**Evolution of Water Consumption in the USA:**

**A Network Approach**

BY

Sk Nasir Ahmad
Bachelor of Science in Civil Engineering
Bangladesh University of Engineering and Technology, 2009

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Civil Engineering
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Sybil Derrible, Chair and Advisor
Thomas Theis
Amid Khodadoust

## ACKNOWLEDGMENTS

Dedicated to my wonderful parents, Abul Kamal Mumajjad Ahmad and Nazira Begum because of their unwavering support throughout my life.

I would like to express my gratitude to my advisor, Dr. Sybil Derrible, for his support, patience, and encouragement to achieve my goal. His technical and editorial advice was essential for the completion of my dissertation successfully.

Then, I would like to thank Dr. Thomas Theis and Dr. Amid Khodadoust for being in the committee and help to overcome the technical difficulties I have encountered.

Finally, I would like to thank my wife J Gulshan Ara Hanif for her untiring support, encouragement and patience.

## TABLE OF CONTENTS

CHAPTER PAGE

## LIST OF FIGURES

FIGURE PAGE

## LIST OF TABLES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| gpd | Gallon Per Day |
| gpcd | Gallon Per Capita Per Day |
| Mgd | Mega-gallon per day |
| USGS | U.S Geological Survey |
| NCDC | National Climatic Data Center |
| USACE | U.S Army Corps of Engineering |
| USBR | U.S Bureau of Reclamation |
| USEPA | U.S Environmental Protection Agency |
| NWS | U.S National Weather Service |
| NRCS | U.S Natural Resources Conservation Service |
| MW | Megawatt |
| TWh | Terawatt-hours |
| kPa | Kilopascal |

## ABSTRACT

Water is essential to life, and knowing the consumption trends in water is paramount if we aspire to become more sustainable. The traditional statistical tools such as mean and standard deviation, can easily fail to capture these trends. For this work, the limits of these traditional statistical tools are first highlighted and a new network approach then developed and applied to better understand trends in water consumption. Data from the United States Geological Survey (USGS) for the years 1985, 1990, 1995 and 2005 were used in gallons of water per-capita per day for all US counties. Essentially, a network is formed between counties when they have consumption values within a certain range, $\pm \xi$, of one another. A giant cluster rapidly emerges, containing more than 80% of the nodes for a $\xi$ of 1%. The counties with the highest number of connections are associated with the mode of distribution, and multi-modal patterns are observed for earlier years. Moreover, the average shortest-path length can be seen as the spread of a distribution. The diameter and density of the networks are also used. Overall, beyond a possible process of homogenization, water consumption patterns do not seem to have evolved much from 1985 to 2005, and no spatial correlation was detected. While the methodology is yet to be formalized, it manages to give meaningful insights while addressing the limitations of traditional statistical analyses.

# 1. INTRODUCTION

The critical nature of water is well known. Indeed, water is the most essential element for any living organism and the presence of water is the most distinctive feature of Earth over the other planets of the universe. In fact, no living organism can survive without water. Although, water covers approximately 70.9% of the earth's surface (Chin 2002), only 1.8% of total water is available for use, not only for human beings but also for every living organism that depends on freshwater. Of this limited quantity of usable water, less than 0.3% is found in lakes and rivers, which is known as surface water, and the rest is found in the form of groundwater and polar ice (Gleick 1993). Water is used extensively in almost every sector including domestic, commercial, agricultural and industrial. More specifically, the United Sates depends to a great extent on surface water and in 2005 surface water withdrawals accounted for 80% of the total water withdrawal in the nation (Kenny et al. 2009). With the growth of population and advancement in human civilization, water consumption has increased significantly but the amount of water available for use remains constant, which makes water one of the most critical resources for mankind. As water is the most important element for every sector, thus water-use systems must be sustainable and able to support human habitation. Water-use systems include water treatment systems, water distribution systems, wastewater collection systems, wastewater treatment systems and irrigation systems. For well-functioned water-use systems, water demand should be measured precisely.

Thus, understanding how water is being used is paramount. In other words, identifying current trends in water consumption is critical if we aspire to live in a more sustainable world. This is

further reinforced by the fact that urbanization requires that large quantities of water are distributed and disposed of in small areas.

Traditional statistical indicators (e.g., mean and standard deviation) are commonly used to study and compare water consumption trends in different regions. These indicators are also used to define and explain density and distribution functions (e.g., probability distribution functions and cumulative distribution functions). A distribution is characterized by its *location*, *spread* and *shape*, where, the *location* indicates an expected value, the *spread* designates a measure of variation and finally *shape* specifies the general shape of a function. Generally, the arithmetic mean is used as the location and standard deviation as the spread of the distribution. Additionally, these traditional indicators are most often used first-hand in models to predict the water demand; such models include the 'per-capita' model, 'extrapolation' model, 'disaggregation' model, 'multiple-regression' model, and the 'land-use' model to name a few (Chin 2002).

Numerous studies have been carried out to study water consumption. A quick search of the keyword 'Consumption of Water' in Scopus revealed 67,287 results. The U.S. Geological Survey (USGS), for instance, has published many reports and publications using these standard indicators to analyze US water consumption trends (Solley, Pierce, and Perlman 1998; Solley, Merk, and Pierce 1988; Solley, Pierce, and Perlman 1993; Hutson 2004; Kenny et al. 2009). While these indicators offer many benefits, the results are somewhat limited and they may even be biased because of errors and outliers contained in the original datasets or because of issues with binning the data in meaningful bins.

The main objective of this dissertation is to partially remediate this problem by developing a novel methodology that can give extra information about water consumption trends

and therefore supplement traditional analyses. The proposed methodology takes a network approach that links two regions when they share similar water consumption levels. The realm of network science has emerged strongly in the past 15 years, and it has been applied in myriad fields, including telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks (Newman 2010; Easley and Kleinberg 2010). Search for the term 'Network Science" in Scopus resulted in 118,066 documents. In this dissertation, adopting this approach can be essentially related to a process of forming "friendship" networks (akin to social networks). Social network is one of the most common network present around us. Moreover, it is also very complex in nature. Every social actor (e.g., individuals or institutions) act as a node and the interactions between them denoted as links. Different complex theories such as "Six Degrees of Separation", "80/20 Rule" are common in social network. In the context of this work, linking two entities that share similar consumption values is most closely related to the concept of *homophily* in social networks; *homophily* essentially relates two individuals that share common attributes such as income for instance. Once formed, these networks have measurable properties that can be analyzed and offer insights about consumption trends. Moreover, this methodology is more effective to determine outliers and errors in the data. In this dissertation, a traditional statistical analysis is first performed, where I point out its limits, and I then introduce the new network methodology and apply it to US water consumption data.

Several federal agencies exist that offer pertinent data on water resources, whether hydrologic and geologic data, or more specifically water consumption data. The major agencies include the National Climatic Data Center (NCDC), the U.S Army Corps of Engineering (USACE), the U.S Bureau of Reclamation (USBR), the U.S Environmental Protection Agency

(USEPA), the U.S Geological Survey(USGS), the U.S National Weather Service (NWS), the U.S Natural Resources Conservation Service (NRCS) (Chin 2002). To achieve the objective of this dissertation, the data was collected specifically from the U.S. Geological Survey (USGS 2014). The data provide estimates of water withdrawals by county, source of water and category of use, in five-year intervals. Of relevance, the USGS provides water withdrawals data for public supply, domestic, irrigation, livestock, aquaculture, industrial, mining and thermo-electric power. The majority of the water is withdrawn for irrigation and thermo-electric power purposes. The total withdrawals by category is shown in Figure 1.

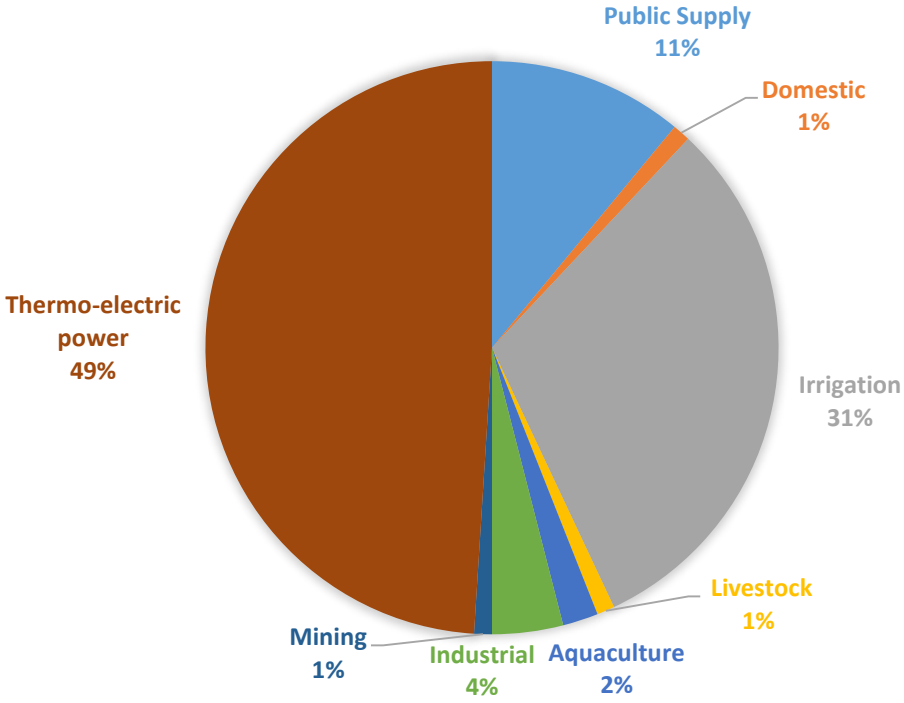



Figure 1 Total Water Withdrawals by Category in 2005 (Source: USGS)

Among the different categories, I focus specifically on public supply, which is defined as: "… water withdrawn by public and private water suppliers that provide water to at least 25 people or have a minimum of 15 connections… [and] … is delivered to users for domestic, commercial, and industrial purposes, and also is used for public services and system losses" (Kenny et al.

2009). Although it only accounts for 11% of the total water withdrawal in the US, it is directly linked to personal demand, which is the focus of this research. Moreover, specifically for this study, the datasets from 1985-2005 are used for 3,109 US counties (i.e. all counties except the counties of Alaska, Hawaii, Puerto Rico and Virgin islands). Moreover, the dataset for the year 2000 was not used since data for Connecticut, Hawaii, Kentucky, Maine, Massachusetts, Ohio, Pennsylvania, Tennessee, Texas , Utah were missing for this year. Finally, to pursue the network analysis, the free python library igra*ph* is used extensively (Csardi and Nepusz 2006).

This dissertation is largely based on the article "Evolution of Public Supply Water Withdrawal in the USA: A Network Approach" that is currently in press for publication.

# 2. LITERATURE REVIEW

## 2.1 Water-Use Systems

As water is one of the key elements for human habitation, water-use systems should be designed, in principal, to meet the demand for water. Water for these systems are collected from lakes, rivers and groundwater. Additionally, the rates of withdrawal should be, in principal, sustainable due to the limited availability of water as highlighted in the Introduction. The design of these systems are dictated by different demographic, economic, commercial and industrial requirements. Some main design facets are given below (Chin 2002). Four main types of water systems exist.

*Domestic water-supply systems*

Domestic water-supply systems extract water from a source, from which it is treated in a water treatment plant and finally distributed to the users. Moreover, these systems are designed in a way that the rate of extraction should meet the requirements for sustainability; treatment ensures sufficient quality and distribution confirms adequate amount and pressure during the peak hour; typically kept at about 240 kPa.

*Domestic wastewater collection systems*

Domestic wastewater collection systems are responsible for the collection and treatment of the wastewater. The treatment plant ensures that the wastewater is clean, that it can be discharged, and that the discharged water is not able to degrade the quality of the receiving water.

*Irrigation systems*

Water requirement of the crops are met by the rainfall and irrigation. To design the irrigation systems, estimation of crops water requirement is necessary. Generally, water is extracted from groundwater for irrigation. Additionally, care must be taken to ensure sustainable extraction rates similar to domestic water-supply systems.

*Hydroelectric power systems*

One of the most important use of water is for hydroelectric power generation. In fact, the largest power plant in the world is the Three Gorges hydroelectric plant in China, producing more than 22.5 MW or 92.2 TWh per year(Bin 2004; Jackson and Sleigh 2000). The open channels used for this purpose typically have flows that vary seasonally and that experience frequent drought, and dams and reservoirs are commonly built to make it economically feasible. Reservoirs are used to store water for continuous supply to human habitation and water released from the reservoirs behind the dams is used to generate electricity.

## 2.2 Estimated Use of Water in the United States

Numerous U.S federal agencies provide information regarding water in the United-States. In their publications, the United-States Geological Survey (USGS) offer significant information on water use, available from 1960 onwards for every 5 years. Estimations at the county level are available from 1985. From 1960 to 1980, data are available on state level. One of the most important variables that affect water consumption is simply population, which essentially drives demand. In the US, the population has doubled from 1950 to 2005. Moreover and symptomatic of the general trend of the past two centuries, the urban population has increased significantly

faster than the rural population. Thus, the spatial patterns of water usage have also changed over this time period. The trends in public supply water withdrawal are given in Table 1. The data comes from the USGS and the measurements were performed at the source (i.e., water treatment plants), which therefore includes any losses due to pipe leakage.

