

Opportunities for Using the Industrial Assessment Center Database for Industrial Water Use Analysis



Kristina O. Armstrong
Susana Garcia
Sachin Nimbalkar

October 30, 2020

**Approved for public release.
Distribution is unlimited.**

DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via US Department of Energy (DOE) SciTech Connect.

Website www.osti.gov

Reports produced before January 1, 1996, may be purchased by members of the public from the following source:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone 703-605-6000 (1-800-553-6847)
TDD 703-487-4639
Fax 703-605-6900
E-mail info@ntis.gov
Website <http://classic.ntis.gov/>

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange representatives, and International Nuclear Information System representatives from the following source:

Office of Scientific and Technical Information
PO Box 62
Oak Ridge, TN 37831
Telephone 865-576-8401
Fax 865-576-5728
E-mail reports@osti.gov
Website <http://www.osti.gov/contact.html>

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Energy and Transportation Science Division

**Opportunities for Using the Industrial Assessment Center Database for Industrial Water
Use Analysis**

Kristina Armstrong (Oak Ridge National Laboratory)
Susana Garcia (Oak Ridge National Laboratory)
Sachin Nimbalkar (Oak Ridge National Laboratory)

Date Published: October 30, 2020

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES	v
LIST OF TABLES	vi
ACRONYMS	vii
ACKNOWLEDGMENTS	ix
1. INTRODUCTION	1
2. IAC DATA SET	2
2.1 PREPARATION OF THE DATA SET	5
2.2 STATISTICAL DESCRIPTION OF THE DATA SET	5
2.3 CLEANING THE DATA	7
2.3.1 Median Absolute Deviation	9
2.3.2 Unit Cost of Water	11
3. REGRESSION ANALYSIS	12
3.1 VARIABLE SELECTION	12
3.2 INITIAL CORRELATION ANALYSIS	13
3.3 STEPWISE REGRESSIONS	14
3.3.1 Covariance	14
4. RESULTS	16
4.1 WATER USE	16
4.2 WATER COST	17
5. DISCUSSION	18
5.1 SUBSECTOR VARIATION	18
5.2 MEAN VS. MEDIAN	19
5.3 UNIT COST OF WATER	19
5.4 NEGATIVE COEFFICIENTS	19
5.5 DATA ENTRY ERRORS	20
5.6 INDUSTRIAL WATER SOURCES	20
5.7 FUTURE WORK	20
6. CONCLUSIONS	22
7. REFERENCES	23
APPENDIX A. WATER USE AND COST TABLES	A-1

LIST OF FIGURES

Figure 1. Map of IAC Assessments in the contiguous United States.	2
Figure 2. Number of IAC assessments conducted each year.	3
Figure 3. Number of IAC assessments conducted in each subsector.	3
Figure 4. Map of IAC assessments with water data.	4
Figure 5. Number of IAC assessments conducted each year with water data.	4
Figure 6. Number of IAC assessments conducted in each subsector with water data.	5
Figure 7. High spread of water use (left) and cost (right) reported.	6
Figure 8. Water use box plots (left) before and (right) after data cleaning.	8
Figure 9. Water cost box plots (left) before and (right) after data cleaning.	8
Figure 10. Water use box plots (left) before and (right) after data cleaning with MAD.	10
Figure 11. Water cost box plots (left) before and (right) after data cleaning with MAD.	10
Figure 12. Skewed but reduced spread of the post-MAD analysis data set.	10
Figure 13. Water stress averaged over each state.	13
Figure 14. Correlations between water use and individual variables.	13
Figure 15. Scatter plots water use vs. the chosen variables for the printing subsector.	14
Figure 16. Plots illustrating the difference between the means of the utilities and food subsectors from other subsectors.	18
Figure 17. Variable means by subsector with shading for the number of other subsectors with significantly different means.	19
Figure 18. Scatter plot of water cost vs. employees for the paper sector.	20

LIST OF TABLES

Table 1. Top 10 IAC assessments in terms of water use and water cost.	6
Table 2. Subsector descriptive statistics.	7
Table 3. Loss of data with cleaning methods.	9
Table 4. IAC database variables	12
Table 5. Water use: model statistics.....	16
Table 6. Water cost: model statistics	17
Table A. Water use: model statistics and coefficients.	A-3
Table B. Water use: VIFs.....	A-4
Table C. Water use: variable p -values.	A-5
Table D. Water cost: model statistics and coefficients.	A-6
Table E. Water cost: VIFs.....	A-7
Table F. Water costs: variable p -values.	A-8

ACRONYMS

DOE	U.S. Department of Energy
IAC	U.S. Department of Energy Industrial Assessment Center
MAD	Median Absolute Deviation
NAICS	North American Industry Classification System
ORNL	Oak Ridge National Laboratory
SIC	Standard Industrial Classification
STATCAN	Statistics Canada
VIF	Variance Inflation Factor

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support and guidance of Joseph Cresko at the U.S. Department of Energy's Advanced Manufacturing Office. The authors also thank Sujit Das, Mini Malhotra, Paulomi Nandy, Kiran Thirumaran, Jennifer Travis, and Thomas Wenning, who assisted in reviewing this report. The authors also thank Prakash Rao and William Morrow of Lawrence Berkeley National Laboratory; Alberta Carpenter and James McCall of National Renewable Energy Laboratory; Sarang Supekar and Nwike Iloeje of Argonne National Laboratory, who reviewed a draft of this report and provided valuable comments. The authors also greatly appreciate the support and participation of their industry partners.

1. INTRODUCTION

The manufacturing sector accounted for approximately 5–6% of total U.S. water use in 2015. Of that amount, 75–80% is self-supplied withdrawal from surface-water and groundwater sources and the remainder is from public water supplies [1], [2]. The cost of source water is a relatively small fraction of total production costs, even in water-intensive industries. Although manufacturing facilities commonly locate in water-scarce areas [3], water scarcity still poses a great risk to the manufacturing sector. Reliable water is necessary for any facility that relies on it for process and comfort cooling, cleaning, employee use, and steam generation. Water is also critical for the food and beverage, refined petroleum, lumber and wood products, paper, chemicals, and electronic equipment industries, which have the highest freshwater intake in the United States [3].

With several acknowledged threats to water security (e.g., higher water intake and disposal costs, changing water and pollution regulations [4]), many barriers to better industrial water efficiency still exist. One of these barriers is the lack of reliable data on overall U.S. industrial water use—how it is used and the quantities required for each sector. If a facility cannot be easily compared with a facility of similar size and sector, knowing if it is effectively using water conservation best practices is difficult.

The last federal government-sponsored U.S. water-use survey was conducted in 1983 [5]–[7], forcing one industry-heavy state, Texas, to mandate its own water survey, with penalties for failure to respond [8], [9]. Conversely, Canada's data collection agency Statistics Canada (STATCAN) collects industrial water use data every two years (since 2005). Using questionnaires developed with data users and respondents, STATCAN collects data on the water intake (including information on the source, purpose, treatment, and possible recirculation of this water), as well as the water discharge and its treatment. Sampling depends on the sector; all facilities are surveyed for some sectors (excluding the lower 5% by size), whereas only a sample is conducted for others. If a survey is received, responding is mandatory [10]. Although these data are available to users and facilities in the United States, they must be adjusted for U.S. weather and economic conditions. These differences make using as a proxy for U.S. water data difficult, leaving researchers looking for other data sources.

One potential source of industrial water use data is the U.S. Department of Energy (DOE)-sponsored Industrial Assessment Centers (IACs). IACs are university-based organizations that provide free audits to small- and medium-sized manufacturing facilities to identify productivity improvement and waste and energy reduction opportunities. To receive an IAC assessment, facilities must have an annual energy spend between \$100K and \$2.5M, gross annual sales less than \$100M, fewer than 500 employees, and be near a facility (150–180 miles, depending on the IAC), although exceptions can be made. The IACs also maintain a database of all the audits conducted, which currently holds more than 19,267 assessments and 145,000 recommendations (as of July 24, 2020) [11]. This database also contains energy utility (electricity, natural gas, and other fuels) and water utility data, making it a potential data source for industrial water use.

This report attempts to create regression models to predict a small- or medium-sized industrial facility's annual water use or cost based on its industrial subsector and several possible relevant variables. Using data collected by IAC assessments, models for several industrial subsectors were generated via stepwise regression techniques to determine which variables (annual sales, number of employees, facility/plant area, annual production hours, and a water stress metric) are relevant. These data were first cleaned of outliers using standard techniques and collinearity among the variables investigated.

2. IAC DATA SET

The IAC database contains data from 40 years of assessments, and at the time the data set was pulled for analysis, it held 19,085 assessments in the database covering North American Industry Classification System (NAICS) codes between 111 and 812 [12], with most being in manufacturing (311–339). For the 10,970 assessments that had no NAICS code reported (conducted before NAICS codes were established), the NAICS code was determined from the reported Standard Industrial Classification (SIC) code via a cross walk. Regarding SIC codes with multiple matching NAICS codes, for this analysis, the lowest NAICS for the SIC was used [13] because all assessments after 2007 have NAICS codes, and this did not affect the water analysis.

Figure 1 shows the distribution of IAC assessments across the United States, with the darker green indicating more assessments conducted in that state. Figures 2 and 3 show the distribution of assessments across time and industrial subsectors (via NAICS code), respectively.

