Forecasting water consumption in towns of different sizes

Adam Piasecki[1*], *Agnieszka* Pilarska[2], and *Radosław* Golba[2]

[1]AGH University, 30-001 Cracow, Poland
[2]Nicolaus Copernicus University, 87-100 Toruń, Poland

Abstract. The aim of the work was to compare water consumption forecasting in two towns of different sizes. The objects of research were the town of Toruń and the town of Żnin in central Poland. Two models were built for each. The models were constructed using the multiple regression method. In constructing the models, explanatory variables determined by Principal Component Analysis (PCA) were used. The set of explanatory variables identified to construct each individual model differed. The models for Toruń obtained better forecast quality assessment criteria values. This was mainly due to the water supply system in the small town (Żnin) being less resilient to sudden, short-term changes in consumers' water use. At the same time, the importance of the location of the meteorological stations from which data was taken to build the model was emphasised.

1 Introduction

The infrastructure for supplying water to urban residents is classed as critical. Therefore, even at the design and construction stage, it is necessary to consider all conditions that may negatively impact its operation. One basic requirement is to quantify water demand in a prospective period, taking into account the needs of all end-users. In practice, this is done using guidelines that provide a normative measure of water demand. Research results show that real water consumption in many towns deviates significantly from the assumed design values [1, 2]. Thus, water consumption forecasts are an important additional element for facilitating the rational operation of water supply systems. The literature has presented many methods and solutions in this domain [3–5]. At the same time, it is impossible to indicate the best method, due to limitations in the ability to compare forecast results (the use of diverse measures of forecast accuracy). Local conditions and the unique characters of objects of study (including size of water supply system) can be decisive in determining results, so are also of significant importance. The aim of the work is to compare water consumption forecasts in two towns of different sizes. This was done using a multiple regression model. A set of explanatory variables was first defined using Principal Component Analysis (PCA).

[*]Corresponding author: adm.piasecki@gmail.com

2 Research area

The objects of study were the town of Toruń and the town of Żnin in the Kujawsko-Pomorskie voivodeship in central Poland. Toruń has over 200,000 inhabitants and a very well developed water and sewage infrastructure. The length of the water supply system is currently 621.2 km, and the sewerage network is 654.2 km long. At present, almost all the town's residents are connected to the municipal water and sewage system [6]. Żnin is much smaller – a town of only 14,000 residents. In recent years, a number of investments have been made in water and sewage infrastructure in the town. This has increased the availability to this infrastructure. Currently, approximately 97% of the town's population is connected to the water supply and 86% to the sewage system. In both towns, industrial use of mains water is proportionally small.

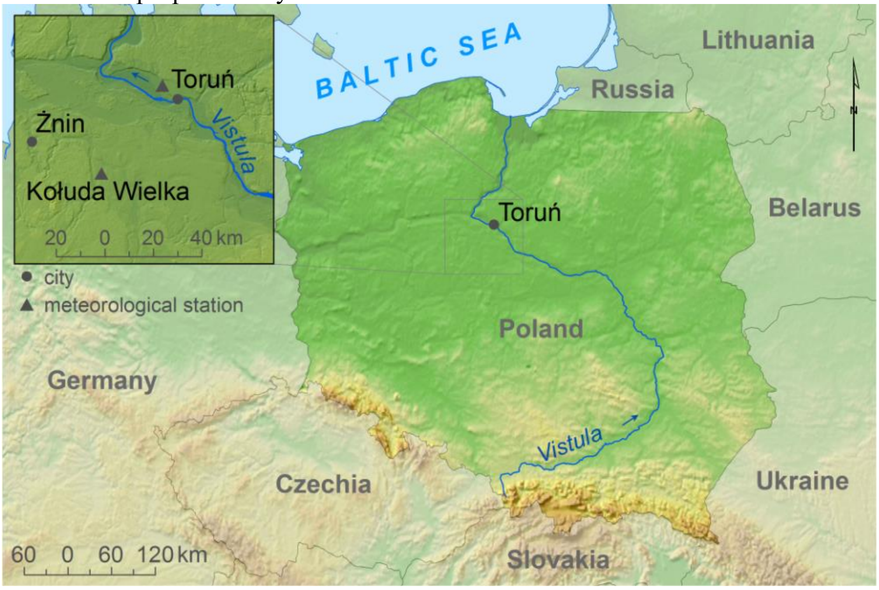


Fig. 1. Location of research objects

3 Materials and methods

The study uses daily water consumption values for the towns being studied. The data was made available by the companies Torun Waterworks Company Ltd and Zakład Wodociągów i Kanalizacji "WIK" Sp. z o. o. The study also used daily values of selected meteorological parameters recorded at the Toruń-Wrzosy and Kołuda Wielka stations. These meteorological stations belong to the Institute of Meteorology and Water Management - National Research Institute (IMWM-NRI). Data was used for the period 2011–2017.

