

WATER USAGE DISTRIBUTION AND ITS RELATION WITH INCOME USING SAS AND ARCGIS

Fariba Nowrouzi Kashan, University of Louisville, Louisville, Kentucky

ABSTRACT

This project aims to visualize different levels of water usage in both residential and commercial areas in different tract divisions of Jefferson County, Kentucky. Also, GIS maps are used to find the relationship between water usage and different socioeconomic areas. Finally, the numeric relationships between water usage, total population, and total income are examined. The databases are from the Louisville Water Company and the Kentucky Data Center. The water usage database contained 1,486,085 clients for 2004, and the census tract information is for the year 2000. SAS and ArcGIS software were used to randomize the database, change the data formats, geocode data, and visualize the relationship between water usage and census tract areas. GIS maps show that the northeast and east end areas of Jefferson County have the highest residential water usage, and have a higher range of income compared to other areas. SAS programming confirms a strong association between residential water usage and income.

INTRODUCTION

The goal of this project was to use ARCGIS to show different levels of water usage in different socioeconomic classes. Data were collected from the Louisville Water Company and from the Kentucky Data Center to examine the relationship between water usage and demographic variables. The database contains 1,486,085 clients for 2004. Each observation contains information about service type (Residential and Commercial), Tax district (Jefferson, Bullitt and Oldham counties), Address, Zip code, billing date, service period, and usage amount. The latest Census Tract data from the Kentucky Data Center that we could access were related to 2000. In this paper, we assume that the differences between census tract data from 2000 to 2004 are not significant and can be ignored.

DATA INFORMATION

Figure (1): the Water Company data distribution in the three counties of Jefferson, Oldham, and Bullitt.

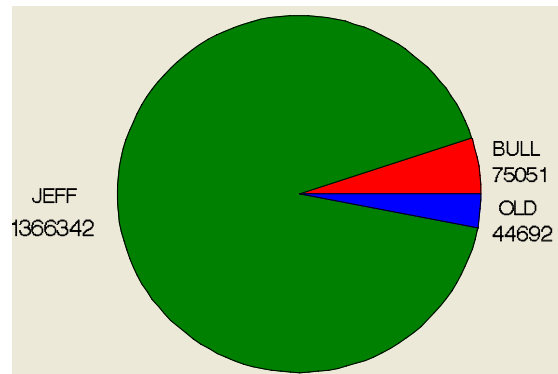


Table (1): Average and standard deviation of water usage

Taxing _ Distinct	Mean	Std Dev	N	Min.	Max.
Bullitt	13.4	47.38	75051	1	3192
Jefferson	17.4	108.93	1366342	1	59075
Oldham	17.6	45.02	44692	1	2397

Table (1) shows the average and standard deviation of water usage in each county. This illustrates that the average water usage in Bullitt County is lower than the two other counties, and Jefferson County has the highest standard deviation.

DATA PRE-PROCESSING

To combine the water usage data and Geography area information, we needed to Geocode the data. According to the GIS dictionary, the steps of the geocoding process are as follows: "The steps involved in translating an address entry, searching for the address in the reference data embedded in an address locator, and delivering the best candidate or candidates. These steps include parsing the address, standardizing abbreviated values, assigning each address element to a category known as a match key, indexing the needed categories, searching the reference data, assigning a score to each potential candidate, filtering the list of candidates based on the minimum match score, and delivering the best match. The process requires reference files, input address records, address locators, and software."(1)

Zip code information can be given in 5- digit or 9-digit form. The database was a mixture of both kinds of zip code. To use the zip code as a match key for geocoding, SAS software was used to transform all zip codes to a 5-digit format. The database was very large, but SAS software made it easy to change the data format by editing data and using a transformation formula. The transformations are done by applying the "Substr(Zipcode,1,5)" formula. For zip code column, the zip code was used as a match key to geocode so that all points with the same zip code were located on top of each other in the center of each zip code area. Since census tract division does not match with the zip code division, using zip code as a match key for geocoding caused a problem when combining the socioeconomic information with water usage for each zip code area. For this reason, addresses were used as a match key. Different counties have the same street names and house numbers. Therefore, all streets files and addresses of the three counties had to be downloaded in different map layers. After separating the database for the three counties, they were geocoded. Geocodes of Oldham and Bullitt Counties illustrated that more than 50% of the addresses in the database were not matched with addresses in the address locator. Investigation showed that the address locator file that we used was more than five years old, and does not contain the new house and business addresses. Therefore, we concentrated on the Jefferson County observations.

