Emma House
DENIN Environmental Scholar Program
Research summary
5/5/2020

Urbanization Effects on U.S. Groundwater Resources – Data Core Study

Abstract:

A data driven analysis led to the intermediate results of several groundwater variables and their relation to socioeconomic variables. Groundwater level (depth to water, feet below land surface) and water use (public supply, million gallons / day) information from the United States Geological Survey (USGS) were analyzed along with population data for 295 U.S. counties for the years 2000 and 2015. Results gave low positive correlation scores for depth to water vs. population (0.063-0.133), high positive correlation scores for water use vs. population (0.721-0.731), and medium positive correlation scores for depth to water vs. water use (0.171-0.296). This suggests that there is a complex system involved in the relationships between these variables, and as a result, there is more data to be examined on the interdependencies of urbanization effects on U.S. groundwater resources.

Introduction and Background:

This is a summary of my research completed this year with the backing of the Delaware Environmental Institute. In cooperation with my research advisor, Dr. Gao in the Geography and Data Science departments, who has strong interests in urbanization and data analysis, I led a study focused on data core. Dr. Michael, professor in the departments of Earth Sciences and Civil and Environmental Engineering, added an aspect of physical science to the project with an emphasis on hydrogeology. The focus of my studies at the University of Delaware thus far has been in Environmental Engineering with a concentration in Water Resources and Water Quality, so it was ideal that the interests of all three of us could be combined into one research project.

Water is described by the United Nations as the crucial link between society and the environment, as it is critical for socioeconomic development, energy and food production, healthy ecosystems, and human survival. Water has been the focus of many policy and development programs throughout history, and today, countries such as the U.S. have created complex water infrastructure systems to provide their population and industries with clean water in sufficient quantities. It is important to study the use of water and effects of water use on the environment so that informed sustainable policy decisions can be made as populations grow.

Groundwater, or water below the land surface, is less frequently studied than its above land surface counterpart, surface water. However, much of the U.S. gets their water from groundwater sources: it makes up 27% of the nation’s water use and increasing, according to the NGWA (National Groundwater Association). Furthermore, extremely important processes occur in the Earth’s groundwater such as storage, recharge, groundwater-surface water interaction, and contaminant transport. Common issues that can occur include groundwater contamination from anthropogenic and natural sources, and groundwater depletion from excessive pumping of the water resource that exceeds recharge rates. We were interested in studying not only groundwater quality and quantity variables but also their relationship with socioeconomic variables that indicate urbanization, such as population and urban land use.

Before any analysis could begin, a big part of this project was to determine the extent to which the study area has been studied. An extensive literature search was performed, identifying previous research projects which had performed analyses on that which we are interested in: groundwater parameter trends and their relation with socioeconomic parameter trends, at a county level and national scale. The overall results of the literary search were that socioeconomic variables could be incorporated into an analysis such as this, as it was oftentimes done with the combination of future climate scenarios and satellite data with water resource and hydrological models. Also notable, many papers studied groundwater on a regional and local level, but few on a national level. Some papers studied cities, others, focused on a watershed approach to analyze present or future water stress in the U.S.

Along with a literature search, as this is a data driven analysis, a database search was performed. Organizations that have extensive groundwater quality and quantity information include the USGS (United States Geological Survey), USEPA (United States Environmental Protection Agency), and many regional and local agencies. Databases identified for groundwater information include USGS NWIS (National Water Information System) Groundwater Historical Instantaneous Data for the Nation, USGS Water Use and Availability reports, USGS Groundwater Watch, EPA Drinking Water Data and Reports, EPA Injection Well Inventory, and the WQP (Water Quality Portal) which has access to many agencies including the NWIS and EPA. Socioeconomic data is more readily available and may be retrieved from the U.S. Census Bureau, National Land Cover Database, or NASA, MODIS, or LiDAR satellite imagery.

Methods and Experimental Design:

Upon the identification of the USGS as the agency with the most extensive groundwater data with a total of 8,000 NWIS groundwater wells in the U.S., a master final dataset was able to be compiled for the years of interest: 2000 and 2015. Aggregated to the county level, results for groundwater parameters depth to water and water use were correlated with a population parameter to determine their relationship and any relevant trends.