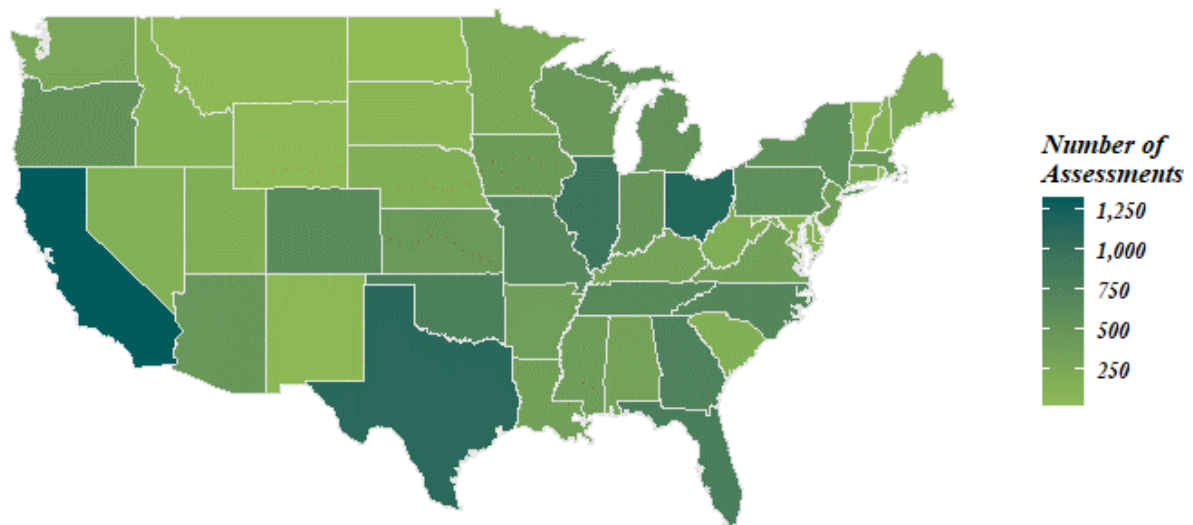


Figure 1. Map of IAC Assessments in the contiguous United States.

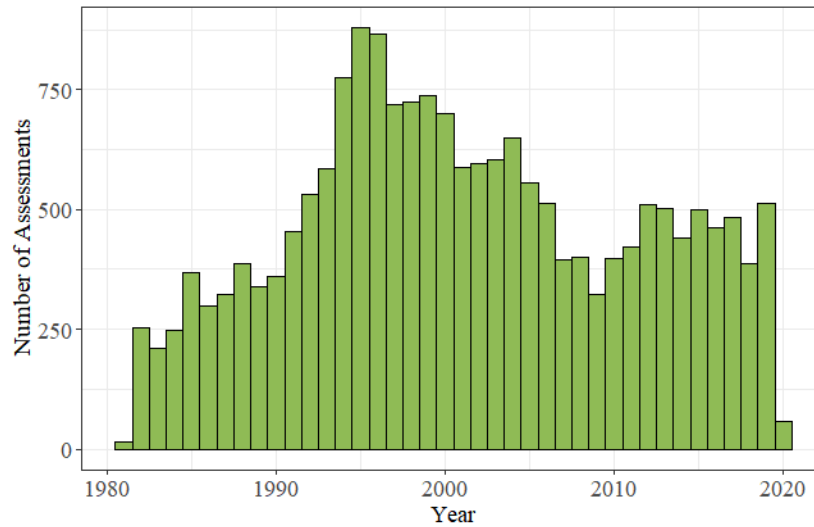


Figure 2. Number of IAC assessments conducted each year.

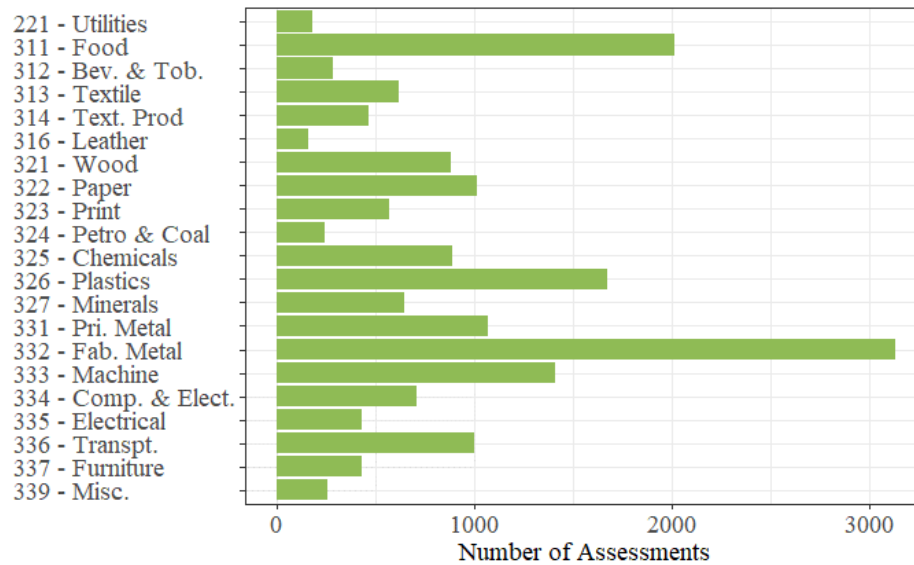


Figure 3. Number of IAC assessments conducted in each subsector.

Despite this abundance of data, water data have only been part of the requested utility data for the past 4 years, with only 644 assessments having any water-use data. Figures 4 through 6 show the distributions of assessments across geographic location, time, and subsector, respectively.

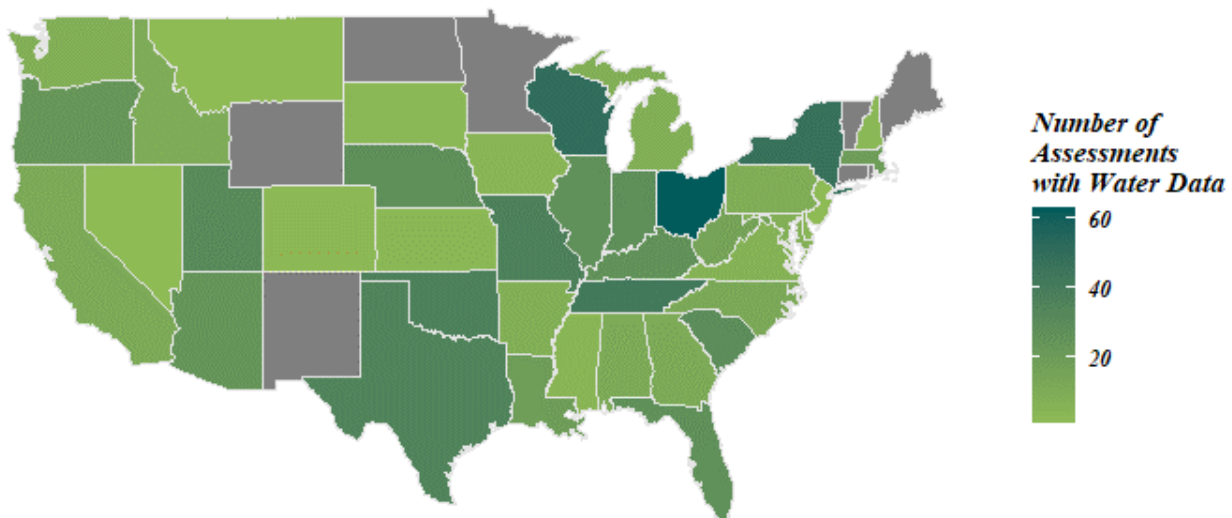


Figure 4. Map of IAC assessments with water data.

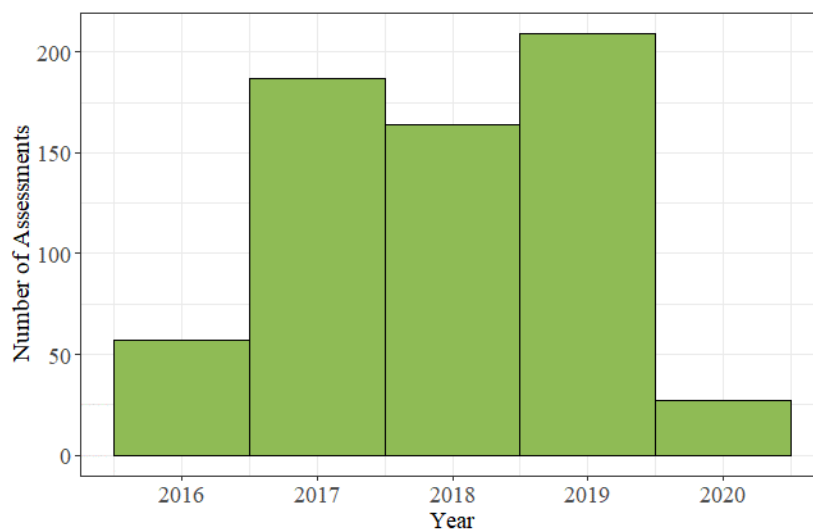


Figure 5. Number of IAC assessments conducted each year with water data.

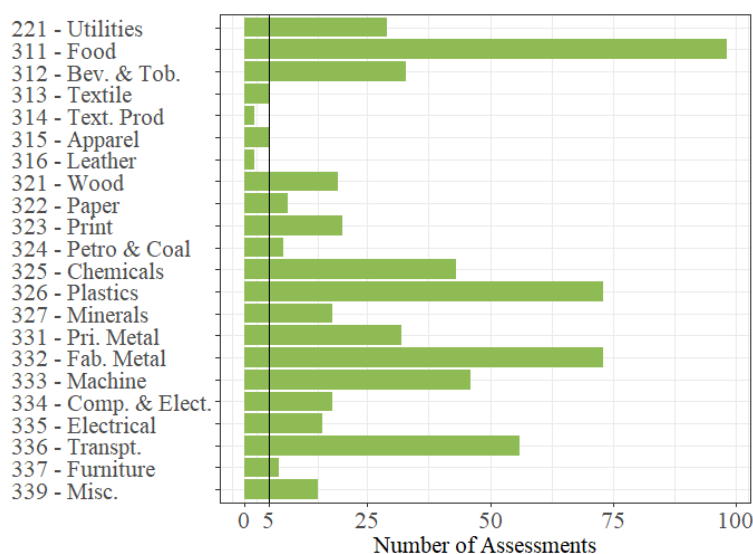


Figure 6. Number of IAC assessments conducted in each subsector with water data.

When beginning an assessment, IACs collect data on the facility demographics (e.g., industrial subsector, location), size of the facility (e.g., square footage, sales, production, number of employees), and utility bills (e.g., electricity, fuels, water, waste streams) [14]. The specific variables collected by the IAC teams are describe in more detail in Section 3, below, (see Table 4)

2.1 PREPARATION OF THE DATA SET

The following steps were taken to prepare the IAC data set for the water use analysis conducted for this report.

- Removed all data without water use or cost observations
- Removed data from all nonmanufacturing NAICS codes (111 and those greater than 400)
- Removed all observations part of a NAICS group with five or fewer assessments (i.e., Textile, Textile Products, Apparel, and Leather)

These steps left 613 assessments for 18 NAICS codes. Figure 6 shows the distribution of assessments across subsectors.