To make GIS software work more efficiently, a 10% randomly selected subset of all databases was used; that is, almost 140,000 records. The database contains both residential and commercial water usage information. Figure 4 shows all zip code areas of Jefferson County, and Figure 5 illustrates how the commercial units in the random subset have been distributed in Jefferson County.

Figure 6 shows all census tracts of Jefferson County. One definition of the census tract that can be found in a website is "A small, relatively permanent statistical subdivision of a county established by the US Census and designed to be homogenous with respect to population characteristics, economic status, and living conditions. Tracts usually have between 2,500 and 8,000 residents" (2). Jefferson County in Kentucky contains 170 tracts. Figure 7 illustrates total census tract divisions and zip code areas. They have different boundaries. Therefore, statistical information about census tract may vary for the data in one zip code area. To avoid this problem, we combined the data from Geocoding with each census tract that contains data in the map.

Figure 2: Kentucky Counties

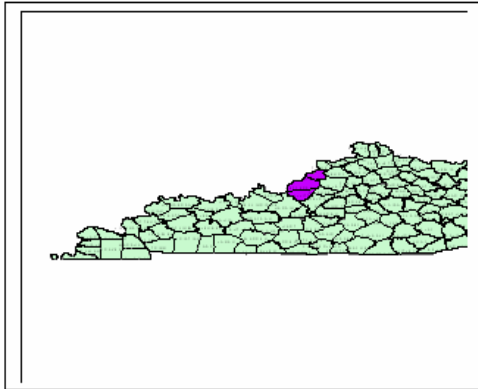
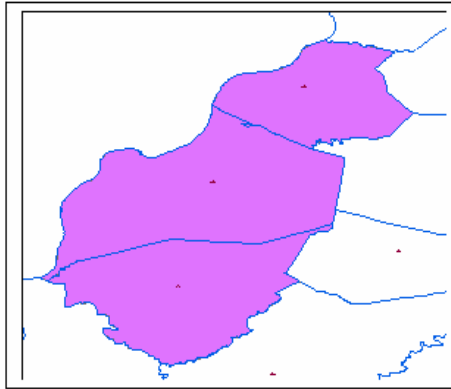
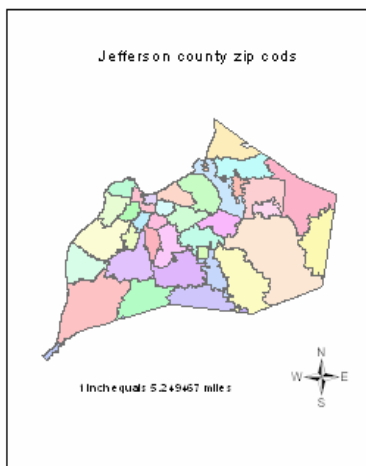


Figure 3: Oldham, Jefferson and, Bullitt Counties



Figures 2 and 3 illustrate all the Kentucky Counties and the location of Oldham, Jefferson and Bullitt Counties. Jefferson County is in the center.

