AI Prediction of Michigan Water Use

Jacob Javier

Springboard Data Science Career Track

AI Prediction of Michigan Water Use

With the human population still increasing, the demand for consumer goods is increasing as well. To keep up with demand, natural resources like water are heavily impacted to produce goods and services. For example, the Great Lakes provides a source of revenue for recreational businesses, helps generate hydroelectric energy, and is used to in industrial manufacturing. Although water levels have been on a steady incline since 2013; there was a 73% decrease in water levels from 2020 to 2023 (U.S. Environmental Protection Agency 2023).

In June 2024, Martusiuk aggregated the Department of Environment, Great Lakes, and Energy's water usage data for the Great Lakes Basin from 2013 to 2022. This decade's worth of data can be used to gain insight into the increasing demands on Michigan's water supply using regression analysis and forecasting.

The primary objective of this project is predicting the future impact on Michigan's water supply based on archived industry and geographical water consumption levels. More specifically, this project will showcase how water consumption changes over time across industries. Population growth data will also be included as an explanatory variable for the increased demand on the industries. With the annual population data, correlations may be identified across the industries to explain changes in demand. Additionally, highlighting the relationships between industries' water consumption will provide insight into any underlying effects on water use.

Data Wrangling

The water use data set (2024) originally contained eight features and 6630 observations. The data ranged from 2013 to 2022 and showed the water use data of seven unique industries that was sourced three different water sources. The seven industries that were evaluated were Commercial-Institutional, Electric Power Generation, Irrigation, Livestock, Public Water Supply, and Industrial-Manufacturing. In addition to the unique categorical values, there was also an aggregate observation that totaled the water use data across all seven industries. Similarly, there were three water source features that were evaluated with an additional aggregate total feature of all three water sources. The featured water sources included Inland, Great Lakes, and Groundwater. All water use data was reported in number of gallons. Lastly, there were 85 counties that were evaluated for this survey. For the purposes of this model, counties were removed to better understand the specific impact the various industries had on water use. Future studies that include this feature may develop more robust models to forecast future water use.

Michigan population data was also collected from the U.S. Census Bureau from two csv files. The datasets each contained over 14 extraneous feature categories that were summed into a total population estimate. Since population was hypothesized to be an explanatory variable to Michigan's water use, the total population estimate was the only feature used. Investigating the relationship between different migration categories falls outside the scope of this model. When concatenating the two datasets, 2020 was accounted for twice because it was the transition year between csv files. The population data for 2020 in the newest file (2020 – 2023) was retained as the latest update. The population and water use data sets were merged using year as the common feature.

Aggregations and reshaping

One of the main goals for aggregating the data was to distinguish the water source and water use features from each other. After the county and previous index feature were dropped, the DataFrame was melted to create a categorical water source feature and a continuous water use feature. All pre-calculated totals for water source and industries were dropped to ensure data was not overly represented.

Water use and population values were also scaled to represent billions rather than individual units. Since both datasets were estimates of their respective features, reducing these values for higher interpretability was appropriate. After removing counties, the remaining data was also transformed by sum aggregating to represent the total water use for each industry's water source. The final data transformation was adjusting year to a datetime object for future analysis and modeling. From box plots and value counts, no outliers or null values were identified within each industry. The resulting DataFrame contained 210 observations across five features: year, population (Bil.), water_source, industry, and water_use (Bil. gal).

Data Exploration

The goal of this project is to forecast Michigan's water use; using industries as a potential explanatory feature. Water use data for each industry will be plotted over the years to understand how it changes over time and how it correlates with population growth. Correlations between industries, water sources, and population must also be assessed to see if there are any underlying relationships between changes over time. Lastly, each feature will be analyzed using the autocorrelation function to show any relationships from the data within itself over time. To better define these goals, three questions were used to guide EDA.

1. How does water use change over time?
2. What relationships exist between industry, water source, and population?
3. Are there annual changes in water usage?

How does water use change over time?

To get a baseline for analysis, water use was sum aggregated and compared against the annual population; this was to confirm if there was any underlying relation between increasing populations and water usage.

The Michigan population increased 1.2% from 9.9 million to 10.0 million people (Fig. 1). Compared to the global growth rate in 2022 (0.83%), this is a major increase in the population. In contrast, the water use since 2013 declined by 22.6% (Fig. 1). Although the result contradicts the initial assumption that an increasing population may increase water use demands, a real relationship between population and water use may still exist that can be confirmed using linear regression analysis.

