STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 153–161, DOI 10.2478/stattrans-2022-0021
Received – 06.12.2021; accepted – 05.04.2022

Regression model of water demand for the city of Lodz as a function of atmospheric factors

Czesław Domański[1], **Robert Kubacki**[2]

ABSTRACT

One of the Sustainable Development Goals (Goal 6) set by the United Nations is to provide people with access to water and sanitation through sustainable water resources management. Water supply companies carrying out tasks commissioned by local authorities ensure there is an optimal amount of water in the water supply system. The aim of this study is to present the results of the work on a statistical model which determined the influence of individual atmospheric factors on the demand for water in the city of Lodz, Poland, in 2010-2019. In order to build the model, the study used data from the Water Supply and Sewage System Company (Zakład Wodociągów i Kanalizacji Sp. z o.o.) in the city of Lodz complemented with data on weather conditions in the studied period. The analysis showed that the constructed models make it possible to perform a forecast of water demand depending on the expected weather conditions.

Key words: water demand, atmospheric factors, regression model.

1. Introduction

As the global climate changes and the urban population continues to grow, water resources in many of the world's cities are likely to be under increasing stress from reduced water supply and increased demand (Bates et al., 2008).

There have been several studies investigating the role of weather and climate variables in municipal water consumption (e.g. Balling and Gober, 2006; Ghiassi et al., 2008).

Previous studies used maximum and minimum temperatures and precipitation as explanatory variables to estimate water consumption. In addition, the interactions among different weather and climate variables that influence water use are not well understood (Praskievicz and Chang, 2009).

[1] Institute of Statistics and Demography, University of Lodz, Poland. E-mail: czedoman@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-6144-6231.

[2] Poland. E-mail: robertkubacki@o2.pl. ORCID: https://orcid.org/0000-0003-0591-9529.

The aim of this study is an attempt to verify the hypothesis as to whether weather factors can better describe the phenomenon of water demand for the city of Lodz, Poland. Water is needed by everyone. Correctly predicting its demand is important to achieve two opposing objectives. Firstly, its quantity should be sufficient to satisfy the city's needs. Secondly, it should not be wasted. The data obtained clearly show that daily water demand varies from day to day, week to week and month to month. This is compounded by trends related to changing behaviour of the population and other users of the water supply system. In this study it was possible to obtain the total number of cubic metres pumped per day into the system. Data on individual consumers (households, industry, education, health) are only available in an aggregated form. Nevertheless, when reading this study, we should be aware that households are responsible for the consumption of 69% of the total volume of water in the city.

2. Regression models

2.1. Multiple regression

A multiple regression model is written as:

$$y_i = \beta_0 + \sum_{j=1}^{d} x_{ij}\beta_j + \varepsilon_i, \ \ i = 1,2,\dots n, x \ \in \boldsymbol{R}^d \qquad \varepsilon_i \sim N(0,\sigma^2), \tag{1}$$

where β_0 corresponds to the intercept, $\beta_1,\dots,\beta_d$ correspond to the model coefficients, x_i to the observation/measurement data, and ε to the residuals.

The objective function for the residual sum of squares is written as

$$\mathcal{L} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - f(x_i;\beta))^2, \tag{2}$$

By plugging in the regression model equation from above we get

$$\mathcal{L} = \sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{d} x_{ij}\beta_j)^2, \tag{3}$$

where n corresponds to the number of observations and d corresponds to the number of features of the data set (Walesiak and Gatnar, 2009).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(pred_i - obs_i)^2}{n}} \tag{4}$$

The RMSE is the square root of the sum of the squared difference between the observed and predicted values, normalized by the number of observations n.

The lower RMSE the better the model fits the data (Géron, 2017).

Overfitting reduces the generalization properties of a model. When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance; hence, the values of the coefficients become huge. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients, this problem is alleviated (Harrington, 2012). Regularization methods constrain the model parameters in some way and thus are suitable to prevent overfitting.

In many regularization models an additional term is added to the optimization function for the optimal parameter estimates $\hat{\beta}_{opt}$.

$$\hat{\beta}_{opt} = arg\min\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda g(\beta) \tag{5}$$

where g is a function of the coefficients β, which encourages the desired properties of β, and λ is a regularization parameter.