2.2 STATISTICAL DESCRIPTION OF THE DATA SET

The IAC water data are very skewed; a few observations with very high water use or cost dominate the spread data set (Figure 7). This imbalance can also be seen from the substantial discrepancy between the mean and median for each NAICS code, indicating that the data are strongly skewed. This skewing is beyond realistic expectations, making it likely that the very high water use and cost values are not just outliers, but inaccurately entered data. This skewing is further seen from the top 10 highest water use and cost entries (Table 1). The IAC assessments with the highest water use are not the same assessments that have the highest annual water cost. While decreasing block rates could be partially responsible for this, because most municipal water costs range from \$1 to \$8 per kgal [15], these are likely data entry errors, especially for the water use data. Another possible explanation for this discrepancy is that the very large water users reported all water use, including self-supplied. More investigation into the data collection process would be necessary to better understand if this was the case.

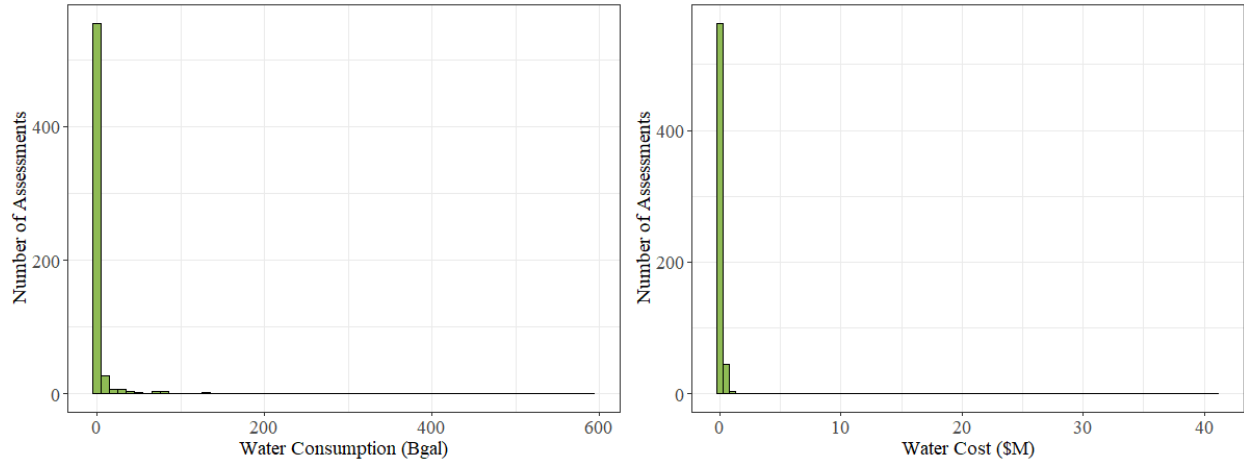


Figure 7. High spread of water use (left) and cost (right) reported.

Table 1. Top 10 IAC assessments in terms of water use and water cost.

Water use				Water cost			
ID	Use (Mgal)	Cost (\$k)	Unit cost (\$/kgal)	ID	Use (Mgal)	Cost (\$k)	Unit cost (\$/kgal)
WM0099	585,200	\$3	\$0.000005	GT0899	13,315	\$41,143	\$3.09
TT0214	326,952	\$683	\$0.002089	UU0163	6,054	\$10,103	\$1.67
GT0908	284,558	\$0.001	\$0.000000	UU0148	11,065	\$1,605	\$0.15
IP0087	129,828	\$153	\$0.001181	UD0966	471	\$1,547	\$3.28
SF0516	127,511	\$319	\$0.002500	SU0464	259	\$949	\$3.66
OK0942	89,662	\$505	\$0.005628	SD0550	350	\$935	\$2.67
MZ0268	84,236	\$451	\$0.005354	AM0782	108	\$794	\$7.36
CL1806	77,900	\$235	\$0.003010	NC0596	223	\$781	\$3.50
MA0778	76,309	\$141	\$0.001854	UA0188	24,013	\$733	\$0.031
IC0230	69,348	\$154	\$0.002215	TT0212	327	\$731	\$2.24

Additionally, the data set has a wide range and high standard deviation for water use and cost: all subsectors had standard deviations greater than their mean (see Table 2). Statistically, this means that there is a 95% chance that the true mean is less than zero water use or water cost for several of these industries (13 subsectors for water use and 11 subsectors for water cost); practically, this provides evidences that the data set may not be useful for this exercise.

Food manufacturing has the highest mean water use and cost, with a water use standard deviation almost five times the size of the water use mean and a water cost deviation more than seven times the mean.

Table 2. Subsector descriptive statistics.

NAICS code	Water use (Mgal)			Water cost (\$K)		
	Mean	Std. dev.	Median	Mean	Std. dev.	Median
221 Utilities	4,674	10,816	35.0	64.6	85.9	35.0
311 Food Manufacturing	14,637	67,975	33.8	550.8	4,146.2	55.4
312 Beverage & Tobacco Manuf.	26.8	41.3	6.73	104.6	164.3	36.7
321 Wood Products Manuf.	1,238	3,825	6.00	25.7	38.3	7.24
322 Paper Manuf.	227	485	4.05	16.3	27.9	7.86
323 Printing & Related Support Activities	1,288	3,011	1.79	11.8	22.7	5.16
324 Petroleum & Coal Products Manuf.	2,805	7,920	7.06	28.9	30.2	25.3
325 Chemical Manuf.	8,882	21,541	18.7	106.9	260.0	16.5
326 Plastics & Rubber Products Manuf.	5,460	38,343	3.02	28.9	84.3	7.04
327 Nonmetallic Minerals Product Manuf.	1,833	4,216	4.24	41.8	65.2	14.6
331 Primary Metals Manuf.	1,169	4,258	18.2	119.9	222.4	17.6
332 Fabricated Metal Products Manuf.	1,465	8,312	3.38	45.7	97.2	8.28
333 Machinery Manuf.	882	2,351	5.63	38.6	116.9	4.82
334 Computer & Electronic Product Manuf.	5,756	21,069	7.26	91.3	172.5	9.79
335 Electrical Equipment Manuf.	1,735	6,640	8.00	29.7	32.2	22.6
336 Transportation Equipment Manuf.	2,158	9,635	7.52	60.8	136.9	14.5
337 Furniture & Related Product Manuf.	1,999	4,970	4.07	25.9	25.0	16.6
339 Misc. Manufacturing	1,303	3,464	12.6	39.7	56.0	8.99

2.3 CLEANING THE DATA

To reduce the spread of the data, all data points at which water usage (20 data points) or water cost (12 data points) were greater than three standard deviations were removed from their respective data set (see Figures 8 and 9).

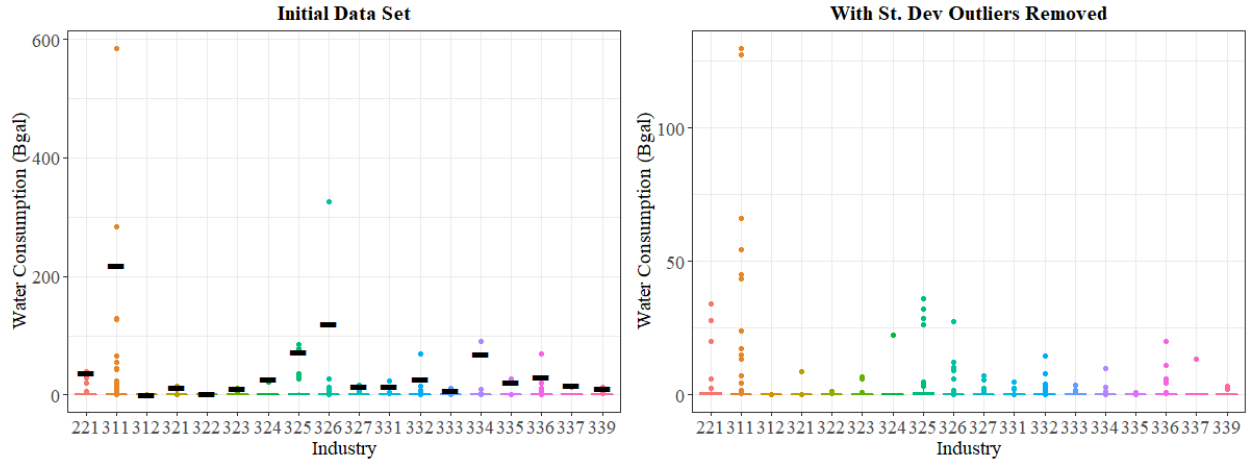


Figure 8. Water use box plots (left) before and (right) after data cleaning. The black bars represent the outlier cutoff points.

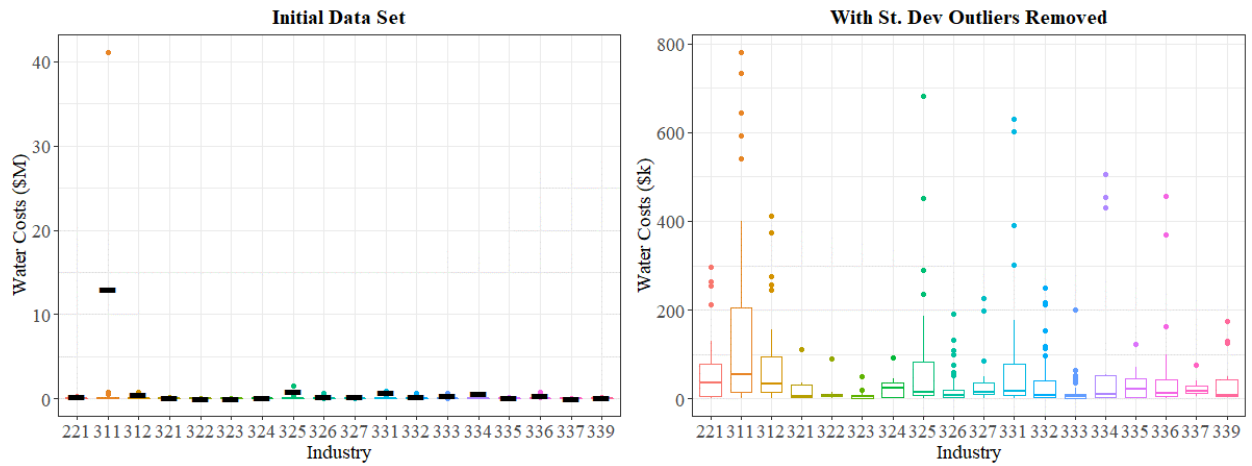


Figure 9. Water cost box plots (left) before and (right) after data cleaning. The black bars represent the outlier cutoff points.

