Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2918 - 2924
Received 25 April 2021; Accepted 08 May 2021.

Forecasting-Mining Prediction of Water Consumption for Residential Sectors

Roger Rozario A.P[1], Antonita Shilpa J[2],

[1]Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India

[2]Karpagam College of Engineering, Coimbatore, Tamilnadu, India

Abstract

Forecasting of water consumption is done to determine the management decisions and choices of investments for urban water management. The data is of a single family's the annual water consumption. The data is cleared of the unusual fluctuations on few days due to external influences using the sampling methods. The various methods used for forecasting the future consumption are, the fuzzy c-means prediction, the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) prediction model and mining - forecasting model. The comparison is based on the correctness of its prediction against the following years. The results highlight that using mining - forecasting techniques predict a more accurate solution compared to the other methods. The mining – forecasting method causes the least deviation of 5%, while the fuzzy c-means predicts with a 58% deviation and BIRCH causes 38%.

Keyword – clustering, curve fitting, prediction, data mining

INTRODUCTION

Around the world the water supply is a vast network. It helps industries, commercial areas, agricultural farms, residential buildings and many others. The requirement of water is uncertain since it is affected by weather making the distribution of water to all the sectors a challenge.[3] This change of demand is due to population changes and hydrological changes which causes severe drought and sometimes flood magnifying uncertainty.[1] The possibility of a warm year increases the dry periods of the same year. Approximately four degrees of global warming will be reached.[2] within the next century. In the recent past the unprecedented water consumption by residential area has caused severe drought conditions. This is due to the change in spatial patterns in storage of water which is always unpredictable.

The supply and demand of water is greatly affected by the dynamics of water surface. The abundant storage of water in a sparsely populated city can be supplied to a densely populated city on demand. The challenge in supplying water for residential demand is a challenge itself.The cost of water transfer is increased when the distance is longer, due to change in water pressure and groundwater supply. There is a need of accurate water consumption forecast to reduce the complexity in water supply which in turncan reduce the cost of water on transport.[4]A common approach is to use small reservoirsto store water in different areas. This reduces water transport. It is the most efficient environmental method to do so. The main contribution of water forecast is to reduce the excess storage and more selectively understand that the difference between supply and demand can be reduced.

In our method we will be using fuzzy c-means clusters, BIRCH Algorithm and an ML method for the same. It is understood that the ML method is more concrete and understands the demand better.

Clustering algorithms are based on the Euclidean distance between the relative data. In fuzzy c-means clustering, the distance is fuzzified then the c means algorithm is performed; later, the data is the defuzzified to its original format.[7] The fuzzification and defuzzification of data is done using curve fitting algorithms. A curve fitting algorithm is selected based on the data dependency; in our case the S-curve is the most suited fitting algorithm.Using this algorithm,the water consumption can be predicted. Here,we use a multivariate fuzzy c means clustering method because there are lots of variables affecting the water consumption for residential sectors.[6]The number of cycles of operation for clustering is fixed and the number of iterations is proportional to the clustercount, hence increasing the time of operation drastically

The next algorithm implementedis the BIRCH algorithm which is a hierarchical based iterative clustering algorithm. The hierarchy is based on the number of clusters. The number of clusters are again clustered into groups,making it a heap structure thereby reducing the number of traversals to reach a specified cluster.This algorithm overcomes the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2918 - 2924
Received 25 April 2021; Accepted 08 May 2021.

time delay caused by the fuzzy c-means clustering algorithm.

LITERATURE SURVEY

The water consumption is based on various factors such as climatic conditions, the usage of the water, the number of members in the household and,etc.Hence, it is a multivariate component. Wen Zhang et al. put forward an approach to consolidate this multivariate component as a single unit using the matrix clustering. This helps convert the explicit information into implicit data; it can be used for the prediction algorithms as a single value which considers all the various factors tothose thateach is subjected to.[8]

Pei Shi et al. recommend that we can perform the same operation of prediction using a nonlinear method for continuous generated data using K-medoids clustering. These nonlinear data have various conditions and assumptions which can slightly change a prediction higher for adopting a prediction method. It is similar to matrix clustering but It is rather time-consuming but achieves a better result.[5]

