TABARI Ethnic Groups Variable Descriptions

Overview: the two key variables in this dataset are **Ethnic_Code** and **Ethnic_Name. Ethnic_Code** provides a unique three letter identifier for each ethnic group found within ISO-language codes, the EPR dataset, or the CAMEO codebook. **Ethnic_Name** provides the most common name for each of these recorded ethnic groups, according to Wikipedia. The additional variables in the dataset are included mainly as references, and record the ISO, CAMEO, and EPR codes and names for each ethnic group (where applicable). **EPR-Group Countries** and **Additional Countries Found in Wikipedia** record the countries where an ethnic group was indicated to exist according to the EPR and Wikipedia, respectively. What follows is a more detailed description of each variable and any associated coding decisions.

Ethnic_Code

Ethnic_Code records three-letter (lower case) abbreviations for each (verified) ethnic group included within the TABARI Ethnic Groups Dataset. The coding scheme for creating **Ethnic_Code** is as follows: Ethnic groups were identified (i) by matching the languages listed in the ISO 639 standard language page to ethnic groups via Wikipedia search; (ii) based on the ethnic groups listed in the EPR 3.1 dataset; and (iii) based on the ethnic and religious actor groups listed in the CAMEO codebook. When an ethnic group matched a specific ISO 639-2 language entry, the corresponding three-letter **ISO 639-2 Code** was used for that ethnic group's **Ethnic_Code**. The ethnic groups that did not match any **ISO 639-2 Codes** were assigned pneumonic **Ethnic_Codes**; subject to the constraint that the assigned pneumonic **Ethnic_Code** could not already be "in use" as a language-code within the **ISO 639-2 Code** database. Note that as a result of the latter constraint: (i) some ethnic groups were assigned a three-letter lower case **Ethnic_Code** that does not match their existing three-letter upper case CAMEO religion/ethnic group code (ii) some ethnic groups were assigned **Ethnic_Codes** that were not the ideal pneumonic abbreviations of that group's **Ethnic_Name** . Also note that some ethnic groups included within the **Ethnic_Code** list nest other **Ethnic_Code** groups. Finally, as noted in the **Irrelevant ISO Entry** variable description section further below, **Ethnic_Codes** are recorded as blank entries for ISO 639-2 languages that do not correspond to existing ethnic groups.

Ethnic_Name

Ethnic_Name records the actual name of the ethnic group that is listed under **Ethnic_Code.** Names were assigned based on the most common English name for a given ethnic group (based on Wikipedia). Note that alternative names for a given ethnic group can be found under **English Name of Language (ISO)** and/or **EPR Ethnic Group Names 3.1** for most ethnic groups. **Ethnic_Names** were recorded as blank entries for **ISO 639-2 Code** language entries that did not correspond to an existing ethnic group.

ISO 639-2 Code

Three-letter (lower case) ISO standard language abbreviations; corresponding to the standard ISO languages listed by the ISO (http://www.loc.gov/standards/iso639-2/php/code_list.php). When **ISO 639-2 Codes** corresponded to specific (living) ethnic groups, they exactly match the three-letter **Ethnic_Codes** abbreviations described above (as they were used to create the **Ethnic_Codes** entries in these instances). Where these **ISO 639-2 Codes** do not correspond to specific (living) ethnic groups (e.g. Phoenician), they are still recorded as entries under **ISO 639-2 Code** for reference purposes, but are recorded as blank entries under **Ethnic_Codes.** Conversely, blank entries on **ISO 639-2 Code** correspond to **Ethnic_Code** entries (i.e. ethnic groups) that have no **ISO 639-2 Code** language-entry. For IS0 639-2 Codes that have both "B" (bibliographic) and "T" (terminology) entries; the bibliography entry is used for **ISO 639-2 Code** here (and for **Ethnic_Code**) as it is usually the more pneumonic of the two.

ISO 639-2 Code Alternate

For ISO 639-2 languages that have both "B" (bibliographic) and "T" (terminology) entries under **ISO 639-2 Code**, the "T" (terminology) entries are recorded here for reference purposes.

ISO 639-1 Code

Two-letter (lower case) ISO standard language abbreviations; corresponding to the standard ISO languages listed here (http://www.loc.gov/standards/iso639-2/php/code_list.php). These are included mainly for reference purposes, and were not at all used in the creation of **Ethnic_Code.**

English Name of Language (ISO)

The (ISO) English names for the languages listed under the aforementioned **ISO 639-2** standard language codes. These language-names were taken from the ISO standard language page: (http://www.loc.gov/standards/iso639-2/php/code_list.php). These names serve both as a general reference and (in many instances) as alternative dictionary names for the ethnic group-names listed under **Ethnic_Name.** Note however there are a number of instances where the (ISO recorded) language spoken by an ethnic group did not match the name of that group.