Even after data cleaning, a high skew level still remained; therefore, other methods to clean the data were attempted. Table 3 shows the effect of the different data cleaning methods on the sample size.

Table 3. Loss of data with cleaning methods. MAD = median absolute deviation.

NAICS code		Original data set	Removal via std. dev.		Removal via MAD		Removal via water unit cost
			Use	Cost	Use	Cost	
221	Utilities	29	28	29	17	25	10
311	Food	98	96	97	74	82	52
312	Bev. & Tobacco	33	32	32	25	27	24
321	Wood	19	18	18	16	17	13
322	Paper	9	9	9	7	8	4
323	Print	20	19	19	14	18	9
324	Petro & Coal	8	8	8	7	8	7
325	Chemicals	43	40	42	25	33	22
326	Plastics	73	72	72	52	60	49
327	Minerals	18	17	18	13	15	11
331	Pri. Metal	32	31	31	24	24	20
332	Fab. Metal	73	72	72	53	55	44
333	Machine	46	43	44	25	38	19
334	Comp. & Elect.	18	17	18	12	13	10
335	Electrical	16	15	16	13	16	8
336	Transpt.	56	55	54	36	48	32
337	Furniture	7	7	7	4	7	5
339	Misc.	15	14	15	9	11	8
TOTAL	Total	613	593	601	426	505	347
	Total Removed		20	12	187	108	266

2.3.1 Median Absolute Deviation

Because outlier identification via mean and standard deviation is not the best method for correcting skewed data (it assumes a symmetric distribution, is highly influenced by sample size, and considers outliers to be part of the normal distribution), an alternative approach was explored. Using the median absolute deviation (MAD) approach [16], outliers were identified by having a high difference between the value (water use or cost) and the subsector median. The criteria for rejecting an observation are based on the median of the difference of the value and median:

$$|x_i - M_{NAICS,i}| > T * M_{|x_i - M_{NAICS,i}|} * b \quad (1)$$

where x_i is the value for the observation, $M_{NAICS,i}$ is the median of those values, T is the stringency criteria (similar to the number of standard deviations from the mean that defines outliers; in this case, 3), $M_{|x_i - M_{NAICS,i}|}$ is the median of the differences between the observations and medians, and b is a constant linked to the normalcy of the data (1.4826). All observations where Eq. 1 was true were removed from their respective data set (Figures 10, 11, and 12). This method does assume that the true data (once the outliers are removed) fit a normal, symmetric distribution, and are not actually a skewed distribution.

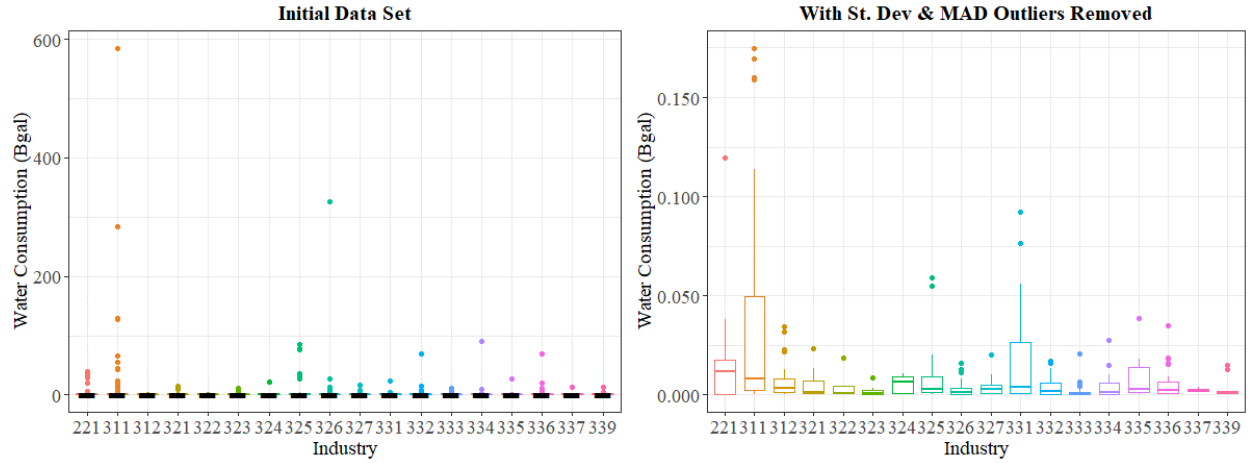


Figure 10. Water use box plots (left) before and (right) after data cleaning with MAD.

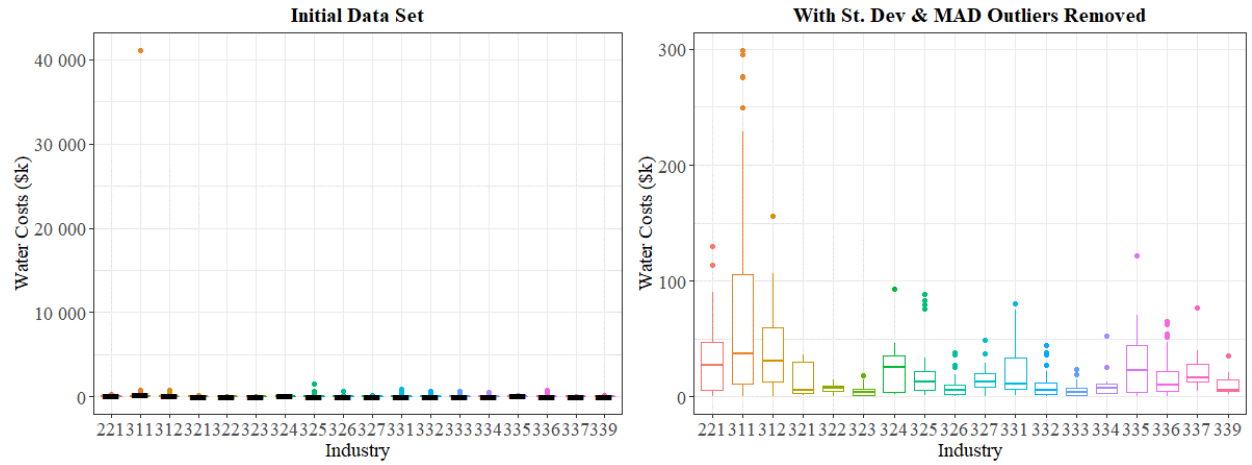


Figure 11. Water cost box plots (left) before and (right) after data cleaning with MAD.

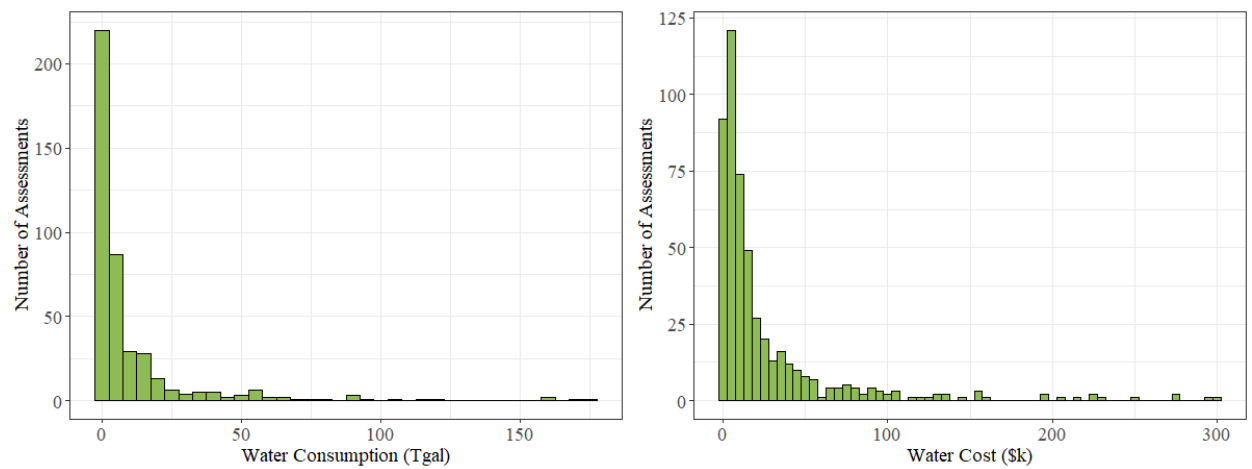


Figure 12. Skewed but reduced spread of the post-MAD analysis data set.

This method removes all 20 use and 12 cost data points removed by the standard deviation approach, plus an additional 176 use and 96 cost data points. The resulting box plots show the reduced spread of data points, and the histograms show a much smoother, though still skewed, data set.

2.3.2 Unit Cost of Water

Calculating the unit cost of water was another attempt to remove potentially bad data from the data set. The cost per 1,000 gal was calculated for each observation, and any observation where this was less than \$1 or greater than \$10 was removed [15].

However, this resulted in a very small sample size, with a still large range of values, so it was also abandoned. Without a reliable method to determine which of the values (use or cost) was the cause of the unrealistic water cost, all values not within the range were removed for both analyses. If a second-step criteria were made to determine which value was inaccurate, alternatives to discarding the entire observation would be available (e.g., replacing the inaccurate value with the mean or median of the subsector, only removing the inaccurate value and leaving the other). This could potentially keep the sample at a reasonable size.

3. REGRESSION ANALYSIS

Using the data cleaned by the MAD technique, the IAC data were explored for suitability in a water use or annual cost prediction model. As stated previously, in addition to water and energy use/cost data, the IACs collect data on sales, facility size, production, operation hours, and more, as shown in Table 4. These variables were qualitatively and quantitatively explored for use in prediction models.