The unsupervised clustering tends to be more accurate than a supervised clustering in a univariate prediction. Ming Tang et al. state that the unsupervised learning enhances the output of the prediction increasing the consistency, reliability and, stability of the predicted water consumption.[9]

Yu Cheng Chien et al. reason that by using the clustering algorithms, it can be identify the various households which fall into the same range of all variated but may differ in the consumption of water and also households with different variates but similar water consumption. From this prediction it is possible to find the inverse document to understand the differences and similarities between families based on water consumption.[10]

Hoang Nguyen et al. state that the hierarchical clustering algorithm proves to be a more accurate solution compared to the fuzzy c-means algorithm which cause a great difference in the performance factors. The hierarchical model turned out to be a better solution because it reduces the number of iterations based on the hierarchy hence, reducing the number of traversals.[2]

Dan Halbersberg et al. evince that by implementing sequential pattern mining and sequence clustering algorithms based on time,it is possible todecide the water consumption based on the distinctive changes compared to previous usage. It enables to predict the seasonal or climatic changes.[11]

Chao Wang et al. describe that the water consumption of a household may only have a slight deviation from every other day usage. When there is a huge change of water consumption during consecutive days, it is easier to predict if the change is seasonal or permanent.[4]

Zhou Xiangyu et al. suggest that linear regression algorithm groups similar users, based on water consumption and estimate them to have same properties. The changes in water consumption are sequentially based on seasonal changes mostly. The users have same consumption of water leading to the same deviation in various seasons and against similar variants.[6]

Shilpa J Antonita et al. propose that the logic selection is done through Gaussian and Markov models leading to a better accuracy and reduced complexity for prediction. The combination of Gaussian and Markov models prove to have better solutions than that of any mathematical model for prediction.[1]

PROPOSED METHOD

The proposed method predicts water usage up to 2050 using the data of change in water usage for the years from 2000 to 2007. The change in water need determines the overall water need deviation, so the anticipation of change must be done first. The prediction cannot be estimated after 2050 because the devices that consume water and the natural resource cannot be predicted. The cause behind the inefficiency in projection is the increase in global temperature and change is climatic change. The components of the proposed system are, gettingdate of projection as input, the estimate of consumption, the water usage distribution, and change in water necessity in the consecutive day.

4.1 Prediction Estimation

The water consumption estimation is splitas, relative increase coefficient, temperature saturation, stepwise increase in usage, and decrease in usage.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2918 - 2924
Received 25 April 2021; Accepted 08 May 2021.

4.1.1 Relative Increase Coefficient

Relative Increase coefficient (RIC) is an approximation using the Taylor's approximations through the inverse of MLE of the water usage data,

$$\text{RIC} = \frac{\sum_{n=a}^{x}(U_n - U_a)}{x-a} \quad \ldots (1)$$

For all forms of approximation, the relative increase coefficient remains the same, for both Taylor series and Linear approximation

4.1.2 TemperatureIncrease

The increase of temperature determines the gradually increases water consumption in household. The gradual increase in temperaturealso affects water consumption. The calculation for temperatureincrease is calculated through Taylors expression. The temperature increase is calculated by solving the logistic curve equation,

$$Ts = T * \left(1 + a\, log_e{}^{-1}(b.t)\right) \quad \ldots (2)$$

where *a*,*b* are constants, and *t* is time. The three characteristic values U_0, U_1, and U_2 leads to three time intervals, t_0, t_1, and t_2.in the equation. Hence the equation 2 becomes,

$$Ts = \frac{2U_0U_1U_2 - {U_1}^2(U_0+U_2)}{U_0U_2-{U_1}^2} \quad \ldots (3)$$

The temperature increase found can be used for further analysis of waterconsumption projection.

4.1.3 Slowfluctuation in temperature

The linear approximation techniques is used for calculating the fluctuation inusageof water till the temperature saturation is reached. They are widely used in the method offinite differencesto produce first order methods for solving or approximating solutions to equations.

Taylor series as linear approximation is,

$$f(x) \approx f(a) + f'(a)(y-a) \quad \ldots (4)$$

This is an optimal approximation for *y*when it is close to*a*; since a curve, when closely examined, will start to look like a straight line. Hence, the equation of the curve after *a* is same as the equation for thetangent lineto the graph of *f*at *(a, f(a))*.

