

Catalogue no. 16-001-M
ISSN 1917-9693
ISBN 978-0-660-75076-7

Environment Accounts and Statistics Analytical and Technical Paper Series

Statistical approaches for estimating industrial water intake in Canada

by Rezvan Taki and Beni Ngabo Nsengiyaremye

Release date: March 27, 2025



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- | | |
|---|----------------|
| • Statistical Information Service | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line | 1-514-283-9350 |

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "[Standards of service to the public](#)."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2025

Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

An [HTML version](#) is also available.

Cette publication est aussi disponible en français.

Statistical Approaches for Estimating Industrial Water Intake in Canada

by **Rezvan Taki** and **Beni Ngabo Nsengiyaremye**

Abstract

The reliable estimation of industrial water use is critical for establishing realistic water conservation goals in Canada's manufacturing sector, mineral extraction industries and thermal-electric power generation sector. To evaluate the predictive accuracy of several statistical models at the national level, this study uses survey data to explore modelling techniques including the eXtreme Gradient Boosting (XGBoost) model, the Thin-Plate Spline (TPSPLINE) model, Multiple Imputation by Chained Equations (MICE), linear regression, partial least squares (PLS) regression, and least absolute shrinkage and selection operator (LASSO) regression. The XGBoost model is found to be the best for predicting water use in the manufacturing sector, while linear regression is most accurate for thermal-electric power generation industry water use. In the mineral extraction industries, PLS regression, the TPSPLINE model and LASSO regression are found to perform well for coal, metal ore mining and non-metallic mineral mining. Finally, the study predicts industrial water use for non-surveyed years (2007 to 2021) to enhance the consistency and quality of national water use data for effective water management planning. The selected models and results provide valuable insights into sustainable industrial water use management in Canada and may be useful for other countries facing similar challenges.

Keywords: Industrial water; regression; Canada.

1 Introduction

Industrial facilities rely on water for their production processes and obtain water from two main sources: groundwater (wells and aquifers) and surface water (lakes and rivers) (Statistics Canada, 2024). The water used in the manufacturing, mining and electrical power generation sectors is referred to as industrial water intake. In mines and thermal-electric power generation plants, water is used to extract raw materials, combine them with other inputs, and cool equipment to produce steam and drive turbines that generate electricity. Similarly, in manufacturing, water is used for various purposes such as cleaning, cooling, sanitation, maintenance and conveying intermediate inputs, and it is embedded in the final product (Bradley, 2017). Mining water withdrawal¹ involves water used for all activities related to mining and extraction; quarrying; and the milling of solids such as coal, metal and non-metallic minerals.

The industrial sector accounted for approximately 19% of the world's water withdrawal, while the agricultural and municipal sectors accounted for 69% and 12% respectively (Food and Agriculture Organization of the United Nations, n.d.). In the breakdown of industrial water use in Canada, thermal-electric power plants are the largest industrial water users, accounting for an average of 84% (from 2005 to 2021), followed by manufacturing firms at 14% and the mining industry at 2%.

However, industrial water use estimation has received less attention than agricultural and domestic water use, potentially because of a lack of data collection in some countries (Kumar, 2004). Data on industrial water use are beneficial for the public, community leaders, policy makers, water managers and other interested stakeholders. Furthermore, the compilation of facilities-based data and information on a range of freshwater issues could help to raise awareness of the overall state of water in Canadian industrial facilities.

Acquiring data through surveys can be expensive. Another option is to model water use based on auxiliary data that are correlated with water intake. However, the quality of modelled data is typically lower and may encompass even higher uncertainties when compared with survey data (Malla et al., 2019). As a result, an economical and

1. The term "water withdrawal" (also known as "water intake") refers to the amount of water extracted from a water resource such as a river, a lake or groundwater for various uses. Water consumption refers to the amount of water permanently removed from a water resource and consumed for production or other activities.

effective approach involves a combination of biennial survey data and the application of modelling techniques during the periods without surveys to fill the gaps. This hybrid method optimizes data compilation and collection costs while ensuring that accuracy levels remain satisfactory.

Water use during periods without survey information can be estimated using various modelling techniques, including a) extrapolation models that utilize past trends of water use, such as simple regression or trend analysis; b) non-economic, multiple-coefficient models that are mathematical functions incorporating auxiliary variables but exclude economic factors, such as water price; c) econometric models that estimate water use based on multiple factors, including weather, economics and demographics; and d) accounting models based on the relationships between different types of water use consumption, water deliveries, discharge and water loss. However, these models can be complex and difficult to address (Templin et al., 1977).

Many variables including water recirculation, energy and labour can serve as auxiliary parameters for industrial water use. The cost of water intake can also be used as an explanatory variable in the modelling of industrial water use. In the econometric modelling of water use, the average cost of water intake can be considered. However, this approach could lead to bias in the regression equation attributable to measurement error or endogeneity, which means the same factor may influence water intake and water cost at the same time. For example, when the price of water increases because of scarcity (Dupont & Renzetti, 2001).

The Industrial Water Survey is crucial in providing reliable and comparable data at a national level in Canada. It serves as a valuable tool for policy and decision makers, as well as for the scientific community. Moreover, obtaining accurate water intake data at the national level is crucial for enhancing the overall quality of global water resources monitoring. The primary objective of this study is to develop a water use estimation method for various industry sectors—including mining (metal, coal and non-metal), thermal-electric power generation and manufacturing—for the years when survey data are not available at the national level. Therefore, this study aims to present appropriate and consistent water-use information for the industries concerned.