Table 1 Water Consumption Trend in the USA

| Year | Population Served in Million | Total withdrawn in Mega-gallon per day (Mgd) | Per Capita Usage in gpcd |
|---|---|---|---|
| 1960 | 136 | 21,000 | 151 |
| 1965 | 153 | 23,600 | 155 |
| 1970 | 165 | 27,000 | 166 |
| 1975 | 175 | 29,000 | 168 |
| 1980 | 186 | 34,000 | 183 |
| 1985 | 200 | 36,500 | 183 |
| 1990 | 210 | 38,500 | 184 |
| 1995 | 225 | 40,200 | 179 |
| 2000 | 242 | 43,300 | 179 |
| 2005 | 258 | 44,200 | 171 |

From the table 1, we can observe that per capita usage kept increasing from 1960 to 1990, reaches to its culmination in 1990, when per capita usage was 184 gpcd. After 1990 per capita usage decreased slightly despite the continuous increment of population and population served.

## 2.3 Consumption of Water

As mentioned earlier, the realm of water and water-use is one of the key elements for human habitation. Therefore, numerous studies have been performed on water consumption. Searching for the keyword “Consumption of Water” in Scopus resulted in 67,324 documents as shown in

Figure 2. The exponential growth in the past 45 years clearly highlights the relevance of the topic in the scientific community.

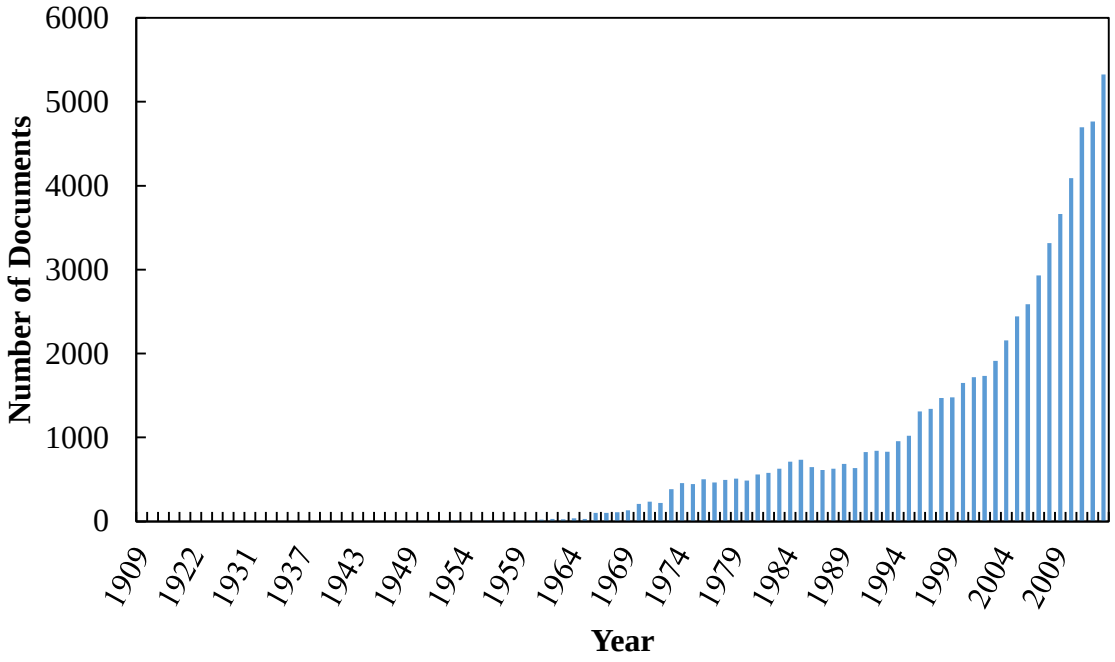



Figure 2 Number of Documents for "Consumption of Water" by year in Scopus

2.4 Water Sustainability

Along with water consumption, the later part of the twentieth century also saw the advent of *sustainability* and *sustainable development.* Among all the natural resources, water is vital because of its extensive use and limited availability. Water and sustainability therefore go naturally hand in hand. Using Scopus once again, the keyword "water sustainability" returned a total of 13,487 documents between 1980 to 2013. The results are shown in Figure 3 that shows a similar trend with Figure 2.

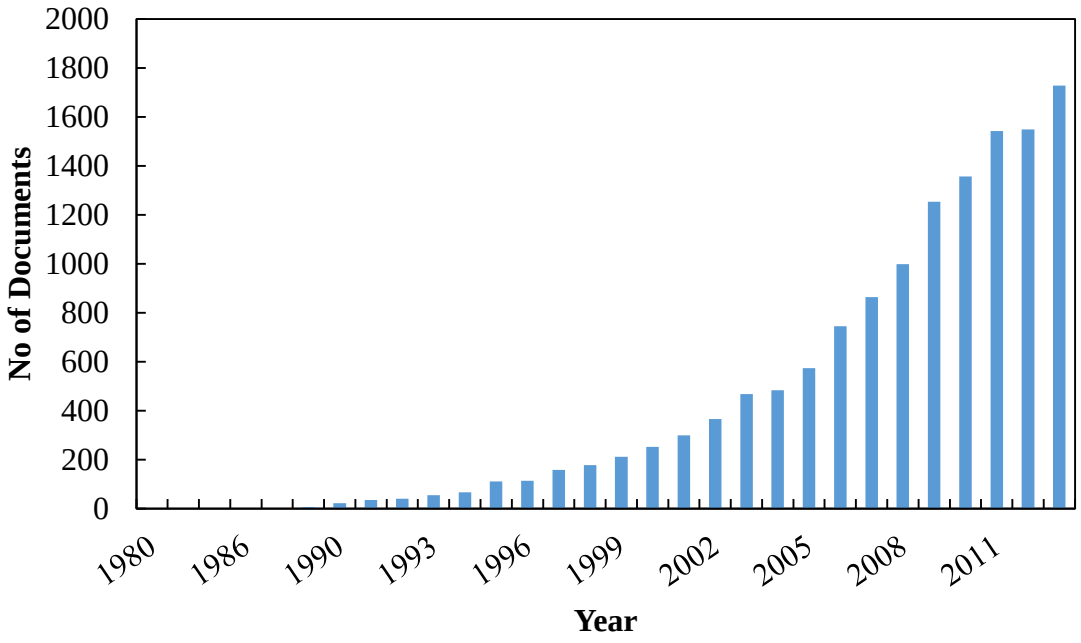



Figure 3 Number of Documents for "Water Sustainability" by year in Scopus

## 2.5 Traditional Statistical Analysis

From the collected USGS data, several traditional statistical measures, including the arithmetic mean, the median, and the standard deviation. The arithmetic mean is arguably the most widely used statistical indicator. It represents an *average* point in the data and it is also used as a measurement of the central tendency either of a probability distribution or a random distribution. The median represents the middle point in a dataset when arranged is ascending order. The standard deviation is a measurement of the spread of the distribution. Although the mean is often seen as a representative value of the entire dataset, this is only true for symmetrical distributions. Skewed distributions such as the lognormal distribution for instance have means that are offset from the *mode* of the distribution; where the mode of a distribution is the most common point or bin in a frequency analysis.

In general in most datasets, some observations are distant from the general trends. These observations are known as outliers. An outlier can be a real observation or may appear due to error. Outliers can easily biased the outcome of the traditional analysis. Thus, they are often excluded from an analysis. Most commonly, outliers are discarded from a dataset using the following equation.

$$Outlier \begin{cases} if\ \ the\, value\, is\, lower\, than\, Q_1 - 1.5IQR \\ if\, the\, value\, is\ greater\, than\, Q_3 + 1.5IQR \end{cases} \qquad (1)$$

where, $Q_1$ represents the lower quartile, $Q_3$ represents the upper quartile and IQR is the inter quartile range.

For a graphical representation of the distribution, histograms are typically used, and it was first introduced by Karl Pearson (Pearson 1895). To construct a histogram, the range of the distribution, which is the difference between the maxima and minima, is first divided into separate bins. Then, the number of observations for each bin is recorded. Finally, the histogram is plotted, where the base of each histogram represents the bin width and the height represents the frequency for that particular bin. Although histograms are ubiquitous and present in all standard statistical textbooks (Scott 1992), finding an appropriate bin width turns out to be a major and highly non-trivial challenge. Indeed, and the selection of different bin width can easily provide different types of distributions.

Mathematically, histograms involve discrete data. When normalized, these histograms are defined as Probability Density Function (PDF):

$$PDF: f(x) = \frac{P(x_i - \frac{dx}{2} < x \le x_i + \frac{dx}{2})}{dx} \qquad (2)$$

where, $P(x_i - \frac{dx}{2} < x \leq x_i + \frac{dx}{2})$ represents the probability of variable *x* to lie in the given range. Thus, a range has to be provided, which refers back to the issue of defining a bin width. Any errors are then transferred in the calculation of the cumulative distribution function (CDF), which is the integral of the PDF between 0 and 1. Naturally, the provision of different ranges or bin widths can generate different PDFs a given dataset.

To further study how these distributions have evolved in the time periods studied, two more measures can be calculated: the Gini coefficient and entropy. These two measures have notably been used in income distribution studies to estimate how "equally" wealth is distributed in a nation. For this study, Gini coefficient and entropy quantify the homogeneity of the data.

This conception of equality was first introduced by Vilfredo Pareto. While gardening, Pareto found that 80 percent of his peas are produced by 20 percent of peapods. He also carefully observed that 80% of the land in Italy is owned by 20% people. This phenomenon is very common in all societies and it is known as the Pareto principle or the 80/20 rule (Barabási 2002). The Gini coefficient was initially developed in 1912 (Ceriani and Verme 2012). It is defined mathematically based on the Lorenz curve (Lorenz 1905), which is the cumulative distribution of a probability or in our case the cumulative consumption of water. The Gini coefficient divides the area above the Lorenz curve by the area under an equivalent 45° line (i.e., that assumes a perfectly uniform distribution and also known as line of equality). The coefficient is therefore bounded by 0 and 1; a low value suggest a more equal distribution, with 0 representing perfect equality, and the opposite is true for high values. A Typical Lorenz curve and line of equality is shown in Figure 4, and the Gini coefficient calculates the ratio of the area A by the area of A + B.

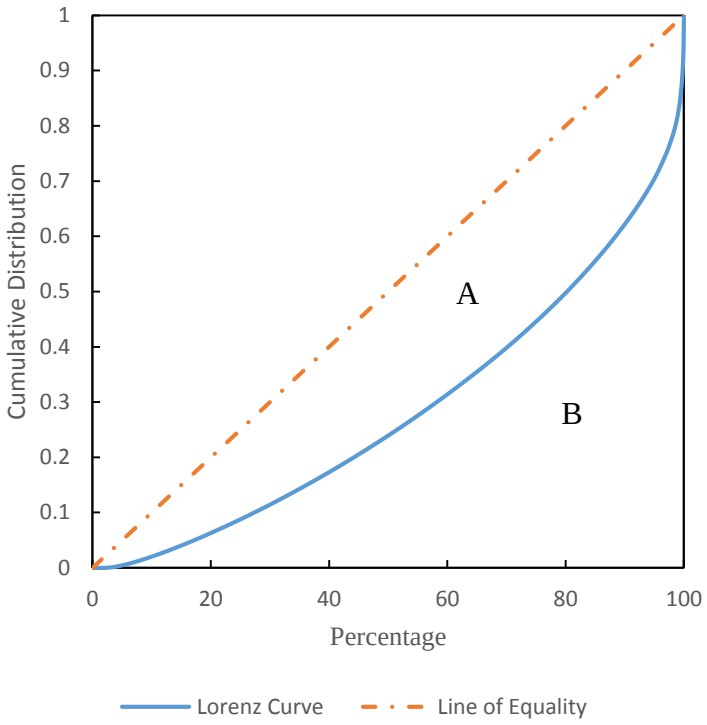



Figure 4 Typical Lorenz Curve and Line of Equality

The concept of entropy was first applied by Theil (1967) to measure inequality in a distribution. Nonetheless, it originates from statistical mechanics developed by Boltzmann (1872) and popularized in information theory by Shannon (2001). Mathematically it is defined as:

$$H(x) = -\sum p(x_i)\log(p(x_i)) \tag{3}$$

where $H(x)$ is the entropy, and $p(x_i)$ is the probability of event $x_i$, Entropy is bounded by 0 and infinity. A low value suggests an unequal distribution, while the large value suggests a more equal distribution unlike the Gini coefficient.