Due to this, the process is also called the tangent line approximation. The equation for usageincrease is as follows,

$$U_{proj} = Uast + B\,(proj - last) \quad \ldots (5)$$

where B $= \frac{U_{last} - U_{first}}{last - first}$ … (5a)

U_{first} = the usage of the leastwater usage

U_{last} = the usage of the highestwater usage

U_{proj} = the usageon the projected date

first = the usage in first day

last = the usage on the last day

proj = the projected date

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2918 - 2924
Received 25 April 2021; Accepted 08 May 2021.

4.1.4 Decrease in usage

When the usageis saturated, there will be a decrease in the usage at some point of the year with respect to the climatic changes.

The equation for usage decline is,

$$U_{proj} = RIC * (Us - Ulast) * cli \ldots (6)$$

where *cli* is change in temperaturedecline throughout the date of projection.

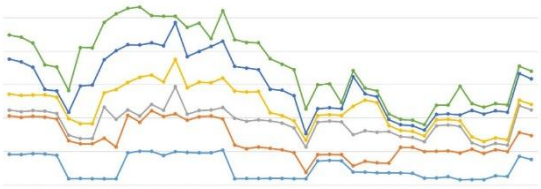
4.2 Water Consumptioncomparisons

The consumption of a watercannot remain constant since it has various factors affecting itThemeasure of water usage is an unstable consumption. The change in water usage is an important impact for water storage and transport. As a result of various studies of water consumption, the anticipation of change in water demand considering the usage can be subdivided into, curve for each household, ratio of total usage to total storage, prediction of total water demand, and classification of demand for each household.

4.2.1 Curve for each household

The water consumption data is available for years from2000 to 2007 for various households. Curve fitting for everyhousehold is essential tofind of deviation water consumption and similarities between them. The curve fitting is done by calculating the ratio between householdsand its usage for every day and the outcome is fit by a Gaussian curve.

Figure 1: Change in water intake for different households



The equation for the Gaussian distribution is,

$$f(u|\ \mu, s^2) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(u-\mu)^2}{2s^2}} \quad \ldots (7)$$

where μ is the mean, s^2 is the standard deviation.

4.2.3 Forecast of total waterusage

The ratio of total usageto daily usageof water for each day is result data, $R_{p/f}$. It fits in a Gaussian s-curve through which the forecastis done. The Gaussian curve is used for the total usage because of itsaccurate ratio rather than using the mean of all ratios. The result of curve fitting leads to the forecast to be least deviated from the actual usage.

4.2.4 Classification of usage for separatehousehold

The classification is done through the partition method. The sum of all individual ratios is the usage water ratio. This method has two calculations; to calculate the individual usage and calculate sum and divide it by the predicted total usage, $R_{tp/ip}$; second, to multiply $R_{tp/ip}$ and the individual forecast. Hence the equation for $R_{tp/ip}$ is,

$$R_{tp/ip} = \frac{\Sigma\{pUsage_j \mid j = individual \quad house\,hold\}}{pUsage_{total}} \quad \ldots (9)$$

where $pUsage_j$ = the usage predicted for household*j*

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2918 - 2924
Received 25 April 2021; Accepted 08 May 2021.

$pUsage_{total}$= the usage predicted for all household

The usage of individual household, d_i is derived from the equation,

$$d_i = \frac{pUsage_i}{R_{tp/ip}} \quad \ldots (10)$$

From equation 10, the usage of individual household can be known with more accuracy. The equations are executed via programming modules written in Java.

RESULTS AND DISCUSSION

The proposed forecasting-mining method was verified with the fuzzy c-means clustering and BIRCH algorithm to predict the water consumption. The highest mismatch in predicting the water usage is caused by fuzzy c-means clustering, the next comes the BIRCH algorithm and, it is found that the forecasting-mining has the least mismatch in predicting the water consumption (in litres).

From table 1, we can understand the increase in storage that is predicted and the cost of transport of water for that measure is very high for fuzzy c-means clustering, comparatively less for BIRCH and significantly less for the forecasting-mining technique.

Table1: Predicting Water Consumption on a same day

	Fuzzy C-means clustering	BIRCH algorithm	Forecasting-Mining technique	Actual Usage
Household 1	223.08	187.33	147.29	143
Household 2	211.2	177.92	133.12	128
Household 3	230.88	223.08	162.24	156
Household 4	273.87	247.02	191.53	179
Household 5	404.88	339.81	260.28	241

From table 2, we can understand the excess amount of water being transported. The fuzzy c-means clustering has the highest transportation exess, comparatively less for BIRCH and significantly less for the forecasting-mining technique.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2918 - 2924
Received 25 April 2021; Accepted 08 May 2021.