The research progressed in three main stages:

1. determining explanatory variables,
2. building the models,
3. forecasting, and evaluating the models.

A set of explanatory variables was determined using Principal Component Analysis (PCA) based on the procedure presented by M. M. Haque et al. [7]. To determine the explanatory variables, data from 2011–2016 were used. The analysis adopted a correlation matrix. A Bartlett sphericity test was conducted and the Kaiser–Mayer–Olkin coefficient (KMO) was determined, which confirmed the validity of using the PCA method for the whole period and for warm periods (April–September). The KMO coefficient of data from the cold periods (Jan–March and October–December) for Żnin was below the threshold value of 0.500(defined by Hutcheson and Sofroniou [8]), and only slightly exceeded it for Toruń (0.502). For this reason cold periods were not subjected to a separate detailed analysis.

After performing the calculations, it was decided that only two components would be taken into account in determining the explanatory variables: PC1 and PC2. Principal component 3 (PC3) was omitted despite having an eigenvalue above 1.000 (the criterion of H. F. Kaiser [9]). A. Balicki [10] emphasises the importance of the ability to interpret components in determining their number. In this case, including PC3 would have prevented the results from being properly interpreted.

According to the procedure presented by M. M. Haque et al. [7], the set of explanatory variables was determined based on the following parameters:
– selecting the variables with the highest correlation coefficients for a given component (variable loadings).
– attempting to avoid the multicollinearity problem in the regression model by selecting mutually uncorrelated variables.
– attempting to ensure diversity in the character of the selected variables, e.g. “previous-day water consumption” is not a meteorological variable, and so was included in the construction of the explanatory models.

The models were constructed using the multiple regression method. In constructing the models, explanatory variables determined by Principal Component Analysis (PCA) were used. Two models were built for each. The first model, “Year”, was built using daily data for 2011–2016. The second model, “Hot Periods”, was built using daily data for the warm months (April–September) of 2011–2016. The models were then used to predict water consumption in 2017. The following indices were used to assess the quality of forecasts:

$$E = 1 - \frac{\sum_{i=1}^{n}(A_t - F_t)^2}{\sum_{i=1}^{n}(A_t - A_{mean})^2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2}$$

$$\mathrm{NRMSE} = \frac{RMSE}{\frac{1}{n}\sum_{i=1}^{n} A_t}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where: A_t – observed values, F_t – predicted values.

Statistical analyses were performed in PS IMAGO 5 software using the IBM SPSS Statistics analytical engine and in Statistica.

4 Results

The Principal Component Analysis results are presented in Tables 1 and 2. In the case of both the town and the town, the two principal components identified for both the entire analysed period and the hot half-year explained a total of 60% of variance. It should also be emphasised that while both components combined explained a similar percentage of the variance in each town, the amount of the variance explained by PC1 was more than double that of PC2 (Table 1).

Table 1. Eigenvalues and percentage of explained variance.

Town	**Żnin**				**Toruń**			
Model	**Year**		**Hot Periods**		**Year**		**Hot Periods**	
PC No.	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**
Eigenvalue	3.333	1.151	2.947	1.347	3.357	1.186	2.994	1.441
Explained variance [%]	47.62	16.45	42.11	19.24	47.96	16.94	42.77	20.58

For the entire period, the highest variable loadings of the PC1 component were attained by variables representing thermal conditions. A high loading was also attained by the "previous-day water consumption" variable. For PC2, the highest loadings were for "daily sum of precipitation" and "average daily relative humidity" (Table 2). Considering all the mentioned criteria for selecting explanatory variables, the following were selected:

– for Żnin: "average daily temperature", "previous-day water consumption", and "daily sum of precipitation".

– for Toruń: "maximum daily temperature", "previous-day water consumption", and "daily sum of precipitation".

Table 2. Correlation of the two distinguished components with the analysed variables in 2011–2016 and in hot periods of 2011–2016.