The programming language R was used through the interface of RStudio, as it is known to have superior ability to work with large datasets, perform statistical computing and analysis functions, and can create user defined plots and graphs. Another service that R can perform that proves useful to this project is with one of its many reusable functions: the *dataRetrieval* package. Developed by data scientists in the USGS, this package has functions to discover, retrieve, and parse hydrologic and water quality data from the USGS databases with the use of several inputs.

Using the *whatNWISdata* function, common parameters for the 8,000 groundwater well stations were found and the top 30 identified by frequency, as identified in Table 1. Then, the daily value service was used with the *readNWISdata* function with inputs of the groundwater wells and most common parameters, stored in R as lists. Other inputs were start and end dates as the first and last days of the years 2000 for the first run and 2015 for the second run. This output large datasets for each year queried: with daily values for each station with data in any parameter. Many of the parameter columns were left with values of "NA", meaning there was no data available for that parameter for that well station and day.

Common Parameter Codes for USGS Groundwater Wells										
Parameter Code	**72019**	**00010**	**00095**	**00940**	**00400**	**00191**	**00945**	**00900**	**00915**	**00925**
# Observations	21797	4237	3940	3311	3241	3123	3105	3049	3022	3020
Parameter Code	**00930**	**70303**	**00931**	**00935**	**00405**	**00955**	**00950**	**00932**	**70301**	**00028**
# Observations	2987	2984	2975	2898	2895	2887	2881	2876	2812	2768
Parameter Code	**01046**	**62611**	**01056**	**70300**	**90095**	**00403**	**62610**	**00300**	**00631**	**00618**
# Observations	2680	2572	2517	2501	2415	2312	2287	2073	2064	1964

Figure 1: The top 30 parameter codes based on number of observations in 8,000 USGS groundwater wells. These include measures of depth to water, temperature, specific conductance, and dissolved oxygen, among others.

To study observational water variables and their trends and correlations with other variables such as population and water use, it is imperative that all datasets are at the same scale. With the *dpylr* package and the use of the pipe operator (%>%), the observation outputs were able to be aggregated from daily values to county level averages. This was done for each year's data by grouping by station code and taking the arithmetic average of daily values for each variable for each station, resulting in a dataset with unique stations as the row headers, with one value per column (if data is available) for each well. Subsequently, the arithmetic average of each station in each county was taken by grouping and averaging by county codes, resulting in a dataset with unique counties as row headers and average values for any parameters available in the columns. Next, to successfully run an analysis on the data for the two years, the counties with data in common would need to be determined. From an initial number of 399 unique counties with data for the year 2000 and 569 counties with data in the year 2015, a merged dataset produced the 301 counties with data from both years. Furthermore, it was determined that only one parameter had extensive data (in more than three counties), thus, 280 counties were identified to have data for the years 2000 and 2015 for the depth to water parameter, and the first variable of the final dataset had been compiled.

