompassed by the aggregation criteria. When these variables are computed, the computations take the mean, minimum, and maximum of the non-missing values on the variable across the groups. Missing cases are ignored in the computations. As one implication, if only 1 data point of those groups and years being aggregated has a non-missing value for a variable, then the maximum, minimum, and average will each have that 1 value.

The aggregated variables will have new, unique names to reflect the varied aggregations. These names will consists of the name in the Minorities at Risk data documentation, with the string "_min", "_max", or "_mean" appended (but see exception 2 below). For instance, if the user chooses the variable "CulDiff" with the state as the level of physical aggregation, the output data set will contain the variables "CulDiff_min" "CulDiff_max" and "CulDiff_mean". If there are three groups within a particular state, with values on CulDiff of 5, 7, and 0, then the output data would contain the values CulDiff_min=2, CulDiff_max=7, and CulDiff_mean=4.

There are two exceptions to these naming conventions.

Exception 1: If a string/alphanumeric variable (like group name) is selected, only a minimum and maximum value will be computed across the cases. No mean will be computed, as it would be nonsensical.

Exception 2: While the statistical analysis program Stata allows variable names to be quite long, SPSS and LIMDEP only allow 8-character variable names. Normally new variables names default to the convention above. However, if the user has requested command files to read the data into SPSS or LIMDEP, the names will be altered and reduced to an 8 character abbreviation. If this is the case, the command file generated by the program will contain a list of variable names as they can be used in analysis (both the original name, and the new minimum/maximum/mean names). Rules for reducing the length of names are as follows:

1. The added "_min", "_max", or "_mean" are shortened to "mn", "mx", or "me" respectively.
2. If the name is still too long, then vowels in the variable name will be sequentially removed by removing all occurrences of "a", all occurrences of "e" (except when "e" appears as part of "me" for mean), occurrences of "i", etc.
3. If the name is still too long, then starting with the 2nd character of the name, characters will be sequentially removed until the name is short enough.

Additional Aggregation Variables

If the user requests a physical aggregation other than the group, or a temporal aggregation unit other than annual, *MARGene* creates a set of new variables to mark key information about the aggregated cases. In particular, the following new variables are created that describe the aggregated case. Note that if a user selects groups, annually, as the aggregation, these variables will not appear in the output data, as no aggregation will be performed. Also note that these variable names may be shortened if the user requests LIMDEP or SPSS command files.

numcode_mean, numcode_min, numcode_max: When multiple groups are aggregated within a case, these three variables contain the mean, minimum, and maximum group number (from variable *numcode*) across the cases that are aggregated. For example, if groups 36501 and 36502 were aggregated into state level information (about state 365), we would obtain group_min=36501, group_max=36502, and group_mean=36501.5.

group_min, group_max: When multiple groups are aggregated within a case, these two variables contain the group names of the minimum numbered and maximum numbered group across the cases that are aggregated.

ccode_mean, ccode_min, ccode_max: When multiple states are aggregated within an output case (if the region or globe is selected as the physical aggregation unit), these three variables contain the mean, minimum, and maximum country code number (from variable *ccode*) across the cases that are aggregated. If the state is selected as the physical aggregation unit, then these variables will each contain the same value, as the mean, minimum, and maximum ccode are identical.

region_mean, region_min, region_max: When multiple regions are aggregated within an output case (if the diaspora or globe is selected as the physical aggregation unit), these three variables contain the mean, minimum, and maximum region number (from variable *region*) across the cases that are aggregated. If the state or region is selected as the physical aggregation unit, then these variables will each contain the same value, as the mean, minimum, and maximum region are identical.

diaspora_mean, diaspora_min, diaspora_max: When multiple diaspora are aggregated within an output case (if the region or globe is selected as the physical aggregation unit), these three variables contain the mean, minimum, and maximum diaspora number (from variable *diaspora*) across the cases that are aggregated. If the diaspora is selected as the physical aggregation unit, then these variables will each contain the same value, as the mean, minimum, and maximum ccode are identical. Note that in version 1, diaspora are not yet coded, and so these variables will not be created.

year_1, year_2: These two variables contain the first and last years potentially covered in the aggregation period. For instance, in a 5 year aggregation period, these values might be year_1 = 1960 and year_2 = 1964. If annual is selected as the temporal aggregation, year_1 and year_2 will be equal.