## 2.6 Network Science

The foundation of Graph Theory, the "father" of Network Science, was laid by Leonhard Euler in 1736 in his famous paper 'Seven Bridges of Königsberg' (Shields 2012). Later, Paul Erdős and Alfréd Rényi developed probabilistic theory in Network Science in eight papers on random graphs (P. Erdös and Rényi 1959); Paul Erdös and Rényi 1960). Much later in the late 1990s, Duncan Watts and Steven Strogatz described the small world problem mathematically in 1998 (Watts and Strogatz 1998), where in a small world network most nodes are accessible through a small number of hops or steps.

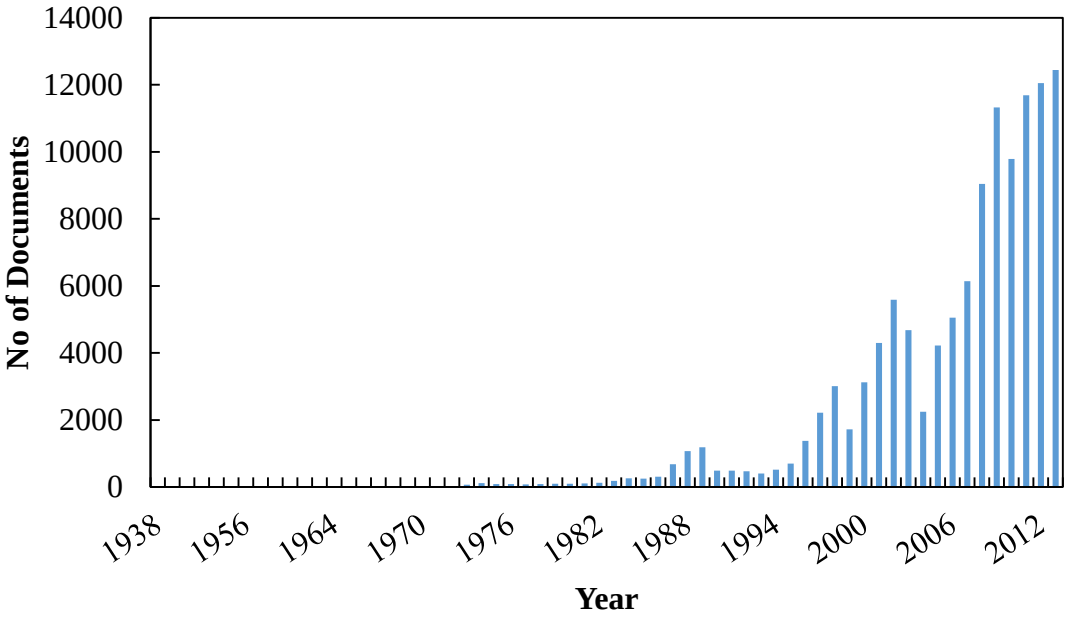



Figure 5 Number of Documents for "Network Science" by year in Scopus

At the same time, Albert-László Barabási's scale free networks (Barabási and Albert 1999) popularize the realm of Network Science, where the degree distribution of a scale free network follows a power law. Since then, many studies have been produced in "Network Science" in

myriad fields. A search in Scopus for the term "Network Science" resulted in 118,066 documents from 1938 to 2013 (shown in Figure 5).

## 2.7 Network Theory and Measures

A network or a graph $G$ is a collection of vertices/nodes $N$ joined by edges/links $L$; $G=\{N, L\}$. While two nodes can be connected by multiple links, or nodes can be connected to themselves (i.e. self-edge or self-loop), these types of nodes and links are not relevant for this work. Moreover, networks can be directed or undirected if a direction is given to a link. A node $n_i$ is connected to another node $n_j$ by a link $l_{ij}$. These connections are typically represented in the form of an adjacency matrix $A_{ij}$ where:

$$A_{ij}\begin{cases}1 \text{ if there is a link between nodes } n_i \text{ and } n_j\\ 0 \text{ otherwise}\end{cases} \quad (4)$$

The first measure that can be calculated is the degree $k$ of a node, i.e., its total number of connections, defined as $k_i = \sum_j A_{ij}$. In fact, the study of the distribution of degrees is common in network analysis, i.e., how many nodes have a degree 1, 2, 3, and so on. Some networks tend to follow normal distributions and they are sometimes called Erdös and Rényi networks (Erdös and Rényi 1959, 1960), while others follow power-law distributions and they are sometimes called scale-free networks (Barabási and Albert 1999). Analyzing the distribution of a property in networks can uncover significant insights about their properties.

Another important aspect of networks relates to *distance*. In particular, the concept of geodesic distance or shortest path-length between two nodes is relevant for us. As the name suggests, this is essentially the least number of "hops" or nodes that must be visited to join any pair of nodes. It follows that average shortest path length is defined as the average of the shortest path lengths of that network. In our context, the average shortest path length is used to analyze

the spread of the network, in a similar philosophy as the standard deviation. Also of relevance is the largest of all these shortest path lengths, which is called the *diameter* of a network. For us, the diameter is conceptually close to the difference between maxima and minima. In other words, the diameter of a network is used to determine the extent of a connected network.

A popular measure in network theory is related to the concept of connectivity. One measure of connectivity is called density. Density $\rho$ basically calculates the number of links in a network divided by the total number of potential links in this network, which can be calculated as ½ $N(N\text{-}1)$. Density therefore takes the form:

$$\rho = \frac{2L}{N(N-1)} \tag{3}$$

A final concept that will be relevant in our study is that of *giant cluster*. Although it was not mentioned previously, all nodes in a network do not have to belong to one single network. This is referred to a connected versus disconnected network. Figure 6 shows two such examples, where network 6b contains three sub-networks.

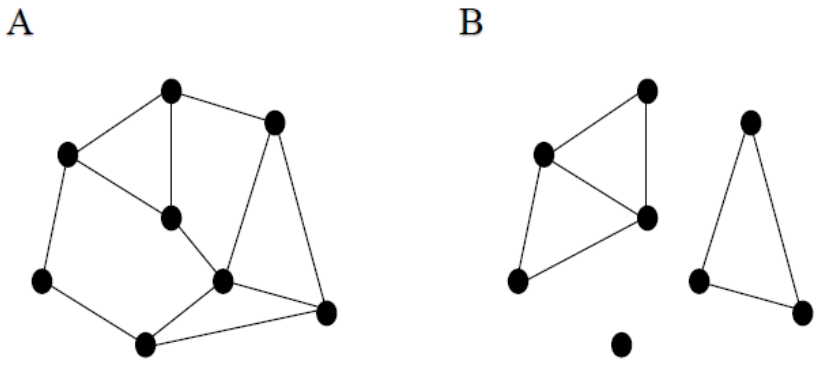



Figure 6 Connected versus Disconnected Networks. A) Connected. B) Disconnected

Some networks, as they grow, i.e., as they accumulate more links, start to have one sub-network that tends to absorb most of the links. This particular sub-network is referred to as a *giant cluster*, and this concept will be used directly in the results section.

One of the most commonly studied types of networks are social network, where each actor of a society is a node and they are connected to others via their interactions. The main conception of social network analysis is to analyze the complete network instead of individual actor and their activities (Wasserman and Galaskiewicz 1994). A social network can be very complex in nature and it therefore often possesses interesting properties. One of the famous theory about social networks is that everyone and everything six or fewer steps away from each other, which is known as "Six Degrees of Separation". It was first introduced by Frigyes Karinthy in 1929 in a short story.

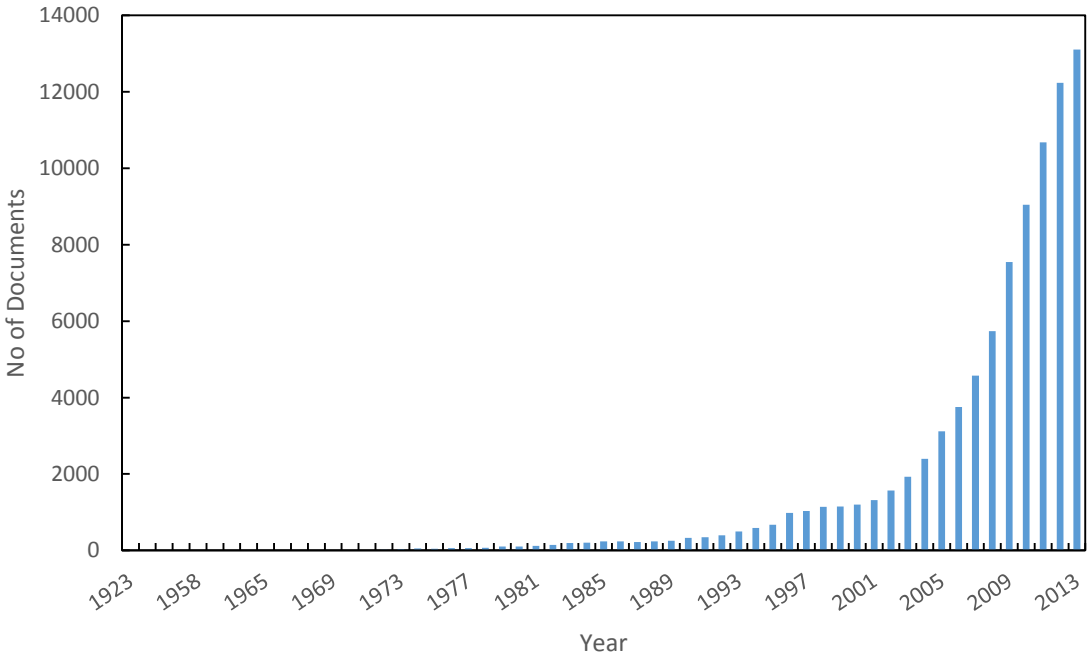



Figure 7 Number of Documents for "Social Network" by year in Scopus

Though that short story was not that popular but it was the first published version of the popular concept of "Six Degrees of Separation". Later it was rediscovered by Stanley Milgram, who was a Professor of Psychology at Harvard. His goal was to find the distance between two random people in the USA (Barabási 2002). Numerous studies have also been produced in the realm of "Social Network", Scopus showing 87,797 documents when searching the term (Figure 7).

## 2.8 Homophily

Homophily is an interesting and prevalent aspect of social network. A propensity for human being to link with similar others is known as homophily. The existence of homophily is observed in myriad network studies. Persons in homophilic relationships share common attributes such as beliefs, education, and social status (McPherson, Smith-Lovin, and Cook 2001). Homophily frequently prompt homogamy, where homogamy is the marriage between people with alike individualities. In our case, we apply this concept of homophily to counties that share similar water consumption properties.

## 2.9 Complexity Science

Complexity science is a relatively new and rapidly developing science discipline. A complex system consists of many individual parts that are interacting with one another and that show emerging behavior. Using tools and techniques from the realm of "Complexity Science", a cornucopia of new and pertinent information can be extracted. Network Science is often associated as a sub-discipline of Complexity Science. Abundant research has been produced in this in the last two decade. A search in Scopus for the term "Complexity Science" resulted in 40,331 documents. The results are shown in Figure 8.

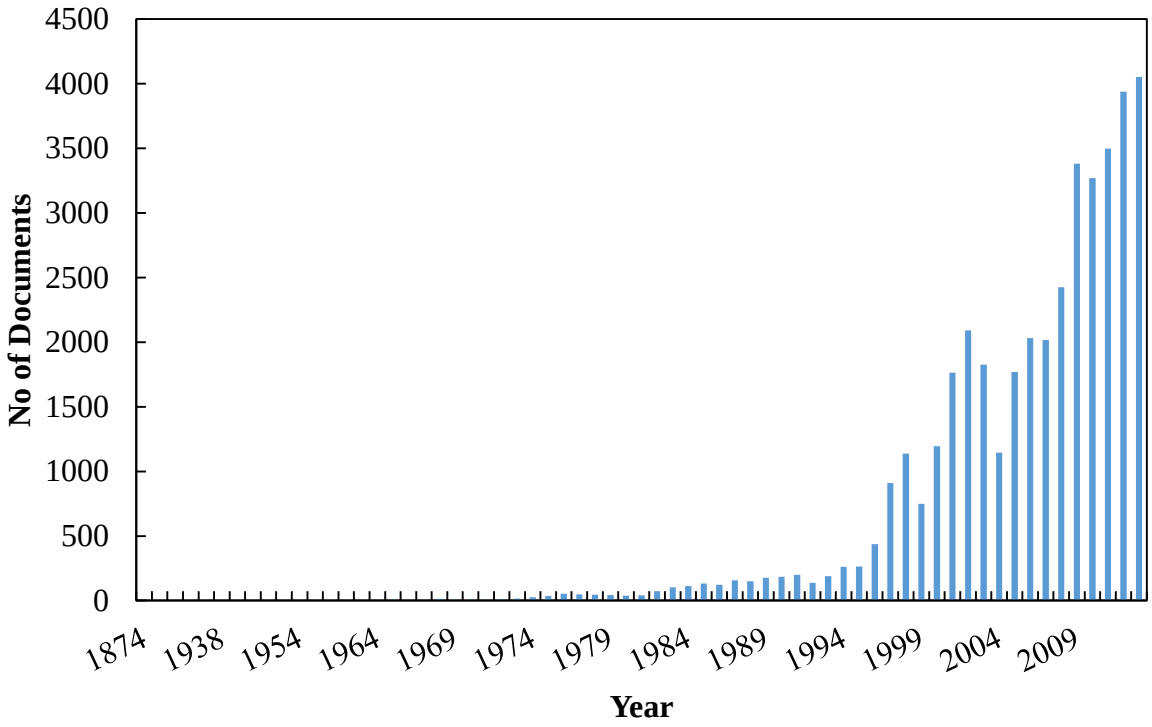



Figure 8 Number of Documents for “Complexity Science” by year in Scopus

# 3. METHODOLOGY

In this section, we first highlight the limits of these traditional statistical tools and we then offer a new network approach to study water consumption.