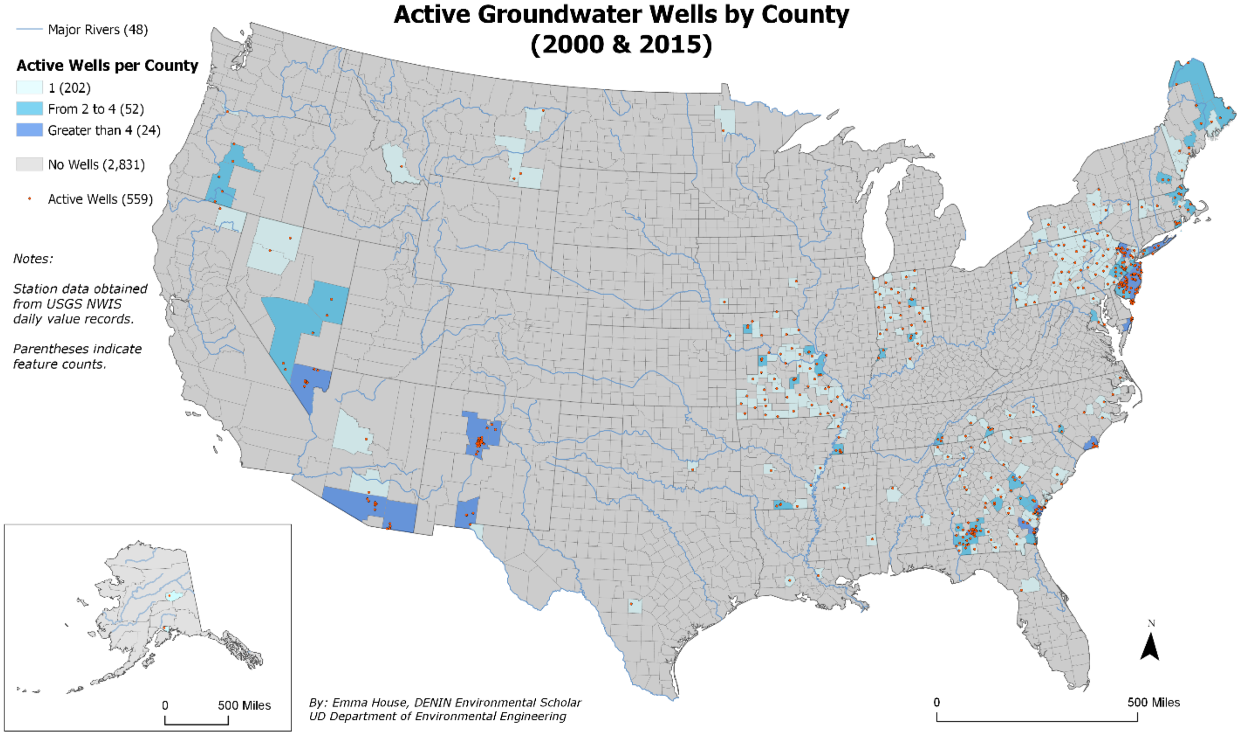

Figure 2: Counties containing active wells for the years 2000 and 2015 resulting from the R data retrieval and data analysis on USGS NWIS daily value observations for U.S. groundwater wells. Densities of well stations are shown in blue scale to give an idea of the reliability of the data.

Population data can provide information on population size, growth, density in urban and rural areas, migration, and urbanization. Human population dynamics such as these are known to have a close relationship with environmental change, such as in the case of this research project: as urbanization occurs and city populations grow, the demand for water increases and there is the potential for anthropogenic pollution and the overuse of water resources. Environmental effects such as these are what we hope to study with the incorporation of socioeconomic variables such as population in this data driven analysis.

There are many population data sources available, one of which is especially relevant to this study: the UN WPP- (United Nation's World Population Prospects) Adjusted Population Count. Consistent with national censuses and adjusted to match UN country totals, this dataset is available in high resolution (30 arc-second, approximately 1 km) global raster format. Upon examining the raster in ArcGIS Pro, it was determined that the pixels fit within the county boundary polygon layer and little to no population would be lost through overlap of the coastline. Aggregating the raster data to county level was done by calculating zonal statistics, with zones being county boundaries sourced from Esri. The counties selected were the 280 of analysis, so population summaries for each county were then extracted to a table, resulting in the second variable input into the final dataset.

A third variable was selected to create a potentially clearer link between the physical and socioeconomic variables. This was a measure of water use. As it is often incorporated into similar studies, water use data is available at a 5-year time step, aggregated by county, from the USGS Water Use and Availability reports. The large dataset has water use separated by source (surface water, groundwater, and fresh and saline for each), as well as by category (including public supply, domestic, irrigation, industrial, etc.) and is accompanied by a report detailing water use trends and data analysis. The third variable for this research project was extracted from this dataset in two measures, freshwater groundwater withdrawals for public supply, and total freshwater groundwater withdrawals.

Finally, to perform an analysis on the variables, county level values are plotted on scatter plots and correlation scores are determined to represent the strength of their relationships.

Results and Discussion:

The result of this data core focused research project was the development of a large dataset including physical and socioeconomic parameters for 295 counties with observational groundwater data. The three variables identified and studied were depth to water level, from USGS NWIS daily value groundwater observations, measured in feet below land surface, water use, from the USGS Water Use and Availability Reports, measured in million gallons per day, and population, from the UN-WPP Adjusted Population Count.