Table 4. IAC database variables

Variable	Used in analysis
Location—U.S. state	Adapted
Facility size	
Total yearly sales (\$)	Yes
Total site employees	Yes
Total plant area (ft ²)	Yes
Total yearly production	
Quantity	No
Types of products (text description)	No
Production quantity units	No
Total yearly hours of operation (hr/yr)	Yes
Total yearly electricity	
Total yearly electricity consumption (kWh) and costs (\$)	Considered
Total yearly electricity demand (kW-month-year) and charges (\$)	No
Total yearly electricity fees	No
Total yearly natural gas consumption (MMBTU) and costs (\$)	Considered
Total yearly other fuel consumption and costs	
Liquified petroleum gas; #1, #2, #4, #6 fuel oil; coal; wood; paper; other gas; other	No
Total yearly water use (kgal) and costs (\$)	Yes
Total yearly water disposal (gal) and costs (\$)	No
Total yearly other disposal (gal) and costs (\$)	
Nonhazardous/hazardous liquids, solids, gases	No

3.1 VARIABLE SELECTION

Initially, many of these variables were considered for use in the analysis as likely to affect water use: other utility use, number of employees, plant area, production hours, production quantity, sales, and location. Several of these factors were quickly eliminated. Electricity and natural gas use were initially considered, but these variables were either not likely to be strongly correlated or would be confounding or colinear with the other variables and were eliminated from the analysis. Production was abandoned early in the analysis because no consistent scale for production (between and within subsectors) exists. Production was reported as production widgets, mass, or other classifications with the unit of measure as a separate reported variable.

As stated previously, location was not used directly but instead was used to determine a location-based water stress score. Using state-level water withdraw and supply estimates from Sandia National

Laboratories [17]–[19], a water stress metric was approximated (loosely based on the standard Water Supply Stress Index) using Eq. 2.

$$Stress_{approx} = \frac{State\ Water\ Withdraw}{State\ Water\ Supply} \quad (2)$$

The result for each state was then matched to each entry’s location, and this was used for the analysis (Figure 13) [20]. The data set has withdraw and supply data at the watershed level (HUC8), and this variable would be more accurate if the IAC location data were at the same level.

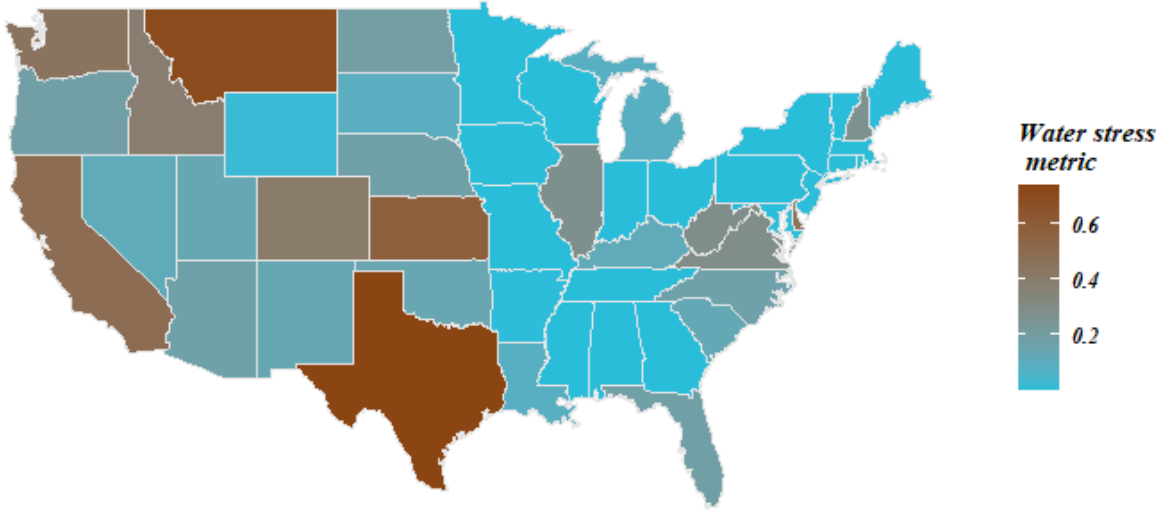


Figure 13. Water stress averaged over each state.

3.2 INITIAL CORRELATION ANALYSIS

As an initial exploration of the data, a simple linear regression was run between water use and each of the remaining variables. By themselves, the single variables rarely had a high adjusted R^2 (greater than 0.5), but most industries had at least one variable with a significant effect on water use. Figure 14 shows the R^2 for any variable that was significant (with a p -value greater than 0.1), and the shading of the cells corresponds to the size of the R^2 value for that correlation (also displayed as text in the cells). Figure 15 shows an example of scatter plots between water use and the variables for a single subsector (323—printing and related activities) where the two significant correlations (Sales and Plant Area) show a slight linear relationship.

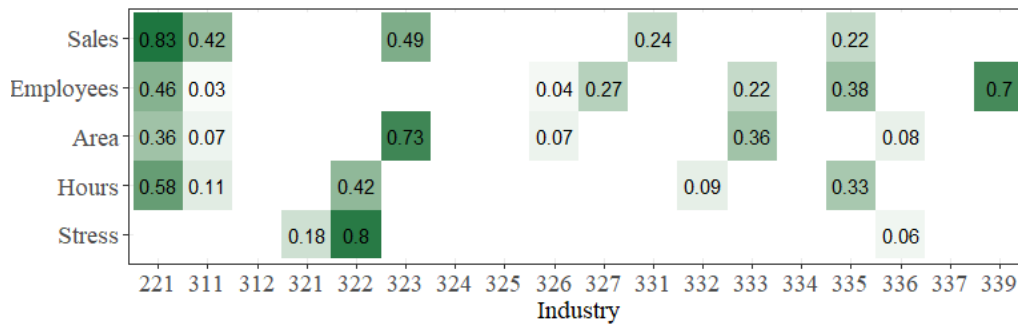


Figure 14. Correlations between water use and individual variables.

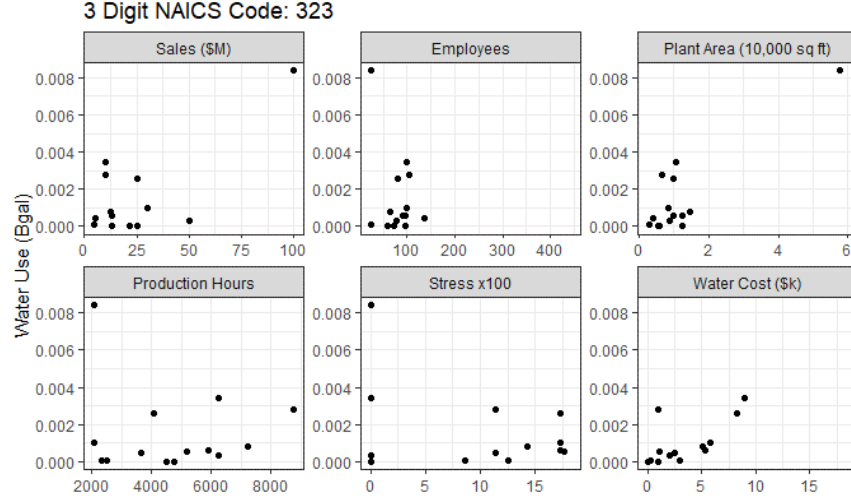


Figure 15. Scatter plots water use vs. the chosen variables for the printing subsector.

3.3 STEPWISE REGRESSIONS

To determine which variables have an effect on water use and cost, two series of stepwise regressions were conducted for each NAICS code. Water use and cost were the dependent variables and sales, employees, area, hours, and water stress were the predictor variables. Stepwise regression is a technique in which regression analyses are conducted for every combination of predictors and dependent variables. For this analysis, only first-order and no interaction variables were used. The model with the highest coefficient of determination (adjusted R^2) that met the acceptance criteria was chosen as the final model.

The acceptance criteria were modeled after the DOE's Superior Energy Performance energy regression acceptance criteria [21]:

- Model p -value < 0.1 ,
- Model adjusted $R^2 > 0.5$, and
- All variable p -values < 0.2 , at least one < 0.1 .

3.3.1 Covariance

Because of the potentially interrelated nature of some of the predictor variables, after the acceptable models were identified, the variance inflation factor (VIF) was used to test for covariance. This method tests the compounding of variance from possibly related variables. The VIF of a predictor variable is determined by conducting a regression with that variable as the dependent variable and the other predictor variables as the predictors. The VIF is calculated using the R^2 value of the regression (Eq. 3). This can be interpreted as the increase in standard error of a model, compared with if the variable was uncorrelated with the other predictor variables.

$$VIF = \frac{1}{1 - R^2} \quad (3)$$

If the VIF is 1, the variables are not collinear, and if the VIF is less than 4 or 5, the collinearity is small and can be ignored. The model's variables should be reconsidered if the VIF is greater than ten. Including variables with VIFs between four and ten is a matter of personal preference and analyst confidence in the model. For the water use models, the initial models for paper manufacturing (322) and petroleum and

coal products (324) had VIFs well above ten for a pair of variable (employees and area, area and stress); however, these models also had unacceptable variable p -values. The new model for paper did not have any problematic VIFs, and no acceptable model was found for petroleum/coal. For the water cost models, again, Plant Area and Stress for petroleum and coal products (324) had high VIF scores, so the next model with the highest R^2 and without one of the two variables was picked as a replacement. Table B and Table E in the appendix show the final models and their VIF scores.

As another test of covariance, the regressions were conducted with mean-centered data. The coefficients did not change, indicating that collinearity is not likely a problem with the final models.

4. RESULTS

4.1 WATER USE

Table 5 shows the results of the regression with water use as the dependent variable. The “Status” column indicates if the models pass or fail all the acceptance criteria. Models in this column marked with “Fail*” indicate if the model has a model p -value lower than 0.1, but does not explain enough of the variance (e.g., has an adjusted R^2 less than 0.5) to fully pass the acceptance criteria. Although these models are not useful for prediction, they do indicate that the variables present in the model have significant relationships to each other and water use or cost. However, other variables affecting water use or cost still need to be accounted for, or else the data have higher inherent unexplained variabilities.

Table 5. Water use: model statistics.