## 3.1 Traditional Analysis

From the water consumption data collected, the per-capita national average, the per-capita arithmetic mean, the per-capita median, and the standard deviation of water consumption in the US from 1985 to 2005 are first calculated. A box plot analysis is also produced for all the years. The per-capita national average is calculated by dividing the total national consumption by the total population served (where we essentially aggregated values from county level data). In contrast, the arithmetic mean is calculated by taking the mean of all values given at the county level. Since those values are already mean per-capita consumption values per county, they carry mean-of-the-mean type errors. Although these types of errors are significant, they are commonly overlooked, often because of data availability issue. Both approaches is pursued because population is not always given in this type of analysis, which can have serious impacts. The median and the standard deviation were also calculated with county level data. Moreover, we pursued this analysis for two different scenarios: (1) with and (2) without outliers.

To get a better sense of the distribution of consumption values, a histogram analysis is also performed for different bin-width. Moreover, from a spatial perspective, we can also visualize how consumption patterns differ across the US for the four years. To further study how these distributions have evolved in the time periods studied, two more measures can be calculated: the

Gini coefficient and entropy. To supplement this initial analysis, the next section introduces a new approach based on network analysis.

3.2 Formal Methodology

For this study, the nodes in this network are the counties, and we therefore have 3,109 nodes. The counties are then linked if they have similar consumption patterns. In other words, a county *i* is linked to county *j* when:

$$\mu_i(1-\xi) \leq \mu_j \leq \mu_i(1+\xi) \quad (4)$$

where, $\mu_i$ is the per-capita consumption of county *i*, $\mu_j$ is the per-capita consumption of county *j*, and $\xi$ is called the cutoff percentage. Different cutoff percentages $\xi$ output different networks, with different network properties that can be analyzed accordingly. Naturally, the different properties introduced above, along with the cutoffs, can also be compared longitudinally, which is relevant here since we have data for the years 1985, 1990, 1995, and 2005. The selection of $\xi$ in the network approach may seem analogous to binning but it is not. Conventionally, a bin represent a range between two values but the selection of $\xi$ in the network mimics a range around each individual value, which is completely different. In fact, this network approach is more similar to a machine learning approach where we increase the cutoff $\xi$ and calculate different network properties to better understand the nature of the relationship. All $\xi$ are therefore used unlike bins.

# 4. RESULTS

## 4.1 Traditional Analysis Result

Table 2 shows the evolution of the per-capita national average, the per-capita arithmetic mean, the per-capita median, and the standard deviation of water consumption in the US from 1985 to 2005. A box plot of all the four years is provided in Figure 9. All values shown are in gallons per-capita per day (gpcd).

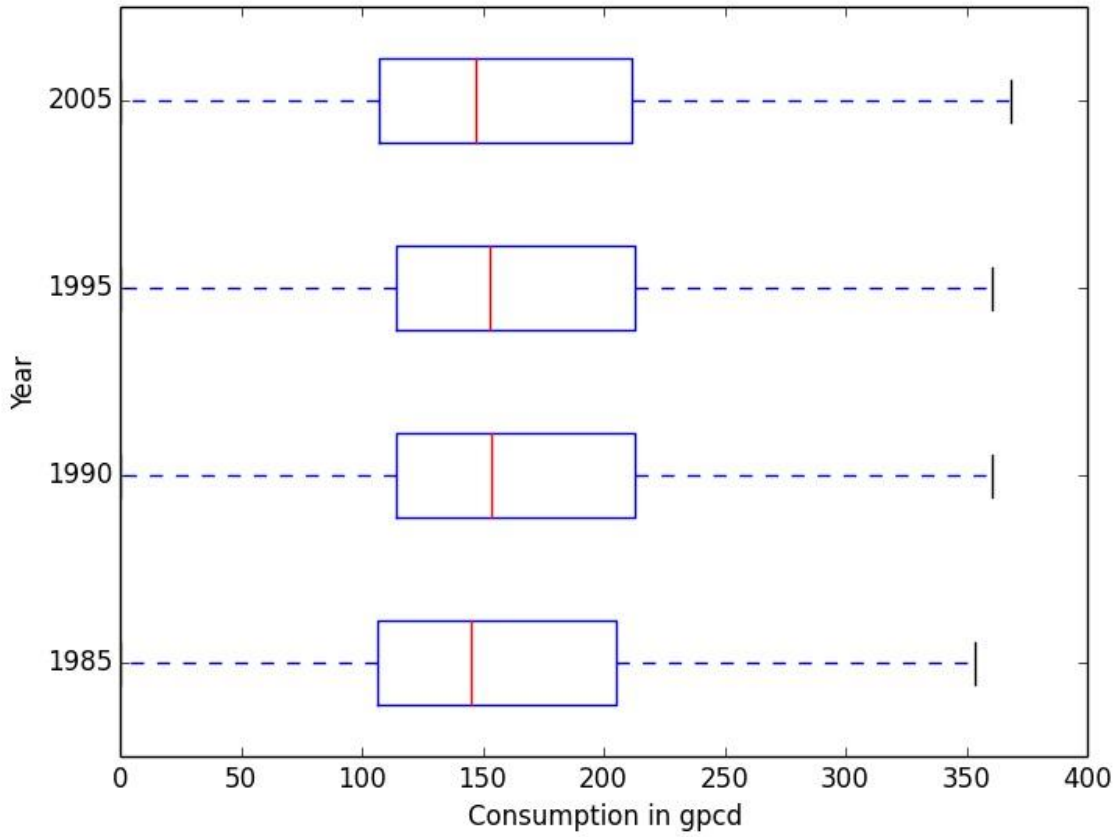



Figure 9 Box plot of water consumption of US counties from 1985 to 2005 (without outliers)

Table 2 Traditional statistical measures for US public supply water consumption

| Year | National Average | Arithmetic Mean | | Median | | Standard Deviation | | Gini Coefficient | Entropy |
|---|---|---|---|---|---|---|---|---|---|
| | | with outliers | without outliers | with outliers | without outliers | with outliers | without outliers | | |
| 1985 | 183.52 | 207.01 | 149.38 | 145.25 | 139.97 | 582.07 | 69.37 | 0.43 | 7.44 |
| 1990 | 184.76 | 212.67 | 157.05 | 153.39 | 148.18 | 580.48 | 72.24 | 0.42 | 7.45 |
| 1995 | 179.67 | 217.82 | 156.26 | 153.02 | 147.44 | 611.90 | 72.03 | 0.44 | 7.4 |
| 2005 | 171.41 | 297.30 | 150.57 | 147.28 | 141.60 | 3871.62 | 75.80 | 0.49 | 6.12 |

Overall, while the national average has decreased in the time period studied, the mean with outliers has significantly increased and the median has stayed relatively constant. The difference between the national average and mean suggests that counties with lower population may tend to consume more water, which is creating this upward bias. This is also reflected in the medians that are lower than both the national average and the mean.

The means without the outliers are much more closely aligned with the medians. Regardless of this phenomenon, the fact these three measures, with and without errors, follow different trends is problematic and it gives us little useful information about the actual consumption patterns.

The standard deviation with outliers stayed relatively constant from 1985 to 1995, and then increased significantly. This rapid increase can suggest that the method of collection may have changed between 1995 and 2005, or that the data collected contains outliers. The latter proposition is slightly more likely since the highest consumption value found for 2005 was 211,610 gpcd compared to the calculated mean of 297.3 gpcd. Moreover, this type of error is also possibly affecting the national average data, which invalidates the trends observed above.

The standard deviation without outliers is much more stable, although it seemed to have increased slightly over the past 30 years.

Figure 10 shows the histogram analysis with a bin-width of 20 gpcd. In this figure, a constant bin width of 20 gpcd was kept up to 500 gpcd, after which all counties were included in one single bin. A histogram analysis is also performed for different bin width for all the four years. Detail results are provided in Appendix section (A2,A3,A4,A5). The problem with data-binning is well known and difficult to solve, despite being presented in all standard statistical textbooks (Scott 1992). As observed in Figure 10, some problems occur in the first bin because several counties have no public use water consumption (they are mostly rural using private wells).

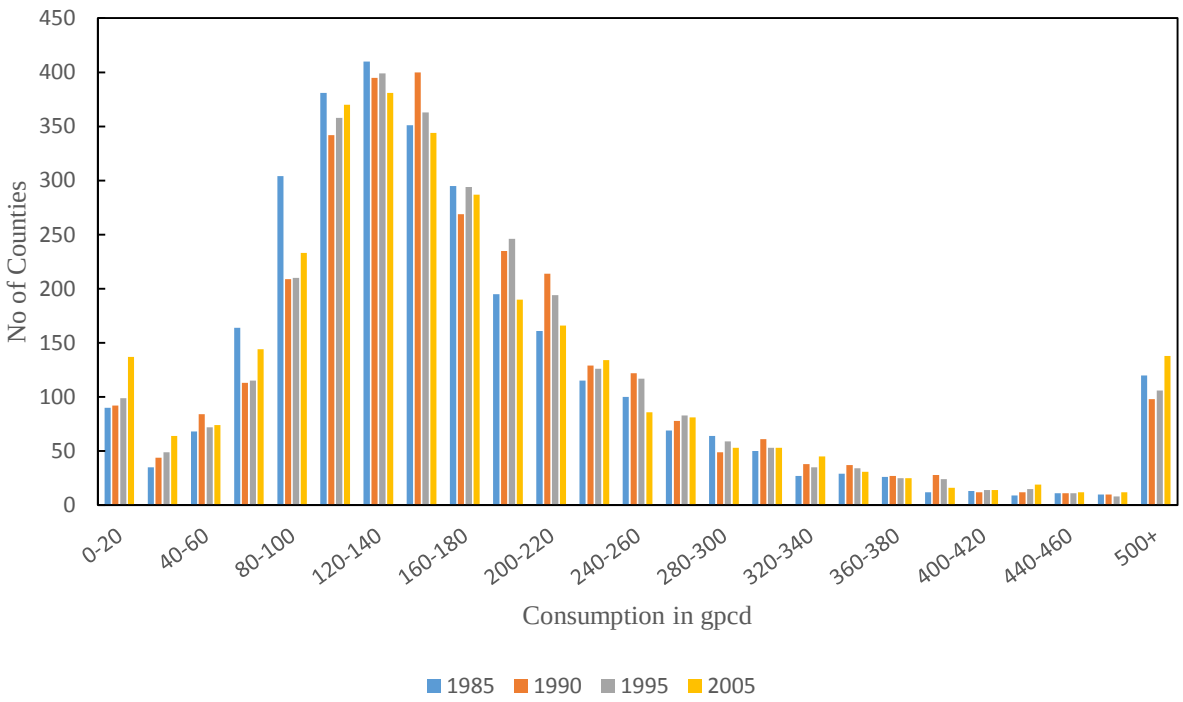



Figure 10 Histogram of Distribution of Water Consumption in US Counties

Despite that, the data tend to best fit lognormal distributions. Lognormal distributions are in fact skewed towards large values, which partially explains the apparent upward bias noted earlier. Indeed, we can see that the averages, means, and medians calculated earlier are systemically located right of the mode of the distribution. In fact, identifying the *mode* (i.e., peak) of the distribution here is critical since it is arguably more representative of overall patterns than the average. On the figure, we observe a single peak in the 120-140 gpcd bin, except for the 1990 data where it may be in the 140-160 gpcd bin.

From a spatial perspective, we can also visualize how consumption patterns differ across the US for the four years. Figure 11 shows the same data in a map form, distributed in seven bins for clarity. Despite the higher resolution, no trends can be identified except the possible observation that counties in Western US may have higher consumption levels.


Figure 11 Spatial Distribution of Water Consumption in US Counties. A) 1985. B) 1990. C) 1995. D) 2005

From Table 2, we see that both the Gini coefficient and entropy stay relatively constant for the three first years. The Gini coefficient then increases significantly for 2005 and entropy follows the opposite trend. While these results would suggest that water consumption is becoming more "unequal", the possible presence of errors in the data prevents us from making any formal conclusion

.

4.2 Network Analysis Result

This section discusses the results from the application of the methodology outlined above.