2.2. Ridge regression

Ridge regression, sometimes referred to as $\mathcal{L}_2$ - regularized regression, is a method to shrink the regression coefficients by imposing a penalty on their size. The Ridge regression uses a squared penalty on the regression coefficient vector β (Patterson and Gibson, 2018).

$$\beta_{RR} = \text{arg }\min\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2 \tag{6}$$

Here, $\lambda > 0$ is a regularization parameter that controls the amount of shrinkage: the larger the value of $\lambda > 0$, the greater the amount of shrinkage. The coefficients are shrunk toward zero but do not reach zero. If $\lambda \to 0$ the parameter estimates β_{RR} approach the parameter estimates of the least-square solution β_{LS}.

$$Case\ \lambda \to 0 : \beta_{RR} \to \beta_{LS} \tag{7}$$

$$Case\ \lambda \to \infty : \beta_{RR} \to \vec{0} \tag{8}$$

We can solve the ridge regression problem using exactly the same procedure as for least squares,

$$\mathcal{L} = \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|^2 = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta^T\beta \tag{9}$$

First, take the gradient of $\mathcal{L}$ with respect to β and set to zero,

$$\nabla\mathcal{L} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\beta + 2\lambda\beta = 0 \tag{10}$$

Then, solve for β to find that

$$\beta_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{11}$$

where **I** corresponds to the identity matrix.

2.3. LASSO regression

The LASSO (least absolute shrinkage and selection operator), also referred to as $\mathcal{L}_1$-regularized regression, is a shrinkage method like the ridge regression, with subtle but important differences. The LASSO estimate is defined by

$$\beta_{LASSO} = \arg\min||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta|| \tag{12}$$

where

$$\lambda > 0, \tag{13}$$

$$||\beta|| = \sum_{j=1}^{d}|\beta_j| \tag{14}$$

The LASSO method performs both regularization and variable selection. During the LASSO model fitting process only a subset of the provided features is selected for the use in the final model. The LASSO forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients. In contrast to the ridge regression, which can be solved analytically, numerical optimization (e.g. coordinate descent) is warranted to find the solution for the LASSO regression (Grus, 2018).

The degree of regularization depends on the regularization parameter λ. Thus, it is useful to evaluate the regression function for a sequence of λ.

3. Reference data

For the water demand analysis, data obtained from the Water Supply and Sewerage Works in the city of Lodz were used. The data from the period 2010-2019 included the amount of water injected into the water supply system each day. The set contains 3652 observations. Weather data obtained from www.ogimet.com were used as explanatory variables. The data contain a summary of the weather condition for all weather stations available on the website. Data from the station closest to the place of water intake for the city of Lodz were used for the study. The features comprising the weather description included: temperature (maximum, minimum, average), dew point temperature, humidity, wind direction, intensity and gust, atmospheric pressure, precipitation, cloud cover, sunshine, horizontal visibility and snow cover. Weather data can be obtained free of charge, but obtaining a complete set of data required writing a program in VBA to retrieve data cyclically after 50 observations.

4. The model estimation

Raw data on pumping volumes and weather factors were combined and subjected to preliminary analysis. Missing data were filled in. Filling in data gaps to preserve as many observations as possible for modelling concerned only weather data.

In addition, this concerned, e.g. the amount of snowfall during holiday periods, which were marked with a "-" and were replaced with a value of 0. From the preliminary observation of the data it can be concluded that the amount of water pumped to the water supply systems in the city of Lodz is decreasing every year. The variable describing YEAR takes the values 1, 2, 3, ..., 10 for successive years of observation 2010, 2011, 2012, ..., 2019. Moreover, it was possible to observe that the amount of pumped water changes depending on the month. The lowest average value of pumped water per month is observed in August. For this purpose, a set of zero-one variables was created for each month of the year with August omitted to prevent collinearity between the variables. Also for the days of the week it was observed that the average amounts of pumped water differ. On Sundays, on average, the least water is pumped into the system. This resulted in dedicated zero-one variables describing the days of the week except Sundays. When observing the outlier variables, it was possible to observe that the lowest amounts of pumped water fall on public holidays. For this purpose, the variable SWIETO was created, which takes the value 1 if the following holidays were celebrated on that day: 1st of January, Easter and Easter Monday (movable holidays), 1st and 3rd of May, 15th of August, 1st and 11th of November, 25th and 26th of December.

Other variables used to build the models are presented in Table 1.