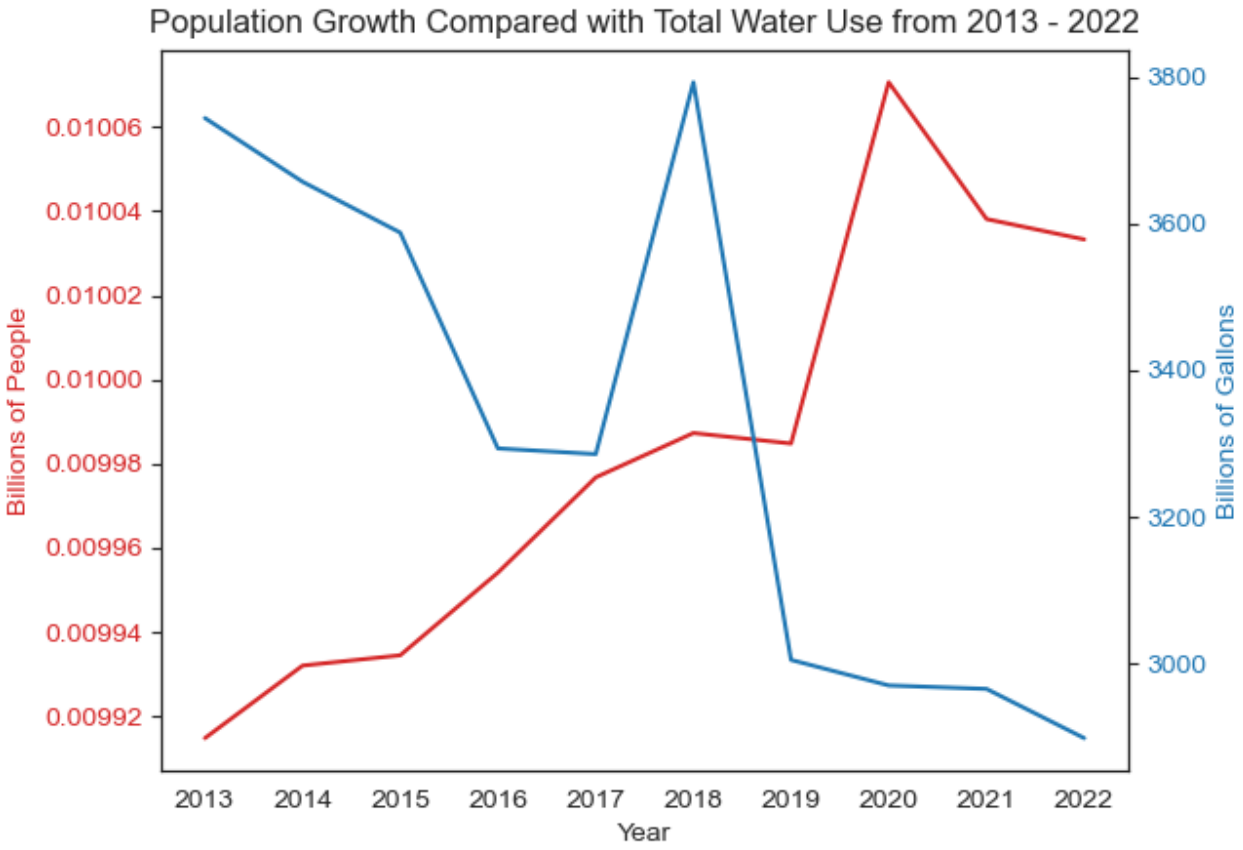

Fig. 1 Temporal distribution of Michigan water use and population estimates.

Looking at the individual industries, only the inland water source for industrial manufacturing and general use; and groundwater for general use increased (Fig. 2). The public water supply from all water sources had the least variability, whereas industrial manufacturing had some of the highest variability over the years. Electrical power generation had the largest water consumption out of all the industries, followed by the public water supply. Both industries were the only ones to be mainly sourced by the Great Lakes. The agricultural industries (irrigation and livestock) pulled most of their water from groundwater and inland sources. This may be due to the vast expanse of low urban areas being further from coastal regions near the Great Lakes.

Despite there being no clear industry contributions (Fig. 2) to the drop in total water use (Fig. 1), there are correlations that will be good to explore. The inverse relationship between population and total water use is of particular interest because it is counter intuitive to human impact on local resources. Additionally, expanding upon the relationship between inland water sources and industrial manufacturing will need correlation tests to confirm the visible trends; similarly, the correlation of groundwater and general use will need to be explored. No annual seasonality between any of the categories have been identified but later autocorrelation will be able to assist with that.

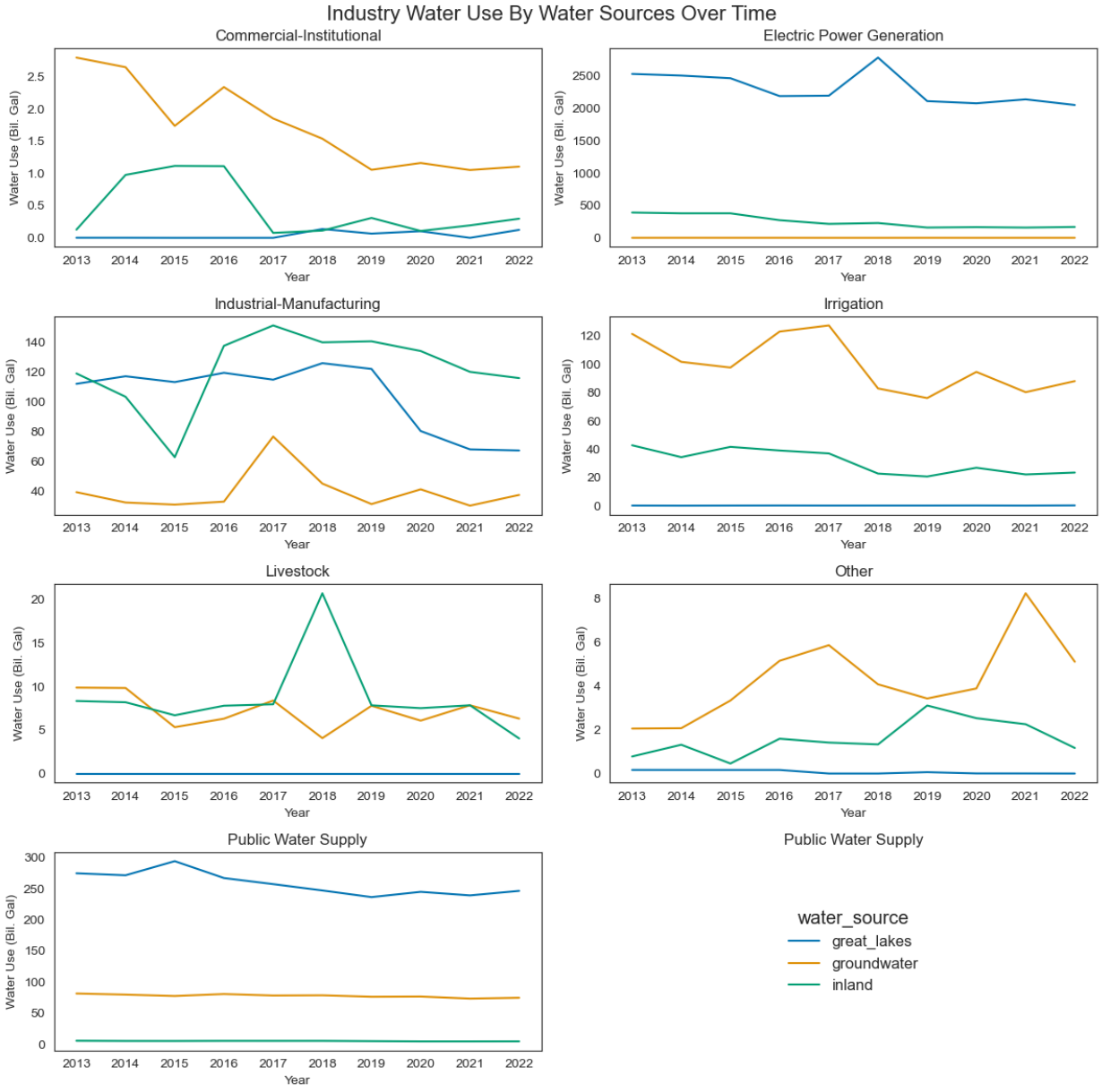

Fig. 2 Temporal distributions of industry water uses distinguished by water sources.