In this paper, section 2 describes the methodology used for each industry, then section 3 present the results. They are followed by a discussion of the applied techniques and concluding remarks.

2 Materials and methods

2.1 Industrial water use data

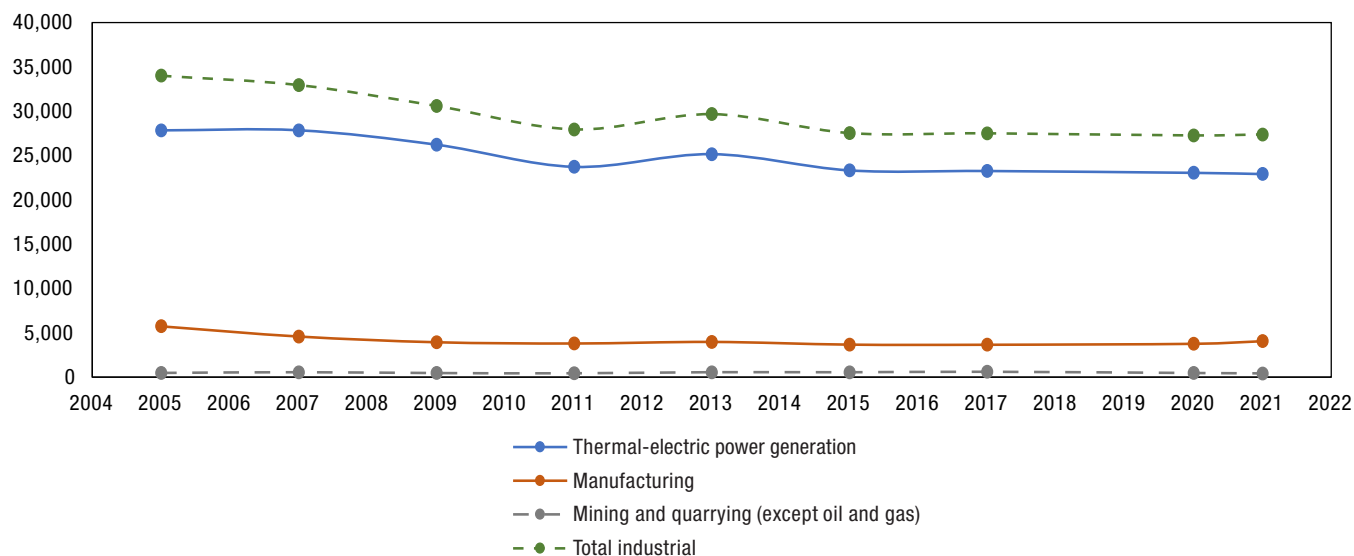
The Biennial Industrial Water Survey (IWS) conducted by Statistics Canada was the primary data source for this study. This biennial survey collects information on water intake, costs, and sources from manufacturing, mining, and power generating facilities across Canada through three individual questionnaires. Since 2005, the survey has been conducted at the national and provincial levels. A census approach is used to obtain water intake data for approximately 100 thermal-electric power generating plants. A probability design is used to sample from the population of 126,431 manufacturing facilities (NAICS 31 to 33) and 871 mining locations (NAICS 2121, 2122 and 2123, excluding 21232).

The IWS data provide valuable information on water use patterns in Canada. According to a recent survey, total industrial water use in Canada was estimated to be 27.36 billion cubic metres in 2021, including all categories of manufacturing, mining and thermal-electric power generating industries. However, water used for the extraction of liquids (such as crude petroleum) and gases (such as natural gas) was not considered in this study.

Chart 1 shows the trend of industrial water intake in the three major industries of mining, thermal-electric power plant and manufacturing. The withdrawals for thermal-electric power generation and other industries have stabilized or decreased because of increasing water use efficiency resulting from technological innovation and recycling efforts over the years (National Research Council, 2002). Although total industrial water withdrawal increased from 27.92 billion cubic metres in 2011 to 29.65 billion cubic metres in 2013, this substantial rise in total water withdrawal was primarily caused by a 6% increase in thermal-electric power generation water intake. This rise during this period was attributable to the return to the operation of nuclear electric power generation facilities in Ontario and New Brunswick. The reduction in thermal-electric power generation water intake after 2013 was because of Ontario's broader strategy to phase out coal-fired power plants (Chart 1).

Chart 1
Industrial water intake in Canada

water intake (million cubic meters)



Source: Table 38-10-0067-01 <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3810006701>.
Table 38-10-0037-01 <https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=3810003701>.

Water intake for the manufacturing industry includes water withdrawal in 18 manufacturing NAICS classifications. The survey was implemented every other year by Statistics Canada, starting in 2005. In this study, available industrial water use data from 2007 to 2021 were used at the national level for the mining and thermal-electric power plant sectors while manufacturing used data from 2005 to 2021. Four major manufacturing industries—primary metal manufacturing, paper manufacturing, food manufacturing, and petroleum and coal product manufacturing—accounted for a significant proportion of the total water intake in the manufacturing sector, averaging approximately 83% annually over the course of the surveyed years from 2007 to 2021. The share of the total water intake attributed to these industries fluctuated slightly across the years. For instance, in 2015, their combined water intake represented around 80% of the total water used in manufacturing. In 2021, the water intake by these four industries was 3.48 billion cubic metres, which equated to approximately 86% of the manufacturing sector's total water intake.