Figures 2-10 depict the nature of the variables relationships as a result of the data retrieval and analysis. Each set of three scatter plots has a figure with 2000 data, a figure with 2015 data, and a figure showing the difference between the 2015 and 2000 data, as expressed in change from 2000 to 2015. In the last plot of each set, positive values represent a positive change in values in the 15-year time.

Correlation scores were calculated using the data for each pair of variables in Excel. The range of values for correlation coefficients is -1.0 to 1.0, where a correlation of -1.0 indicates a perfect negative correlation, a value of 0 indicates that there is no relationship between the variables, and a correlation of 1.0 indicates a perfect positive correlation. In the case of this data, the correlations are never perfect correlations, but most are exhibiting low to medium positive relationships. However, it is important to note that correlation among variables does not always imply causation.

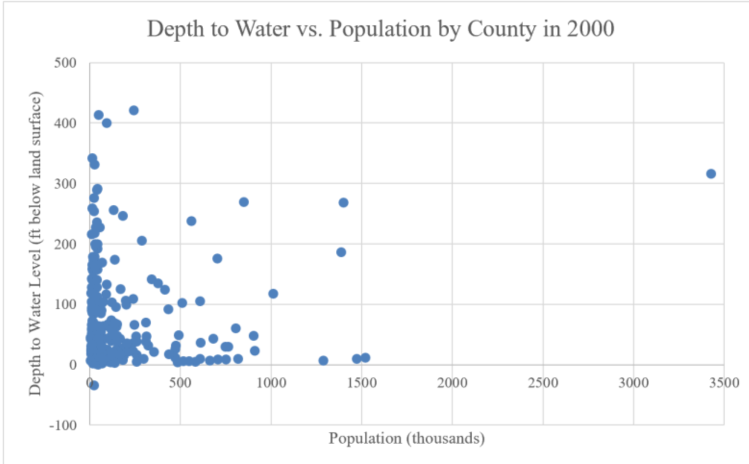

Figure 3: Depth to water level vs. population for the year 2000, with a correlation coefficient of **0.133**.

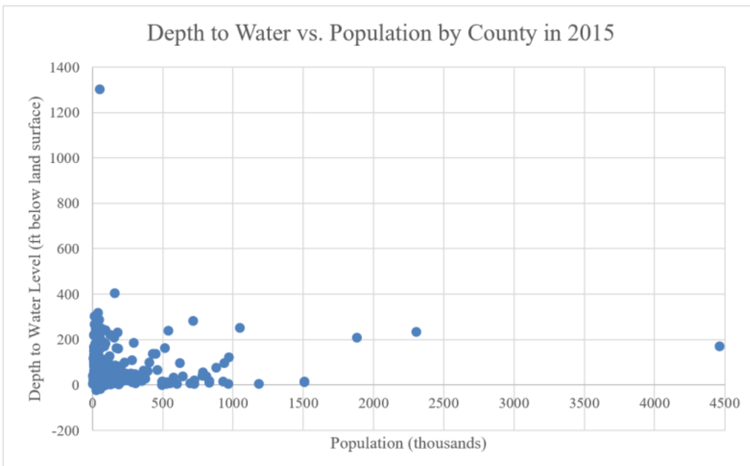

Figure 4: Depth to water level vs. population for the year 2015, with a correlation coefficient of **0.063**.

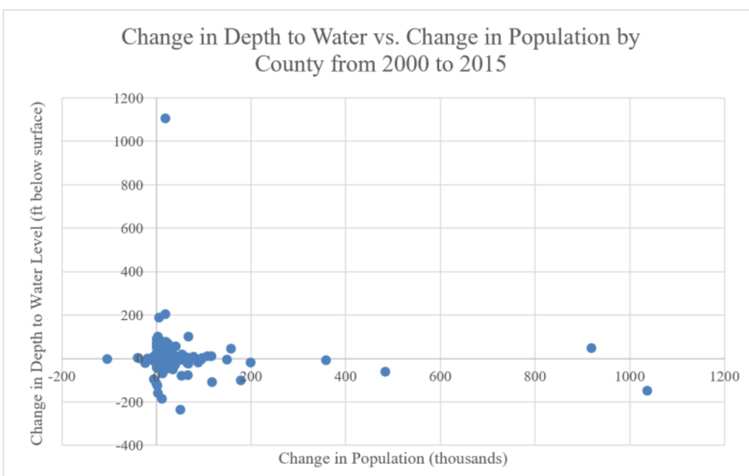

Figure 5: Change in depth to water level vs. change in population from 2000 to 2015, with a correlation coefficient of **-0.084**.