What relationships exist between industry, water source, and population?

For this question, five sub questions were identified to segment analysis. Population was assumed to be a key explanatory feature which was supported with the visual linear trends; so, it was compared against each other feature individually. To see if there was any internal relationships between the categorical features, correlation was also tested within each categorical feature.

1. How does population relate to general water use?

2. How does population relate to each industry?
3. How does population relate to each water source?
4. How do the different industries relate internally?
5. How do the different water sources relate internally?

How does population relate to general water use?

Population was compared against annual water use using Ordinary Least Squares regression to determine the weight of the population coefficient and its significance levels. From this analysis a linear slope of -5398153 billion gallons per billion people was identified. Additionally, Pearson's correlation coefficient was calculated to confirm how much of the water use variability was being explained by population.

As expected, there is a significant (p-value = 0.007) negative linear relationship (adjusted r-squared = 0.571) between the population and total water use. The relationship indicates that for every one new person in Michigan, the total number of gallons used drops by about 0.005 billion. The results were further confirmed using Pearsons correlation coefficient (stat = -0.787, p-value = 0.007) suggesting a strong negative relationship between the two features.

The sample size is low, so more archived data and experimentation would be able to reveal any causal relationship. With the total water use relationship established, the different industries should show a similar trend.

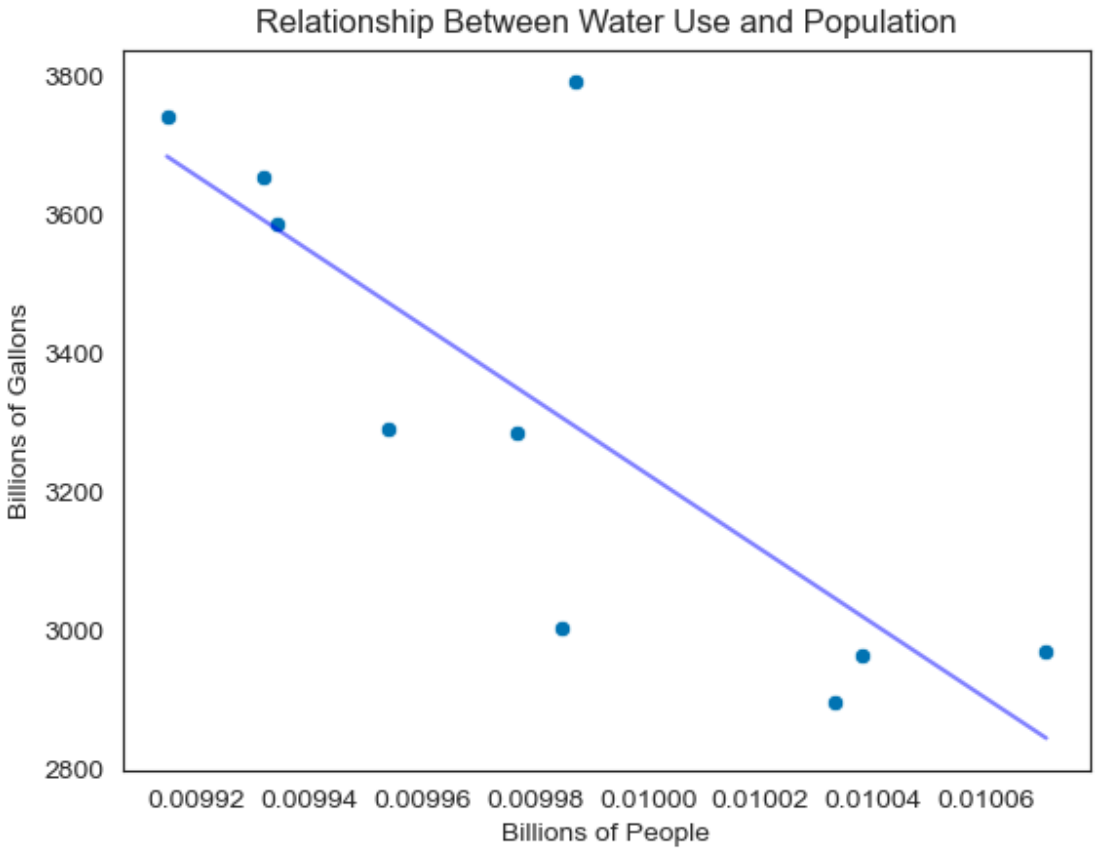

Fig. 3 Linear relationship between water use and population estimates.

How does population relate to each industry?

Population was compared to each industry's water use using Ordinary Least Squares regression to determine the weight of the population coefficient (Table 1). Additionally, Pearson's correlation coefficient was calculated with each industry to confirm how much of the water use variability was being explained by population (Table 1).

Table 1. Summary correlation statistics between population and each industry.

	industry	slope	intercept	R-squared	Adj. R-squared	Prob (F-statistic)	Pearson Coef.	Prob (Pearson)
0	Commercial-Institutional	-1.475124e+04	149.468799	0.695470	0.657404	0.002708	-0.833949	0.002708
1	Public Water Supply	-3.271474e+05	3606.327297	0.673922	0.633162	0.003602	-0.820927	0.003602
2	Electric Power Generation	-4.598570e+06	48464.624618	0.544615	0.487692	0.014819	-0.737980	0.014819
3	Other	2.904343e+04	-283.958503	0.470636	0.404466	0.028498	0.686029	0.028498
4	Irrigation	-3.191146e+05	3316.240230	0.379757	0.302227	0.057793	-0.616244	0.057793
5	Livestock	-2.042918e+04	219.821621	0.068883	-0.047506	0.463817	-0.262456	0.463817
6	Industrial-Manufacturing	-1.471843e+05	1735.161808	0.029759	-0.091521	0.633671	-0.172507	0.633671

In relation to an increasing population (Table 1), commercial (Pearson = -0.833, p-value = 0.003), public water supplies (Pearson = -0.821, p-value = 0.004), and electric power (Pearson = -0.738, p-value = 0.015) all decrease in water use. General use (Pearson = 0.686, p-value = 0.028) is the only industry that scaled with population growth. The agricultural industries' variations were not significantly explained by the population growth (Table 1). This may be due to an increased reliance on those industries creating a weak negative linear relationship. Industrial manufacturing also had a weak relationship with population. Industrial manufacturing may account for demand out of state, and so may not be as heavily influenced by the growing population.