First, as the cutoff percentage increased, rapid emergence of a giant cluster is observed that tends to link all the nodes in the network, indicating that the majority of the value in this data are homogenious. This becomes evident in Figure 12A that shows the proportional size of the giant cluster (i.e. ratio of number of nodes in the giant cluster to the total number of nodes). Indeed, with a cutoff of 1%, more than 80% of the counties are present in one single network.

The growth of this giant cluster, however, is not gradual. In fact, it undergoes sudden jumps that can be refer to as phase transitions. For all years, the size of the giant cluster increases sharply for cutoffs $\xi$ between 0.1% and 0.5%, it then increases mildly betweenn 0.5% and 1.0%, before finally increaseing steadily after 1.0% (not shown here). During the initial and sharp increase, large phase transitions can also be singled out significantly. In 1985, 1990 and 1995, this happens between 0.20 and 0.30%. In 2005 two major phase transitions are observed between 0.20 and 0.25% and between 0.35 and 0.40%.

Regarding the spread of these distributions, the average shortest-path length and diameter of the network are analyzed in Figures 12B and 12C respectively. Essentially, the average shortest path-length is conceptually close to the standard deviation since it offers a measure of the spread of the distribution, while the diameter is conceptually close to the the difference

between maxima and minima. By increasing the cutoffs $\xi$, similar hasty changes, akin to the growth of the giant cluster, are observed for the average path length and the diameter. For 1985, 1990 and 1995, the average path length reaches its peak at $\xi$ = 0.4% and in 2005 this happens when $\xi$ = 0.45%. The diameter of the giant cluster reaches its peak in 1985 at $\xi$ = 0.4%; in 1990 at $\xi$ = 0.4%; in 1995 at $\xi$ = 0.55%; in 2005 at $\xi$ = 0.45%. Despite these phase transitions, we can see that all four years follow similar kinds of patterns. In fact, for a cutoff of 1%, with the potential exception of 1995, the average path length of each year remains almost unchanged, which suggests that water consumption is neither becoming more equal or less equal. A similar observation can be made for the diameter.

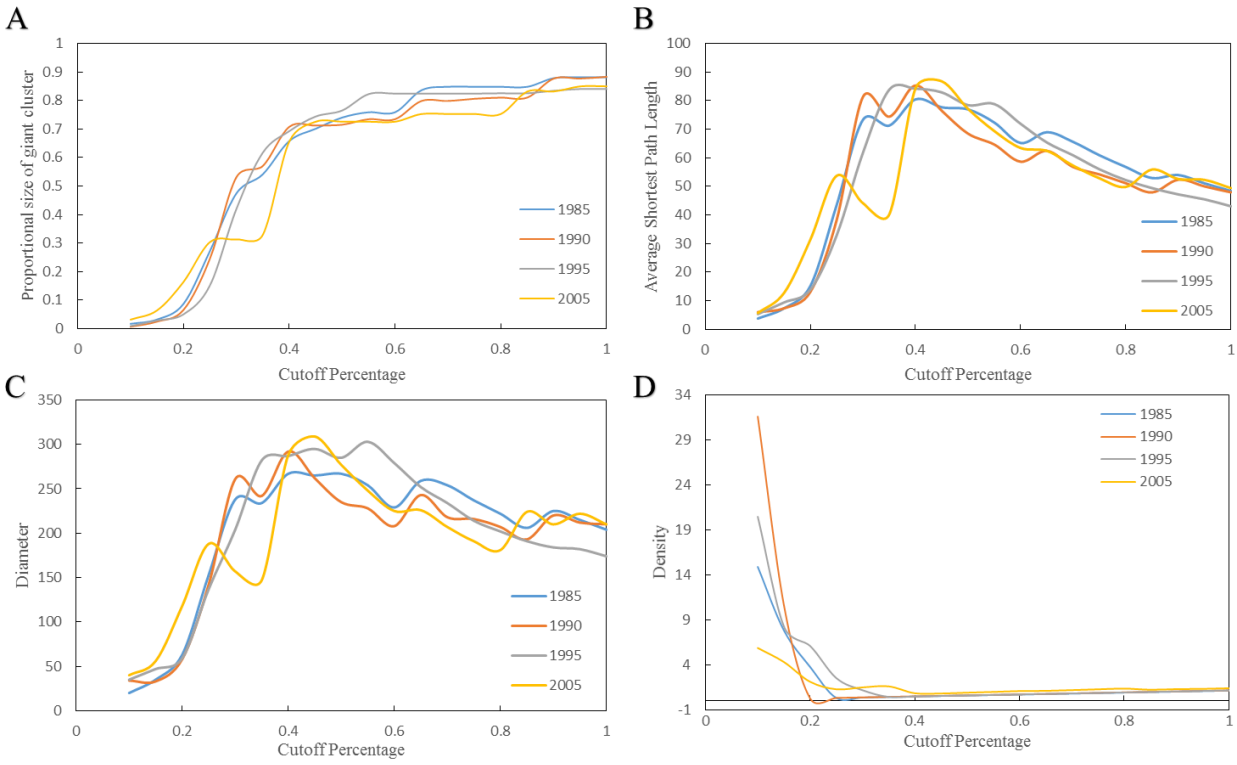



Figure 12 Change in A) Proportional size of giant cluster B) Average shortest-path length C) Diameter D) Density with cutoff percentage

Moreover, to emphasize on connectivity in the giant cluster specifically, the densities $\rho$ at different cutoffs are calculated and plotted on Figure 12D. From the figure, the giant clusters possessess high-density values for low cutoffs, but they drop very rapidly with an increase in $\xi$. This essentially relates to the fact that the giant cluster is becoming bigger. If the giant cluster increases by $N$ nodes, the potential number of links increases by $N^2$ (equation 4). Despite that, density becomes stable at around $\xi = 0.4\%$ for all years, after which it grows slowly. Here again, the fact that the same trend is present for all four years suggest that the system, or here water consumption trends, have not evolved much over the time period considered.

Furthermore, the degree distribution of the networks can offer significant insights. Figure 13 shows the degree distribution of our networks for a cutoff percentage of 1%. As mentioned, studying the frequency distribution of degrees is common in the network literature. Here however, the degree of a node is simply plotted against its water consumption value. This is therefore closer to a scatter plot than of a density function despite appearances. This is relevant and particularly insightful in our context. Indeed, nodes with higher degrees can be seen as more "representative" since their consumptions are more "similar" than other nodes. This notion directly reflects the idea of the *mode* of a distribution, which is arguably more useful as discussed ealier, although no specific data binning procedure is used.

Overall, these distributions tend to follow lognormal patterns, echoing the results from Figure 8, but again no data-binning process was used. Here, more detailed information can be observed. Indeed, the results for 1985 and 1990 show two peaks (which partially explains the significant difference in the mean, median, and average from Table 1); the peaks are at 123.44 gpcd (Mower County, Minnesota) and 170.61 gpcd (Lewis County, New York) for 1985, and at

143.51 (Wayne County, Illinois) and 160.32 gpcd (Ross County, Ohio) for 1990, with degrees around 70. The highest peaks in 1995 and 2005 occur at 136.47 gpcd (Giles County, Virginia) and 141.06 gpcd (Harrison County, West Virginia) respectively, with degrees also around 70. This multi-modal feature is interesting and may be a manifestation of more complex behaviors especially since they seem to converge towards one single peak. More crudely, these results also suggest that per-capita water consumption has stayed relatively stable since 1985, but not in a uniform fashion. Despite that, it is interesting to note the maximum and mimum values of consumption for each year in the giant sub graph.

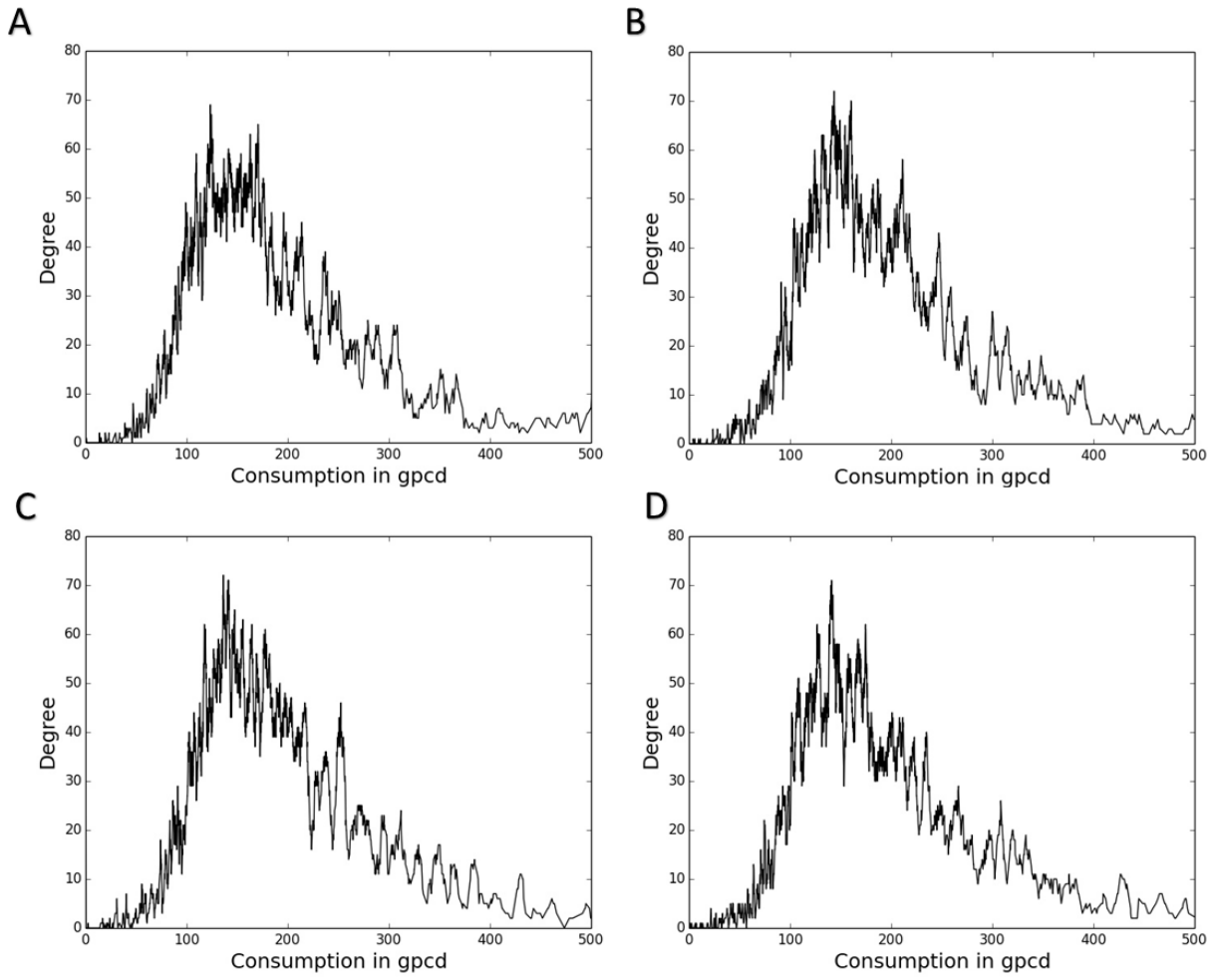



Figure 13 Degrees versus consumption. A) 1985 B)1990 C)1995 D)2005

While minima are consistant across the four data sets, maxima vary. For 1985 and 1990, maxima are around 400 gpcd; this decreases to about 340 gpcd for 1995 and increases to 440 gpcd for 2005. This partially explained why the average path length of 1995 was slightly smaller than other years as seen above.

Spatially, to visualize the counties that belong to the giant cluster before and after these major phase transitions, Figure 14 shows maps of the US for cutoffs $\xi$ of 0.25% and 1% for all four years.