year_mean, year_min, year_max: These three variables contain information concerning the actual aggregated data within the aggregation period. So if Kenya is recognized as a state in 1963, within the 5 year observation period from 1960 to 1964, we would obtain year_min=1963, year_max=1964, year_mean=1963.5. If there is a discrepancy between year_min and year_1, it is an indication that data was not available in year_1 for the selected physical aggregation unit. Similarly, a discrepancy between year_max and year_2 indicates that data was not available in year_2 for the selected physical aggregation unit.

agg_cnt: When multiple groups or years are aggregated within an output case, this variable counts the number of group-years that are aggregated into the output case. This variable counts *only* group years that are actual cases within the MAR data set, and *not* any artificially created control cases created if the user chooses to include non-MAR states/groups. As an example, if the user selects “annual” as the temporal aggregation, and “state” as the physical aggregation, a value of “3” on agg_cnt on some output case would indicate that data on 3 groups were aggregated in the creation of that data line. Similarly, if the user selects “5 year” as the temporal aggregation, and “group” as the physical aggregation, a value of “3” on agg_cnt on some output case would indicate that data on 3 years were aggregated in the creation of that data line. If for some reason the value of agg_cnt is missing, it will be displayed with a value of “0” (0 is the missing value code).

Reading Data into Other Software Programs

After *MARGene* has created data, you will want to read it into other software for analysis. *MARGene* can automatically create command files to read data into SPSS, Stata, or LIMDEP. To create command files, when specifying entries on the "Output Options" tab, under the "Output Destination" section, be certain to send your data to an output file, and then check one or more of the boxes to create command files. After your data is generated, look in the destination directory for a file with the same name as the data file you created, but with an extension of ".sps" ".do" or ".lim". This file will be a "command file" to read the data into SPSS, Stata, or LIMDEP respectively. For example, if you create a data set and name it "OUTPUT.DAT", and you choose to create an SPSS command file, you will find in the directory where you saved the data a command file with the name "OUTPUT.DAT.SPS". To use these command files, run SPSS, Stata, or LIMDEP and read the output file into the program. Then, following instructions provided with that software, select and run the commands in the file.

Note that there is an option on the "Output Options" tab to include a header line in the output. If this option is selected, the output file will contain as its first line a list of variable names for each variable. Note that if you plan to use a command file to read your data set into one of these files, you do NOT need *MARGene* to output a header line on the "variables" output page. If you do output a header line, your data set will contain one case with every variable appearing as missing data.

If you want to examine your output data in a spreadsheet (such as Microsoft Excel), simply be sure to specify that you want to create data separated by Tabs, check the option to include a header line with variable names, and then use the "File | Open" command within the spreadsheet program to open the data file. The data should be read into separate columns automatically. If you specified a header line on the "variables" output page, the column headers (first line) will be the names of the variables.

MARGene outputs flat ASCII text data, which can be easily read by other software. You can use the created command files as templates to create command files to read the data into other programs such as SAS. Or if you want to examine your output data in a word processor, simply use the "File | Open" command in the word processor and select the output file.

Exceptions: There are a few minor exceptions that occur when selecting command files for certain programs, as follows:

Exception 1: LIMDEP command file name. An exception to the command file naming convention occurs when creating command files to be processed by LIMDEP. LIMDEP sometimes has trouble with Windows 95 style long filenames, that is files more than 8 characters in length or with multiple periods. When you create a LIMDEP command file, the name of the command file will be shortened to use just the first 8 characters of the name of your data file if it is longer. So for example a data set named "LONGFILENAME.DAT" will have an associated LIMDEP command file of "LONGFILE.LIM". Also, to fit LIMDEP requirements, *MARGene* will output a maximum of 500 characters per line when it creates a data set if you have specified a LIMDEP command file.

Exception 2: LIMDEP and SPSS variable names. LIMDEP and SPSS only allow variable names up to 8 characters in length. When command files for SPSS or LIMDEP are created, variable names will be shortened to accommodate the 8 character variable name limit of these programs. See details in the above section "Aggregated MAR Variables."

Exception 3: String variables in LIMDEP. LIMDEP does not allow any string variables at all to be read and used in the program. As a result, if the user selects any string variables (for instance, group names) and also checks LIMDEP as a command file to create, *the string variables will be deselected and will not be written to the output file*. A warning message to this effect will be displayed to the user.