How does population relate to each water source?

Similar to industry and general water use, population was compared to each water source use using Ordinary Least Squares regression to determine the weight of the population coefficient (Table 2). Additionally, Pearson's correlation coefficient was calculated with each industry to confirm how much of the water use variability was being explained by population (Table 2).

Groundwater had the weakest correlation with population growth (Table 2, Pearson = -0.374, p-value = 0.286) which may be represented by the less publicly accessible industries like commercial irrigation. Inland sources showed the strongest relationship (Pearson = -0.907, p-value = 0.000) which may be explained by the geographic distribution of the population. Only so many people can aggregate near the Great Lakes so taking advantage of rivers and lakes across the state will be a good project to investigate further.

Table 2. Summary correlation statistics between population and each water source.

	water_source	slope	intercept	R-squared	Adj. R-squared	Prob (F-statistic)	Pearson Coef.	Prob (Pearson)
0	inland	-1.630446e+06	16697.765387	0.822980	0.800853	0.000290	-0.907183	0.000290
1	great_lakes	-3.544232e+06	38047.915600	0.449927	0.381168	0.033750	-0.670766	0.033750
2	groundwater	-2.234743e+05	2462.004882	0.140219	0.032747	0.286388	-0.374459	0.286388

How do the different industries relate internally?

To better understand how the different industries impact each other's water use, a heat map (Fig. 4) was used to initially visualize some underlying correlations. From that initial assessment, seven pairs of industries were identified to have a strong correlation with one another. These seven pairs were then analyzed using Ordinary Least Squares regression and Pearson's correlation coefficient.

Some of the strongest relationships (Fig. 4) that were identified were between irrigation and general (corr = 0.999), livestock and public water supply (corr = -0.993), irrigation and commercial (corr = 0.997),commercial and general (corr = 0.991), electric power and livestock (corr = -0.966), and electric power and public water supply (corr = 0.928).

Despite the strong correlations, only irrigation and general industries had explanatory power (Adj. R-squared = 0.995, p-value = 0.033; Pearson = 0.999, p-value = 0.033) towards the other. The general industry has a much more noticeable effect on the change of irrigation water use (slope = 23.594) than the inverse (slope = 0.042). This may just be due to the scale difference between the two features.

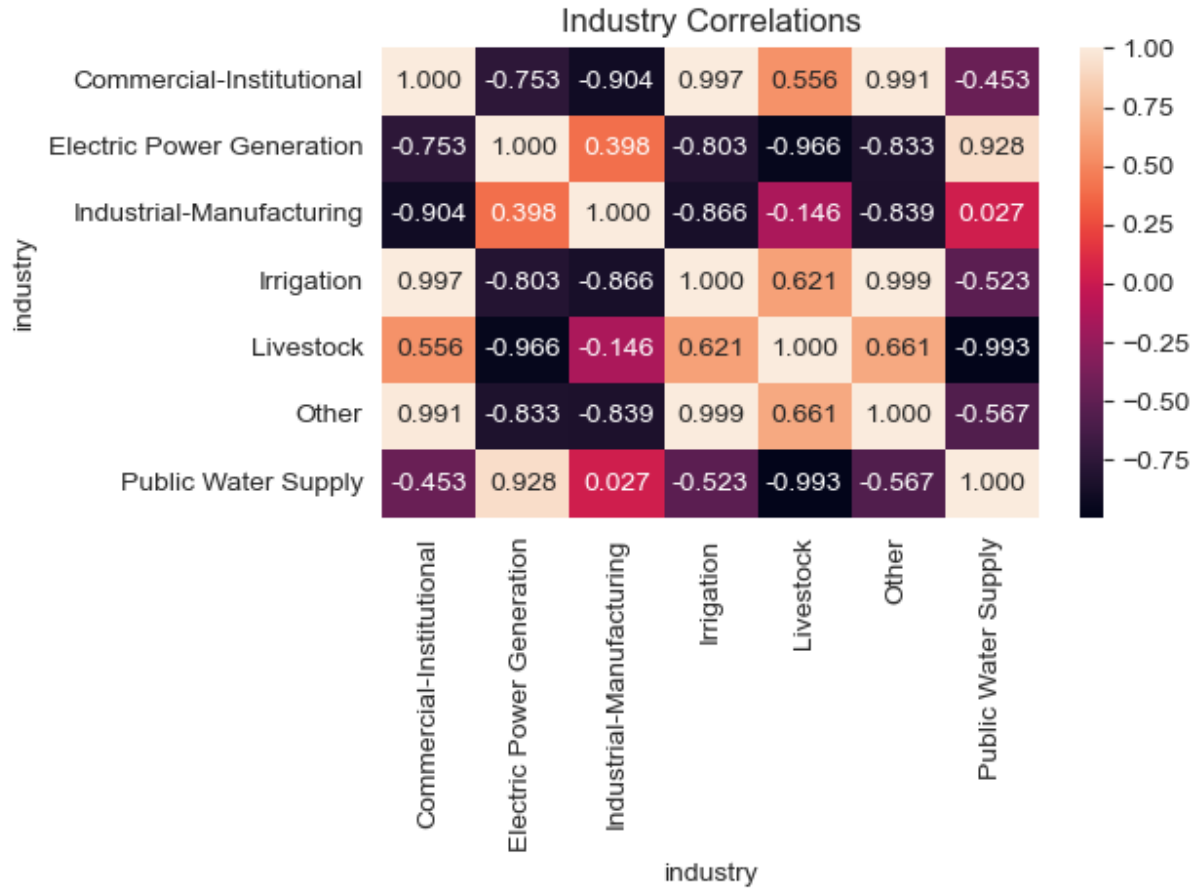

Fig. 4 Heatmap of internal industry correlations of water use.

All other correlated features identified with the heat map had more than a 5% chance of observing as strong of a correlation coefficient as they produced. Irrigation, for example, almost had a significant (Pearson = 0.997, p-value = 0.051) positive (slope = 57.431) relationship to the commercial industry.

How do the different water sources relate internally?

Lastly, the water sources were compared against each other to see if impacts on one water source may correlate with other water sources. The water sources were analyzed using Ordinary Least Squares regression and Pearson's correlation coefficient after plotting a heatmap (Fig. 5) of the correlations.

Fig. 5 Heatmap of internal water source use correlations.