The IWS response rate, defined as the proportion of the total sample that provides usable data for the survey, was high for thermal-electric power plants, ranging from 84% to 100% depending on the year. For the manufacturing sector, the response rate was lower, ranging from 62% to 84%, while the response rate for the mineral extraction industries ranged from 65% to 79% (Government of Canada, 2022).

To model water intake estimates for the non-surveyed years, explanatory variables from different annual surveys were used to estimate Canada's industrial water intake over the non-surveyed years of the 2007-to-2021 period. Summary statistics for all explanatory variables used in the estimation of each industry subgroup are presented in Table 1. In the manufacturing sector, for simplicity, the water intake values from four NAICS classifications—315, 316, 323 and 337—were combined in a single NAICS classification and the estimation was performed for the combined NAICS classifications.

Table 1
Statistics for variables in estimation models for each industry

Industry	Explanatory Variable	Mean	Standard Deviation	Minimum	Maximum
Manufacturings	Energy consumption Gigajoules)	104,629,956	165,657,892	145,068	776,211,157
	Energy use, final demand (kilotonnes)	1,799	392	1,038	2,290
Mining (Coal)	GDP (million dollars)	4,560	798	2,569	5,988
	Energy use (Terajoules)	980,503	16,391	70,172	125,049
Mining (Ore metal)	GDP (million dollars)	13,648	1,672	10,571	16,034
	Kerosene, net supply (Megalitres)	477	187	261	863
Mining (Non-metallic mineral)	light fuel oil, net supply (Megalitres)	2,804	992	1,450	4,531
	Electricity produced (Megawatt hours)	230,229,055	10,260,128	209,685,459	249,705,324
Thermal-electric power plant	Fuel cost (Uranium)(thousands dollars)	342,577	114,246	167,487	454,750

Note : The values listed in the table correspond to the dataset used for modeling each sector from 2005 to 2021. Data that are unavailable or suppressed to meet confidentiality or data quality standards are excluded from the calculated statistics. The mining industry variables contain all variables utilized for estimating coal, ore metal, and non-metallic industries. Specific statistics are obtainable from the author.

Sources: Statistics Canada. Table [25-10-0025-01](#) Manufacturing industries, total annual energy fuel consumption in gigajoules, 31-33. Table [36-10-0434-03](#) Gross domestic product (GDP) at basic prices, by industry, annual average (x 1,000,000). Table [25-10-0030-01](#) Supply and demand of primary and secondary energy in natural units. Table [25-10-0030-01](#) Supply and demand of primary and secondary energy in natural units. Table [38-10-0096-01](#) Physical flow account for energy use. (Retrieved December 20th, 2023). Table [38-10-0109-01](#) Energy use, by sector. Table [25-10-0020-01](#) Electric power, annual generation by class of producer. Table [25-10-0018-01](#) Electric power generation, annual cost of fuel consumed by electric utility thermal plants, inactive (x 1,000). (Retrieved December 18th, 2022).

2.2 Methods

Various statistical models were fitted to each sector and subsector dataset, with model selection based on the complexity of the dataset and the specific characteristics of each industry. For the manufacturing sector, which consists of multiple industries each categorized by a specific NAICS, the water intake modelling proved challenging because of the varying nature of the data across these industries. Therefore, we focused on using MICE and XGBoost, a powerful machine learning algorithm, to model water intake across the sector. This approach was selected because of its ability to handle non-linear relationships and high-dimensional features efficiently, which are common in the manufacturing sector.

For other industries such as thermal-electric power plants and mineral extraction, we used different modelling approaches: linear regression as well as LASSO, PLS and TPSPLINE models. These models were selected based on the simpler structure of these sectors' datasets and the nature of the relationships between the variables, which were better suited for linear or semi-parametric modelling techniques.

The analysis was conducted using statistical packages in R version 3.5.1 and the TPSPLINE model was developed in SAS 9.4. The following sections provide more details about these modelling approaches.

2.2.1 eXtreme Gradient Boosting

The XGBoost model is a machine learning method that predicts manufacturing water intake by using a gradient boosting framework with a decision-based ensemble method. The model is built on making regression trees, which consequently minimizes model error in the way that the new regression tree adjusts the previous regression tree. The final prediction is calculated as the integration of the ensemble model and can be defined as follows:

$$\hat{y}_i = \phi(X_i) = \sum_{(k=1)}^K f_k(X_i), f_{(k)} \in F, i = 1, \dots, n \quad (1)$$

XGBoost improves the objective optimization function by optimizing the loss function and complexity penalty, which are referred to as $\sum_{i=1}^n l(y_i, \hat{y}_i)$ and $\sum_{k=1}^K \Omega f_k$. (Chen & Guestrin, 2016)

2.2.2 Multiple imputation

The MICE algorithm has been applied in many studies to impute all features of a database many times based on a prediction algorithm (Van Buuren, 2007). This method assumes that improving the output of the prediction models is possible by chaining the input variables together. Chaining is defined as the iterative process of using previously imputed variables as input to predict the next variable with missing data in order to improve the prediction accuracy

and to obtain more accurate imputations for all variables in the dataset (Hallam et al., 2022). Hence, the imputation prediction keeps changing until it converges toward a stable solution with the lowest bias (Azur et al., 2011). One requirement of this technique is that the data should be missing at random.