NAICS code	Number of assessments	Status	Model p -value	Adj. R^2
221 Utilities	17	Pass	0.0000	0.87
311 Food Manufacturing	74	Fail*	0.0000	0.45
312 Beverage & Tobacco Manuf.	25	Fail	0.2913	0.02
321 Wood Products	16	Fail*	0.0553	0.18
322 Paper	7	Pass	0.0103	0.98
323 Printing & Related	14	Pass	0.0001	0.73
324 Petroleum & Coal Products	7	Fail	0.2538	0.59
325 Chemical Manuf.	25	Fail	0.1582	0.08
326 Plastics & Rubber Products	52	Fail*	0.0316	0.07
327 Nonmetallic Minerals Prod.	13	Fail*	0.0381	0.27
331 Primary Metals Manuf.	24	Fail*	0.0097	0.34
332 Fabricated Metal Products	53	Fail*	0.0141	0.09
333 Machinery Manuf.	25	Fail*	0.0010	0.36
334 Computer & Electronic Prod.	12	Fail	0.1808	0.23
335 Electrical Equipment	13	Fail*	0.0198	0.45
336 Transp. Equipment	36	Fail*	0.0489	0.08
337 Furniture & Related Prod.	4	Pass	0.0953	0.97
339 Misc. Manuf.	9	Pass	0.0052	0.92

* Indicates a model with significant p -values, but too low of an Adjusted R^2 to be considered valid.

Only utilities (221), printing and related activities (323), furniture and related products (337), and miscellaneous manufacturing (339) had models that passed all the criteria—four of the 18 industry subsectors with enough data points for the analysis. Food (311), wood products (321), paper (322), plastics and rubber products (326), nonmetallic minerals products (327), primary metals (331), fabricated metal products (332), machinery (333), electrical equipment (335), and transportation equipment (336) manufacturing all had significant models but with too low of an adjusted R^2 . This suggests that the identified variables significantly impact water use but that additional, unaccounted for variables have more of the variation or too much noise in the data to account for all the variability. Identifying and including more variables may allow for better predictive models in these subsectors. Table A through Table C in the appendix describe the model statistics and variable coefficients, the VIF values, and the coefficient p -values for each model.

4.2 WATER COST

Table 6 shows the results of the regression with water cost as the dependent variable. Again, the “Status” column indicates if a model passed, was significant but had too low of adjusted R^2 , or failed to pass the criteria.

For water cost, only two of the 18 subsectors generated valid models — printing and related activities (322) and petroleum and coal products (324). Most of the subsectors (12) — utilities (211), food (311), beverage and tobacco (312), paper (322), chemical (325), plastics and rubber (326), primary metals (331) fabricated metals (332), machinery (333), electrical equipment (335), transportation equipment (336), and miscellaneous (339) manufacturing—had significant models but with too low of an adjusted R^2 to be considered valid. Therefore, the relationship between the variables and water cost is significant, but it does not explain enough of the variance to be a useful predictive model. Again, identifying and including additional variables may allow for the identification of predictive models for these subsectors. Table D through Table F in the appendix describe the model statistics and variable coefficients, the VIF values, and the coefficient p -values for each model.

Table 6. Water cost: model statistics

NAICS code	Number of assessments	Status	Model p -value	Adj. R^2
221 Utilities	25	Fail*	0.00	0.18
311 Food Manufacturing	82	Fail*	0.05	0.11
312 Beverage & Tobacco Manuf.	27	Fail	0.10	0.22
321 Wood Products	17	Fail*	0.09	0.47
322 Paper	8	Pass	0.00	0.57
323 Printing & Related	18	Pass	0.00	0.74
324 Petroleum & Coal Products	8	Fail*	0.04	0.10
325 Chemical Manuf.	33	Fail*	0.02	0.09
326 Plastics & Rubber Products	60	Fail	0.66	-0.06
327 Nonmetallic Minerals Prod.	15	Fail*	0.00	0.48
331 Primary Metals Manuf.	24	Fail*	0.00	0.28
332 Fabricated Metal Products	55	Fail*	0.02	0.12
333 Machinery Manufacturing	38	Fail	0.26	0.09
334 Computer & Electronic Prod.	13	Fail*	0.04	0.23
335 Electrical Equipment	16	Fail*	0.00	0.20
336 Transp. Equipment	48	Fail	0.30	0.05
337 Furniture & Related Prod.	7	Fail*	0.02	0.43
339 Misc. Manufacturing	11	Fail*	0.00	0.00

* Indicates a model with significant p -values, but too low of an Adjusted R^2 to be considered valid.

5. DISCUSSION

5.1 SUBSECTOR VARIATION

Despite the variation in mean water use among the subsectors, the large standard deviation indicates that there might not be a significant difference in the water use or cost among the NAICS codes. A Tukey test was conducted to compare in-group and out-group variance between each combination of subsectors. A Tukey test examines the null hypotheses that the difference between the means of any two subgroups (i.e., industrial subsectors) is greater than zero. Failing this test implies that the difference within a group is greater than between the group, and therefore, the groups are not significantly different. Most combinations of subsectors failed this test at a 95% confidence level, as shown in Figure 16, indicating no difference between the groups. However, food manufacturing (311) was significantly different from 11 other subsectors—beverage and tobacco (312), wood products (321), printing (323), chemical (325), plastics and rubber (326), nonmetallic mineral (327), fabricated metal products (332), machinery (333), computer and electronics (334), electrical equipment (335), and transportation equipment (336).

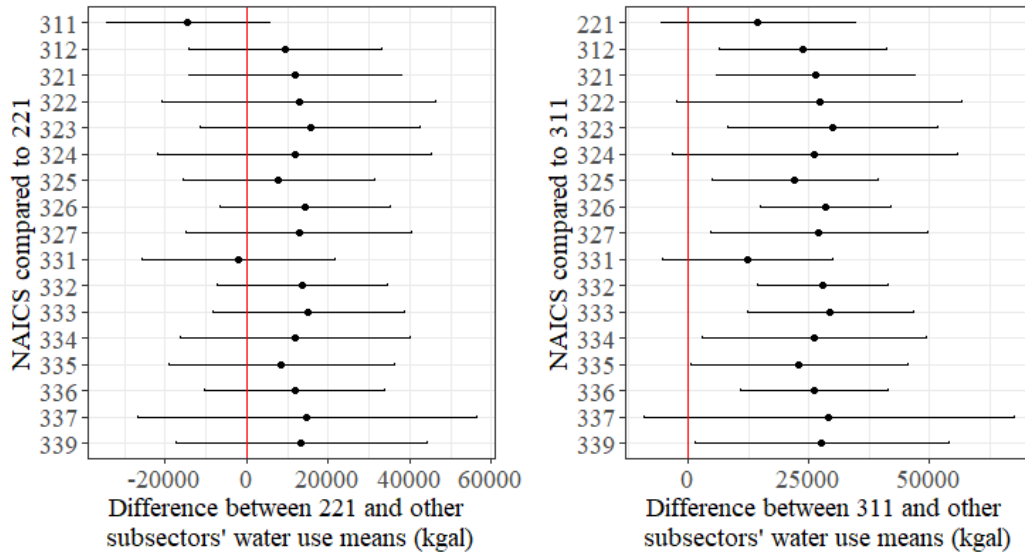


Figure 16. Plots illustrating the difference between the means of the utilities and food subsectors from other subsectors. The dots indicate the difference between the means and the horizontal lines represent the 95% confidence interval for the means.

A Tukey test around water cost showed similar results, with the food manufacturing water cost being significantly different from 15 subsectors, the 11 subsectors that were different for water use plus utilities (221), paper (322), primary metals (331), and miscellaneous manufacturing (339). The beverage and tobacco subsector was also significantly different from three additional subsectors: plastics and rubber (326), fabricated metal (332), and machinery (333). Figure 17 provides the subsector means for each variable; darker tiles indicate that more subsectors have significantly different means. For example, the utilities (221) subsector is significantly different from eight other subsectors in terms of plant area, and no subsectors are significantly different in terms of water stress. Overall, there are very few significant differences among the subsectors for any variables in the analysis.

	221 - Utilities	311 - Food	312 - Bev. & Tob.	321 - Wood	322 - Paper	323 - Print	324 - Petro & Coal	325 - Chemicals	326 - Plastics	327 - Minerals	331 - Pri. Metal	332 - Fab. Metal	333 - Machine	334 - Comp. & Elect.	335 - Electrical	337 - Transpt.	339 - Furniture	Misc.
Use (Mgal)	17.1	31.5	7.68	5.12	4.2	1.51	5.3	9.38	2.91	4.36	19.2	3.6	2.02	5.38	8.58	5.39	2.4	3.76
Cost (\$)	33.8	72.8	40.7	13.8	7.14	5.04	28.9	22.1	8.03	16.3	21.6	8.74	5.71	11.3	29.7	18.4	25.9	10.7
Sales (\$M)	22.1	87.7	58.5	56.5	40.8	27.1	30.2	66.7	43.2	31.6	85.1	61.5	58.3	63.2	100	230	83.5	62.3
Employees	52.1	198	120	168	218	104	69.2	139	191	144	176	154	199	236	215	420	432	273
Production Hours	8760	6500	5440	4620	6270	4790	6730	6710	6580	6430	6690	4960	4930	5490	6360	5960	4420	6180
Plant Area (10,000 sq ft)	109	2.29	1.79	10.2	3.12	1.5	4.29	4.64	2.03	4.41	2.4	3.28	1.93	1.35	2.43	3.22	3.54	1.33
Stress x100	9.2	17.1	22.5	20.1	19.4	8.18	13.6	7.44	14.1	13.9	8.85	15.4	9.11	16	14.7	9.78	9.71	19.1

Figure 17. Variable means by subsector with shading for the number of other subsectors with significantly different means.

The lack of differences between subsectors is possibly because all IAC assessments were conducted at small- to medium-size facilities. Because of their size, these facilities could have more commonalities than differences, especially with respect to how they use water (e.g., a larger proportion of water is used by landscaping or employees than by the processes that would make them unique among the subsectors). Additionally, some industry subsectors (chemicals and foods) are so broad in their products that too much variation exists within the subsector to notice a difference outside the subsector.

5.2 MEAN VS. MEDIAN

Another indicator about the quality of the data is the large difference between the mean and median within each subsector. The mean is substantially higher than the median, indicating substantially skewed data. An attempt was made to adjust the data set to account for this difference (see Section 2.3.1); the resulting data remain skewed, but dramatically less so.