Missing Values

For most variables, a "-9" in the output file the *MARGene* creates indicates that the value is missing. However, the value representing "missing" varies by variable, as defined in the Minorities at Risk data codebook. These values will always be integer values. A few variables – including year, ccode, region, diaspora, group name, and group number – will never take on missing values. It is incumbent on the user to ensure that when data is read for subsequent analysis, missing value codes such as -9 are coded as missing for the appropriate variables.

The Stata, SPSS, and LIMDEP command files that *MARGene* creates to read the data contains commands to convert missing values appropriately.

Program Files

MARGene uses a variety of files for input, logging, and saving results.

Log File

Given appropriate input files and a correct installation, *MARGene* should operate without errors. However, few complicated programs are ever distributed bug-free, and problems with input files and data could lead to errors. *MARGene* generates a log file containing error messages generated by most internal errors. If an error should occur and you cannot figure out why, this log file may be useful to either the user or to *MARGene*'s programmers in figuring out what happened. This file is normally called "error.log" and is saved in the main *MARGene* program directory. A complete path to the location of the error log may be specified in the initialization file "MARGene.ini".

Input and Configuration Files

Raw input data and configuration information are included in three critical files that must be available to *MARGene* upon initialization. It is important for the user not to attempt to alter these files, or to be very careful when doing so! Alterations to the configuration file may lead to *MARGene* being unable to find necessary input files, either initially upon startup or later as procedures are executed. Any alterations performed on the main Minorities at Risk data file may render it

unreadable unless the user is careful to maintain spacing, commas, and tabs as in the original files. Of course, changes to the data file will render analysis of that data nonreplicable unless carefully documented and reported.

Configuration Information – file "MARGene.ini:

This file is contained in the main directory where *MARGene* is installed, typically "C:\Program files\MARGene". The file contains file names and directory paths to other necessary input files and certain other information necessary for program initialization. Lines beginning with ";" are treated as comments (i.e. they are ignored). Each line consists of a key name plus a file name, path, or numeric value enclosed in quotation marks. Recognized key values and their definitions are as follows.

error_file_name: full path to the file where *MARGene* will record any error messages. Default name is "MARGeneErrors.log"; without any path information, this file will be saved under the main directory where *MARGene* is installed.

mar_interface_file_name: relative path to and name of the interface file for the raw Minorities at Risk data set. Usually the file will be stored within the “Data” subdirectory under the directory where *MARGene* is installed.

mar_raw_file_name: relative path to and name of the raw Minorities at Risk data set. Usually the file will be stored within the “Data” subdirectory under the directory where *MARGene* is installed.

mar_linked_doc_file_name: relative path to and name of the documentation file for the Minorities at Risk data set. Usually the file will be stored within the “Documentation” subdirectory under the directory where *MARGene* is installed.

nation_list_file_name: relative path to and name of the interstate system membership list from the Correlates of War project. Normally this path and file will be "Data\states.csv". NOTE in v1.0 of MARGene. At this time, the Correlates of War system membership data officially runs to 1997. This data has been extended to 2000 in the unofficial file “states2000.csv” by the simple expedient of making all exit dates 2000. This temporary state membership data set should be replaced by an updated version as soon as one is released by COW.

major_list_file_name: relative path to and name of the interstate system major power membership list from the Correlates of War project. Normally this path and file will be "Data\majors.csv". At this time, the Correlates of War major power data officially runs to 1997. This data has been extended to 2000 in the unofficial file “majors2000.csv” by the simple expedient of making all exit dates 2000. This temporary state membership data set should be replaced by an updated version as soon as one is released by COW.

first_nation_year : first year any state is a nation according to interstate members list to be read by *MARGene* (currently, this state membership list is read from the file states.csv).

last_nation_year : last year any state is a nation according to interstate members list to be read by *MARGene*.

Saved_Settings: name of the file where user settings may be saved or loaded. Normally this file is "settings.bin".

Minorities at Risk data set – file "MAR.csv" :

This file, normally contained in the "data" subdirectory below the directory where *MARGene* is installed, contains the 2003 conversion of the original Minorities at Risk data set into the pooled time-series format standard for most data analysis. This dataset codes the characteristics of and conditions faced by various at-risk minority groups in various years. The raw data file here is in comma separated variable format, and contains variables as defined by the data set interface file (below).

