An Investigation into Identifying Factors and Building Models for Prediction of Water Usage in Regional Australia

MEHRYAR NOORIAFSHAR
mehryar@usq.edu.au

TEK NARAYAN MARASENI
w0007649@mail.connect.usq.edu.au

University of Southern Queensland, Toowoomba, Australia

Abstract

This paper is based on a research project with the aim of developing a suitable model for future water consumption in Toowoomba, Queensland, Australia. The project's main aims were to, systematically, investigate the contributory factors in water usage and then build a mathematical model for prediction and performing sensitivity analysis. Water is without any doubt the most important resource used in farming, industrial and domestic applications. Hence, this project is timely and very appropriate in terms of meeting the needs of the community in and around Toowoomba.

The paper demonstrates how the most suitable multiple regression models were built in a progressive manner. For instance, a systematic investigation into the accuracy of models has revealed that by incorporating three dummy variables and using those in conjunction with either the city population or the number of dwellings in the city would produce the most accurate results. These dummy variables represented the presence or absence of tariff, restrictions rebate and dry weather.

Key words: Water Consumption, Multiple Regression, Dummy Variables

Introduction

Water is probably the most important natural resource. Approximately 70% of the surface of our planet is covered by water which is a main source for the natural supply and distribution of water in the form of rain and melted snow. The underground water resources are also the supply sources. Efficient management and distribution of this valuable resource however, remains to be a challenge. At a global level, the complexities associated with water usage are probably increasing at an exponential rate as the population and pollution do. At a regional level, matching the supply with the demand is certainly a complex issue for the authorities. There is no doubt that we have a responsibility to properly manage and preserve water for future generations. Hence, there is a global awareness about methods of dealing with water usage.

This paper presents the findings of a research project on investigating the most promising methods of predicting the water consumption in Toowoomba, Queensland, Australia. The project was funded by the Faculty of Business, the University of Southern Queensland and its main objectives were:

1. To identify the most suitable factors which contribute to water usage in Toowoomba.
2. To build a series of mathematical models using the selected variables.
3. To test and validate these models.
4. To choose and recommend the most suitable approach for predicting water consumption in Toowoomba.

Toowoomba is Australia's largest inland city and is situated in the headwaters of the Murray Darling Basin. The population of the Greater Toowoomba is around 135,000 and 95,000 of those people live in the City Council boundaries. There has been a decline in the average rainfall over the dam catchments in the area over the past 20 years Toowoomba's total annual metred water usage is approximately 11,000 Megalitres (ML). The average water usage for the residential sector is estimated to be around 240 litres/person/day (*Water Futures Toowoomba* briefing paper by Toowoomba City Council, 2005).

Building the Model

The model chosen for this project was multiple regression. Multiple regression is, possibly, one of the most commonly used tools for identifying links between different variables. Once a suitable model is developed, the analyst can either determine a response variable for a given values of explanatory variables or predict some future values. The techniques of multiple regression have been used in similar studies into water resources managenet (Davis, 2003; Havlak, 2004; Rolf, 2004). The availability of user friendly computer software programs (such as ForescastX^{TM}) has certainly helped the popularity of multiple regression. The automatic generation of various accuracy measures, degree of explained variation and relationship significance also contribute to this popularity (Wilson *et al*., 2001; Pensiero and Nooriafshar, 2005). As it is shown in this paper, the major part of the model development is not merely a quick data entry into the computer software and pressing the solution buttons. The most important steps are the selection of appropriate variables. The choices are not always obvious and straight forward. As this paper demonstrates, a number of models should be developed and tested systematically. This process would most certainly entail inclusion and omission of variables. It can also lead to more innovative choices as demonstrated in this paper.

After identifying the potential explanatory variables, we tried to include them in the initial models, using the available data, and then run the models. The software however, prompted us that the number of independent variables had exceeded the limit with the given time series data of 15 years. It should be noted that due to the very challenging nature of obtaining data for this project, the size of the time series was limited to 15 years. Therefore, in order to screen and select only the highly potential explanatory variables, we had to determine the correlation coefficients, one by one, of all independent variables with dependent variable (water consumption). The list is given below as Table 1.

Table 1 – Correlations between the response and the potential explanatory variables

Description	**Water consumption**
Dwellings	0.52
Home units	*0.17*
Flats	0.20
Vacant land	**-0.27**
Boarding	0.40
Commercial	0.20
Hotels	*0.07*
Mixed business	*0.01*
Motels	0.59
Sporting ground	**-0.06**
Worship	**-0.45**
TCC properties	**-0.15**
TCC-Prop-Leased	**-0.24**
Count	0.32
Population	0.58

The following points were considered in selecting the most relevant explanatory variables:

1. **Negative correlation**: It is obvious that all explanatory variables should have positive relations to water consumption as all of them are contributing factors. Therefore, those explanatory variables whose correlation with dependent variable is negative (shown in boldface) were removed from the model. Therefore, we rejected five explanatory variables of Vacant land, Sporting ground, Worship, Toowoomba City Council (TCC) property and TCC properties lease land.