2.2.3 Partial least squares

The PLS regression technique is a standard constructed predictive model when highly collinear explanatory variables exist (Geladi & Kowalski, 1986). In this technique, the relationship between a matrix of predictor variables (X) and the response variable (Y) is explained by latent variables or X-scores (T). The X-scores can explain the maximum amount of variability in X and Y (Gelaye et al., 2023). The equation is as follows:

$$X = \tau P' + \varepsilon \quad (2)$$

in which, τ and P' are the score matrix and loading matrix and ε is the matrix of the X-residuals.

τ can also be calculated by the transformed PLS weights matrix as below:

$$\tau = XW^* \quad (3)$$

Finally, the response (Y) is computed by the Y-weight matrix (C^*) and the related residuals (F).

$$Y = \tau C + F \quad (4)$$

The water intake model for the applicable sector was implemented in RStudio version 2022.07.2+576.

2.2.4 Least absolute shrinkage and selection operator regression

The LASSO regression technique is a penalized linear model that uses shrinkage of data values toward a central point like a mean (L1 regularization). The equation of LASSO regression is as follows:

Residual sum of squares + λ^* (the sum of the absolute value of the magnitude of coefficients)

$$\sum_{i=1}^n (y_i - \sum_j x_{(ij)} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (5)$$

in which λ represents the computed amount of shrinkage while minimizing the residual sum of squares (Owen, 2007).

2.2.5 Multivariate thin-plate spline model

A TPSPLINE model is a non-parametric regression that uses the PLS method to fit a model to the data. Calculation of the thin plate smoothing spline estimates involves selecting one design point from the group and treating all observations as replicates of that design point. The function f , with a penalizing least squares estimate, can be calculated by minimizing the following quantity:

$$\frac{1}{n} \sum_{k=1}^n (y_k - f(x_k))^2 + \lambda J(f) \quad (6)$$

where the first term evaluates the goodness of fit and the second term evaluates the smoothness of f . For more information about thin-plate spline-based fitting algorithm, readers referred to an article by Meinguet (Meinguet, 1979).

2.3 Statistical analysis

To evaluate the performance of predictive models in each economic sector we used the leave-one-out cross-validation technique. In this technique, we put aside sample data (one year from the surveyed years from 2007 to 2021) and train the model on all the remaining data, then we examine the model performance metrics on the excluded data. The best model was obtained using the above methods based on the lowest sum of the squared estimate of error (SSE) and mean absolute percentage error (MAPE).

$$MAPE = \frac{1}{n} \sum_{t=0}^n \left| \frac{A_t - F_t}{A_t} \right| \quad (7)$$

$$SSE = \sum_{t=0}^n (A_t - F_t)^2 \quad (8)$$

Where:

N = number of sample size

A_t = surveyed water intake

F_t = predicted water intake

3 Results and discussion

3.1 Manufacturing sector

Our analysis compared several methods, including multiple imputation and the gradient boosting technique (XGBoost), across 18 NAICS classifications in the manufacturing sector. Using a leave-one-out cross-validation approach for the 2007-to-2021 period, we found that XGBoost consistently outperformed the other multiple imputation. Table 2 summarizes the prediction errors (MAPE and SSE) over the NAICS of the manufacturing industry for the validation years from 2007 to 2021.

Table 2
Comparison of MAPE of the models over the years of 2007-2021 of manufacturing sector

Year	XGBoost		Multiple imputation	
	SSE (MCM) ²	MAPE percent	SSE (MCM) ²	MAPE percent
2007	33,256	31	280,534	51
2009	14,711	98	289,803	109
2011	15,038	85	235,659	98
2013	16,830	46	265,753	55
2015	61,210	28	239,300	69
2017	47,161	48	236,936	56
2020	40,295	71	151,492	167
2021	167,268	59	243,809	138

Note : Abbreviations: SSE sum of squared error; MAPE, mean absolute percentage error; (MCM)², million cubic meters squared.

Source: Authors' computations.

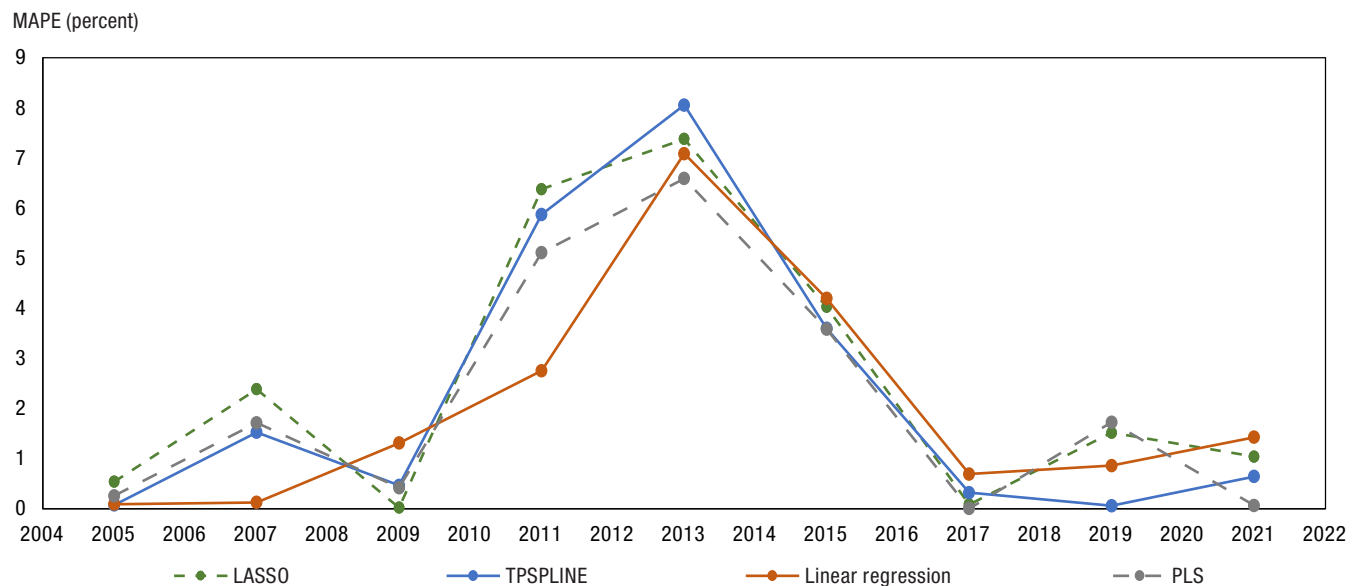
The average MAPE for XGBoost over the years was 58%, compared with 93% for the multiple imputation method. Additionally, XGBoost demonstrated superior performance in terms of model accuracy, as it produced the lowest total SSE over the years. The SSE for XGBoost was consistently lower than that of the multiple imputation method, indicating its better fit to the data.

