DERIVATION OF FINAL DATASET

AUGUST 5, 2015

There were two datasets used to construct the final dataset. These included the following:

1. Aquastat.csv which is the most recent available data on total volumes of municipal, industrial, and agricultural water use by country. The data was extracted via a query to the Food and Agriculture Organization’s AQUASTAT database which can be found at [www.fao.org/nr/water/aquastat/data](http://www.fao.org/nr/water/aquastat/data).
2. Data\_Extract\_From\_World\_Development\_Indicators\_Data.txt is a dataset that contains the year 2014 information on population GDP, as well as a number of development indices. The data was originally generated via a query to the World Bank’s World DataBank which can be found at databank.worldbank.org/data/.

The steps used to construct the final dataset from the aforementioned datasets are as follows:

1. The following changes were made to the AquaStat data described in A.:

* The first row of observations were ignored, which contained incomplete header information
* Observations 201 and above were removed, which contained meta data information for the dataset
* The csv file was imported into SAS and only the following variables were kept upon import:
  + Column 2 and the variable created was Country\_Name
  + Column 3 and the variable created was Global\_Region
  + Column 5 and the variable created was Agr\_Withdrawl
  + Column 9 and the variable created was Ind\_Withdrawl
  + Column 13 and the variable created was Muni\_Withdrawl
  + Column 17 and the variable created was Total\_Withdrawl
* The variable Agr\_Withdrawl is defined as the annual quantity of self-supplied water withdrawn for irrigtation, livestock, and aquaculture purposes.
* The variable Ind\_Withdrawl is defined as the annual quantity of water withdrawn for industrial uses.
* The variable Muni\_Withdrawl is defined as the municipal water withdrawal as a percentage of total water withdrawal.
* Total\_Withdrawl is defined as the annual quanitity of water withdrawn for agricultural, industrial and municipal purposes.

At the end of this process the dataset, AquaUsage, had 200 observations and 6 variables.

1. The following changes were made to the World Development Indicators dataset described in B.:

* The dataset was imported into SAS and the following variables were kept within import:
  + Obscure values within the dataset such as ‘..’ were changed to the numeric value 0
  + The data was repositioned utilizing the TRANSPOSE functionality so that each column contained a variable and each row contained an observation. This was done by the variables Country\_Name and Country\_Code with the ID being Series\_Code and the VAR being STAT\_VALUE
  + The variables Country\_Code and Country\_Name were kept as well as the ones below which were kept and renamed, the rest were disposed of:
    - SP\_POP\_TOTL was kept and renamed to Total\_Population
    - NV\_AGR\_TOTL\_ZA was kept and renamed to ValueAdd\_Agri
    - NY\_GDP\_MKTP\_CD was kept and renamed to GDP
    - NY\_GDP\_PCAP\_CD was kept and renamed to Percapita\_Income
  + The ValueAdd\_Agri variable contains agriculture, value added (% of GDP). Agriculture corresponds to ISIC division 1-5 and includes forestry, hunting, and fishing, as well as cultivation of crops and livestock production. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs.
  + The GDP variable contains Gross Domestic Product at purchases price which is the sum of the gross value added by all resident producers in the economy plus and product taxes and minus any subsidies not included.
  + The Total\_Population variable is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship – except for refugees not permanently settled in the country of asylum, who are generally considered part of the population of their country of origin.
  + The Percapita\_Income variable is GDP per capita, or the gross domestic product divided by midyear population.

At the end of this process, the dataset PopStat had 215 observations and 6 variables

1. The datasets produced in steps 1 and 2 were merged based on the values of Country\_Name. All observations from both data sets were kept in the merged dataset. If a value a value of Country\_Name was only present in one dataset, values of the variables merged from the other dataset were missing
2. The datasets produced above in steps 1 and 2 were merged together based on the values of Country\_Name. Prior to the merge, both datasets were sorted individually by the variable Country\_Name. All observations from both datasets were kept upon merging. If missing values were present in only one dataset, the values of the variables merged from the other dataset were missing.
   * Four new variables were created with the merged dataset. These variables include the following:
     + Total\_Withdrawl\_Capita which is the Countries total population divided by their total water withdrawl.
     + Agr\_Withdrawl\_Capita which is the Countries agricultural withdrawal divided by the Countries total population.
     + Ind\_Withdrawl\_Capita which is the Countries industrial withdrawal divided by the Countries total population.
     + Muni\_Withdrawl\_Capita which is the countries municipal withdrawal divided by the Countries total population.

At the end of this process, the dataset final\_combined\_data had 246 observations and 15 variables.