5.3 UNIT COST OF WATER

Another attempt to remove potentially invalid data was made by approximating the unit cost of water and removing all the observations that did not fit within the range of \$1/kgal to \$10/kgal. This removed too many observations, reducing the data set to 347 observations, and was not included in the final analysis (see Section 2.3.2).

5.4 NEGATIVE COEFFICIENTS

For some models, the direction of the relationship was not what it was expected to be, implying that decreasing the stress predictor variable (or increasing any of the other predictor variables) would decrease the dependent variable (i.e., water use or cost). This unexpected difference indicated that the data may have a problem. Visual examination of some of these instances showed that outliers potentially skewed what may be otherwise “flat” or even positively related data to appear negative and significant, where it might realistically be insignificant. For example, Figure 18 shows water use and sales for the paper subsector (322); the original analysis showed a negative correlation between these variables (black line)

and water use and employees. If the highest observation is removed from the data set (performing a regression on data points with water use less than 5,000 Mgal), the correlation becomes positive (red line) and the correlation between water use and employees is no longer in the model. A model with the same variables as the original (sales, employees, production hours, and stress) resulted in insignificant variable p -values and an invalid model. Attempts were made to find models that did not use variables with “wrong direction” coefficients without further reducing the data set, but that led to most alternative models being rejected for too low of an adjusted R^2 .

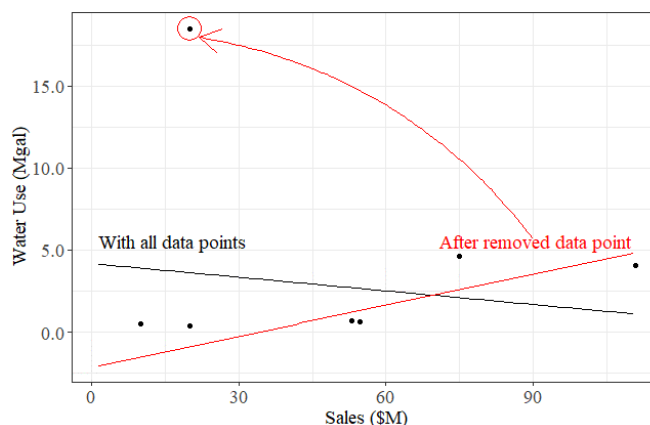


Figure 18. Scatter plot of water cost vs. employees for the paper sector. The black line shows the regression equation generated by analysis; the red line shows the regression equation with the high-water use observation removed from the data set.

5.5 DATA ENTRY ERRORS

The database’s requested units for water use were 1,000 gal (kgal or Tgal), whereas all other entries were “single” units (e.g., gallon, kilowatt-hour, \$). This could have led to confusion during data entry and errors in the data set.

5.6 INDUSTRIAL WATER SOURCES

Finally, it was noted that most (75–80%) industrial water use is self-supplied withdrawal from surface-water and groundwater sources and only a small portion from public water supplies [1], [2]. The IAC assessments only request utility (public water supply) water data, potentially leaving out a large fraction of a facility’s water use from the data set and this analysis. The effect of this missing data can be seen by examining the mean water use of known high-demand industries. As an extreme example, the paper facilities assessed by IACs have a mean water use of 227 Mgal (which was 4.2 Mgal after outlier removal). This amount is much lower than all other industries, despite being considered a high-water-use industry. This discrepancy is likely due to its high intake of non-municipal water, which was not taken into account by the IACs. It is also possible that some of the assessment *did* include self-supplied water, which could explain some of the unrealistically low water unit costs seen in Table 1. More guidance from the IACs to the facilities may be necessary to remove any confusion about what data is supposed to be collected.

5.7 FUTURE WORK

This analysis assumes that water use (and cost) for set of facilities has an underlying normal distribution for the purpose of detecting outliers during exploratory statistical analysis. It also assumes that the industrial subsector to which each facility belongs is the main variable that explains differences in

patterns in water use and cost. Future work may explore different underlying assumptions and examine additional explanatory variables that can further explain more complex patterns of water use and cost. The large variance within the dataset suggests the presence of an underlying highly skewed distribution where a small number of facilities is responsible for using large volumes of water (or having an associated high water expenditure) and a large number of facilities using small volume of water (or having low water expenditures). Assuming an underlying highly-skewed, peaked distribution, such as log-normal, may provide more insights on the relationship of water use or cost and the corresponding explanatory variables. A logarithmic transformation and scaling or standardization methods may be performed to better detect outliers with techniques that assume symmetrical distributions such as the median absolute deviation (MAD).

Since several industrial subsectors have a small number of data points (see Table 3) with high variance, future analysis may focus on analyzing all subsectors combined, and identifying clusters of facilities with similar water use and cost. Unsupervised machine learning algorithms for detection of clusters could provide meaningful insights on the main drivers of water use and the controls of water cost.

IAC associated microdata, currently not available to the public, may be requested to obtain information at a higher spatial resolution. This information may be useful to refine the analysis of spatial variability of water use and cost, and potentially match facilities with their corresponding municipal water utility services.

6. CONCLUSIONS

Overall, an opportunity exists to use water data collected during the IAC assessments for a water use or cost-predictive analysis. However, increasing the amount of data collected and improving the data collection quality is necessary for a useful analysis. The analysis on the water data from 613 IAC assessments could not generate acceptable models for all NAICS codes, and some that were created had negative coefficients. Additionally, several problems were observed in the data set before and after analyses were conducted, further supporting the conclusion that the current data set is not suitable for predictive analysis with the assumption of a normal distribution. The authors understand that the primary focus of IACs is to conduct energy assessments and collect energy use data. However, this study shows the tremendous value in collecting water use data and the opportunity to use the IAC data to conduct national-level industrial water use analyses.

This analysis identified opportunities to improve the quality of water data collected during IAC assessments. Although an opportunity exists to improve data collection quality, acknowledging typical water data collection challenges in the manufacturing industry is important. There are several challenges identified by the authors based on their interactions with industry contacts.

- Facilities do not always meter or track self-supplied water from on-site wells, surface water sources, and so on.
- Lack of metering and submetering on the system level makes acquiring system-level water use difficult.
- Individual system-level discharges are difficult to meter because they usually drain to the facility outlet by gravity via underground channels that are hard to access.
- Water infiltrations and precipitation are difficult to estimate and cause errors to water balancing because they are typically drained to the same facility outlet.
- In certain cases, data gaps can be filled by installing portable strap-on flow meters. However, finding the right location to install strap-on portable flow meters (e.g., ultrasonic flow meters) for spot measurement is challenging. They require a clean pipe with a straight mounting section and a length of at least 20–25 times its diameter, and they must be away from any constriction points for accurate readings.

If these barriers could be addressed and water use more fully monitored within industry, predictive analyses for industrial water use would be more feasible.

7. REFERENCES

- [1] C. A. Dieter *et al.*, “Estimated use of water in the United States in 2015,” 2018. doi: 10.3133/cir1441.
- [2] P. Rao, D. Sholes, W. R. Morrow, and J. Cresko, “Estimating U.S. manufacturing water use,” *2017 ACEEE Summer Study on Energy Efficiency in Industry*, 2017, 104–115, [Online]. Available: <http://aceee.org/files/proceedings/2017/data/64395-aceee-1.3687710/t001-1.3687829/f001-1.3687830/a001-1.3687905/0036-0053-000025-1.3687918.html>.
- [3] W. M. Hanemann, “The economic conception of water,” in *Water Crisis: Myth or Reality?*, P. P. Rogers, M. R. Llamas, and L. Martínez-Cortina, Eds. London, UK: Taylor & Francis, 2006, 61–91.
- [4] P. P. Rogers, M. R. Llamas, and L. Martínez-Cortina, *Water crisis: Myth or reality?* London, UK, 2006.
- [5] R. A. Becker, “Water Use and Conservation in Manufacturing: Evidence from U.S. Microdata,” 2015.
- [6] R. A. Becker, “Researching water use in U.S. manufacturing and mining,” *Global Water Forum*, 2016. <http://www.globalwaterforum.org/2016/10/24/researching-water-use-in-u-s-manufacturing-and-mining/> (accessed Jul. 02, 2018).
- [7] R. A. Becker, “Water use and conservation in manufacturing: Evidence from U.S. microdata,” *Water Resour. Manag.*, vol. 30, no. 12, 4185–4200, 2016, doi: 10.1007/s11269-016-1414-7.
- [8] K. Kluge, “Texas Municipal and Industrial Water Use Survey,” 2016, [Online]. Available: <https://water.usgs.gov/wausp/wudr/wu-forum-files/Kluge-TWDB-20160427.pdf>.
- [9] Texas Water Development Board, “Industrial Water Use Survey for End of Calendar Year December 31, 2017,” 2017. http://www.twdb.texas.gov/waterplanning/waterusesurvey/survey/doc/2017/2017_Industrial.pdf?d=4654.02 (accessed Jul. 02, 2018).
- [10] Statistics Canada, “Biennial Industrial Water Survey (IWS),” 2018. <http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5120> (accessed Sep. 27, 2019).
- [11] Industrial Assessment Centers, “Industrial Assessment Centers,” 2020. <https://iac.university/> (accessed Jul. 24, 2020).
- [12] Industrial Assessment Centers, “Industrial Assessment Center Database,” 2018. <https://iac.university/download> (accessed Dec. 06, 2018).
- [13] NAICS Association, “2017 - 2022 SIC to NAICS Crosswalk,” 2017. <https://www.naics.com/wp-content/uploads/2014/10/SIC-to-NAICS-Crosswalk.pdf> (accessed Apr. 08, 2019).
- [14] M. B. Muller, “IAC Assessment Database Manual Version 10.2,” 2011. [Online]. Available: https://iac.university/technicalDocs/IAC_DatabaseManualv10.2.pdf.
- [15] U.S. Department of Energy Federal Energy Management Program, “Water and Wastewater Annual Price Escalation Rates for Selected Cities across the United States,” 2017. doi: DOW/EE.1670.
- [16] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median,” *J. Exp. Soc. Psychol.*, 4–6, 2013.