Minorities at Risk data set interface file – file "MAR.mdf":

This file, normally contained in the "Data" subdirectory below the directory where *MARGene* is installed, contains information that defines the format and variables in the Minorities at Risk raw data file. This file will normally have the extension "mdf" standing for "*MARGene* data file." Within this file, lines contained with square brackets [] represent comments and are ignored in processing. The file begins with six key variable lines, followed by a varying number of lines, one for each variable within the Minorities at Risk raw data file.

Each key line consists of a key name plus a file name, path, or numeric value enclosed in quotation marks. The initial 6 key variable lines are defined as follows:

data_file_name: relative path and file name from the main directory where *MARGene* is installed to the raw Minorities at Risk data file.

first_year_possible: first year that contains any data within the Minorities at Risk data file.

last_year_possible: final year that contains any data within the Minorities at Risk data file.

label_line_in_data_file: "true" if the first line of the Minorities at Risk data file contains a list of variable names; "false" if it does not (and so the first line of the file contains an actual observation).

number_of_variables: number of variables within the Minorities at Risk data file. *MARGene* will report an error in processing if it encounters less than this exact number of variables on a line when reading the data file.

number_of_cases: number of cases within the Minorities at Risk data file. *MARGene* will report an error in processing if it encounters more or less than this exact number of cases when reading the data file.

For the *n* variables within the MAR data set (where *n* equals the value on the number_of_variables key in the interface file), the interface file then contains one line each, beginning with "variable =" followed by a string in quotations containing the following information in order:

Variable Name: name of the variable as it appears in the MAR codebook, and as it will appear for selection within *MARGene*. Variable names in the MAR data file must be no more than 8 characters in length.

Variable Type: one of the following three values: "integer" if the variable takes on integer values only (e.g. 0, 1, 2); "real" if the variable may take on real number (decimal) values (e.g. 2.3, 17.32); "string" if the variable is string of alphanumeric characters without mathematical interpretation (e.g. "Bulgaria").

Time Availability: the frequency with which the variable is measured. Possible values are "1" for variables measured annually; "2" if measured biennially; "5" if measured every 5 years; "10" if measured every decade; and "0" if the value is constant for the group over time (as in country codes, group names, etc.).

Tab Number: the number of the sub-tab under the main "Variables" tab where this variable should appear within *MARGene*.

Tab Label: the name of the sub-tab under the main "Variables" tab where this variable should appear within *MARGene*.

Variable Hint: the hint phrase that will pop-up when the user keeps their cursor over the name of the variable for approximately 1 second. *NOTE that this hint MAY NOT contain any commas or quotation marks, as these separators indicate the beginning of the next field to MARGene.*

Documentation Link: a one-word name that refers to a key location in the Minorities at Risk codebook/documentation suitable for pulling up documentation within *MARGene*. Currently, these links are not implemented, and so placeholder values are inserted in the interface file here.

Missing value: the integer value that is considered to mark missing data on this variable. If a variable does not have a missing value (so all cases should be defined), a '-99' or other unused value must still be placed in the variable definition.

A sample variable definition line, then, looks like this:

```
variable = "Numcode,integer,1,1,Identity,Unique Group Number, grVar,-9"
```

In this example, the variable named "Numcode" is an integer measured annually, which will appear on tab # 1 under the variable tab, which is the tab labeled "Identity", with the pop-up hint "Unique Group Number", with a link in the documentation to the keyword "grVar". The variable is missing if the value is a -9.

The interface file is set up to be a flexible way to allow modifications of and additions to the Minorities at Risk data file. Using the interface file, variables and observations may be added to or deleted from the data set, and as long as the interface file is updated, *MARGene* will automatically accommodate the new variables and cases. However, note that 5 particular identifying variables in particular within the interface file must have the following specific names so that *MARGene* can read and recognize key identifying variables. These variables also cannot have any missing values in the data set. [The names can only be changed if the program code (in unit MARTypes of the Delphi source code) is changed by a programmer to match the new names within the data set.] These names and the corresponding variables are:

Group number identification variable: must be named 'numcode' in the .mdf file. This identification variable is a number that concatenates the ccode number of the "host" state with an integer numbered 1..n for the number of the group within that state.

Group name identification variable: must be named 'group' in the .mdf file.

Time/year identification variable: must be named 'year' in the .mdf file.

Country code of "host" country: must be named 'ccode' in the .mdf file.

Region where group/country is located: must be named 'region' in the .mdf file.