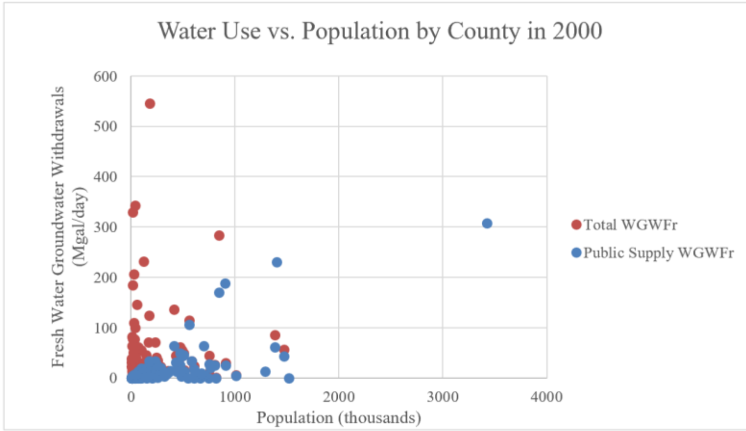

Figure 6: Water use vs. population for the year 2000, with a correlation coefficient of **0.731** for public supply and a correlation coefficient of **0.150** for total supply.

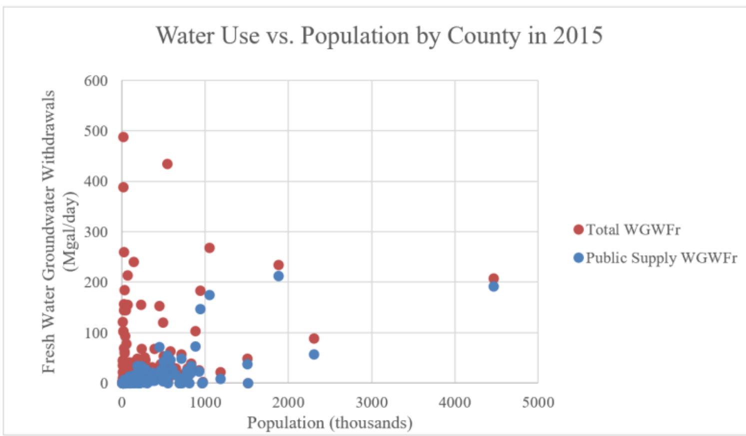

Figure 7: Water use vs. population for the year 2015, with a correlation coefficient of **0.721** for public supply and a correlation coefficient of **0.282** for total supply.

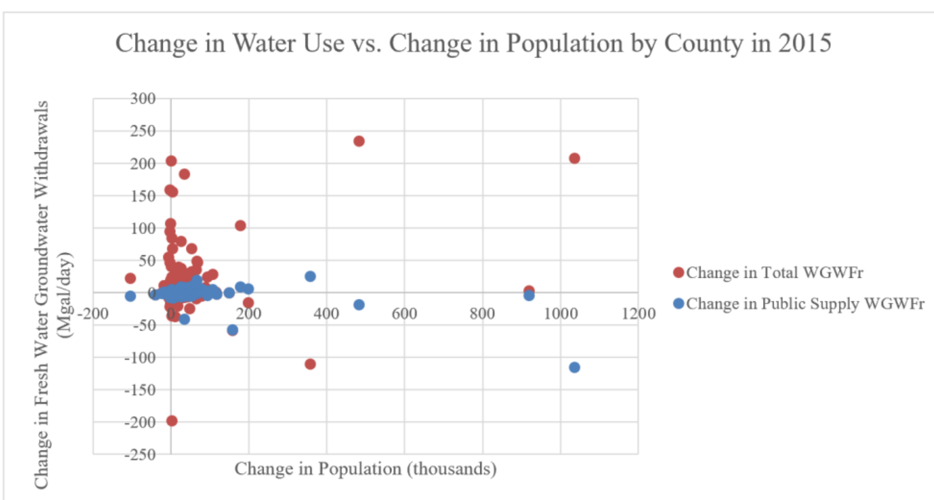

Figure 8: Change in water use vs. change in population from 2000 to 2015, with a correlation coefficient of **-0.548** for public supply and a correlation coefficient of **0.485** for total supply.