The inland water source had the most noticeable relationship (slope = 8.012, Adj. R-squared = 0.756, p-value = 0.007; Pearson = 0.893, p-value = 0.007) with the Great Lakes. No other sources had any significant nor strong enough (corr > |0.893|) to compare with inland and Great Lake water sources. Many of the inland sources may be sourced or connected to the Great Lakes so changes in one may be more noticeable in the other. For industries more reliant on the Great Lakes or inland water sources (Fig. 2) like electric power generation, industrial manufacturing, and the public water supply; as water use increases from one source, the use also increases with the other. This is important information to plot out future planning for water supplies.

Are there annual changes in water usage?

By understanding how the data correlates with itself, preliminary predictions can be later modeled and compared. Seasonality was identified up to lag 2 for population (Fig. 6, CI = ± 0.620), total water use (Fig. 7, CI = ± 0.620), commercial industries, public water supplies, and inland water sources. The lag for the auto-correlation function represents periods of years. Confidence intervals for the AFC were calculated with an alpha of 0.05.

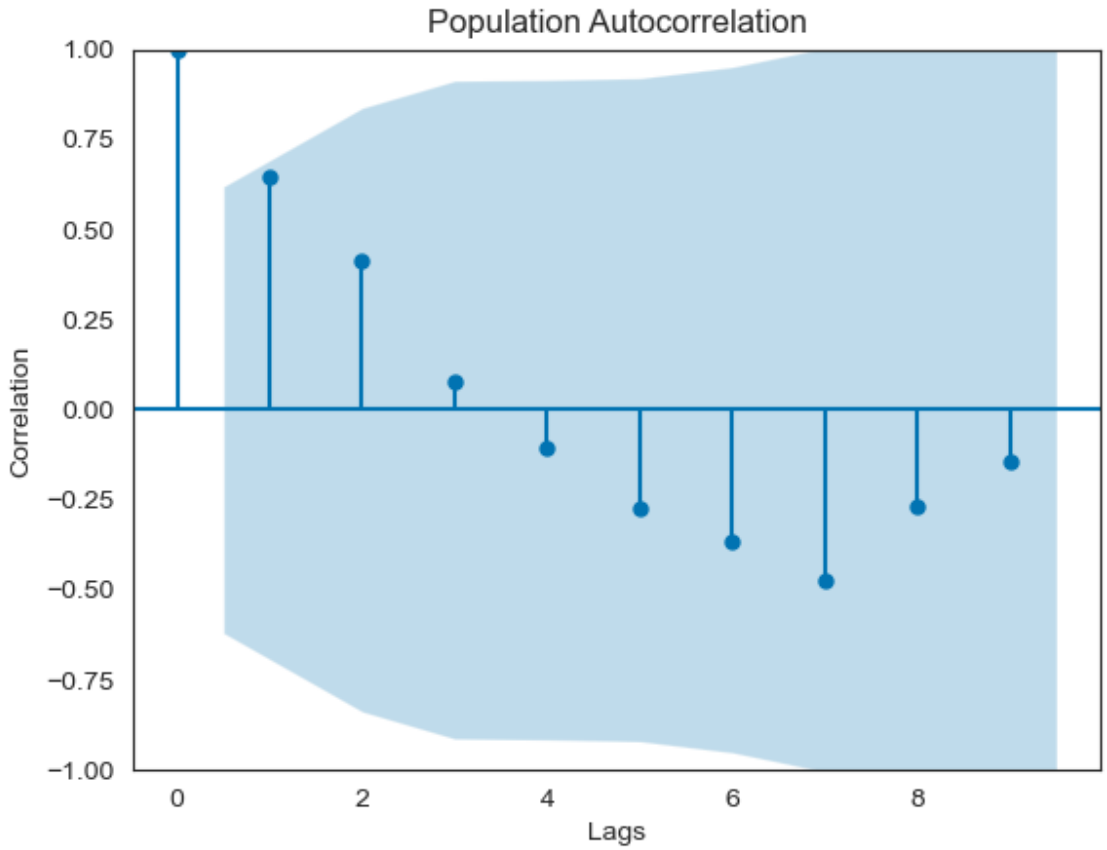

Fig. 6. Auto-correlation function of annual population estimates.

One of the weakest auto-correlations with livestock's annual water use (lag_1 = 0.047, CI = ± 0.620). The Great Lakes (lag_1 = 0.190, CI = ± 0.620) and ground water (lag_1 = 0.183, CI = ± 0.620) sources also expressed very weak auto-correlations making their data best represented by other regressive features. All other industries expressed a lag correlation up to lag 1.

EDA Summary

Water use in Michigan has had a steady decline from 2013 to 2022 which negatively correlates with the state's population growth during that time. If a causal relationship is identified, population growth has the biggest impacts on water used from the Great Lakes and inland sources in the commercial, public, electrical generation, and general industries. This indicates that encouraging population growth for Michigan may reduce the usage of water in these industries, leaving resources available for new organizations and innovations. The data is limited by not being able to account for the available water in each of these sources and efficiency advancements for industry demands.

Groundwater sources seem to be the most inaccessible water source which may account for it being used, at most, 130 billion gallons in any one year by a single industry (Fig. 2). Agricultural

and industrial manufacturing industries must have confounding features that drive their water usage. These industries may have a majority of their goods exported to other regions that influence their demand on local resources; or their demand for water is minimal compared to the other resources required to produce their goods and services.

Inter-feature correlations also indicate that irrigation and general use industries are positively correlated with one another (Fig. 4). These two industries account for over 99% of the variation for each other. This means that instilling regulations to limit or promote the use of resources in one may result in a change in the other. Similarly, the use of the Great Lakes and surface-level water sources are linked to one another. If the water uses impact one source too much, regulations on the other water sources may mitigate the damages.

Population, total water use, commercial industries, public industries, and inland water sources all displayed strong autocorrelations. The confidence intervals for each auto-correlation function is large indicating that more data may alleviate some of the uncertainty behind any predictions that could be gained from this exploratory analysis. The autocorrelations are consistent with other correlations between industries, water sources, and the population growth identified earlier. This analysis will be an excellent basis to compare the models to.