Diaspora identifier: must be named 'diaspora' in the .mdf file.

Note that these 5 key identifying variables will always be included in output data sets generated by MARGene.

Known Bugs and Problems

None at this time.

Updating MARGene with new MAR data

MARGene is designed to be updateable as new versions of Minority at Risk data become available. 3 steps must be followed to prepare MARGene to handle new data.

1) **Output data to flat text file**. The MAR data was delivered to Scott Bennett (and is probably maintained by MAR) in Stata format. First, load the new version of the MAR data (in Stata format) into Stata (because of the size of the data file, you may need to issue the Stata command "set memory 30m" before loading the data set). Then, run the following two commands on the data set within Stata to convert it to a format appropriate for MARGene:

```
mvencode _all, mv(-99)
outsheet using MarDataV8.994.csv, comma nolabel replace
```

The first commands converts all internal missing value codes in Stata to the text value "-99". The second command creates a flat text, comma-separated data file that can be read by MARGene. It is very important that the file be comma separated, and that raw values (rather than labels) are written to the flat data file. The options on the Stata command ensure that this is the case. Be sure to change the name in the second command to have the correct file name / version number.

Note that the name of the data file created should match the name of the interface file (except that the interface file will have a .mdf extension), and the name of the data file and interface file must be updated in the initialization file, margene.ini (see below).

2) **Update interface file (.mdf) with new information**. When the data set is updated with either additional cases, new variables, variable name changes, changes in missing values, or changes in desired pop-up hints, the interface file must be updated to correspond to the new data. The structure of the interface is detailed above in the section "Minorities at Risk data set interface file – file "MAR.mdf"**:**." Depending on what changes are made in the data set revision, some or all of the following lines may need to be changed:

data_file_name = "Data\MarDataV8.997.csv"
first_year_possible = "1940"
last_year_possible = "2000"
label_line_in_data_file = "true"
number_of_variables = "452"
number_of_cases = "6835"
variable = "numcode,integer,1,1,Identity,Unique Group Number, grVar,-9"

- You will almost certainly need to update the "**data_file_name**" entry to match the name of the revised data file.
- If the data set is expanded temporally (earlier or later), you will need to update the "**first_year_possible**" and "**last_year_possible**" entries.
- If variables are added or deleted, or cases are added or deleted, then the corresponding entries will need to be updated.
- You should not need to update the "**label_line_in_data_file**" entry. However, if you change the stata outsheet command in such a way that it does not output a line of labels in the data file, then the "true" on this entry will need to be changed to "false."
- If variables are deleted, you will need to delete corresponding "**variable = "…**"" lines in the file.
- If variables are added, you will need to add corresponding lines. Note that the lines must be added in the interface file in the corresponding location to their appearance in the data file! See the section discussing variable lines above. *Note that there is an excel spreadsheet designed to make this part of the updating easier!* Currently this file is named MARGeneInterfaceFilev8.994.xls. This file can be used as a master file for updating variables, hints, etc. The last column of this file creates the

"variable=…" lines that must appear in the .mdf interface file! The easiest way to update the interface file in its .mdf format will be to update the .xls spreadsheet, and then copy and paste the final column of information into the .mdf file. Be sure to paste as plain text so that MARGene can read the information.

3) **Update initialization file (margene.ini) to refer to new interface file**. The name of the interface file must be updated in the initialization file, margene.ini. The "margene.ini" file is located in the main directory where MARGene is installed. Open the file. In that file, look for the line that looks something like

mar_interface_file_name = "Data\MarDataV8.997.mdf"

This line gives MARGene the name and file path to the interface file. Update the name on the file to match the new interface file name. Since both the data and interface file should be kept in the "data" subdirectory, be sure to keep the relative path "Data\" in the command line.