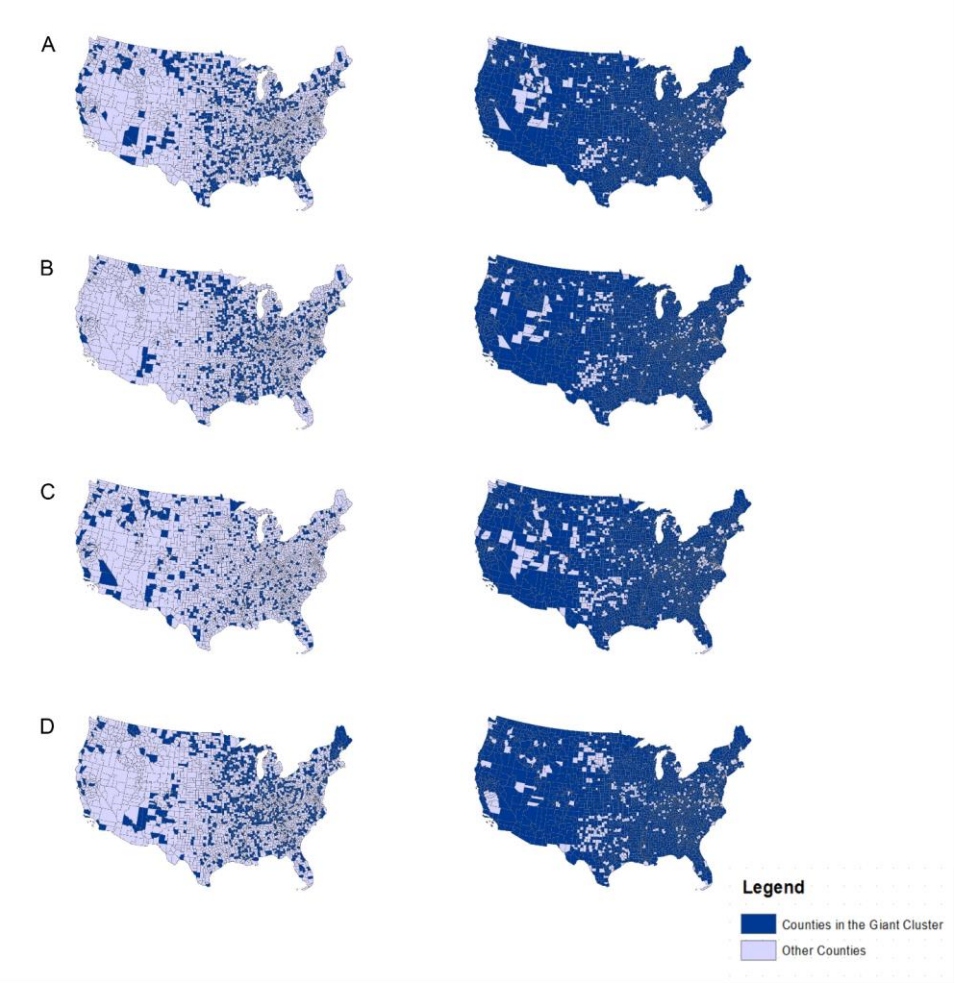



Figure 14 Number of counties in the giant cluster for cutoffs of 0.25% (left) 1.0% (right).

A)1985 B)1990 C)1995 D)2005

In this figure, counties can either be part of the giant cluster or not. While we intially expected to find that counties with similar properties (i.e., climate, population, and so on) would be spatially correlated, this does not seem to be the case. Indeed, except for several large counties in Western US located in the Rocky Mountains that are mostly rural, no inference can be made. This is surprising since water consumption is often thought to be correlated to climate. We also plotted water consumption vs. population density for all counties, and no statistical correlation could be established here either (not shown). Moreover, this lack of correlation also highlights that water consumption may not depend on whether or not a county is located close to a large body of water. For instance, many counties in the American Southwest are facing significant challenges, with groundwater aquifers reaching seriously low levels, and yet their water use remains part of the giant cluster. To further highlight the benefits of this network methodology, counties which have similar consumption with Cook County in 2005 are depicted in the Appendix A6. A full JavaScript visualization is also presented on the Complex and Sustainable Urban Networks (CSUN) Lab's website at http://csun.uic.edu.

These findings are discussed in the next section.

## 5. DISCUSSION

Despite being common and nearly ubiquitous, traditional statistical tools and techniques are limited and they can often fail to capture complex phenomena. Without accounting for outliers for instance, the Gini coefficient and entropy measures suggested that the water consumption trends were becoming less "equal", while this was a pure artifice of the data set. These errors are actually significant and in fact increasingly likely in this era of "Big Data". Although, this problem can be partially remediated by removing outliers, the method often end up sending mixed and sometimes contradictory signals about a phenomenon. One of the notable benefits of introducing a new network methodology is in fact the ability to select a sub-sample from the dataset that automatically identifies outliers. Another benefit is the fact that no binning process is required to study the data, which can be highly subjective and further bias the results. Furthermore, the topological properties calculated from these networks are not affected by the presence of outliers, which corroborates the robustness of the methodology. For comparison change in average shortest path length and diameter of the giant cluster is provided in supplementary section S7.

Essentially, this network topology can be compared to the construction of a social network, where two counties are linked if they share similar consumption patterns. By increasing the value of the cutoff percentage $\xi$, we rapidly see the emergence of a giant cluster. This is essentially part of the process of selection of a sub-sample. The evolution of this giant cluster grows through abrupt changes, referred to as "phase transitions", before becoming more stable for $\xi$ between 0.4% and 0.5%.

This type of evolution accompanied with phase transitions is not atypical in the social network literature. In our case, at 1%, the giant cluster is comprised of more than 80% of all counties. This does not suggest that the data for 20% of the counties is erroneous, but it does suggest that including these other counties might bias the results and provide wrong insights into consumption trends.

In our context, we find that comparing the degree of a node with its consumption can help determine important trends, akin to the mode of a distribution. This notably helped us identify the multi-modal feature of the data in earlier years. Despite that, the plots seem to evolve towards a single peak, which may suggest a process of "homogenization" in consumption. In other words, water consumption may in fact become more "equal" despite the results in Gini coefficient and entropy. Because of the transition from multi-modal to uni-modal, no definite conclusions can be made from considering the values beyond the fact that the consumption values for the highest degrees evolved from 136 gpcd to 141 gpcd between 1995 to 2005. The difference between the two values is too small to actually determine whether consumption trends have increased or stayed relatively constant.

Moreover, here, the average path length and diameter can be used as proxies to the standard deviation and the difference between maximum and minimum. After stabilizing the value of the cutoff $\xi$, these indicators are fairly close for all years although maxima may have had an impact. Indeed, the maximum for 2005 was higher than all other years. Therefore, although we witness a single peak in the consumption trends of 2005, we can also witness a wider base. Studying the evolution of density, we find that it decreases rapidly before settling for a cutoff of 0.4. This result suggests that values for cutoffs smaller than 0.4 may not have a meaning, hence the need to test various cutoff percentages for any application.

Finally and despite early intuition, no spatial patterns were found in the data by mapping the members of the giant clusters at cutoffs $\xi$ of 0.25 and 1%. In particular, climate does not seem to have an impact of water consumption. This is therefore an area that requires further studying.

# 6. CONCLUSION

The main goal of this dissertation was to analyze the evolution of water consumption in the USA. To achieve this goal, I used per-capita daily consumption data at the county scale for the years 1985, 1990, 1995, and 2005 from the USGS. Identifying current trends is paramount if we aspire to develop effective policies to become more sustainable. Traditional statistics, however, can fail to capture these current trends, and we first highlighted the limits of traditional statistics by calculating several measures and noting conflicting and biased results, notably due to errors in the dataset.

Having these limitations in mind, I formulated a new approach based on network theory. Essentially, two counties were linked if they showed similar consumption values, within plus or minus a cutoff percentage $\xi$. Immidiately, we observed the formation of a giant cluster, containing more than 80% of the counties with a cutoff $\xi$ of 1%. We then looked at the evolution of the average shortest path length, diameter, and density, and we found that the networks tend to stabilize after a cutoff $\xi$ of 0.4%. More importantly, we found that these measures do not evolved significantly over the 20 year periods. Looking at the degree of each nodes versus consumption, we made a parallel between the nodes with the highest degrees and the mode of a distribution. We also observed multi-modal features in earlier years, perhaps capturing complex patterns, while the data showed uni-modal distributions for 1995 and 2005, with the highest degree node showing a consumption of 141 gpcd in 2005.

It is, however, impossible at this point to determine whether consumption trends are increasing or staying constant, which is in itself a clearer and more consistent message than the conflicting patterns found in the traditional analysis. Finally, the spatial analysis of the giant cluster showed no relationship between water consumption and geography. In other words, the data cannot corroborate a relationship between water consumption and climate for instance.

Overall, considering the limitations of traditional statistical tools to capture current trends, a novel and more robust method is necesary. In this article, we used a network approach to study consumption patterns, and while the methodology itself remains to be formalized, and it still succeeded in providing meaningful insights. Future work should therefore concentrate on formalizing this method and applying it to many other resources, thus effectively analyzing current trends so as to eventually develop policies contributing to a more sustainable world.

## 7. REFERENCE


Barabási, Albert-László. 2002. Linked: The New Science of Networks. Perseus Books Group New York.

Barabási, Albert-László, and Réka Albert. 1999. “Emergence of Scaling in Random Networks.” Science 286 (5439): 509–12.

Bin, Tian. 2004. “1, Dai Huichao~ 2, Wang Shimei~ 1 (~ 1School of Civil Engineering, China Three Gorges University, Yichang 443002 China)(~ 2China Yangtze Three Gorges Project Development Corporation, Yichang 443002 China); STRENGTH CHRACTERISTICS OF SOIL IN SLIDE ZONE AND DETERMINATION OF ITS PARAMETERS [J].” Chinese Journal of Rock Mechanics and Engineering 17.

Boltzmann, L. 1872. “Weitere Studien Über Das Wärmegleichgewicht Unter Gasmolekülen, Sitzungs. Akad. Wiss. Wein 66 (1872), 275–370; English: Further Studies on the Thermal Equilibrium of Gas Molecules.” Kinetic Theory 2: 88–174.

Ceriani, Lidia, and Paolo Verme. 2012. “The Origins of the Gini Index: Extracts from Variabilità E Mutabilità (1912) by Corrado Gini.” The Journal of Economic Inequality 10 (3): 421–43.

Chin, David A. 2002. Water-Resources Engineering (3rd Edition). 3 edition. Prentice Hall.

Csardi, Gabor, and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” InterJournal, Complex Systems 1695 (5). http://www.necsi.edu/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf.

Easley, D., and J. Kleinberg. 2010. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press. http://books.google.com/books?id=atfCl2agdi8C.

Erdös, P., and A. Rényi. 1959. "On Random Graphs I." Publ. Math. Debrecen 6: 290–97.

Erdös, Paul, and A. Rényi. 1960. "On the Evolution of Random Graphs." Publ. Math. Inst. Hungar. Acad. Sci 5: 17–61.

Gleick, Peter H. 1993. Water in Crisis: A Guide to the World's Fresh Water Resources. Oxford University Press, Inc.

Hutson, Susan S. 2004. Estimated Use of Water in the United States in 2000. Vol. 1268. Geological Survey (USGS).

Jackson, Sukhan, and Adrian Sleigh. 2000. "Resettlement for China's Three Gorges Dam: Socio-Economic Impact and Institutional Tensions." Communist and Post-Communist Studies 33 (2): 223–41.

Kenny, Joan F., Nancy L. Barber, Susan S. Hutson, Kristin S. Linsey, John K. Lovelace, and Molly A. Maupin. 2009. Estimated Use of Water in the United States in 2005. US Geological Survey Reston, VA.

Lorenz, Max O. 1905. "Methods of Measuring the Concentration of Wealth." Publications of the American Statistical Association 9 (70): 209–19.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." Annual Review of Sociology, 415–44.

Newman, Mark. 2010. Networks: An Introduction. 1 edition. Oxford University Press, USA.

Pearson, Karl. 1895. “Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material.” Philosophical Transactions of the Royal Society of London. A, 343–414.

Scott, D.W. 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley Series in Probability and Statistics. Wiley.

Shannon, Claude Elwood. 2001. “A Mathematical Theory of Communication.” ACM SIGMOBILE Mobile Computing and Communications Review 5 (1): 3–55.

Shields, Rob. 2012. “Cultural Topology: The Seven Bridges of Königsburg, 1736.” Theory, Culture & Society 29 (4-5): 43–57.

Solley, Wayne B., Charles F. Merk, and Robert R. Pierce. 1988. Estimated Use of Water in the United States in 1985. CIR - 1004. United States Geological Survey.

Solley, Wayne B., Robert R. Pierce, and Howard A. Perlman. 1993. Estimated Use of Water in the United States in 1990. CIR - 1081. United States Geological Survey.

———. 1998. Estimated Use of Water in the United States in 1995. CIR - 1200. United States Geological Survey.