Variable	**Model Year**				**Model Hot Periods**			
	Żnin		**Toruń**		**Żnin**		**Toruń**	
	PC1	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**
Minimum daily temperature [°C]	0.934	0.184	0.900	0.257	0.900	0.132	0.827	0.397
Maximum daily temperature [°C]	0.972	0.074	**0.978**	0.078	0.953	-0.034	0.952	0.042
Average daily temperature [°C]	**0.976**	0.120	0.976	0.137	**0.984**	0.030	**0.969**	0.136
Average daily relative humidity [%]	-0.561	0.603	-0.651	0.552	-0.182	**0.833**	-0.405	**0.785**
Daily sum of precipitation [mm]	0.161	**0.695**	0.109	**0.684**	0.196	0.464	0.086	0.635
Average daily wind speed [m/s]	-0.080	-0.359	-0.100	-0.357	-0.218	-0.483	0.084	-0.294
Water use (previous day)	**0.465**	-0.351	**0.438**	-0.440	**0.376**	-0.429	**0.535**	-0.396

For warm periods, the variables that met the adopted criteria for PC1 were "average daily temperature" and "previous-day water consumption", while only "mean daily relative humidity" met the criteria for PC2 (Table 2). As a result, for the construction of the explanatory model using the multiple regression method, the same variables were adopted for both towns: "average daily temperature", "previous-day water consumption" and "average daily relative humidity".

Table 3 presents the water consumption models that were constructed and the selected statistics characterising them. The highest R^2 coefficients were obtained by model II, and the lowest by model IV. In all models the most important variable was variable x_1 ("previous-day water consumption"). It explained 54–66% of the variance of the dependent variable. The remaining variables were of minor importance, except variable x_4 in models II and IV.

Table 3. Models of water consumption and their selected statistics (x_1 – previous-day water consumption, x_2 – maximum daily temperature, x_3 – daily sum of precipitation, x_4 – average daily relative humidity, x_5 – average daily temperature).

Model	Variable	Statistics				
		R^2variance	Partial correlation	Semipartial correlation	Tolerance	p
M1 (Toruń – year)	$W_{T\_year} = 7640.52+0.779x_1+31.307x_2-79.938x_3$					
	x_1	0.650	0.142	0.083	0.892	0.000
	x_2	0.007	-0.158	-0.093	0.991	0.000
	x_3	0.007	0.787	0.740	0.899	0.000
M2 (Toruń – hot period)	$W_{T\_hot} = 18984.14+0.65x_1-94.94x_2+45.18x_3$					
	x_1	0.606	0.715	0.575	0.788	0.000
	x_4	0.074	-0.415	-0.257	0.821	0.000
	x_5	0.003	0.096	0.054	0.848	0.001
M3 (Żnin – year)	$W_{Z\_year} = 329.94+0.792x_1+1.956x_2-3.748x_3$					
	x_1	0.661	0.793	0.749	0.896	0.000
	x_5	0.003	0.115	0.067	0.877	0.000
	x_3	0.004	-0.103	-0.059	0.977	0.000
M4 (Żnin – hot period)	$W_{Z\_hot}= 749.60+0.687x_1-3.103x_2+2.187x_3$					
	x_1	0.542	0.704	0.650	0.896	0.000
	x_4	0.025	-0.229	-0.155	0.938	0.000
	x_5	0.002	0.071	0.047	0.927	0.018

Models M1 and M2 built for Toruń have the best forecast accuracy according to the selected evaluation criteria (Table 4). M2 was assessed as having forecast quality better than M1,butonly according to criterion E. At the same time, a slight difference in value (between M1 and M2) for the other criteria should be noted. All the criteria indicated much weaker forecast quality in the M3 and M4 models.

Table 4. Forecast quality indices according to the selected criteria.

Model	Criterion			
	E	RMSE	NRMSE	MAPE [%]
M1 (Toruń – year)	0.418	1748.295	0.050	3.542
M2 (Toruń – hot period)	0.513	1886.063	0.053	4.099
M3 (Żnin – year)	0.289	141.115	0.099	6.958
M4 (Żnin – hot period)	-0.085	166.968	0.111	7.979