COW and MAR Linkages and Country Code Compatibility

MARGene relies on the Correlates of War state system membership data to determine what country-years are acceptable as input and output cases. In most cases, the original Minorities at Risk data country codes match the Correlates of War country codes. In a few cases, however they did not (for instance, the MAR data set used separate code numbers for Russia and the USSR). Several cases where the original MAR data did not match have been updated or changed to more properly mesh with COW, as follows:

- The Nordic Saami have been deleted (this group had been assigned a new non-COW country number because they are a small group spread across multiple states).
- "Germany" from the original MAR data has been split into East and West for the appropriate years.
- There is now only 1 Yugoslavia in the database and with the appropriate COW code. NOTE: the previous values for the Croats for Yugoslavia as 347 and 345 were different. These are now listed as separate groups within ccode 345. The values for the Croats formerly listed in the MAR data set under ccode 345 are now labeled as Croat A. The values for the Croats formerly listed in the MAR data set under ccode 347 are now labeled as Croat B.
- USSR/Russia is now only 365 (the COW code) and is labeled USSR pre-1992 and Russia post-1992.
- There is now only 1 Ethiopia using the COW code but retaining all ethnic groups.
- Burma was listed as Myanmar only for the observations for the Hill (or Highlands) people-- it is now labeled Burma for every group.

In a few other cases, discrepancies remain between the COW listing of state membership dates and data in the MAR data set. *In all such cases, the COW state listing takes precedence over the data in the MAR set.* That is, if the MAR data set has information about a group in a country which COW does not consider a state, then that data is not

included in MARGene's output. These cases (where the MAR data includes an entry but COW does not consider the entity to be a state, and so where the data will not be output) include the following:

- COW considers East and West Germany (country codes 265 and 260 respectively) to become states as of 1954 and 1955 respectively. MAR observations on these states in 1950 will not be included in MARGene output.
- From 1958 through 1961, Syria (ccode 652) formally joined with Egypt as a single state; COW considers these as one entity under the Egypt country code during this period. MAR observations for Syria in 1960 are not included in MARGene output.
- COW considers Pakistan to become a state as of 1947. MAR observations for Pakistan in 1940 are not included in MARGene output.
- COW considers North and South Vietnam (ccodes 816 and 817) to become states as of 1954. MAR observations for Vietnam from 1940 to 1950 are not included in MARGene output.

Internal Details for Programmers

Programmers looking to understand the internals of *MARGene* are free to examine the source code distributed with the program. If code is found that is in error, please notify the MAR Coordinator at minpro@cidcm.umd.edu. WHEN POSSIBLE, we will also try to answer questions about how the program was developed and on internal details of the program that will allow extensions and program verification.

Note 1: In version 1, the diaspora variable is expected to be undefined (i.e. missing) within the Minorities at Risk data set. However, code for handling the variable (aggregating, etc.) is complete within the program. If the selection of diaspora as a physical aggregation unit is enabled, then selecting it should result in the program working just fine, but with the new option. Note when dealing with the diaspora variable, it is generally treated simply as an integer value, and it will be grouped as such. The allowable range of the variable is currently set to "missing_value" through "1000". If more than 1000 diasporas are present, then the code in MARTypes.pas must be updated. In contrast, the discrete "region" variable is treated specially in the internals of code as an ordinal variable with specific regions identified in the MARTypes unit within Delphi.

Legal Notice

Copyright

Conditions of Use

MARGene, a software program to facilitate data set creation using the Minorities at Risk data set, is neither shareware nor in the public domain. It is copyrighted. The program is distributed as freeware by the authors. YOU MAY NOT FURTHER

REDISTRIBUTE ANY PART OF THIS PROGRAM, INCLUDING FULL DATA FILES, SOURCE CODE, HELP AND DOCUMENTATION FILES, AND THE EXECUTABLE PROGRAM. Delphi source code for *MARGene* is included as part of distribution for review. You may use the program for research purposes and to distribute data sets associated with resulting publications, with proper citation. However, no commercial distribution is allowed.

Disclaimer of Warranty

This software and manual are distributed "as is" and without warranties as to performance of merchantability or any other warranties whether expressed or implied. Because of the various hardware and software environments into which this program may be put, no warranty of fitness for a particular purpose is offered. The user must assume the entire risk of using the program. Any liability of the author will be limited exclusively to product replacement.

MARGene was developed using Delphi, version 6, by Borland International, Inc., and Inprise.

Bibliography

CIDCM/MAR 2003. *MARGene* v1.0. Software. Website: http://www.cidcm.umd.edu/inscr/mar/

Bennett, D. Scott, and Allan Stam. 2000. "*EUGene*: A Conceptual Manual." *International Interactions* 26:179-204.

Gurr, Ted Robert (with chapters by Barbara Harff, Monty G. Marshall, and James R Scarritt). 1993. Minorities at Risk: A Global View of Ethnopolitical Conflict. Washington, DC: United States Institute of Peace Press.

Index