Table2: Excess in Water Transported

	Fuzzy C-means clustering	BIRCH algorithm	Forecasting-Mining technique	Actual Usage
Household 1	80.08	44.33	4.29	143
Household 2	83.2	49.92	5.12	128
Household 3	74.88	67.08	6.24	156
Household 4	94.87	68.02	12.53	179
Household 5	163.88	98.81	19.28	241

Table3: Percentage deviation in Water Consumption

	Fuzzy C-means clustering	BIRCH algorithm	Forecasting-Mining technique	Actual Usage
Household 1	56%	31%	3%	143
Household 2	65%	39%	4%	128
Household 3	48%	43%	4%	156
Household 4	53%	38%	7%	179

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2918 - 2924
Received 25 April 2021; Accepted 08 May 2021.

Household 5	68%	41%	8%	241
Deviation Average	58%	38%	5%	

From table3, we can understand that for fuzzy c-means clustering 58% of excess water is transported, for BIRCH 38% of water is excessively transported and for the forecasting-mining technique only 5% of excess water is transported. The excess transportation cost along with storage cost for the excess water along with maintenance has to be spent to maintain the excess water which is unnecessary.

Conclusion

Water is one of the most important essentials in human life but it is also the host for multiple other organisms. Transportation of water, maintenance of water along with the precautionary measures for transporting water, balancing the dynamics of hydraulics is a huge challenge. Hence, using a better prediction algorithm can help u reduce the expenditure to a great level. The limitation of the system is that, since the temperature of every place is different, we ought to apply a different curve-fitting algorithm for every place. The system doesn't take intoaccount the humidity and the rainfall levels of the place which could be important impact factors. Similar to these there are many more influencing factors that may not be considered here.

REFERENCES

[1] Shilpa, J. Antonita, and V. Bhanumathi. (2019) Projection of Population and Prediction of Food Demand Through Mining and Forecasting Techniques, International Conference on Artificial Intelligence, Smart Grid and Smart City Applications, Springer, Cham,
[2] Adalgiza del Pilar Rios et al. (2017) Taylor Series Approximation of ZIP Model for On-line Estimation of Residential Loads' Parameters
[3] Stephen B. Duffull et al., (2017) Assessing robustness of designs for random effects pa-rameters for nonlinear mixed-effects models, Journal of Pharmacokinetics Pharmacody-namics.
[4] Youn-Kyou Leea et al.,(2017) Analytical representation of Mohr failure envelope approx-imating the generalized Hoek-Brown failure criterion, International Journal of Rock Me-chanics and Mining Sciences.
[5] Divya Anand;et al. (2017), 'Building an intelligent integrated method of gene selection for facioscapulohumeral muscular dystrophy diagnosis', Int. J. of Biomedical Engineering and Technology, Vol.24, No.3, pp.285 - 296
[6] Lavanya, A, et al.. (2016) 'Inverse maximum likelihood-based edge detection for segmen-tation of breast lesion using active contour', Int. J. of Biomedical Engineering and Tech-nology, Vol.22, No.3, pp.272 – 283.
[7] F. Coşkun, et al. (2017) A statistical-based examination on wind turbines' bi-static scatterings at 1 GHz frequency, Journal of Electromagnetic Waves and Applications, DOI: 10.1080/09205071.2017.1391127
[8] Iliadis, Stergios Skopianos et. al, (2010) A Fuzzy Inference System using Gaussian distri-bution curves for forest fire risk estimation, IFIP International Conference on Artificial In-telligence Applications and Innovations.
[9] Bassant Selim et al., (2016) Modeling and Analysis of Wireless Channels via the Mixture of Gaussian Distribution, IEEE Transactions on Vehicular Technology.
[10] World Population and Human Capital in the Twenty-First Century by Wolfgang Lutz, et al., Oxford University press, 2014.
[11] A Primer of Real Analytic Functions by Steven G. Krantz, Harold R. Parks, Second Edi-tion Springer Science + Business Media, LLC.