□ **Figure 4: Jefferson Zip code**



□ **Figure 5: Commercial Geocoding**

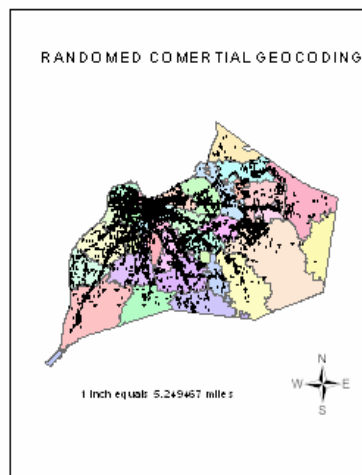


Figure 6: Jefferson Census Tract division

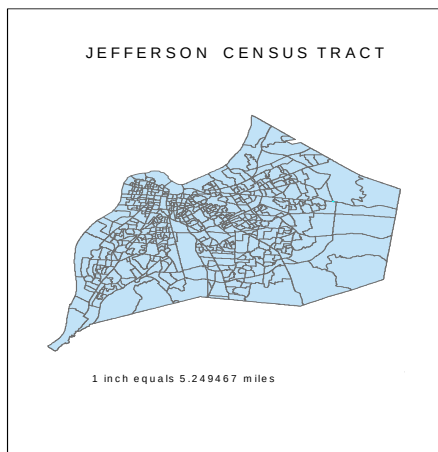
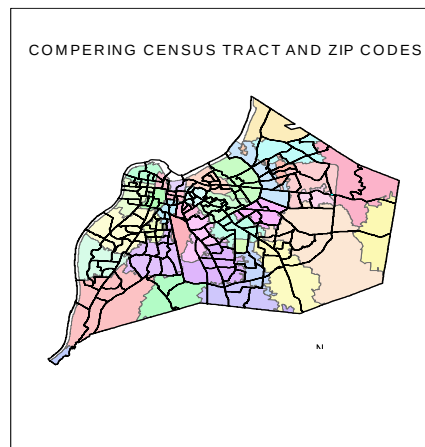


Figure 7: Census tract & Zip Code



In figure 8, the pattern of commercial water usage along the tracts shows which census tracts have the highest amount of commercial water usage. The yellow parts represent the minimum amount range and dark brown areas represent the highest range of commercial water usage. In general, the commercial water usage has been divided into a five-range area. From light to dark, the color explains the area from minimum amount to maximum amount of commercial water usage. It is obvious that area and population are varying in the different tracts. Figure 9 represents the average of commercial water usage with respect to the population. Figures 8 and 9 have a similar pattern. Only in a few tracts in the east area is the average usage in the lower level compared to the whole usage. Figure 10 confirms that the number of households is high in those areas. (It may be better to examine the relation between the commercial water usage and the number of employees in the tracts).