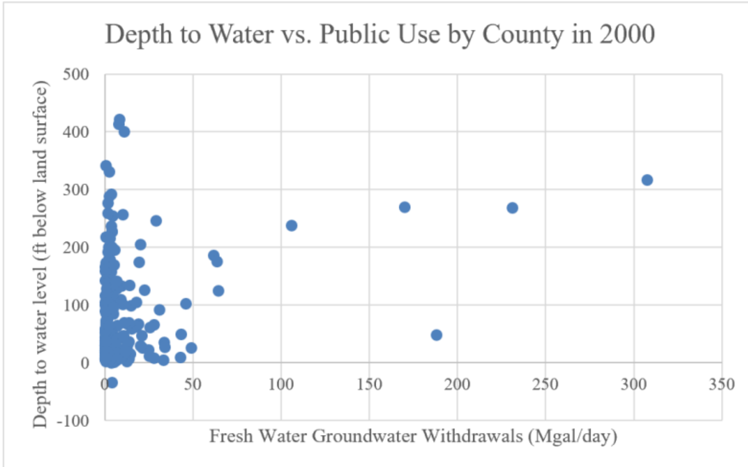

Figure 9: Depth to water level vs. public supply water use for 2000, with a correlation coefficient of **0.296.**

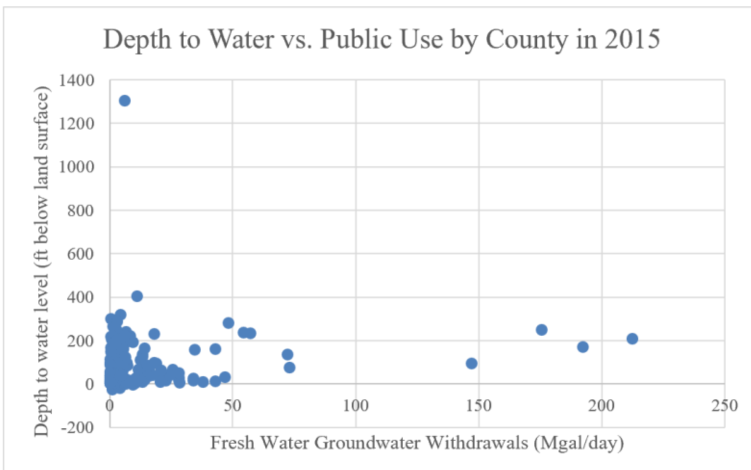

Figure 10: Depth to water level vs. public supply water use for 2015, with a correlation coefficient of **0.171.**

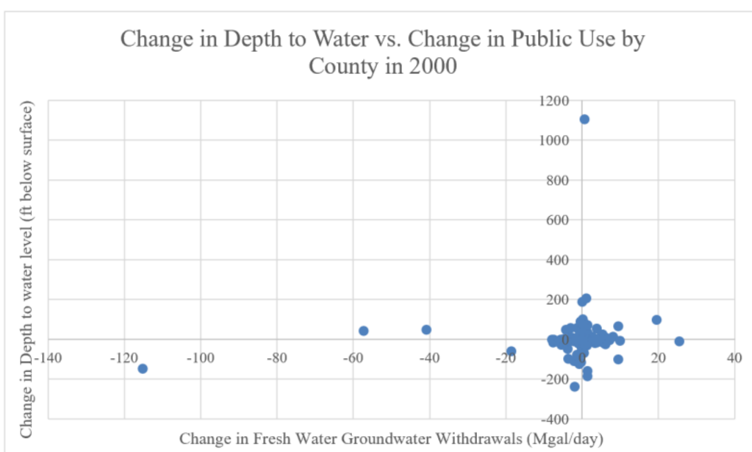

Figure 11: Change in depth to water level vs. change in public supply water use from 2000 to 2015, with a correlation coefficient of **0.094.**