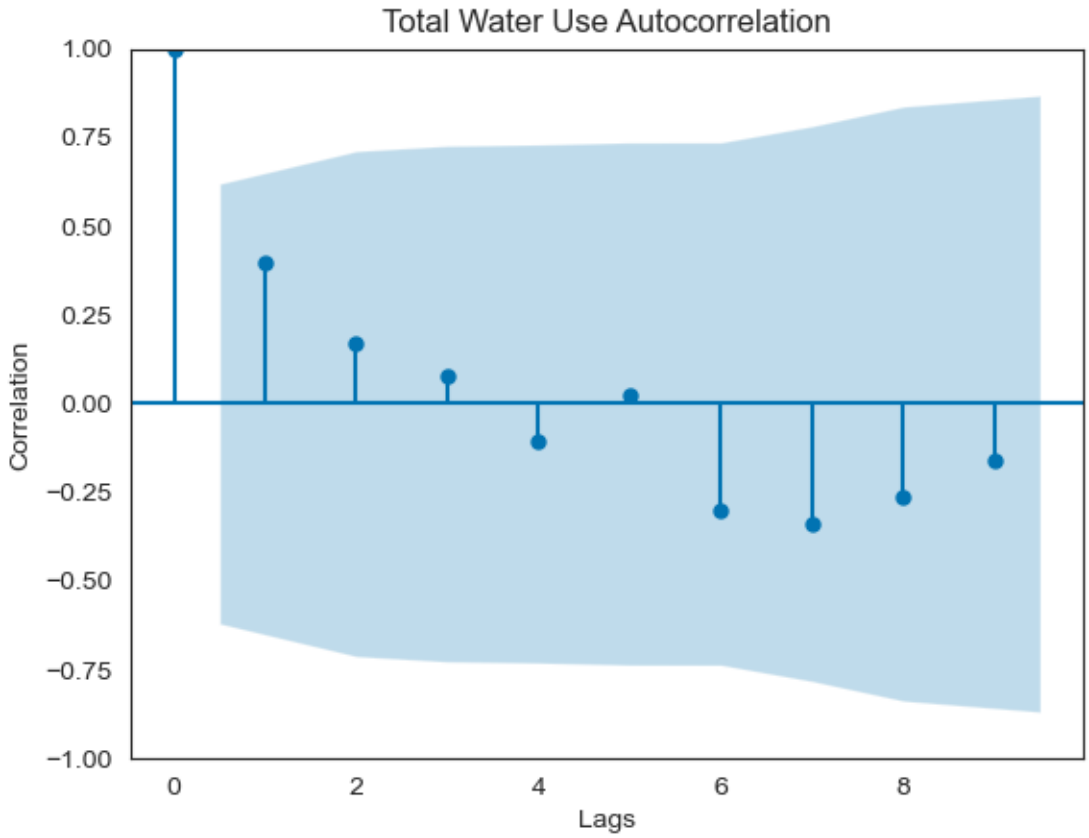

Fig. 7 Auto-correlation function of annual total water use estimates.

Modeling

Three models were used to assess forecasting capabilities for this dataset: ARIMA, linear regression, and K-nearest neighbors. These regressive models were chosen because they were supervised learning techniques and could forecast water use levels. The data was split with a 20% test size. The industry and water source features were aggregated out initially with the intention of running models post-model selection because the total water use showed strong auto-correlation and correlation with the primary explanatory feature (population).

The ARIMA model was tuned for stationarity, autocorrelation, and partial autocorrelation. The initial augmented Dickey-Fuller Tests (ADF) indicated that there was very little stationarity for both population (ADF = 0.862, p-value = 0.993) and total water use (ADF = -2.404, p-value = 0.141). After differencing the data to a single order, water use was testing for stationarity (ADF = -3.451, p-value = 0.009). Population was just outside of stationarity classification (ADF = -2.401, p-value = 0.142), but since there was high correlation found during EDA, it was included as the main exogeneous variable.

Because the data was previously auto-correlated the expected range for the lag order was between 1 to 2. The training data was replotted using the ACF and PACF; then supported using Bayesian Information Criterion to confirm the results. Both the ACF and the PACF supported a lag order of 0 (Fig. 8); which was later confirmed testing for the lowest BIC (BIC_{ACF0} = 104.756, BIC_{PACF0} = 104.756). From hypertune modeling, the data was best modeled as a random walk for ARIMA. This may be due to the significantly small dimensionality of the data.

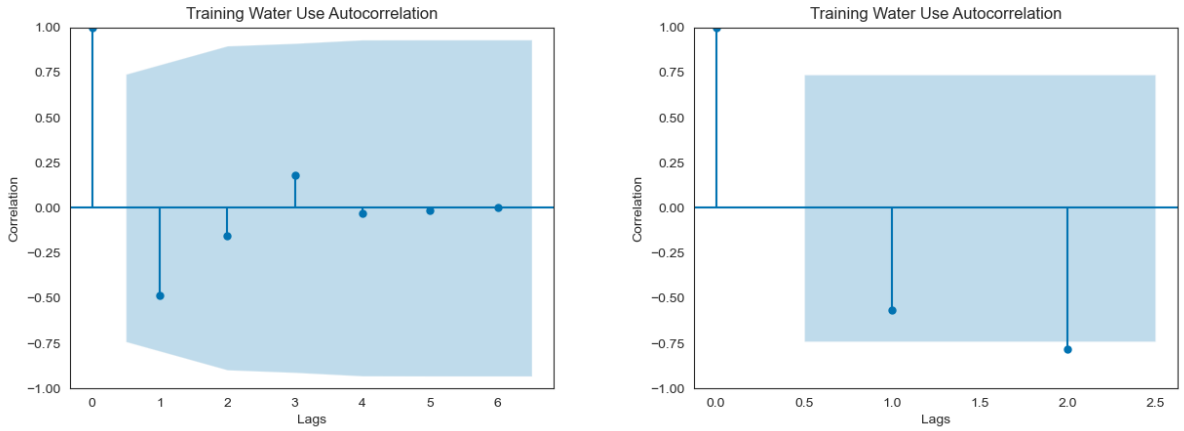

Fig. 8 Total water use of the training data plotted with ACF (left) and PACF (right).

The simple linear regression (lr) model was also selected to model the relationships because of the high correlation and autocorrelations identified earlier. The lack of multiple exogeneous features reduced the need for either ridge or lasso regression methods. No hyperparameters were present that required tuning.

K-nearest neighbors (knn) was also hypertuned for evaluation. The knn model was hypertuned to the number of neighbors, the weight distribution, and the algorithm using a

GridSearchCV. The model was hypertuned to optimize the root mean squared error as the evaluation metric.