WATER CONSUMPTION USING ARCGIS

Figure 8: Total Commercial Water Usage

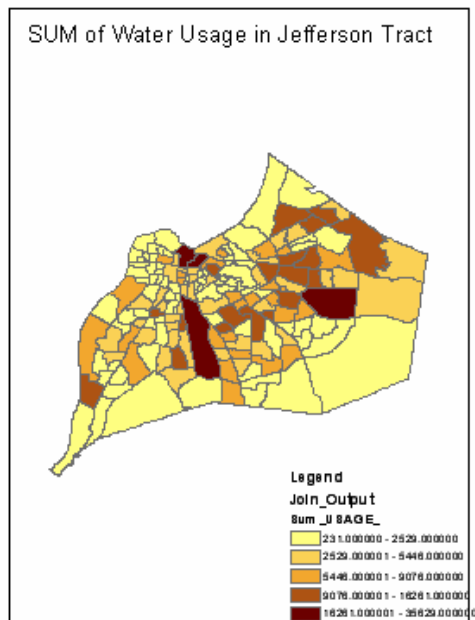


Figure 9: Average Commercial Water Usage With Respect to Population

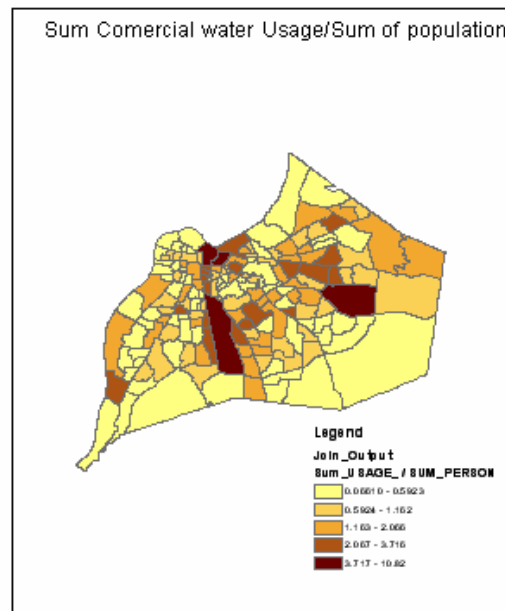


Figure 10: Total Households in Tract

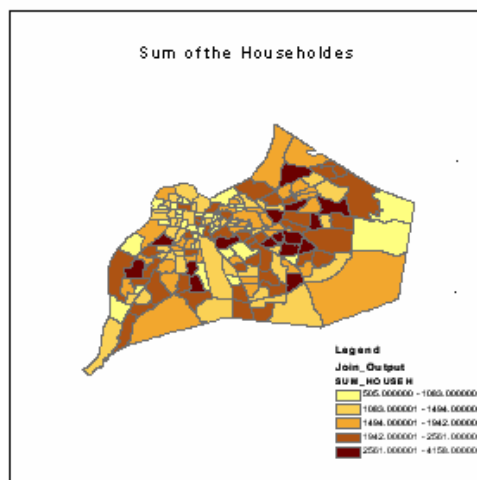


Figure 12 confirms that the northeast and east end areas of Jefferson County have the highest residential water usage. Figures 13 and 14 are almost identical and show that the average of water usage for each person and household in the east end and the northeast are higher compared to other areas. The average size of residential property in the east end and northeast is larger than in the other areas. Therefore, we assume that landscaping causes the higher water usage in those areas. Sprinkler systems are more prevalent. The houses in the east end and northeast are more expensive compared to the west and south. Therefore, there is a direct relationship between water usage and income. Figure 15 illustrates the distribution pattern of median income in the Jefferson County area that is very similar to the distribution pattern of water usage.