- [17] Sandia National Laboratories, "Data Modeling & Analysis - Water Availability, Cost, and Use Data," 2019. <https://energy.sandia.gov/programs/energy-water/data-modeling-analysis/> (accessed Jul. 24, 2020).
- [18] V. C. Tidwell *et al.*, "Mapping water availability, projected use and cost in the western United States," *Environ. Res. Lett.*, vol. 9, no. 6, 2014, doi: 10.1088/1748-9326/9/6/064009.
- [19] V. C. Tidwell, B. D. Moreland, C. R. Shaneyfelt, and P. Kobos, "Mapping water availability, cost and projected consumptive use in the eastern United States with comparisons to the west," *Environ. Res. Lett.*, vol. 13, no. 1, 2018, doi: 10.1088/1748-9326/aa9907.
- [20] C. A. Dieter *et al.*, "Estimated Use of Water in the United States County-Level Data for 2015 (ver. 2.0, June 2018): U.S. Geological Survey data release," 2018. doi: 10.5066/F7TB15V5.
- [21] U.S. Department of Energy, "Superior Energy Performance 50001TM Program Measurement & Verification Protocol : 2019," 2019. [Online]. Available: https://betterbuildingssolutioncenter.energy.gov/sites/default/files/attachments/SEP_50001_MV_Protocol_2019.pdf.

APPENDIX A. WATER USE AND COST TABLES

APPENDIX A. WATER USE AND COST TABLES

Table A. Water use: model statistics and coefficients.

NAICS code	Number of assessments	Model <i>p</i> -value	Adj. <i>R</i> ²	Intercept	Sales (10 ⁶)	Employees	Facility area (10 ³)	Production hours	Water stress (10 ⁻³)
221 Utilities	17	0.000	0.868	5,306	302.3	89.5	---	---	---
311 Food	74	0.000	0.454	-18,253	287.1	---	---	4.54	---
312 Bev. & Tobacco.	25	0.291	0.025	1,915	60.7	---	---	0.81	---
321 Wood Products	16	0.055	0.183	1,743	---	---	---	---	14.8
322 Paper	7	0.010	0.985	-13,937	-27.7	-10.2	---	2.94	11.1
323 Printing & Related	14	0.000	0.730	-260	---	---	14.6	---	---
324 Petroleum & Coal	7	0.254	0.591	-7,937	---	45.8	46.8	1.48	-121.1
325 Chemical.	25	0.158	0.077	8,076	175.6	-66.4	---	---	---
326 Plastics & Rubber	52	0.032	0.071	1,586	---	---	7.8	---	---
327 Nonmetallic Minerals.	13	0.038	0.275	1,513	---	24.2	---	---	---
331 Primary Metals	24	0.010	0.342	-42,134	282.8	---	28.2	5.55	---
332 Fabricated Metal	53	0.014	0.095	495	---	---	---	0.64	---
333 Machinery	25	0.001	0.356	-1,326	---	---	18.2	---	---
334 Comp. & Electronic	12	0.181	0.228	3,408	132.4	---	-100.0	---	48.9
335 Electrical Equip.	13	0.020	0.452	-16,718	43.3	---	---	3.06	---
336 Transp. Equipment	36	0.049	0.083	1,982	---	---	14.2	---	---
337 Furniture & Related	4	0.095	0.973	10,403	55.6	-46.5	---	---	---
339 Misc. Manuf.	9	0.005	0.915	517	-79.2	21.7	56.3	-0.74	---

Note: all variables are in the units specified in **Table 4**

Table B. Water use: VIFs.

	NAICS code	Sales	Employees	Facility area	Production hours	Water stress
221	Utilities	1.50	1.50	---	---	---
311	Food	1.05	---	---	1.05	---
312	Bev. & Tobacco.	1.00	---	---	1.00	---
321	Wood Products	---	---	---	---	1.00
322	Paper	1.62	2.04	---	2.24	3.00
323	Printing & Related	---	---	1.00	---	---
324	Petroleum & Coal	---	1.54	14.57	1.72	14.65
325	Chemical.	2.19	2.19	---	---	---
326	Plastics & Rubber	---	---	1.00	---	---
327	Nonmetallic Minerals.	---	1.00	---	---	---
331	Primary Metals	1.00	---	1.00	1.00	---
332	Fabricated Metal	---	---	---	1.00	---
333	Machinery	---	---	1.00	---	---
334	Comp. & Electronic	3.35	---	3.77	---	1.24
335	Electrical Equip.	1.06	---	---	1.06	---
336	Transp. Equipment	---	---	1.00	---	---
337	Furniture & Related	6.95	6.95	---	---	---
339	Misc. Manuf.	2.26	1.21	2.21	1.20	---

Table C. Water use: variable p -values.

NAICS code	Intercept	Sales	Employees	Facility area	Production hours	Water stress
221 Utilities	0.103	0.000	0.033	---	---	---
311 Food	0.148	0.000	---	---	0.018	---
312 Bev. & Tobacco.	0.680	0.232	---	---	0.313	---
321 Wood Products	0.441	---	---	---	---	0.055
322 Paper	0.019	0.143	0.061	---	0.017	0.031
323 Printing & Related	0.559	---	---	0.000	---	---
324 Petroleum & Coal	0.198	---	0.178	0.209	0.114	0.154
325 Chemical.	0.102	0.062	0.248	---	---	---
326 Plastics & Rubber	0.051	---	---	0.032	---	---
327 Nonmetallic Minerals.	0.412	---	0.038	---	---	---
331 Primary Metals	0.067	0.008	---	0.155	0.083	---
332 Fabricated Metal	0.716	---	---	---	0.014	---
333 Machinery	0.248	---	---	0.001	---	---
334 Comp. & Electronic	0.455	0.059	---	0.092	---	0.135
335 Electrical Equip.	0.083	0.092	---	---	0.040	---
336 Transp. Equipment	0.342	---	---	0.049	---	---
337 Furniture & Related	0.048	0.094	0.069	---	---	---
339 Misc. Manuf.	0.831	0.023	0.001	0.014	0.107	---

Table D. Water cost: model statistics and coefficients.

NAICS code	Number of assessments	Model <i>p</i> -value	Adj. <i>R</i> ²	Intercept	Sales (10 ⁶)	Employees	Facility area (10 ³)	Production hours	Water stress (10 ⁻³)
221 Utilities	25	0.000	0.568	31,558,154	---	---	16.1	-3,600.5	---
311 Food	82	0.000	0.184	-32,984	136.0	---	---	14.9	---
312 Bev. & Tobacco.	27	0.047	0.114	7,446	---	---	---	6.8	---
321 Wood Products	17	0.103	0.223	-526	209.3	-25.3	---	2.0	---
322 Paper	8	0.087	0.474	18,632	113.6	---	---	-2.8	---
323 Printing & Related	18	0.002	0.571	-1,407	91.7	---	27.5	---	13.6
324 Petroleum & Coal	8	0.004	0.741	8,350	---	---	47.8	---	---
325 Chemical.	33	0.044	0.097	-9,095	---	---	---	4.7	---
326 Plastics & Rubber	60	0.024	0.092	-472	---	---	13.2	1.0	---
327 Nonmetallic Minerals.	15	0.655	-0.060	17,355	---	---	---	---	-7.9
331 Primary Metals	24	0.001	0.476	-4,743	95.7	78.4	34.7	---	---
332 Fabricated Metal	55	0.000	0.277	3,331	70.7	15.1	---	---	---
333 Machinery	38	0.020	0.118	3,995	31.6	---	---	---	---
334 Comp. & Electronic	13	0.256	0.086	12,066	281.2	-82.4	---	---	---
335 Electrical Equip.	16	0.036	0.225	-20,116	---	---	---	7.8	---
336 Transp. Equipment	48	0.001	0.202	6,250	---	---	57.0	---	---
337 Furniture & Related	7	0.298	0.055	14,671	---	26.0	---	---	---
339 Misc. Manuf.	11	0.018	0.426	2,830	---	32.1	---	---	---

Note: all variables are in the units specified in **Table 4**

Table E. Water cost: VIFs.

NAICS code	Sales	Employees	Facility area	Production hours	Water stress
221 Utilities	---	---	1.000805	1.000805	---
311 Food	1.084735	---	---	1.084735	---
312 Bev. & Tobacco.	---	---	---	1	---
321 Wood Products	2.071578	1.847849	---	1.188313	---
322 Paper	1.932391	---	---	1.932391	---
323 Printing & Related	1.216106	---	1.228858	---	1.220534
324 Petroleum & Coal	---	---	1	---	---
325 Chemical.	---	---	---	1	---
326 Plastics & Rubber	---	---	1.05907	1.05907	---
327 Nonmetallic Minerals.	---	---	---	---	1
331 Primary Metals	1.006257	1.038595	1.043672	---	---
332 Fabricated Metal	1.658665	1.658665	---	---	---
333 Machinery	1	---	---	---	---
334 Comp. & Electronic	3.400559	3.400559	---	---	---
335 Electrical Equip.	---	---	---	1	---
336 Transp. Equipment	---	---	1	---	---
337 Furniture & Related	---	1	---	---	---
339 Misc. Manuf.	---	1	---	---	---

Table F. Water costs: variable p -values.

NAICS code	Intercept	Sales	Employees	Facility area	Production hours	Water stress
221 Utilities	0.002	---	---	0.000	0.002	---
311 Food	0.210	0.153	---	---	0.000	---
312 Bev. & Tobacco.	0.674	---	---	---	0.047	---
321 Wood Products	0.945	0.102	0.303	---	0.238	---
322 Paper	0.012	0.060	---	---	0.038	---
323 Printing & Related	0.518	0.194	---	0.001	---	0.249
324 Petroleum & Coal	0.281	---	---	0.004	---	---
325 Chemical.	0.560	---	---	---	0.044	---
326 Plastics & Rubber	0.895	---	---	0.103	0.073	---
327 Nonmetallic Minerals.	0.001	---	---	---	---	0.655
331 Primary Metals	0.557	0.001	0.083	0.044	---	---
332 Fabricated Metal	0.049	0.009	0.192	---	---	---
333 Machinery	0.001	0.020	---	---	---	---
334 Comp. & Electronic	0.215	0.128	0.314	---	---	---
335 Electrical Equip.	0.390	---	---	---	0.036	---
336 Transp. Equipment	0.141	---	---	0.001	---	---
337 Furniture & Related	0.322	---	0.298	---	---	---
339 Misc. Manuf.	0.448	---	0.018	---	---	---