In both approaches, primary metal manufacturing (NAICS 331) and paper manufacturing (NAICS 322) were the largest contributors to the total sum of squared errors (SSE). In the XGBoost model, NAICS 331 explained 69% of the total SSE, while NAICS 322 accounted for 16%. In the multiple imputation analysis, NAICS 331 explained 25% of the SSE, and NAICS 322 explained 68%. These findings indicate that these two industries are major sources of variability in the water intake data, highlighting the need for further exploration to better understand their impact on the modelling results.

3.2 Thermal-electric power generation industry

For the thermal-electric power generation industry, we developed models that included generated electricity and the cost of fuel (uranium) consumed in electric thermal-electric power plants² as two explanatory variables. For each model (TPSPLINE, PLS, LASSO and linear regression) the MAPE values show a general trend of low error in the earlier years (from 2005 to 2009), with higher error in some later years like 2011 and 2013 (Chart 2). However, the results revealed that TPSPLINE exhibited very high SSE values (up to 4.1 (MCM)²) in 2011, 2013 and 2015. By contrast, LASSO regression presented moderate SSE values without the extreme errors that showed in TPSPLINE and PLS (see Chart 3). Linear regression emerged as the best-performing model based on the analysis of MAPE and SSE. It had the lowest mean MAPE at 2.00% and the lowest total SSE from 2005 to 2021, indicating high accuracy and a good fit to the data.

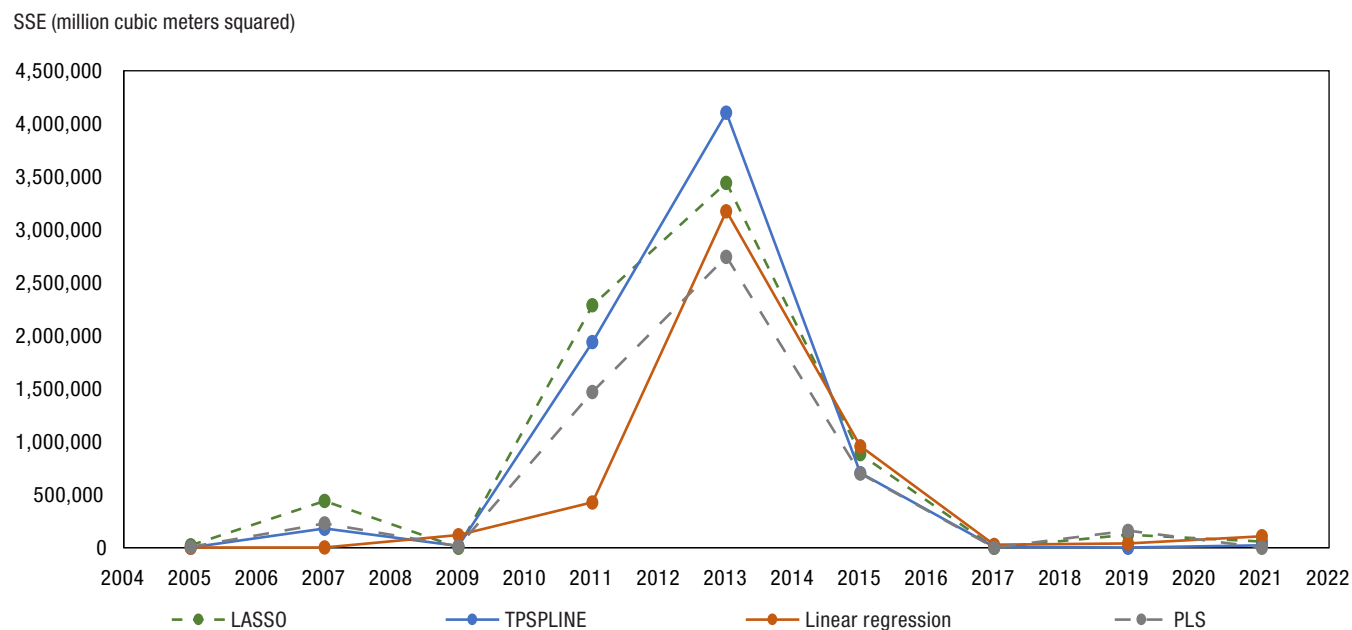
Chart 2
Mean absolute percentage error (MAPE) values for different models in predicting the water intake of thermal-electric power plants



Source: Authors' computations.