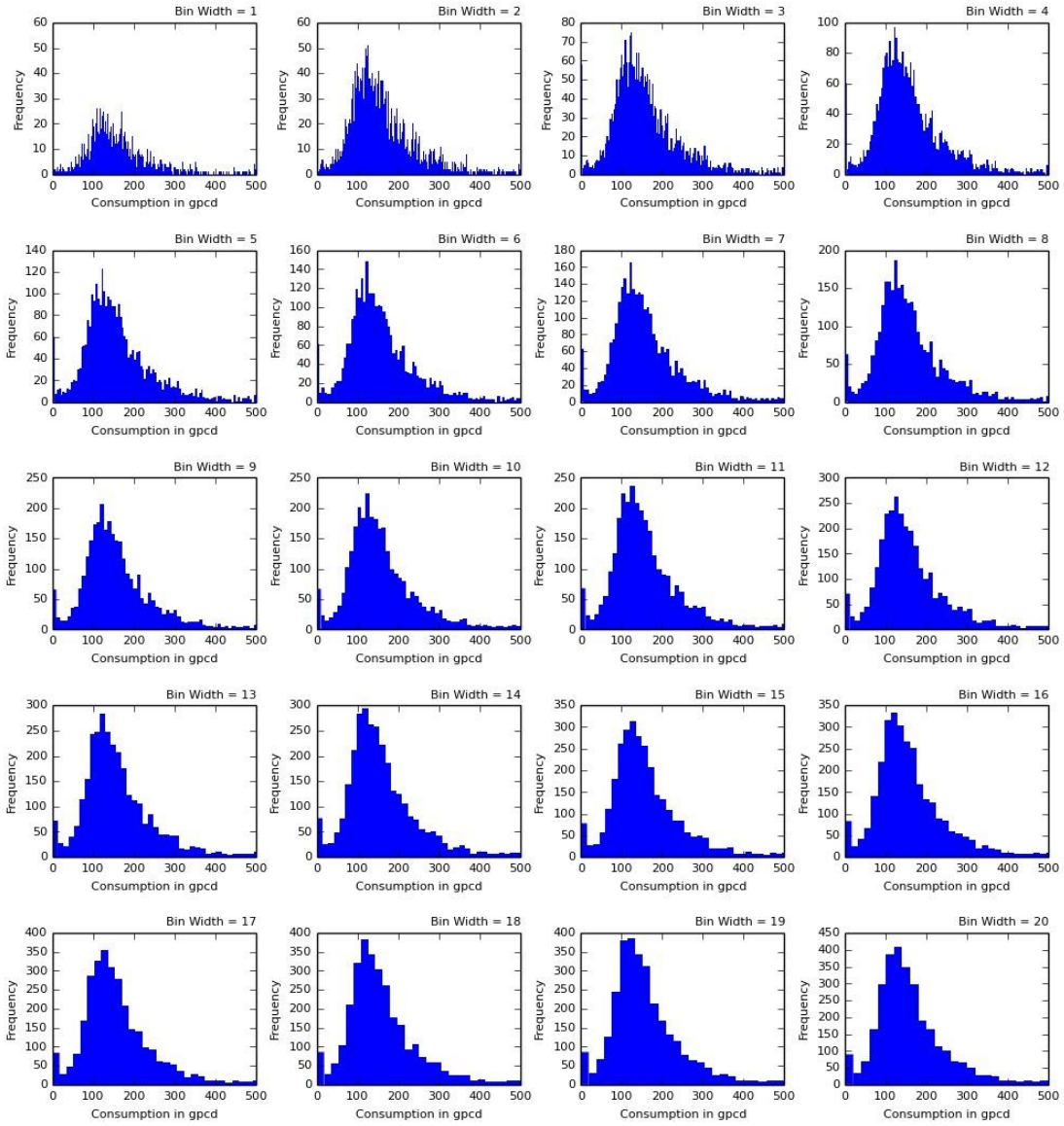
Theil, Henri. 1967. Economics and Information Theory. Vol. 7. North-Holland Amsterdam.

Wasserman, Stanley, and Joseph Galaskiewicz. 1994. Advances in Social Network Analysis: Research in the Social and Behavioral Sciences. Vol. 171. Sage Publications. http://books.google.com/books?hl=en&lr=&id=mB11AwAAQBAJ&oi=fnd&pg=PP1&dq=Social+Network+Analysis+in+the+Social+and+Behavioral+Sciences%22&ots=TIjrLjAf3L&sig=UL0A1fhfAJ62eKnuy5R9csYgKpI.

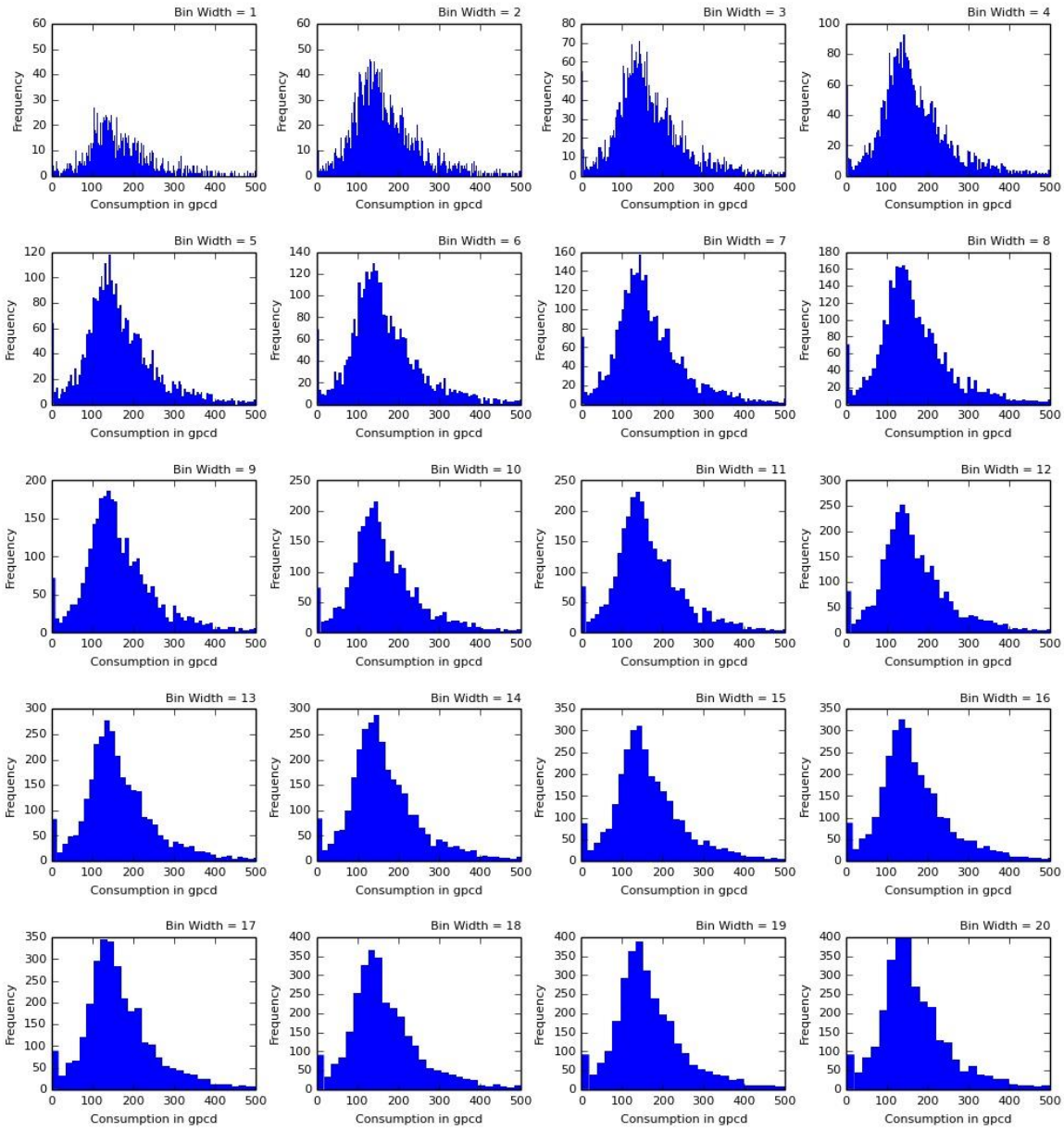
Watts, Duncan J., and Steven H. Strogatz. 1998. “Collective Dynamics of ‘small-World’networks.” Nature 393 (6684): 440–42.