Irrelevant ISO Entry

This binary variable flags **ISO 639-2 Code** entries that do not correspond to a given ethnic group. 1's denote irrelevant (i.e. non-ethnic group) **ISO 639-2 Codes** (and accordingly correspond to blank entries under **Ethnic_Code and Ethnic_Name)** whereas 0's denote relevant (i.e. ethnic group) **ISO 639-**

2 Codes. An **ISO 639-2 Code** can be irrelevant for several reasons. Several of the languages listed by **ISO 639-2** are extinct or are only used in scriptures (e.g. "Ugaritic: uga" or "Samaritan Aramaic: sam"), and thus no longer correspond to any living ethnic group of peoples. Several languages listed by **ISO 639-2** are constructed/artificial languages (e.g. "Esperanto: epo" or "Klingon: tlh") and therefore do not match a living ethnic group. Finally, several of the languages listed by **ISO 639-2** represent very broad language families, and were thus deemed too general to correspond to a single ethnic group (e.g. "Austronesian languages: map" or "Baltic languages: bat").

EPR Ethnic Group Names 3.1

This variable records the ethnic groups that are listed within the Ethnic Power Relations database (http://www.icr.ethz.ch/research/epr). Because the EPR ethnic-group names were originally recorded at the country-group-year level, they were first collapsed to the group level for use in the TABARI ethnic group dataset. Next, the resultant EPR groups were matched by hand to the ethnic groups derived from **ISO 639-2 Codes**. For those EPR groups that did not have an **ISO 639-2 Code** match, they were then added to **Ethnic_Code** as new ethnic group entries. Note that because the most common ethnic group name (via Wikipedia) was used for **Ethnic_Name**, the name entries for **Ethnic_Name** and **EPR Ethnic Group Names 3.1** do not always match. Accordingly, the **EPR Ethnic Group Names 3.1** entries can be thought of as alternative dictionary names for the groups listed under **Ethnic_Name**. Missing entries on **EPR Ethnic Group Names 3.1** correspond to **ISO 639-2 Codes** that do not have a matching **EPR Ethnic Group Name 3.1**.

EPR-Group Countries

For **Ethnic_Codes** that have a matching **EPR Ethnic Group Names 3.1** entry, the **EPR-Group Countries** variable lists the countries that the EPR 3.1 database (http://www.icr.ethz.ch/research/epr) has listed and recorded as active countries for that ethnic group. For each group, the listed countries are separated by a semi-colon-space.

Major Religion

Major Religion is a binary variable that flags all **Ethnic_Code** entries which correspond to a specific religious group (=1, or 0 otherwise). The EPR 3.1 codes a number of religious groups as ethnic groups. Examples include a number of Muslim, Jewish, and Christian denominations as well as lesser-known religions such as Baha'i and Zoroastrian. For consistency, these religious groups (along with any additional religious groups already coded by CAMEO) were included as entries within **Ethnic_Code**. **Major Religion** was then added in order to allow future researchers to identify and remove/separate religious groups from the TABARI ethnic groups dataset, if needed.

CAMEO Religion

This variable records any matching **Ethnic_Code**-to-CAMEO religious group actor names, as based on the CAMEO codebook. CAMEO religion-group names are included here only for those **Ethnic_Code** religious groups that have a matching CAMEO-actor entry. Entries (on **CAMEO Religion**) are left as missing otherwise (including for **Ethnic_Code** religious groups that have no CAMEO-actor match). This entry is intended to serve mainly as a reference.

CAMEO Religion Code

This variable records any matching **Ethnic_Code**-to-CAMEO religious group actor codes, based on the CAMEO codebook. CAMEO religion-group codes are included here only for those **Ethnic_Code** religious groups that have a matching CAMEO-actor entry. Entries (on **CAMEO Religion code**) are left as missing otherwise (including for **Ethnic_Code** religious groups that have no CAMEO-actor match). This entry is intended to serve mainly as a reference. Also note that there are several instances where the three-letter upper case CAMEO religion code does not match that group's assigned three-letter lower case **Ethnic_Code** (because the three-letter CAMEO code was already in-use for a different language or group within **ISO 639-2 Code**).

CAMEO Ethnic Group

This variable records any matching **Ethnic_Code**-to-CAMEO ethnic group actor names, based on the CAMEO codebook. The existing CAMEO ethnic-group names are listed here only for those **Ethnic_Code** groups that have a matching CAMEO-actor entry. Entries (on **CAMEO Ethnic Group**) are left as missing otherwise (including for **Ethnic_Code** groups that have no CAMEO-actor match). This entry is intended to serve mainly as a reference.

CAMEO Ethnic Code

This variable records any matching **Ethnic_Code**-to-CAMEO Ethnic group actor codes, based on the CAMEO codebook. **CAMEO Ethnic Codes** are listed here only for those **Ethnic_Code** groups that have a matching CAMEO-actor entry. Entries (on **CAMEO Ethnic Code**) are left as missing otherwise (including for **Ethnic_Code** groups that have no CAMEO-actor match). This entry is intended to serve mainly as a reference. Also note that there are several instances where the three-letter upper case **CAMEO Ethnic Code** does not match that ethnic group's assigned three-letter lower case **Ethnic_Code** (because the three-letter CAMEO code was already in-use for a different language or group within **ISO 639-2 Code**).

Additional Countries Found in Wikipedia

This variable records any additional countries (i.e., any countries not already listed under **EPR-Group Countries**) that have a sizable ethnic group population (for a given **Ethnic_Code** group) according to Wikipedia. Note: some ethnic groups' Wikipedia entries were very minimal, so the coverage of **Additional Countries Found in Wikipedia** is uneven at best. For each entry, the listed countries are separated by a semi-colon-space.