Figure 11: Residential Units Geocoding

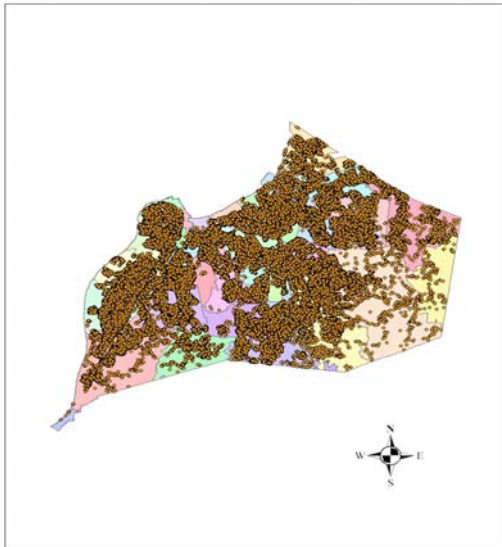


Figure12: Total Residential Water Usage

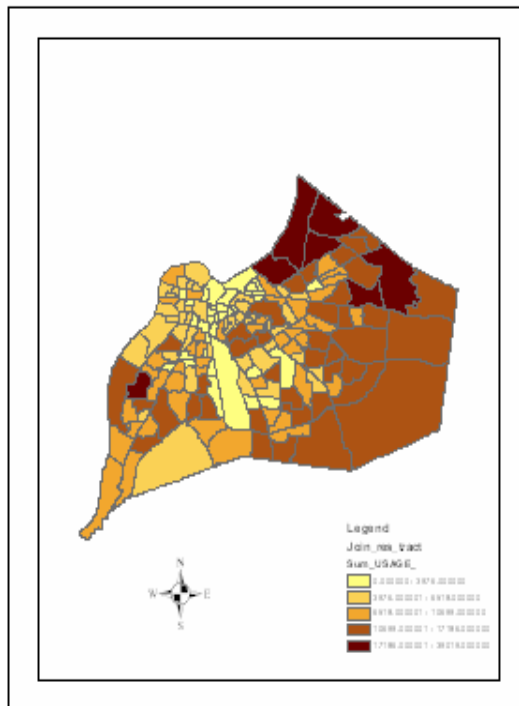


Figure 13: Average Residential Water Usage with Respect to Population

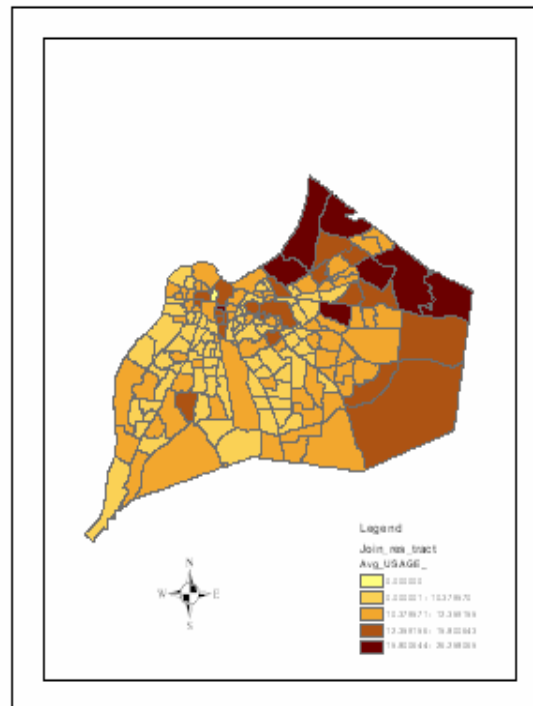


Figure 14: Average Residential Water Usage With Respect to Household

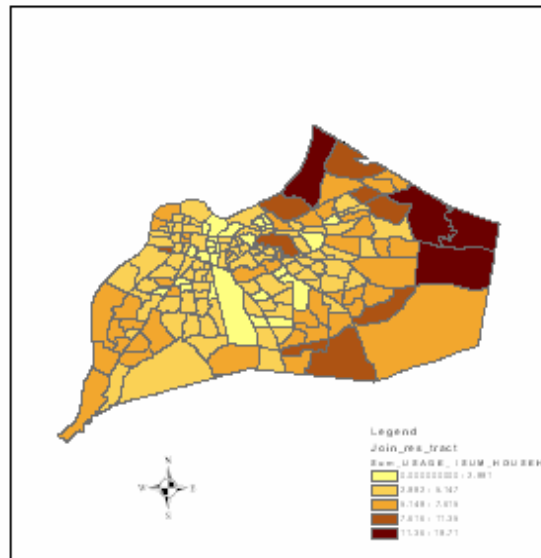
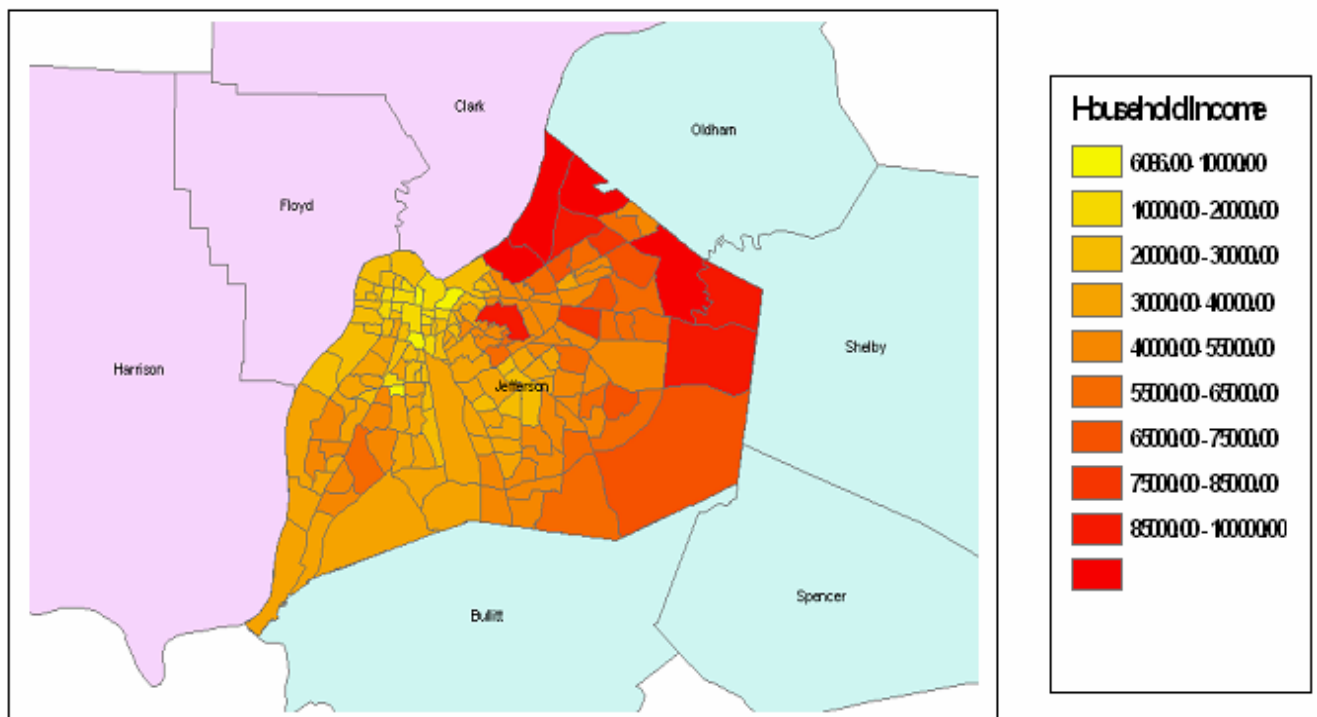