Model Results

The models were evaluated based on their root mean squared error (RMSE) and r-squared (R2) scores. The ARIMA model performed best overall with an RMSE of 62.1 billion gallons and an R2 score of -2.4. The ARIMA model was three times better than the lr model and five times better than the knn model in terms of RMSE. All three models performed poorly in R2 score indicating that the data is no better than an average linear prediction. When comparing the predictions, the ARIMA model was clearly the closest to predicting the actual results; followed by linear regression (Fig. 9).

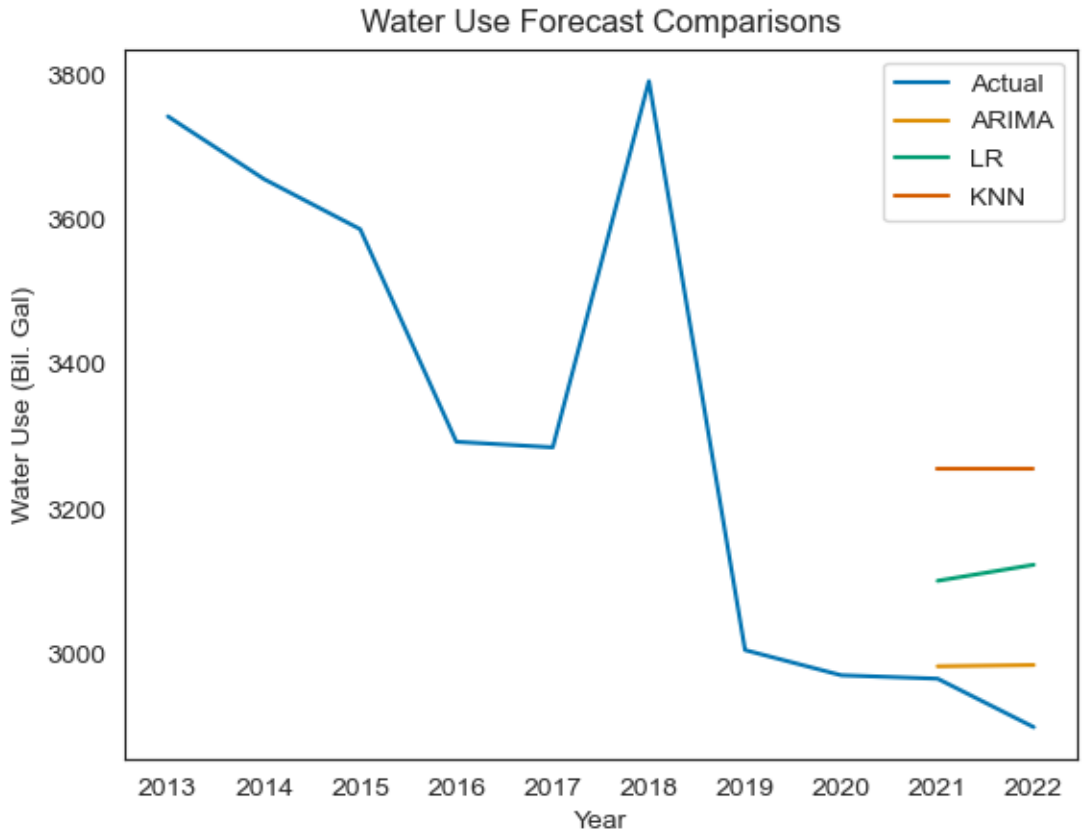

Fig. 9 Water use forecast predictions between ARIMA, Linear Regression, and K-Nearest Neighbors.

Bootstrap Resampling

Before any individual models could be conducted on any of the industries, a well fit model must be achieved. The data was block bootstrap resampled of 10,000 times and modeled using the same ARIMA model parameters. The blocks for the bootstrap were set to 2 to capture the autocorrelation of the data that was discovered earlier.

Both evaluation metrics were strongly skewed so five percent of the data was capped off. The distribution of RMSE for the bootstrap samples was right skewed and weaker than the original modeling (Fig. 10, mean = 280.814, median = 232.681). The R2 distribution for the bootstrap samples were even more variable with a strongly left skewed distribution (Fig. 10, mean = -439.157, median = -1.121). Despite being a marginally better fit model than the original data, the predictions were still no better than an average linear prediction. For these reasons, modeling with this data set further with other response features was determined to be ineffective.