2. In the IWS, nuclear power plants are categorized as thermal-electric power plants.

Chart 3
Sum of squared error values for different models in predicting the water intake of thermal-electric power plants



Source: Authors' computations.

3.3 Mineral extraction industries

To estimate water intake in the mineral extraction industries including coal mining, metal ore mining and non-metallic mineral mining, we applied the explanatory variables explained in Table 1 over different approaches including linear regression, PLS regression, TPSPLINE and LASSO regression.

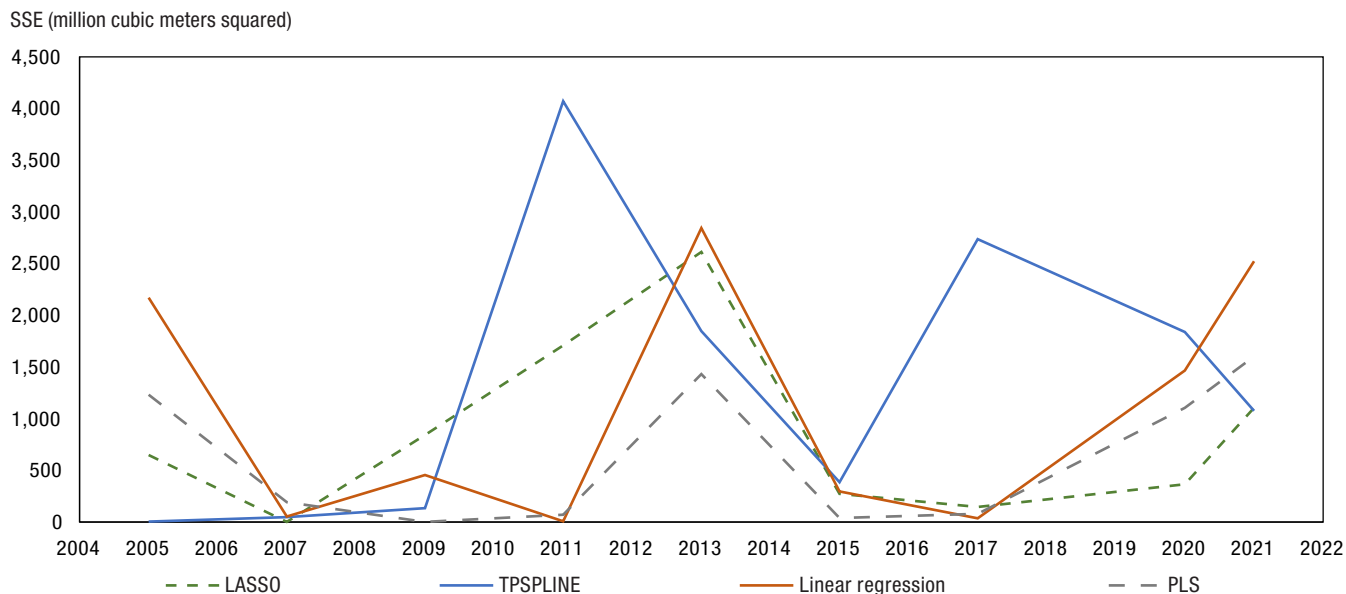
In coal mining modelling, the results show that TPSPLINE produced the highest SSE values in several years—especially in 2011 and 2017—which may be as a result of overfitting or capturing more noise than necessary during these periods. However, LASSO regression exhibited a noticeable increase in SSE values over time, mainly in the later years. For instance, in 2021, the SSE value reached 1,104.8 (MCM)². Linear regression also showed poor performance when there were strong nonlinearities or changing relationships over time. However, PLS showed more stable performance and moderate SSE values in the years with no extreme fluctuations (Chart 4).

In metal ore mining, TPSPLINE generally offered more reliable predictions even though the performance varied by year. The total sum of the SSE value and the average of the MAPE over the years in this method outperformed the other approaches. Additionally, linear regression showed a variable performance pattern. More specifically, it performed poorly in the early years (2005, 2007 and 2009) with very high SSE values. By contrast, linear regression showed significantly improved performance in the later years (2011, 2013, 2015, 2020 and 2021), with SSE values that are much lower and, in some years (like 2013) even perfect. LASSO regression showed moderate to high SSE values in the early years of 2005, 2007 and 2009, but it demonstrates a poor, inconsistent performance in the later years of 2017, 2020 and 2021. PLS regression performed well in some years such as 2007 and 2009, but it struggled in other years, especially in 2017 with the highest SSE value of 28937 (MCM)² (see Chart 5). The average of the MAPE in the three methods of LASSO, linear and PLS regression was similar for the years of 2007 to 2021 at about 24%.

For the non-metallic mineral mining industry, the results showed that LASSO regression performed well compared with other methods, with the lowest total SSE values of 2224.20 (MCM)² and the average MAPE of 5% across all years (Chart 6). Conversely, the TPSPLINE showed significant fluctuations in the MAPE values with 24% and