Based on the correlations of the three variables found, it may be determined with some measure of uncertainty that depth to water level is not very closely related to population, but it is more closely related to groundwater use for public supply. This may be due to increased water use of groundwater resources leading to more drawdown in the water table. Population is definitely a factor here, as water use does show a strong positive correlation with population, but the relationship must encompass other variables as well.

Results of the R *dataRetrieval* from NWIS groundwater wells only consisted of one real extensive variable, depth to water level. This was less than what we had hoped for in terms of water variables, so it was important to ensure that no data was lost along the way. To do this, a test county was identified and studied more in depth. By using grouping and pipe coding in R to sort the 295 counties by occurrences of wells, test counties were identified based on the number of wells, population and city area, and change in the depth to water level variable over the 15-year period. Bernalillo, NM was determined to have the most wells (44), and contains the city of Albuquerque, so I chose this as the county to look further in depth into. Other counties I looked into were Burlington, NJ, as it has a good number of wells, is located near the city of Trenton, and the average depth to water level changed significantly, also, Glynn, GA, as it has the second largest number of wells (27), is coastal and riverine, and the average depth to water behaved irregularly in the time period studied, not only decreasing but going into the negative as to surpass the land surface.

Searching for data for Bernalillo, NM manually on the USGS NWIS web database resulted in the findings of a large number of wells, but many of them are not available for the period of interest, or are duplicates that seem to have inconsistent data. After exporting the county data and running some analysis using R, it was determined that the active wells found with the initial analysis for the county did match the active wells available on the web service database. Also, the lack of parameters was also confirmed, with depth to water being the most extensive among wells by far, and temperature coming in a distant second.

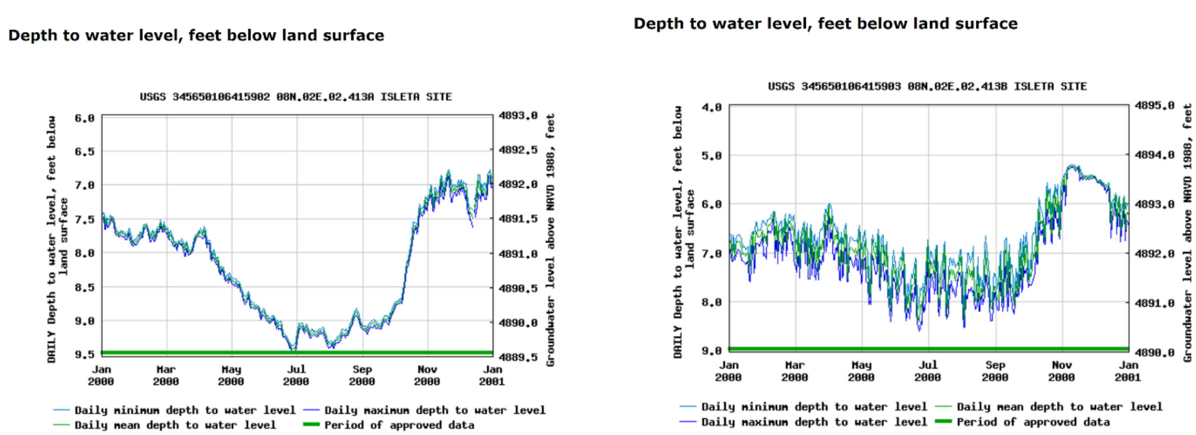

Figure 12: Example outputs from the .A and .B wells of a station in Bernalillo, NM showing daily values of depth to water level for the year 2000.