Figure 15: Median Household Income



SAS ANALYSIS

To examine the relationship between residential water usage and variables such as total population, total households, aggregate household income, and area, SAS software was used. The first step was combining the socioeconomic census tract information with the database from the Water Company and the Geocoded data.

Table 2: Statistics Information

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Sum_USAGE_	170	8236	5293	1400104	0	39019
Total_Population	170	4080	1672	693604	1452	10267
Total_Households	170	1689	674.71195	287133	510.00000	4158
Aggregate_Household_Income	170	90037828	63547327	1.53064E10	8631500	346282400
AREA	170	65359065	117715398	1.1111E10	5662794	1206700000

Table 2 gives a data summary about the five variables in the one hundred and seventy tracts of Jefferson County. The minimum amount of total residential water usage is zero. It confirms that there are no residential units in some of the tracts.

Table 3: Correlation Coefficients

Pearson Correlation Coefficients, N = 170 Prob > r under H0: Rho=0					
	Sum_USAGE_	Total_Population	Total_Households	Aggregate_Household_Income	AREA
Sum_USAGE_	1.00000	0.71844	0.60266	0.86171	0.38179
Sum_USAGE_		<.0001	<.0001	<.0001	<.0001
Total_Population	0.71844	1.00000	0.95025	0.74647	0.16339
Total population	<.0001		<.0001	<.0001	0.0333
Total_Households	0.60266	0.95025	1.00000	0.72554	0.08577
Total Households	<.0001	<.0001		<.0001	0.2661
Aggregate_Household_Income	0.86171	0.74647	0.72554	1.00000	0.27606
Aggregate Household Income	<.0001	<.0001	<.0001		0.0003
AREA	0.38179	0.16339	0.08577	0.27606	1.00000
AREA	<.0001	0.0333	0.2661	0.0003	

The Pearson Correlation Coefficient, r , between two variables shows the strength of the linear association between them. Table 17 shows that the total residential water usage has the strongest linear association with total household income and has the weakest linear association with area in Jefferson County tracts. The scatter plots in Figures 18 and 19 confirm the numeric relations in Table 3.