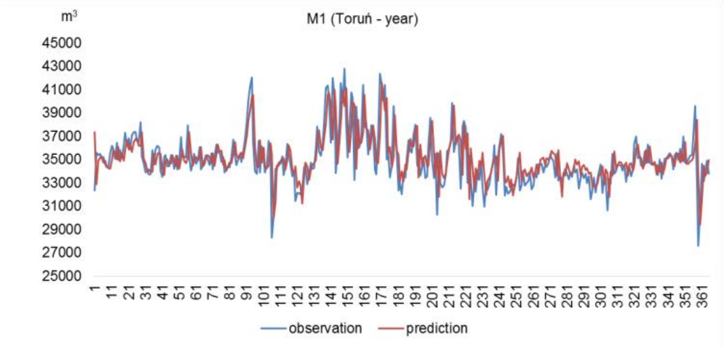
5 Discussion

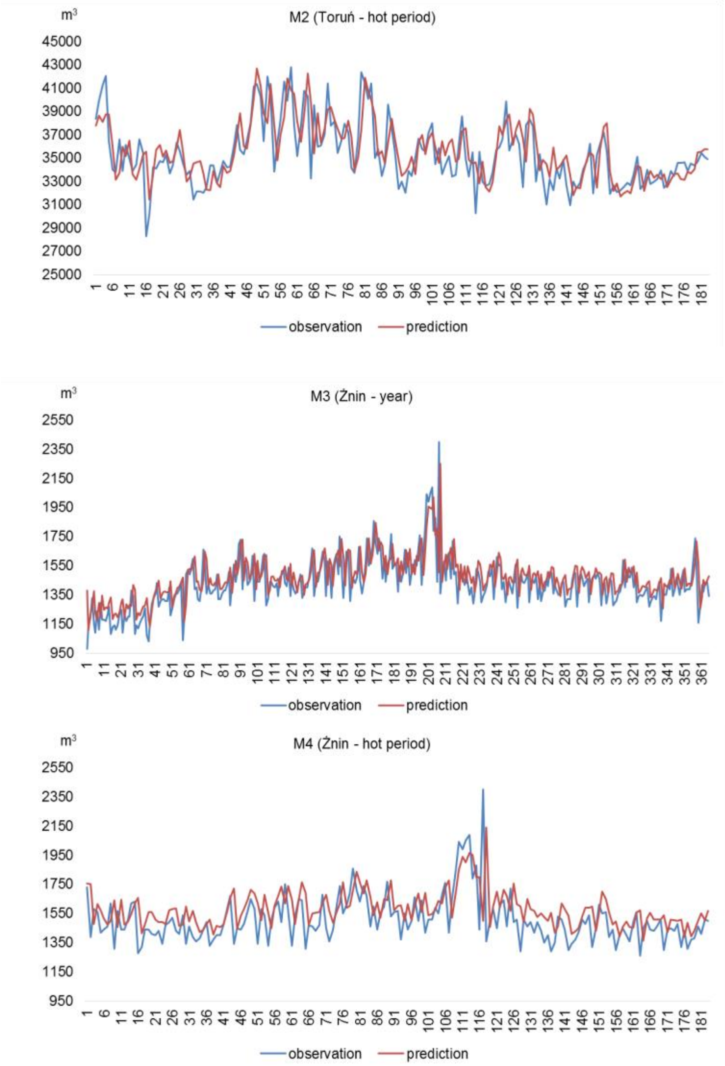
The set of explanatory variables identified to construct each individual model differed. The differences between the variables in the year model and the warm-period model seem understandable. In models for the warm period, relative humidity is more important, while atmospheric precipitation is less. The reasons can be found in the high variability of the second of these parameters. This applies particularly to warm periods, when intense rainfall usually occurs on one, or a few, days a month. The relative humidity parameter indirectly provides information on both rainfall (high values) and evaporation (low values). To some extent it indicates plants' water needs. As shown by numerous studies, the watering of gardens and lawns is the main source of increased water consumption in the warm period [11–12].

The differences between the variables in the year models for the selected towns appear to be interesting. This may be due to the sizes of the two towns, and to their local conditions. However, this is somewhat countered by the compatibility in the sets of variables for their warm-period models. This is thus most likely a result of the location of the meteorological station from which data for Żnin town was obtained. Despite the station's proximity (about 20 km from Żnin), some of the parameters it records may not entirely reflect the town's prevailing conditions. Such parameters may include the sum of precipitation (a very spatially variable parameter, especially in the warm period) and maximum temperature (which often depends on local conditions).

The models constructed differed in forecast accuracy. The models for Toruń obtained much better forecast quality assessment criteria values. This may be in part due to the aforementioned problem with the meteorological data for Żnin. However, what seems more important is the specificity of local conditions and associated water consumption characteristics in a small town. Smaller water supply systems are more susceptible to variability in water consumption, as it can be caused by a small group of consumers. This means that forecasting water consumption is more difficult in a small town than in a large one.