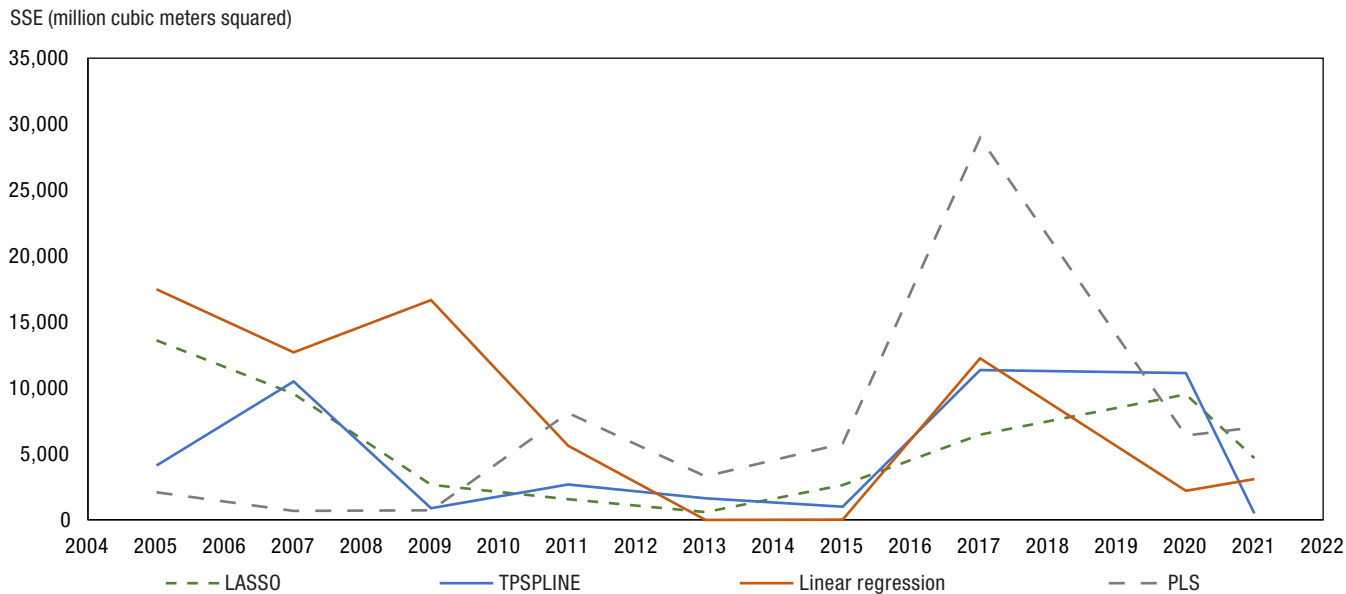
the highest total SSE values across the years (10,959.78 million cubic metres). Linear regression showed variable performance, with relatively low MAPE in some years (e.g., 69% in 2005), but larger errors in others. Its SSE indicates it has fewer large errors than TPSPLINE or PLS, but it still performed worse than LASSO regression.

Chart 4
Sum of squared error (SSE) values for different models in predicting the water intake of mineral extraction industries (coal mining)



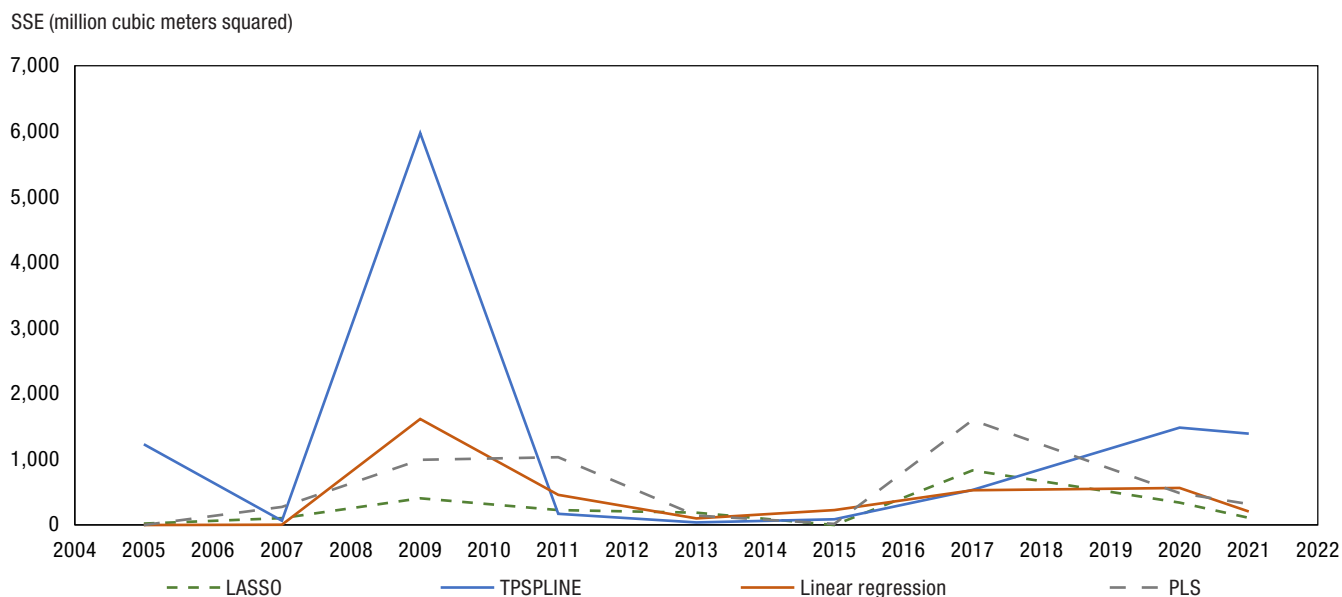
Source: Authors' computations.

Chart 5
Sum of squared error (SSE) values for different models in predicting the water intake of mineral extraction industries (metal ore mining)



Source: Authors' computations.

Chart 6
Sum of squared error (SSE) values for different models in predicting the water intake of mineral extraction industries (non-metallic mineral mining)



Source: Authors' computations.