Figure16: Scatter plot between Total Water Usage and Total household income

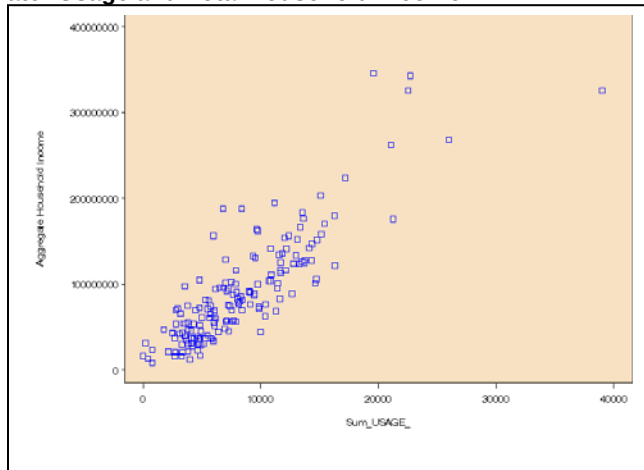


Figure 17: Scatter Plot between Total Water Usage and Area

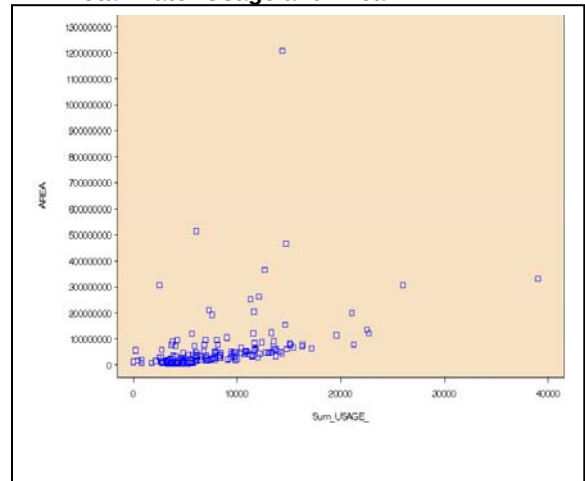


Table 4 shows the result of the estimation of a linear model with total residential water usage considered as a dependent variable and total income as an explanatory variable.

T able 4:

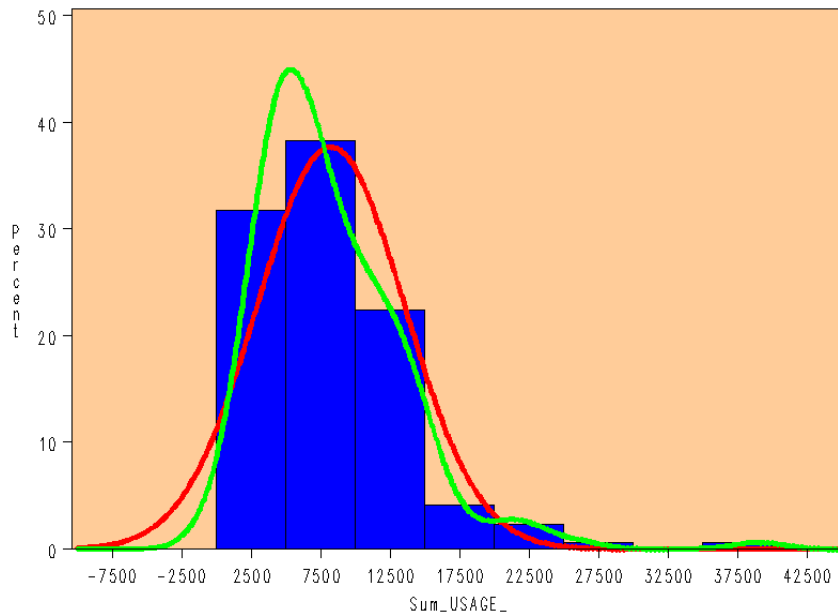
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3516153802	3516153802	484.52	<.0001
Error	168	1219178505	7257015		
Corrected Total	169	4735332306			

Root MSE	2693.88472	R-Square	0.7425
Dependent Mean	8235.90588	Adj R-Sq	0.7410
Coeff Var	32.70903		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS	Standardized Estimate
Intercept	1	1773.14183	359.01559	4.94	<.0001	11531124770	0
Aggregate Household Income	1	0.00007178	0.00000326	22.01	<.0001	3516153802	0.86171

R-square = 0.7425 so that about 74 percent of the variability of the total water usage can be explained by the linear model (total residential water usage) = 17773.142 + 0.000072* (total income). Figure 18 demonstrates that the total residential water usage distribution is not normal. This violates an assumption of estimating a linear model. To see the effect of this violation on the model, we consider the residual plot of the predicted value on figure 19 and residential distribution of total usage on figure 20. The distribution is close to a normal distribution, so the the violation can be ignored.

Figure 18: Residential Water Usage Distribution with kernel Density, Green, and Normal curves, Red:



In figure 18, the green curve, kernel density, fits a nonparametric, smooth distribution for residential water usage. The red curve fits normal density with the two parameters, mean and standard deviation of the data. When the kernel density distribution is closer and similar to normal density curve, it is more significant to assume that the distribution is normal. Considering these two curves does not confirm that water usage distribution is normal.