2. **Poor correlation**: Those explanatory variables whose correlations with the dependent variables were less than 0.20 were regarded as having weak correlations. Based on this assumption, we removed further three independent variables of home units, mixed business and motels) from the model. As a result, the number of our explanatory variables reduced to seven (Table 2).

Table 2 – Correlation between the response and the most influencing explanatory variables

Description	**Water consumption**
Dwellings	0.52
Flats	0.20
Boardings	0.63
Commercial	0.45
Motels	0.59
Count	0.32
Population	0.58

3 **Multicollinearity**: Provided there is not a strong collinearity between dwellings, boardings, motels and population, then the model employing these four variables could most probably be the most promising one. It should be noted that

population is not only responsible for an increase in the number of dwellings but also for boardings and motels. As Table 3 indicates, there is a strong multicollinearity amongst some of the variables. Hence, it would be rather difficult to develop a reliable model using these variables.

Table 3 - Correlation matrix to determine Multicollinearity between explanatory variables

Description	Water consumption	Dwellings	Flats	Boarding	Commercial	Motels	Population	Count
Water consumption	1.00	0.52	0.20	0.63	0.45	0.59	0.58	0.32
Dwellings	0.52	1.00	0.61	0.97	0.97	0.85	0.97	0.93
Flats	0.20	0.61	1.00	0.56	0.52	0.29	0.66	0.54
Boarding	0.63	0.97	0.56	1.00	0.93	0.90	0.96	0.88
Commercial	0.45	0.97	0.52	0.93	1.00	0.86	0.93	0.94
Motels	0.59	0.85	0.29	0.90	0.86	1.00	0.84	0.73
Population	0.58	0.97	0.66	0.96	0.93	0.84	1.00	0.90
Count	0.32	0.93	0.54	0.88	0.94	0.73	0.90	1.00

Table 4 provides details of the models based on the above variables which were developed and tested.

Table 4 – Multiple Regression models and their accuracy measures

Regression Models	Accuracy measures			F
	A-R^2	RMSE	U	
WC = 3,798,360.35 + ((Dwellings) * 250.99)	21.70	688,913	0.79	4.88
WC = 9,224,338.49 + ((Dwellings) * 308.60) + ((Flats) * -6,491.38)	18.00	677,313	0.78	2.54
WC = 21,814,688.48 + ((Dwellings) * -742.76) + ((Flats) * -2,495.50) + ((Boarding) * 549,644.31)	42.55	542,816	0.60	4.46
WC = 18,192,725.22 + ((Dwellings) * -209.33) + ((Flats) * -6,153.77) + ((Boarding) * 482,813.77) + ((Commercial) * -3,109.75)	40.56	526,448	0.59	3.39
WC = 19,945,291.73 + ((Dwellings) * -259.09) + ((Flats) * -6,964.38) + ((Boarding) * 521,528.37) + ((Commercial) * -2,847.59) + ((Motels) * -26,253.28)	34.16	525,617	0.59	2.45
WC = 30,099,526.69 + ((Dwellings) * -298.67) + ((Flats) * -12,881.85) + ((Boarding) * 754,281.48) + ((Commercial) * 3,208.27) + ((Motels) * -243,488.21) + ((Count) * -332.05)	53.99	414,252	0.47	3.74
WC = 12,485,943.03 + ((Dwellings) * -763.28) + ((Flats) * -24,298.99) + ((Boarding) * 659,956.03) + ((Commercial) * 5,300.21) + ((Motels) * -343,979.10) + ((Population) * 530.13) + ((Count) * -430.03)	69.72	314,336	0.33	5.61
WC = 12,078,626.06 + ((Dwellings) * -956.52) + ((Boarding) * 530,098.43) + ((Motels) * -12,668.84) + ((Population) * 152.88)	38.48	535,569	0.59	*3.89*
WC = -4,147,398.65 + ((Population) * 165.43)	28.52	658,209	0.75	*6.59*
WC = 3,798,360.35 + ((Dwellings) * 250.99)	21.70	688,913	0.79	*4.88*
WC = -12,696,521.05 + ((Population) * 367.18) + ((Dwellings) * -349.78)	26.04	643,259	0.73	*3.46*
WC = 14,853,695.89 + ((Population) * -108.39) + ((Boarding) * 238,202.14) + ((Motels) * 13,898.29)	24.82	620,935	0.71	*2.54*
WC = 930,292.08 + ((Population) * 80.02) + ((Motels) * 82,593.07)	26.93	639,364	0.72	*3.58*
WC = -20,428,875.48 + ((Count) * -227.09) + ((Population) * 436.38)	47.12	543,938	0.58	*7.24*
WC = 8,628,451.58 + ((Dwellings) * -625.22) + ((Boarding) * 461,663.09) + ((Commercial) * -1,875.48) + ((Motels) * 18,143.60) + ((Population) * 133.54)	33.02	530,131	0.58	*2.38*
Note: WC, water consumption in kilo liter, A-R^2 is Adjusted Coefficient of Determination (%), RMSE is Root Mean Square Error, U is Theils U-Statistics value and F is F-Statistics				