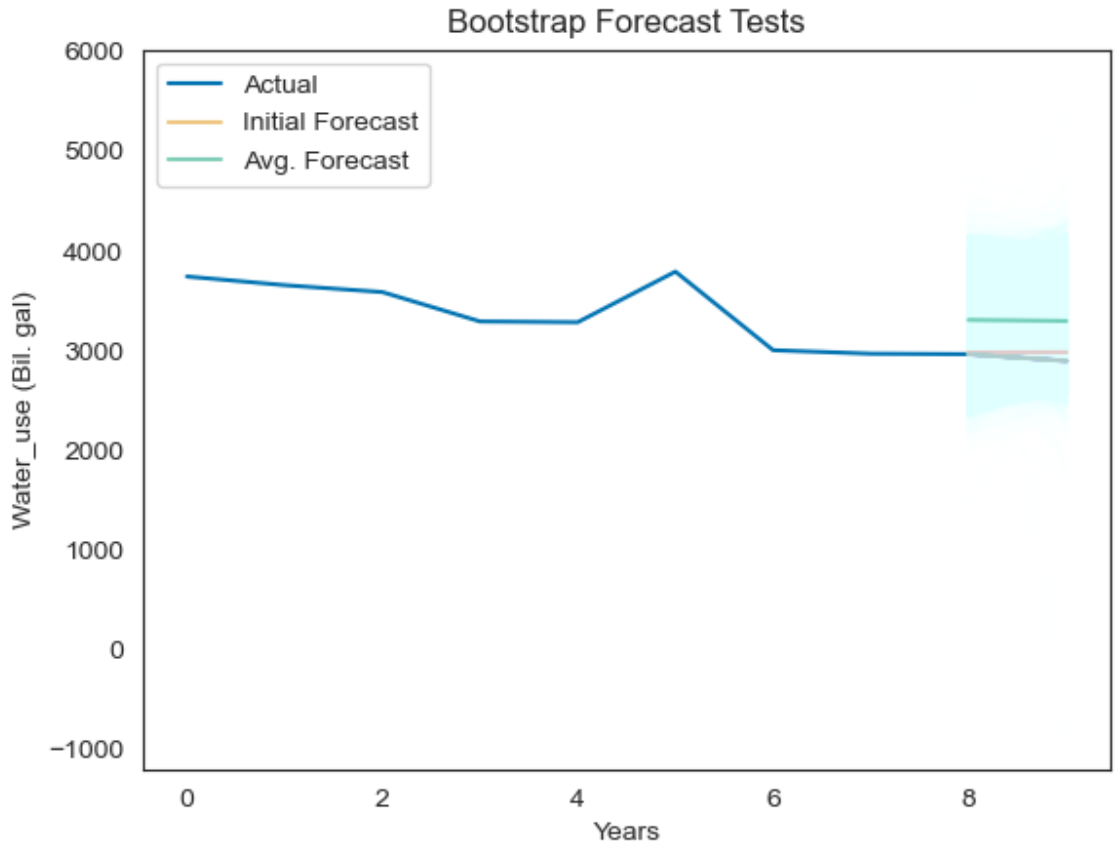

Fig. 10 Bootstrap ARIMA forecast predictions.

Conclusion

On average, the ARIMA model forecasts that water use should increase by 3300 billion gallons in the next ten years (Fig. 11); however poor evaluation metrics indicate that this prediction may be inaccurate. More data surveyed across multiple decades or records of weekly water use would increase the dimensionality for the model to forecast accurate predictions.

Strong linear relationships were established between population, general water use, Great Lake and inland sources, commercial industries, public supply, electrical generation, and general sectors. Although population had a strong correlation during the analysis, it did not contribute to any of the models' predictions. Supporting features do exist within the data, more

observations are just needed to make the model more robust. There existed a strong autocorrelation with a lag between 1 - 2 years, but with such a large confidence interval, the actual lag may have been optimized differently. For the continuation of this project, segmenting the data to weekly water use and population estimates would be easier to determine the lag periods.

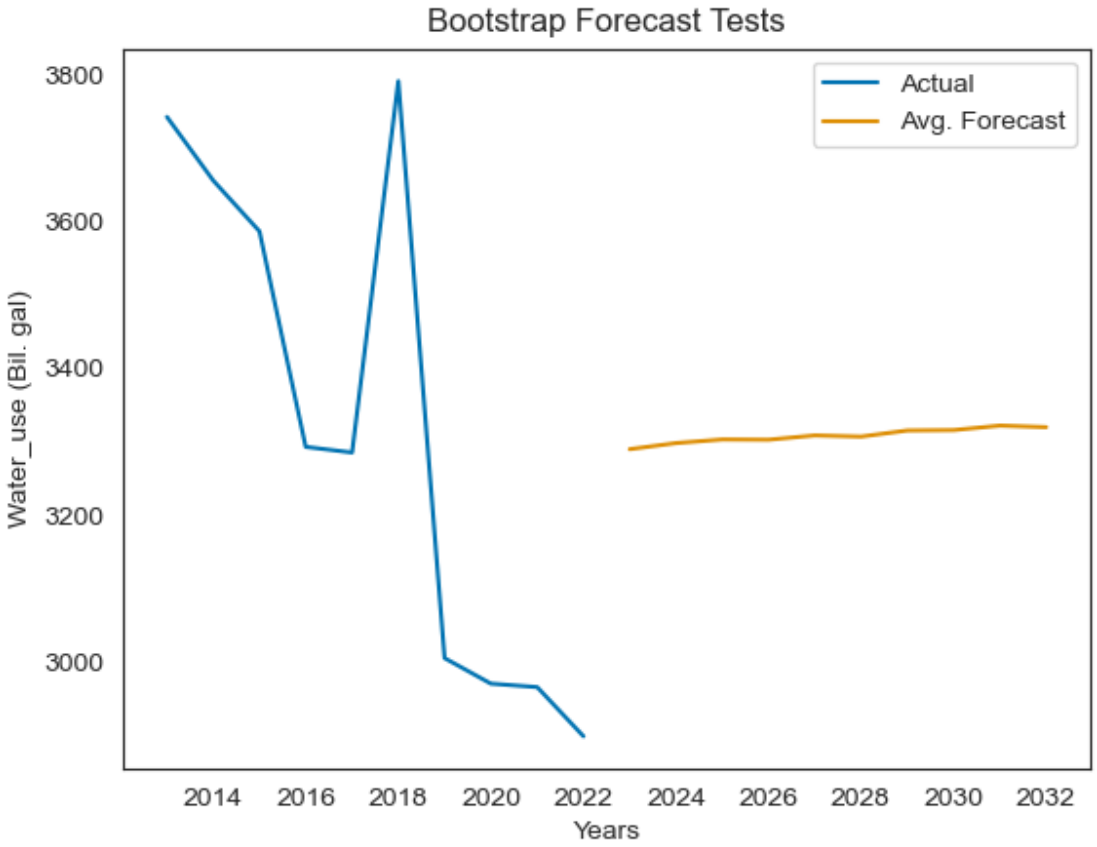

Fig. 11 Average bootstrap forecast of the next 10 years of water use in Michigan.

Future studies will need to archive more than 10 years worth of data to create more robust models. Additionally, creating multivariate models to identify specific industry contributions towards the response variable will be necessary in the future. Multivariate ARIMA techniques will need to be explored more to better aid in this area. Finally, from this model, companies may be able to investigate the relationship between population and the individual industries further to better minimize their impact on water use.

Policy makers should be encouraged to support this work because there were strong correlations that can be investigated further. For example, electrical power

generation was the largest user of water sourced from the Great Lakes. By putting limitations on one of those resources, the other may also become limited. Similarly, the data showed that inland water sources are an underused resource with how little water is being accessed for all industries except industrial manufacturing. By identifying these underlying relationships, water resource monitoring could encourage innovations in the industries of Michigan.

References

Martusiuk, O. (n.d.). Michigan water use data (2013 to 2022) [Data set]. Kaggle. https://www.kaggle.com/datasets/oleksiimartusiuk/michigan-water-use-data-2013-to-2022/data

U.S. Census Bureau. (n.d.). State population totals: 2010-2020 [Data set]. U.S. Census Bureau. https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/state/totals/

U.S. Census Bureau. (n.d.). State population totals: 2020-2023 [Data set]. U.S. Census Bureau. https://www2.census.gov/programs-surveys/popest/datasets/2020-2023/state/totals/

U.S. Environmental Protection Agency. (2023). *Climate change indicators: Great Lakes* [Data set]. https://www.epa.gov/climate-indicators/great-lakes