## 8. APPENDIX

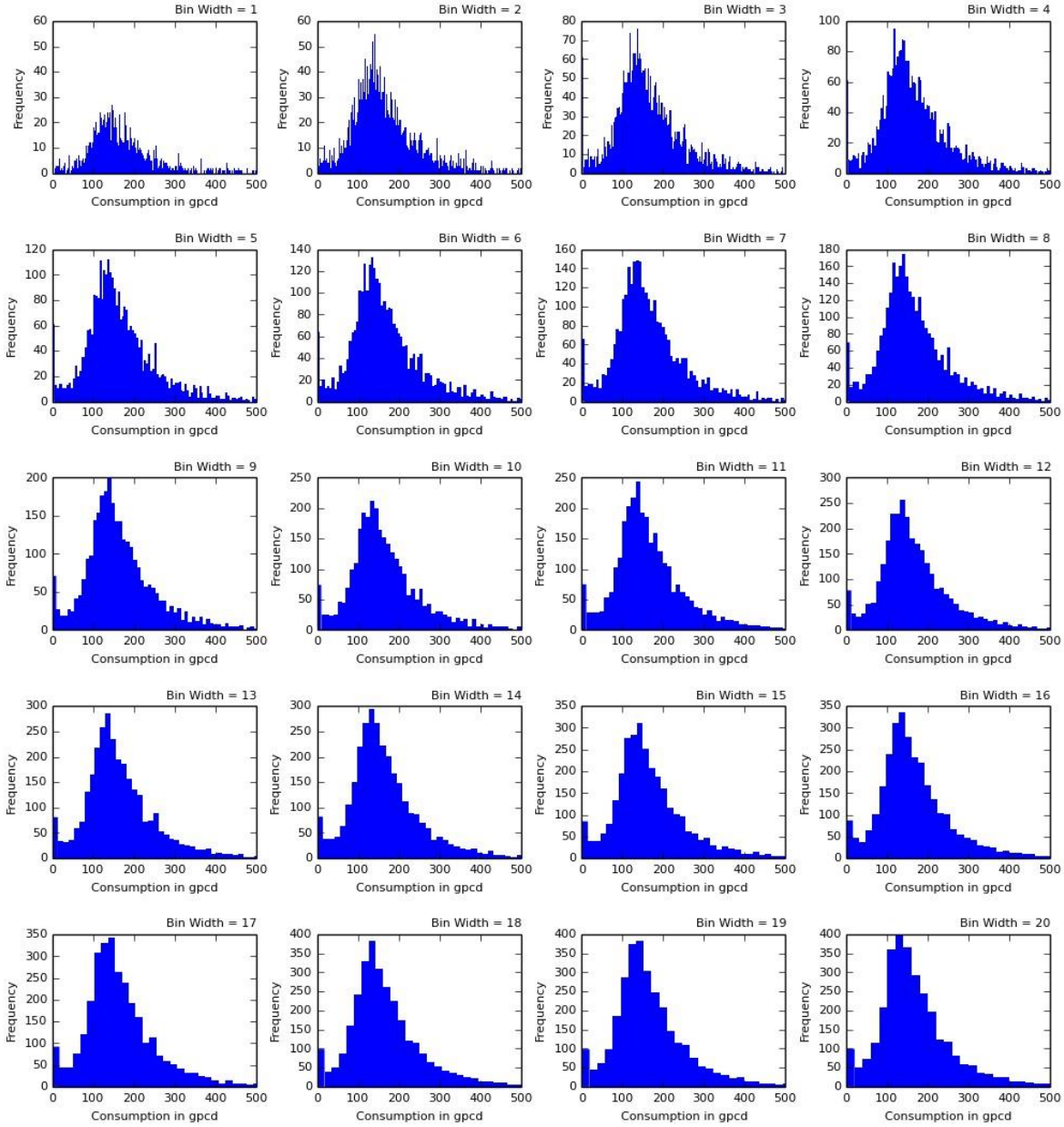
### A. Additional Figures



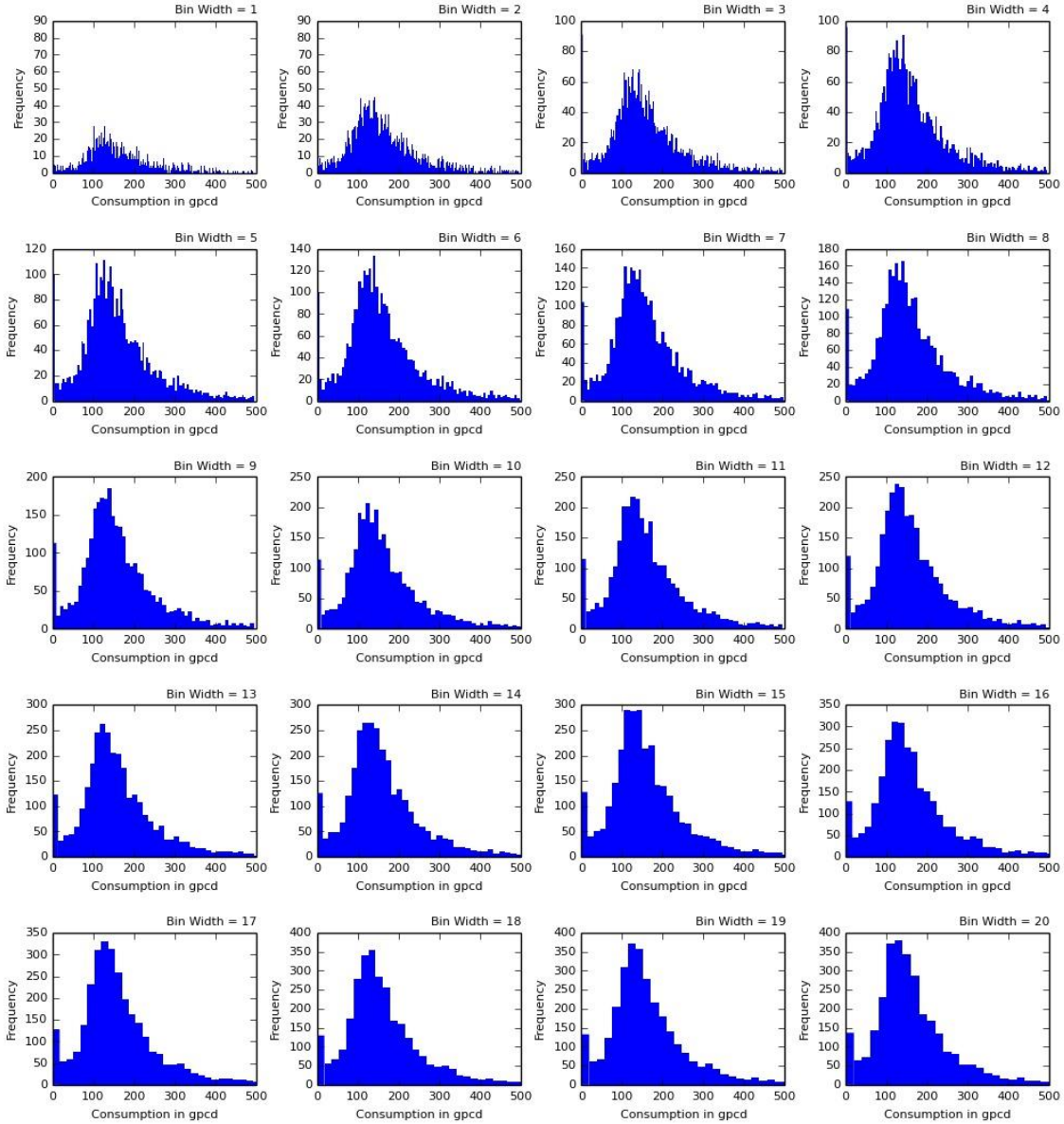

A1 Histogram of Distribution of Water Consumption in US Counties in 1985



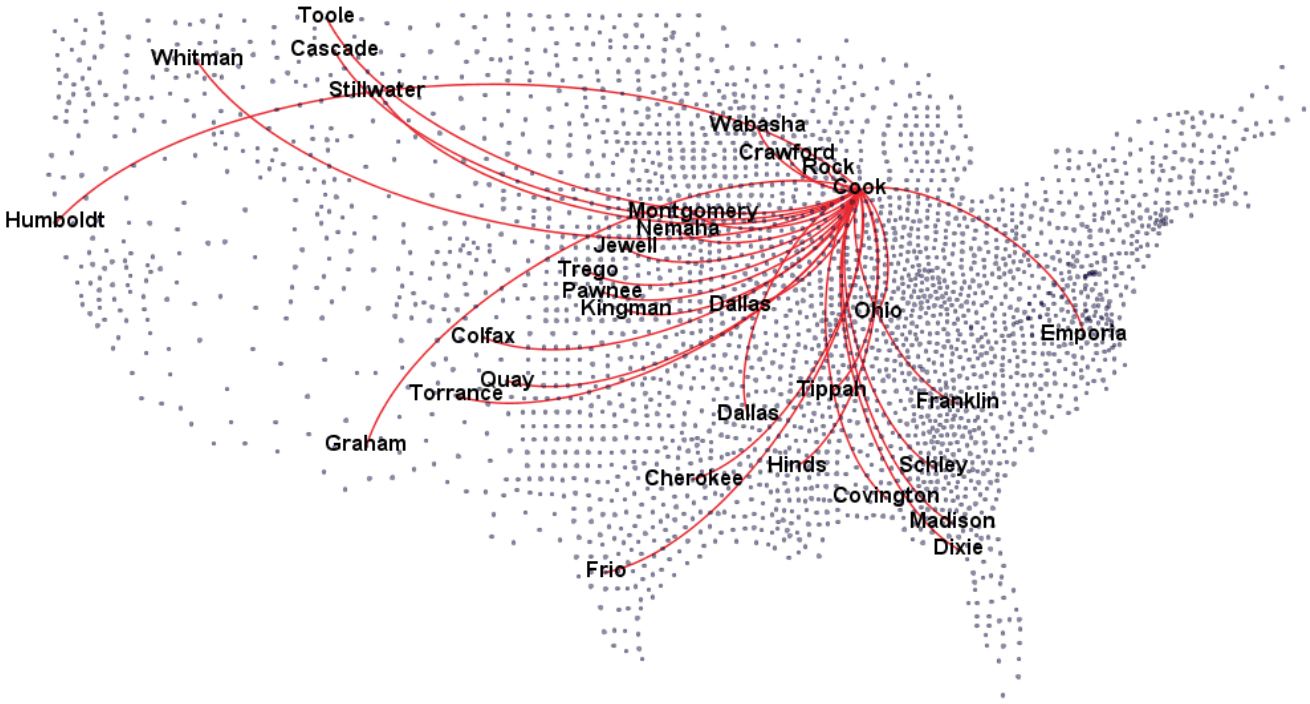

A2 Histogram of Distribution of Water Consumption in US Counties in 1990




A3 Histogram of Distribution of Water Consumption in US Counties in 1995




A4 Histogram of Distribution of Water Consumption in US Counties in 2005




A5 Counties whose Consumptions are Similar to Cook, Illinois in 2005

B. Python Code Used for Computation

```
from igraph import*
import csv
import time
from math import sin, cos, sqrt, atan2, radians
R = 6371.0
start=time.time()
File_Name=raw_input("Name of file :")
out=open(File_Name+".csv",'rb')
data=csv.reader(out)
geoid=[]
consumption=[]
latitude=[]
longitude=[]

for row in data:
    geoid.append(row[0])
    consumption.append(eval(row[1]))
    latitude.append(eval(row[2]))
    longitude.append(eval(row[3]))
out.close()
g=Graph()
g.add_vertices(len(geoid))
g.vs["geoid"]=geoid
g.vs["consumption"]=consumption
g.vs["lat"]=latitude
g.vs["long"]=longitude
#g.vs["label"]=geoid
g.es["weight"]=[]
```

```
summary(g)
a= float(input("enter the cutoff percentage : "))
b= a/100
i=0
k=0
for i in range (0,len(geoid)):
        for k in range(0,len(geoid)):


                if g.vs["consumption"][i]-
(g.vs["consumption"][i]*b)<=g.vs["consumption"][k]<=g.vs["consumption"][i]+(g.vs["consumption"][i]*
b) and i!=k and g.are_connected(i,k)==False:
                        g.add_edges((i,k))
                        lat1 = radians(g.vs["lat"][i])
                        lon1 = radians(g.vs["long"][i])
                        lat2 = radians(g.vs["lat"][k])
                        lon2 = radians(g.vs["long"][k])


                        dlon = lon2 - lon1
                        dlat = lat2 - lat1
                        d = (sin(dlat/2))**2 + cos(lat1) * cos(lat2) * (sin(dlon/2))**2
                        c = 2 * atan2(sqrt(d), sqrt(1-d))
                        distance = R * c
                        g[i,k]=distance
                        k+=1


        i+=1
        if i==int(len(geoid)*.10):
                print "10 percent completed..."
                continue
        if i==int(len(geoid)*.20):
                print "20 percent completed..."
```

```
                continue
        if i==int(len(geoid)*.30):
                print "30 percent completed..."
                continue
        if i==int(len(geoid)*.40):
                print "40 percent completed..."
                continue
        if i==int(len(geoid)*.50):
                print "50 percent completed..."
                continue
        if i==int(len(geoid)*.60):
                print "60 percent completed..."
                continue
        if i==int(len(geoid)*.70):
                print "70 percent completed..."
                continue
        if i==int(len(geoid)*.80):
                print "80 percent completed..."
                continue
        if i==int(len(geoid)*.90):
                print "90 percent completed..."
                continue
        if i==int(len(geoid)*.95):
                print "95 percent completed..."
                continue
#g.simplify(multiple=True,loops=True,combine_edges=None)
summary(g)
#print g.strength(2,mode=ALL,loops=False,weights="weight")
#print g.es["weight"]
c1=g.clusters(2)
```

```
c2=g.get_edgelist()
c3=c1.giant()
out=open(File_Name+"_{}".format(a)+".csv","wb")
new=csv.writer(out)
new.writerow(['Summary'])
new.writerow(['Vertices', g.vcount()])
new.writerow(['Edges', g.ecount()])
new.writerow(['Diameter',g.diameter()])
new.writerow(['Density',g.density()])
new.writerow(['Average path length',g.average_path_length()])
if g.is_weighted()== True:
        new.writerow(["Graph edges are Weighted"])
else:
        new.writerow(["Graph edges are not Weighted"])
new.writerow(["geoid","consumption","Degree","Cluster_id","strength"])
for i in range(0,len(geoid)):
        new.writerow([g.vs["geoid"][i],g.vs["consumption"][i],g.degree(i),c1.membership[i],g.strength(i,
mode=ALL,loops=False,weights="weight")])
new.writerow(["Cluster_id","Vertex_list","No of vertices"])
for i in range (0,len(c1)):
        new.writerow(["{}".format(i),c1[i],len(c1[i])])


new.writerow(['Summary_Giant'])
new.writerow(['Vertices', c3.vcount()])
new.writerow(['Edges', c3.ecount()])
new.writerow(['Diameter',c3.diameter()])
new.writerow(['Density',c3.density()])
new.writerow(['Average path length',c3.average_path_length()])
#new.writerow(['Maximum_consumption',max(c3.vs["consumption"])])
new.writerow(["Edge_list","source_node","Target_node","Type","ID","Label","Weight"])
```

```
for i in range (0,g.ecount()):
        new.writerow([c2[i],c2[i][0],c2[i][1],"Undirected","{}".format(i),"{}".format(i),g.es["weight"][i]
])
new.writerow(["Processing_time ",(time.time()-start)/60,"minutes"])
out.close()
# print g.es["weight"]
# if g.is_weighted()== True:
#         print "true"
# else:
#         print "false"
# print g.density()
virtual_style={}
virtual_style['vertex_size']=12
layout=g.layout("circle")
plot(g,File_Name+"_{}".format(a)+".png",layout=layout,**virtual_style)
# print "it took",(time.time()-start)/60,'minutes'
```

# VITA

**Name**

SK NASIR AHMAD

**Education**

Ph.D. Candidate, Civil and Materials Engineering Department, University of Illinois at Chicago (UIC), August 2012 to Current.

Master of Science in Civil Engineering, Civil and Materials Engineering Department, University of Illinois at Chicago (UIC), August 2012 to December 2014.

B.Sc. in Civil Engineering, Civil Engineering Department, Bangladesh University of Engineering and Technology (BUET), March 2004 to March 2009.

**Professional and Research Positions**

**University of Illinois at Chicago (UIC) Jan 2014 to Current**

Civil and Materials Engineering Department

Research Assistant

**University of Illinois at Chicago (UIC) Aug 2012 to Dec 2013**

Civil and Materials Engineering Department

Teaching Assistant

**Department of Public Health Engineering Jan 2012 to Jul 2012**

Bangladesh

Assistant Engineer

**Bangladesh House Building Finance Corporation Oct 2010 to Dec 2011**

Bangladesh

Assistant Engineer

**Publication**

Ahmad, N. & Derrible, S., 2014, “Evolution of Public Supply Water Withdrawal in the USA: A Network Approach”, in press.

**Conference Abstracts & Technical Reports**

Ahmad, N., & Derrible, S., 2014, "Evolution of Water Consumption in the USA: A Network Approach", Illinois Water 2014, October 14-15, Urbana, Illinois

Ahmad, N., Kermanshah, A., Peiravian, F., & Derrible, S., 2014, "Network Science: A Potential Tool for Analyzing Water Consumption in the USA", NetSci 2014 International School and Conference on Network Science, June 2-6, Berkeley, California

Kermanshah, A., Peiravian, F., Ahmad, N., & Derrible, S., 2014, "Investigating Transportation Network Resilience to Extreme Events", NetSci 2014 International School and Conference on Network Science, June 2-6, Berkeley, California

Peiravian, F., Ahmad, N., Kermanshah, A. & Derrible, S., 2014, "Questioning Box-Counting Method as a Tool for Fractal Characterization of Physical Networks", NetSci 2014 International School and Conference on Network Science, June 2-6, Berkeley, California

**Languages**

Bangla- Native

English- Fluent

**IT Skills**

Microsoft Windows, MS Office (Word, Excel, PowerPoint, Publisher),Python, ArcGIS, AutoCAD, Gephi, R.

**Fellowship**

**NTU Winter School Fellowship 2014**

Workshop on Introduction to Complexity and Complexity Science

Nanyang Technological University

Singapore

**Teaching**

CME 410 Design of Prestressed Concrete Structures

Fall 2012, 43 Students

Teaching Assistant

CME 403 Hydraulic Design

Spring 2013, 42 Students

Teaching Assistant

CME 396 Senior Design I

Fall 2013, 58 Students

Teaching Assistant

**Membership**

Member, Complex and Sustainable Urban Networks (CSUN)

Member , Sustainable Remediation forum (SURF)