One problematic issues in the built models is the very large significance of the x_1 variable ("previous-day water consumption"). It results in a clear shift in the graph showing forecasted values in relation to observed values (Fig. 2). In all the constructed models, the largest forecasting errors were found on days of sudden changes in water consumption. As a result, the models are ineffective for forecasting extreme situations. To improve forecast quality, it appears necessary to expand the model with other variables that will balance out that weighting. This is confirmed by the rather low determination index value. Research results [13–14] indicate that one such variable could be a "day of the week" variable. The variation in amount of water consumed on particular days of the week stems from the normal cycle on which society functions and the activity of economic entities [15]. The main difference is between water consumption on Monday to Saturday versus consumption on Sundays and public holidays. With the "day of the week" variable being qualitative, the information is usually coded binarily. This work did not use that variable due to the PCA method's limitation in taking into account binary coded data.





M2 (Toruń - hot period)

M3 (Żnin - year)

M4 (Żnin - hot period)

Fig. 2. Matching of forecasted values to observed water consumption in 2017 in Toruń and Żnin.

6 Summary

Comparative analysis of the forecasting models for water consumption in the towns of different sizes showed that the models built for the larger of the two (Toruń) had better forecast accuracy. This was mainly due to the water supply system in the small town (Żnin) being less resilient to sudden, short-term changes in consumers' water use. In addition, the meteorological data source used in constructing the prognostic model was shown to have been important. Using meteorological data from a station located outside the town can significantly affect model quality. In the case of towns without meteorological stations, it is in the interest of water and sewage companies to have their own measurement system. Modern technical solutions make building and maintaining such a system relatively inexpensive. This is especially true when we consider the benefits of well constructed forecasting models, which apply particularly to supporting decision-making in the design, development and maintenance of water supply networks, and in the implementation of procedures to optimise the operation of pumping stations, water treatment plants and sewage treatment plants [16].

The work demonstrated the effectiveness of the PCA method in determining meteorological variables for water consumption forecasting models. At the same time, the method was shown to be limited in terms of its inability to handle qualitative data. As a result, the constructed models have fairly low forecasting effectiveness for days with sudden, large changes in water consumption. Their practical use is thus limited. The quality of forecasts would probably be improved by including a "day of the week" variable.

References

1. Z. Hiedrich, J. Jędrzejkiewicz, Environment Protection **29(4)**, 29–34 (2007).
2. H. Hotloś, Environment Protection, **32(3)**, 39–42 (2010)
3. G. Bárdossy, G. Halász, J. Winter, Journal of Water Supply: Research and Technology - AQUA, **58(3)**, 203–211 (2009).
4. J. Adamowski, H. Fung Chan, S.O. Prasher, B. Ozga-Zielinski, A. Sliusarieva, Water Resources Research, **48(1)**, (2012).
5. A. Piasecki, J. Jurasz, B. Kaźmierczak, Periodica Polytechnica Civil Engineering, **62(3)**, 818-824 (2018).
6. A. Piasecki, J. Jurasz, W. Marszelewski, Environment Protection, **38(2)**, 17-22 (2016).
7. M.M. Haque, A. Rahman, D. Hagare, R.K. Chowdhury, Water, **10**, 419 (2018).
8. G. D. Hutcheson, N. Sofroniou, The Multivariate Social Scientist : an introduction to generalized linear models (SAGE Publications, London-Thousand-Oaks-New Delhi, 1999)
9. H. F. Kaiser, EPM, **20**, 1 (1960)
10. A. Balicki, Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne (Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk, 2009)
11. U. Kępa, L. Stępniak, E. Stańczyk-Mazanek, Annual Set The Environment Protection, **15(3)**, 2546-2562 (2013).
12. W. Żuchowicki, R. Gawin, Annual Set The Environment Protection, **15(1)**, 924-929 (2013).
13. S.L. Zhou, T.A McMahon, A. Walton, J. Lewis, Journal of hydrology, **259(1-4)**, 189-202 (2002).
14. A.Studziński, K. Pietrucha-Urbanik M. Dąbek, Journal of Civil Engineering, Environment and Architecture, **61(1)**, 323-332 (2014).
15. A. Piasecki, Ł. Górski, Infrastructure and Ecology of Rural Areas, **4(1)**, 973-984 (2018).
16. J. Stańczyk, J. Kajewska-Szkudlarek, J. Łomotowski, P. Lipiński, P. Rychlikowski, T. Konieczny, Gaz, Woda i Technika Sanitarna, **10**, 372-377 (2018).