Table 1. Other variables used to estimate the models

Variable name	Description
T_MAX	maximum temp. obs. over a 24h period for a given weather station
T_MIN	minimum temp. obs. over a 24h period for a given weather station
T_AVG	average temp. obs. over a 24h period for a given weather station
DEW_POINT	dew point – temp. below which water vapour starts condensing. Expressed in degrees Celsius
HUMIDITY	humidity of the air; it takes values from 0 to 100
WIND_SPEED	wind speed (km/h)
WIND_GUST	wind gusts (km/h)
ATM_PRESSURE	atmospheric pressure, at sea level (hPa)
PRECIPITATION	total precipitation in the last 24 hours (mm)
CLOUD_COVER	total cloud cover
CLOUD_LOW	low cloud cover
SUNSHINE	number of hours of sunshine in the last 24 hours (hours)
VISIBILITY	visibility expressed in km
SNOW	total snowfall in centimetres in the last 24 hours
T_MAX4	zero-one variable taking value 1 for a max. temperature greater than 29 degrees Celsius
HOLIDAY_M1	zero-one variable with value 1 if the day before was a holiday
HOLIDAY_M2	a zero-one variable with value of 1 if there was a holiday two days before

Source: own calculations.

The dataset was split into two subsets. The 2019 data were left as a test set (it was not used in any of the model building stages). Data from 2010-2018 were used for model estimation.

Five competing predictive models were built. The first containing only intercept. The second model with explanatory variables produced only from calendar and holiday variables. The third model was enriched with weather variables. The fourth model used type-one regularization (ridge regression) and the fifth model with type-two regularization (lasso regression). The use of regularisation methods still ensures an easy interpretation of the results while reducing the variance of the random component.

All estimated models were compared with a common measure of RMSE.

The results for the first model with intercept are presented in Table 2.

Table 2. Estimated parameter of model (1) with intercept

Parameter	Estimate	Std. error	P(>|t|)
(Intercept)	111 150	159.3	<2E-16

Source: own calculations.

All estimated parameters in other models are statistically significant and the sign of the estimate is as expected.

RSME measure was used to compare the forecasting performance of the models. The calculated RMSE values for the learning set and the test set for all models are shown in Table 3.

Table 3. The calculated RMSE values for the learning set and the test set in the constructed models

Model	Train RMSE	Test RMSE
Model 1 (Intercept)	9127	8318
Model 2 (Calendar & Holiday)	6083	8899
Model 3 (Weather)	5608	7892
Model 4 (Ridge regression)	5326	7258
Model 5 (Lasso regression)	5329	7294

Source: own calculations.

Comparing the data presented in Table 4, we can conclude that model 1 predicts water demand better than model 2. The inclusion of weather variables in model 3 improves RSME on both the learning set and the test set. Even better results are obtained when using regularization methods (lasso and ridge). Finally, model 4 (ridge regression) was selected as the best model.

The results for the fourth model (ridge regression) are presented in Table 4.

Table 4. Estimated parameters of model (4) – ridge regression

Parameter	Description	Estimate
(Intercept)	(Intercept)	80767.520942
YEAR	Year	-2147.757928
JANUARY	January	9010.219430
FEBRUARY	February	10145.548916
MARCH	March	10397.937947
APRIL	April	8995.271739
MAY	May	7495.161123
JUNE	June	8566.800711
JULY	July	1604.157143
SEPTEMBER	September	5924.705100
OCTOBER	October	8775.879775
NOVEMBER	November	9511.791982
DECEMBER	December	10101.703019
HOLIDAY	Holiday	-11239.859751
MO	Monday	4687.873426
TU	Tuesday	5558.797385
WE	Wednesday	5992.229451
TH	Thursday	6114.855399
FR	Friday	5087.399054
SA	Saturday	4227.928486
T_MAX	Max. temperature	-209.509809
T_MIN	Min. temperature	-215.081234
T_AVG	Avg. temperature	1098.767247
DEW_POINT	Dew point	-554.396231
HUMIDITY	Humidity	70.395127
WIND_SPEED	Wind speed	-26.091345
WIND_GUST	Wind gust	-7.264865
ATM_PRESSURE	Atmospheric pressure	20.991359
PRECIPITATION	Precipitation	-31.693193
CLOUD_COVER	Cloud cover	-474.172728
CLOUD_LOW	Low Cloud cover	46.187129
SUNSHINE	Sunshine	-14.184110
VISIBILITY	Visibility	121.005270
SNOW	Snow	259.730182
T_MAX4	1 if temp. exceeds 29 Celsius degrees	4266.155406
HOLIDAY_M1	Holiday (day before)	-6052.783226
HOLIDAY_M2	Holliday (2 days before)	-3259.093499

Source: own calculations.