After a close investigation of the data, it was revealed that several peaks and troughs had been occurring in the water usage over the years (Figure 1). A further analysis indicated that tariff, restriction rebates and dry weather were the main causes of these increases and drops in water usage. Hence, it was decided to represent the presence and absence of these peaks and troughs with dummy variables in each case.

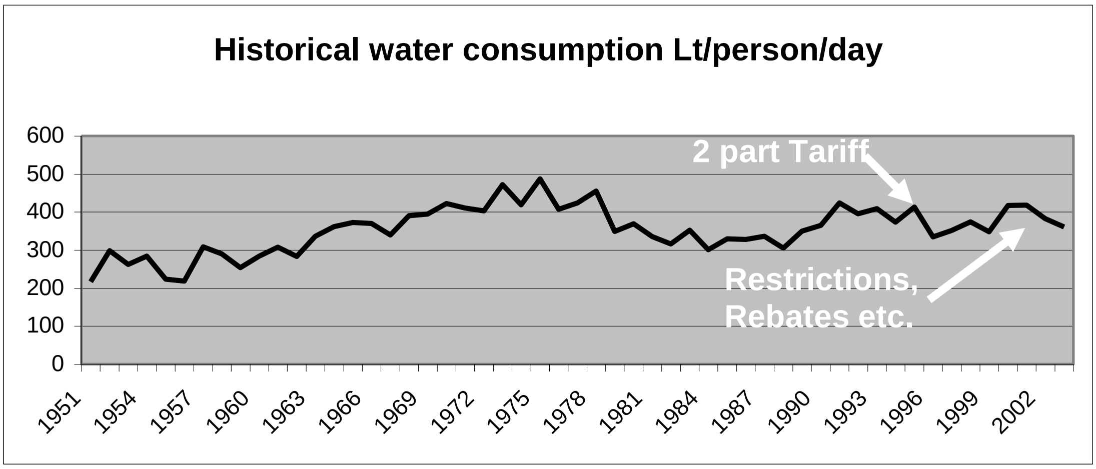

Figure 1 - Peaks and trough of water consumptions after application of tariff and restriction rebates (Source, Meulen *et al,* 2004)

For example, existence of tariff was denoted by ‘1’ and non-existence by ‘0’; having restriction rebates was denoted by ‘1’ and no rebates by ‘0’; and dry weather was denoted by ‘1’ and normal weather by ‘0’. The time series of the most appropriate explanatory variables (used in the recommended models), the corresponding dummy variables and the response variable are shown in Table 5.

Table 5 – Time series showing the response, selected explanatory and corresponding dummy variables

Year	Dependent variable	Independent and dummy variables				
	Water consumption (KL)	Population	Dwellings	Tariff	Restriction rebates	Weather
90	9234417	83574	23570	0	0	0
91	9999605	84614	24002	0	0	0
92	10912507	85107	24292	0	0	1
93	9903269	85612	24791	0	0	0
94	10706136	85848	25038	0	0	0
95	10845777	85878	25502	0	0	0
96	9127337	86569	25,590	1	0	0
97	9543500	86694	25591	1	0	0
98	10222524	86781	25587	1	0	1
99	9887615	89651	27148	1	0	0
00	9778348	90368	27538	1	0	0
01	11172216	91090	27831	1	1	0
02	11553411	91820	28176	1	1	0
03	11987306	92555	28546	1	1	0
04	10520937	93295	28916	1	1	0

As the accuracy measures and F values indicate the most reliable models are those (shown in Table 6) which incorporate the three dummy variables with either population or the number of dwellings as an explanatory variable. The procedures explained above illustrate that different models with different explanatory variables were systematically analysed and tested until the most promising model was found.