Figure 19: Water Usage Predicted Residual Plots

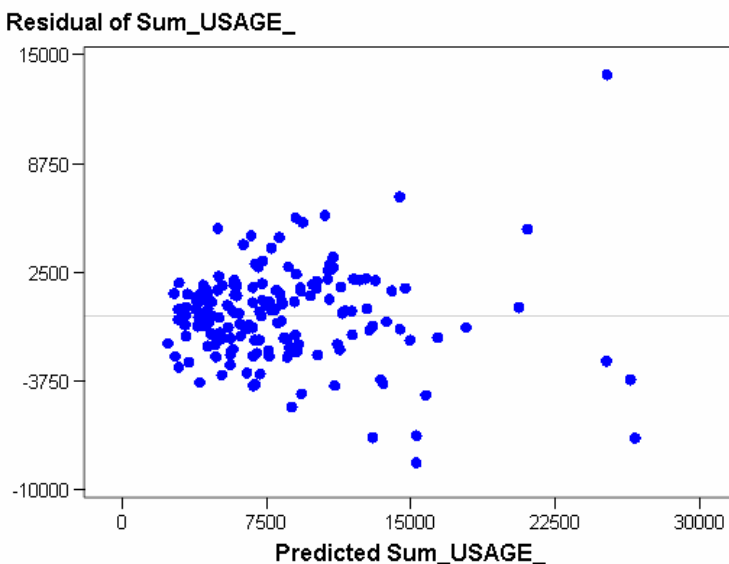
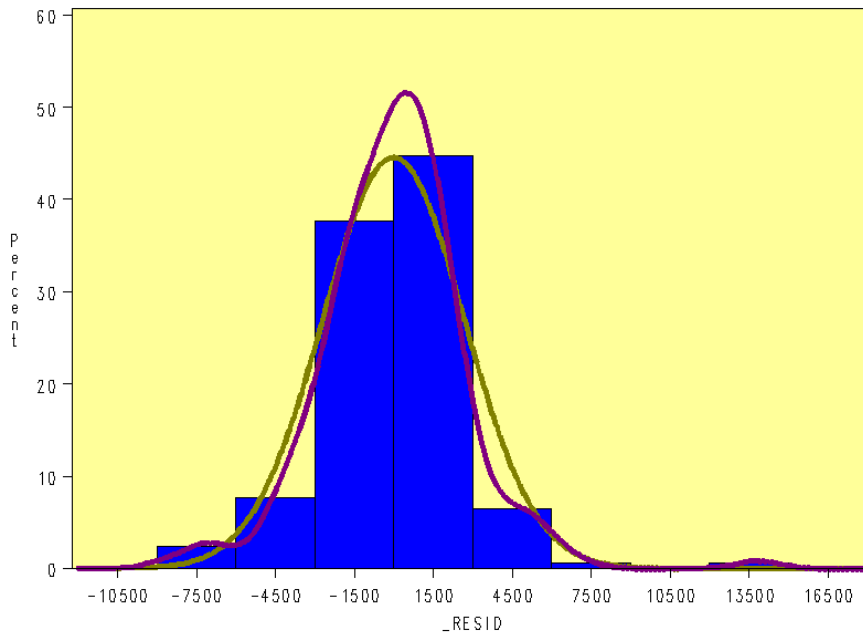


Figure 20: Histogram Distribution, Normal and Kernel Density (Red curve as kernel & yellow as normal)



Comparing the two curves in figure 20, kernel density, the red curve, is almost symmetric around the mean and it is not too far from normal density, the green curve. Therefore, we can assume that the distribution is normal.

Conclusion:

At this point, according to the database and other information that was accessible, the most important relationship of water usage in residential units in Jefferson tracts is household income level. The Water Company needs to provide more water for the improving area in the east end and northeast of Jefferson County. The houses on those areas are bigger and require more water for private swimming pools and sprinkler systems. To estimate a model for Commercial water usage, we need more information such as the number of employees and the kind of production.

References:

1. <http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&searchTerm=geocoding%20process>
2. http://www.google.com/search?hl=en&lr=&rls=RNWE,RNWE:2005-52,RNWE:en&defl=en&q=define:Census+Tract&sa=X&oi=glossary_definition&ct=title

Contact Information:

Fariba Nowrouzi Kashan
 University of Louisville
 Department of Mathematics
 328 Natural Sciences BLDG.
 Louisville, Kentucky 40292
 Phone: (502) 852-6826
 Fax: (502) 852-7132
 E-mail: f0nowr01@louisvill.edu or faribanowrouzi@yahoo.com