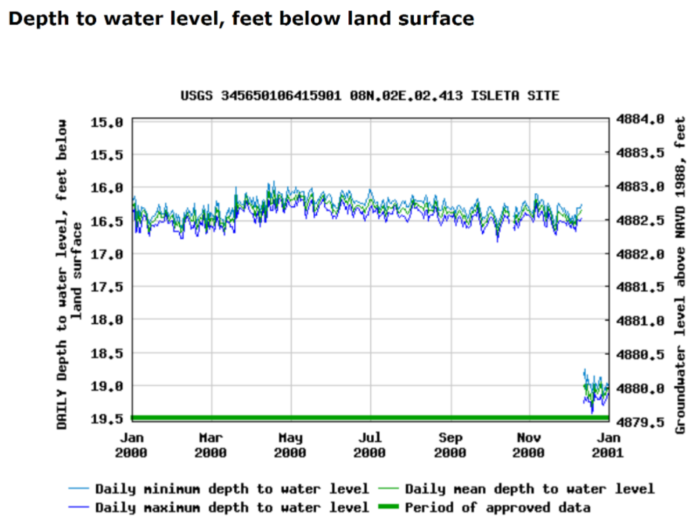

Figure 13: Example output of the same site number with no character afterwards.

Another interesting finding of the deeper inspection of the USGS NWIS data source was that some sites only have data available for certain statistic types, like as in the well in Figure 14. In limiting the R retrieval to only mean values, this may have been excluding data. However, incorporating something like a minimum value for a parameter into the variable correlations would require additional data analysis.

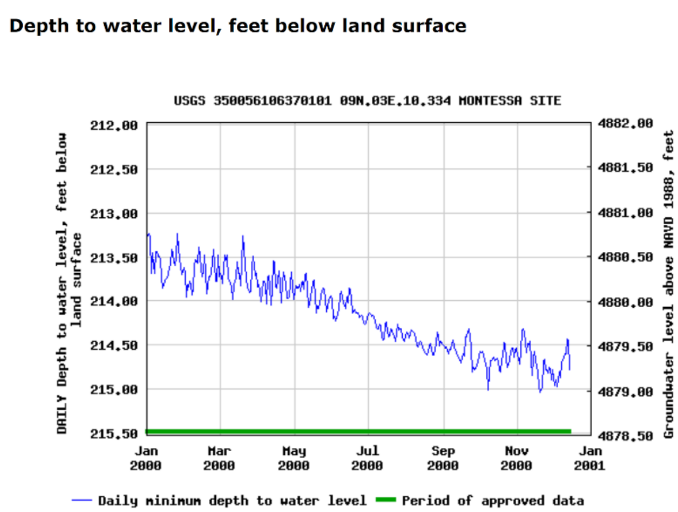

Figure 14: Example output of a well station with only one statistic type available.

This project is continuing into the coming year. With the amalgamation of additional databases, the study will incorporate measures of groundwater quality and urban land use, as well as potentially weather and seasonal variability components, saltwater intrusion at the coast, and / or effects of subsurface geological materials on groundwater, in order to effectively study the effects of urbanization on groundwater resources in the U.S.

References:

Center for International Earth Science Information Network - CIESIN - Columbia University. (2018). Gridded Population of the World (GPW), v4. Retrieved from https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-adjusted-to-2015-unwpp-country-totals-rev11

Cheryl A Dieter. (2018, June 19). Retrieved from https://www.sciencebase.gov/catalog/item/get/5af3311be4b0da30c1b245d8

Drinking Water Data and Reports. (2019, November 12). Retrieved from https://www.epa.gov/ground-water-and-drinking-water/drinking-water-data-and-reports

USGS Parameter Code or Name Search. (n.d.). Retrieved from https://help.waterdata.usgs.gov/code/parameter_cd_nm_query?parm_nm_cd=72019&fmt=html

Sanford, W. E., & Selnick, D. L. (2012, December 26). Estimation of Evapotranspiration Across the Conterminous United States Using a Regression With Climate and Land-Cover Data. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/jawr.12010

UN. (n.d.). Water. Retrieved from https://www.un.org/en/sections/issues-depth/water/

USGS Groundwater Historical Instantaneous Data for the Nation: Build Time Series. (n.d.). Retrieved from https://waterdata.usgs.gov/nwis/uv/?referred_module=gw

Water Resources of the United States. (n.d.). Retrieved from https://groundwaterwatch.usgs.gov/usgsgwnetworks.asp