The best performance of the objective function in the ridge regression model was obtained for parameter $\lambda = 0.1$. R^2 coefficient in this model is 0.6594.

Best λ estimation, which minimizes the residuals (difference between observations and predicions), was achieved by recalculating 100 models with different values of λ.

In Figure 1 we observe calculated mean square error values for selected $\log(\lambda)$ values.

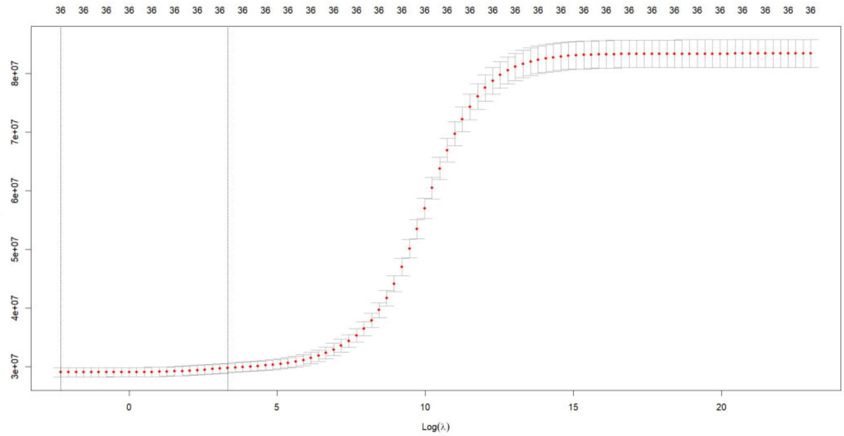

Figure 1. Mean square error values for $\log(\lambda)$ values used in the ridge regression model

5. Conclusions

This study examined the relation between daily weather variables and water use in the city of Lodz, Poland. Similar to previous studies, we found that maximum daily temperature is a good predictor of water demand. We also found that holidays are significant in decreasing the water demand. Moreover, like wind speed is a good predictor of water demand. It is likely that higher wind speed increases evaporation of water, which induces a cooling effect and thus decreases daily water consumption. Together, all these variables explain between 65% of the variations in the city of Lodz. Relatively similar results (up to 61% of the variations explained) were achieved by other authors using ARIMA model (Praskievicz and Chang, 2009).

Further models will also incorporate non-climatic variables such as sociodemographic, prices or structural variables (Zhang and Brown, 2005), which provide our models with greater explanatory power.

Acknowledgement

This paper was presented on the 39th International Scientific Conference “Multivariate Statistical Analysis” (MSA 2021), which financed its publication.

References

Balling, R. C., Jr. Gober, P., (2006). Climate variability and residential water use in the city of Phoenix, Arizona. *Journal of Applied Meteorology and Climatology*, Vol. 46, pp. 1130–1137.

Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P., (2008). Climate Change and Water: *Technical Paper of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC Secretariat.

Géron, A., (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow, Boston: O'Reilly.

Ghiassi, M., Zimbra, D. K. and Saidane, H., (2008). Urban water demand forecasting with a dynamic artificial neural network model. *Journal of Water Resources Planning and Management*, Vol. 134, pp. 138–146.

Grus, J., (2018). Data science od podstaw, Katowice: Helion.

Harrington, P., (2012). Machine Learning in Action, Shelter Island: Manning.

Patterson, J., Gibson, A., (2018). Deep Learning. *Praktyczne wprowadzenie*, Katowice: Helion.

Praskievicz, S., Chang, H., (2009). Identifying the Relationships Between Urban Water Consumption and Weather Variables in Seoul, Korea, *Physical Geography,* Vol. 30, pp. 324-337.

Walesiak, M., Gatnar, E., (2009). Statystyczna analiza danych z wykorzystaniem programu R, Warszawa: PWN.

Zhang, H. H. Brown, D. F., (2005). Understanding urban residential water use in Beijing and Tianjin, China. *Habitat International*, Vol. 3, pp. 469–491.