Table 6 – Multiple Regression models and their improved accuracy measures

Regression Models	Accuracy measures			F
	A-R^2	RMSE	U	
After including two dummy variables and removal of last year data (2004)				
WC = 8,396,107.02 + ((Dwellings) * -1,288.54) + ((Flats) * -39,823.36) + ((Boarding) * 294,373.57) + ((Commercial) * -6,167.42) + ((Motels) * 46,474.22) + ((Count) * 1,692.86) + ((Population) * 352.11) + ((Tariff) * -8,651,818.65) + ((Restriction rebates) * -861,226.90)	80.24	205,889	0.23	*6.87*
WC = 1,372,856.82 + ((Dwellings) * 472.75) + ((Motels) * -101,383.90) + ((Tariff) * -1,230,862.76) + ((Restriction rebates) * 1,619,566.54)	69.15	385,955	0.45	*8.28*
WC = -23,595,526.23 + ((Population) * 416.04) + ((Motels) * -57,928.07) + ((Tariff) * -1,480,026.78) + ((Restriction rebates) * 688,489.24)	77.92	326,488	0.39	*12.47*
WC = -22,994,948.49 + ((Population) * 390.83) + ((Tariff) * -1,507,425.87) + ((Restriction rebates) * 407,513.00)	78.66	338,315	0.40	*16.98*
WC = -1,300,175.62 + ((Count) * 404.95) + ((Tariff) * -3,115,128.53) + ((Restriction rebates) * 1,081,739.27)	77.23	349,456	0.41	*15.70*
WC = 1,417,497.11 + ((Dwellings) * 360.72) + ((Tariff) * -1,189,347.55) + ((Restriction rebates) * 1,176,070.47)	68.56	410,658	0.47	*10.45*
After including three dummy variables				
WC = -2,285,015.06 + ((Dwellings) * 504.85) + ((Tariff) * -1,476,097.93) + ((Restriction rebates) * 1,103,107.17) + ((Dry weather) * 999,764.29)	**87.21**	**248,491**	**0.29**	***23.16***
WC = -23,780,148.50 + ((Population) * 398.62) + ((Tariff) * -1,550,904.17) + ((Restriction rebates) * 525,622.22) + ((Dry weather) * 735,146.93)	**89.34**	**226,829**	**0.26**	***28.25***
Note: WC, water consumption in kilo liter, A-R^2 is Adjusted Coefficient of Determination (%), RMSE is Root Mean Square Error, U is Theils U-Statistics value and F is F-Statistics				

Conclusions

A number of multiple regression models based on different combinations of explanatory variables were built and tested in a systematic manner. The main purpose was to identify the most accurate model for predication of water consumption in the city of Toowoomba. Further analysis revealed that by incorporating three dummy variables in conjunction with either population or number of dwellings, the most reliable models could be developed. As a result, the accuracy (Adjusted R^2=89% and Theil's U-Statistic=0.26) measures increased dramatically in the final model.

At its initial phase, this research project has focused on Toowoomba and the immediate surroundings. It is envisaged to develop a user-friendly Excel based system and present it to Toowoomba City Council. The authorities may use this system for predicting future water usage and performing what-if-analysis scenarios.

References

Davis M. (2003), *Will the U.S. have enough water in the years to come? Water demand forecasting seeks an answer,* Retrieved May 12 2005 from http://www.siu.edu/~perspect/03_fall/water.html.

Havlak R. (2004), *Predicting Water Use in Urban Residential*, Retrieved May 12 2005 from http://twri.tamu.edu/usgs/2003-04/havlak.pdf.

Meulen, J. d., Camody, I., Nichols, R., Mortleman, T., Buckly J. and Flanagan, T. (2004), Toowoomba Water, Water Demand Management Strategies for 2004-05. "Let's *Slow* the Flow" Water Efficiency Plumbing Regulation, Power Point Presentation, City Council, Toowoomba.

Pensiero, D. and Nooriafshar, M. (2005), *MGT 2101 Business Forecasting-Study Book*, University of Southern Queensland, Toowoomba, Queensland, Australia

Rolf J. (2004), "Assessing demands for irrigation water in North Queensland", *Agribusiness Review*, Vol. 12.

Toowoomba City Council (2005), *Water Futures Toowoomb,* Briefing Paper, Retrieved July 19 2005 from http://www.toowoombawater.com.au/images/stories/briefing_paper_1july.pdf

Wilson, J. H., B. Keating, et al. (2001), *Business Forecasting with Accompanying Excel-Based ForecastX^{TM} Software*, Fourth Edition, McGraw-Hill Companies, Inc, New York, USA.

Acknowledgements

The authors would like to acknowledge Karen Van Staden of the Faculty of Business, University of Southern Queensland for her valuable contribution in assisting with data collection.