4 Conclusion

In this paper, we analyzed the accuracy of several statistical models in predicting national water use across various industrial sectors in Canada. Our results indicate that in the manufacturing sector, the Extreme Gradient Boosting method (XGBoost) was the most accurate model for predicting water intake. The significant contribution of primary metal manufacturing and paper manufacturing to the total sum of squared errors (SSE) across the XGBoost and multiple imputation models highlight the critical role these industries play in driving variability in water intake data. The high SSE values associated with these sectors suggest that industry-specific factors or complexities in water use may be influencing model accuracy. Addressing these sources of variability could improve the precision of predictive models. Further research focusing on the exclusive characteristics of these industries is essential for refining water use predictions and enhancing model reliability.

For the thermal-electric power generation industry, the linear regression—based on the consistency, low SSE and low MAPE values—provides stable performance and good accuracy across the years without the extreme variability observed in other models. Even though it assumes a linear relationship, its simplicity and efficiency make it a reliable choice. While LASSO regression showed slightly higher MAPE than linear regression, it performs well in high-dimensional data and avoids overfitting attributable to regularization. It would be the best choice when dealing with datasets that have many features or where feature selection is important. PLS and TPSPLINE can be useful for nonlinear data or high-dimensional problems, but both show more variability and larger errors in some years, making them less consistent and less reliable overall.

For the mineral extraction industries, including coal, ore mining and non-metallic mineral mining, PLS, TPSPLINE and LASSO techniques achieved superior performance compared with the other models, respectively.

Although machine learning demonstrated high accuracy in predicting industrial water use, future research is needed to validate this approach at finer temporal and spatial resolutions, such as at the monthly, provincial or city levels. Moreover, knowledge of water use at the regional monthly scale would allow us to identify trends and variations of water use at lower levels and enhance data management for this crucial natural resource. Furthermore, applying

different techniques, such as econometric approaches, may be necessary to compare the models and examine the structure of industrial water use in greater detail. Future studies may consider the influence of factors such as water price sensitivity and technological innovation over time, because these could gradually affect the modelling results.

Acknowledgement: We sincerely thank Ibrahima Aida Ousmane, Daniel Hurtubise and Martin Hamel for their invaluable contributions. Special thanks to Jenny Watt and Avani Babooram for their editing and feedback, greatly refining this paper. We deeply appreciate Michael Schimpf for defining the project and his support throughout.

References

- Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). [Multiple imputation by chained equations: What is it and how does it work?](#) *International Journal of Methods in Psychiatric Research*, 20(1), 40–49.
- Bradley, M. W. (2017). *Guidelines for preparation of state water-use estimates for 2015*. US Department of the Interior, US Geological Survey.
- Chen, T., & Guestrin, C. (2016). (2016). Xgboost: [A scalable tree boosting system](#). Paper presented at the *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Dupont, D. P., & Renzetti, S. (2001). The role of water in manufacturing. *Environmental and Resource Economics*, 18, 411–432.
- Food and Agriculture Organization of the United Nations. (n.d.). [Water use](#). Retrieved September 29, 2024, from <https://www.fao.org/aquastat/en/overview/methodology/water-use>
- Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1–17.
- Gelaye, K. K., Zehetner, F., Stumpp, C., Dagnew, E. G., & Klik, A. (2023). [Application of artificial neural networks and partial least squares regression to predict irrigated land soil salinity in the rift valley region, ethiopia](#). *Journal of Hydrology: Regional Studies*, 46, 101354. <https://doi.org/10.1016/j.ejrh.2023.101354>
- Government of Canada. (2022). [Surface freshwater use in canada's manufacturing industry, 2017](#). <https://www150.statcan.gc.ca/n1/pub/16-508-x/16-508-x2022001-eng.htm>
- Hallam, A., Mukherjee, D., & Chassagne, R. (2022). Multivariate imputation via chained equations for elastic well log imputation and prediction. *Applied Computing and Geosciences*, 14, 100083.
- Kumar, M. D. (2004). [Roof water harvesting for domestic water security: Who gains and who loses?](#) *Water International*, 29(1), 43–53.
- Malla, R., Sapkota, A., & Prajapati, P. (2019). [Estimation of water use coefficient for assessing industrial water demand of various industries of kathmandu valley](#). *Journal of Environment Science*, 5, 21–26.
- Meinguet, J. (1979). [Multivariate interpolation at arbitrary points made simple](#). *Zeitschrift Für Angewandte Mathematik Und Physik ZAMP*, 30(2), 292–304.
- National Research Council. (2002). *Estimating water use in the united states: A new paradigm for the national water-use information program*. National Academies Press.
- Owen, A. B. (2007). [A robust hybrid of lasso and ridge regression](#). *Contemporary Mathematics*, 443(7), 59–72.
- Statistics Canada. (2024, -03-18). [Industrial water survey, 2021](#). Retrieved September 20, 2024, from <https://www150.statcan.gc.ca/n1/daily-quotidien/240318/dq240318d-eng.htm>
- Templin, W. E., Herbert, R. A., Stainaker, C. B., Horn, M., & Solley, W. B. (1977). [National handbook of recommended methods for water-data acquisition](#). U.S. Government Printing Office. Geological Survey (U.S.). Office of Water Data Coordination.
- Van Buuren, S. (2007). [Multiple imputation of discrete and continuous data by fully conditional specification](#). *Statistical Methods in Medical Research*, 16(3), 219–242.