International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:7, No:12, 2013

Time Series Regression with Meta-Clusters

Monika Chuchro

***Abstract*—**This paper presents a preliminary attempt to apply classification of time series using meta-clusters in order to improve the quality of regression models. In this case, clustering was performed as a method to obtain subgroups of time series data with normal distribution from the inflow into wastewater treatment plant data, composed of several groups differing by mean value. Two simple algorithms, K-mean and EM, were chosen as a clustering method. The Rand index was used to measure the similarity. After simple meta-clustering, a regression model was performed for each subgroups. The final model was a sum of the subgroups models. The quality of the obtained model was compared with the regression model made using the same explanatory variables, but with no clustering of data. Results were compared using determination coefficient (R2), measure of prediction accuracy- mean absolute percentage error (MAPE) and comparison on a linear chart. Preliminary results allow us to foresee the potential of the presented technique.

***Keywords*—**Clustering, Data analysis, Data mining, Predictive models.

I. Introduction

THE multiple regression is one of the most widely used methods of time series estimation and prediction. The advantage of the multiple regression models is the ability to interpret obtained mathematical function to assess the impact of various independent variables influencing the variability of the modeled phenomenon. What is more, those models allow us to explicitly control for many factors that simultaneously affected the modeled time series. In addition, the quality of regression models is comparable with more complex models in many cases.

Like the most statistical methods, multiple regression relies on certain statistical assumptions. The assumptions change with the research objective, which can be split into three groups: the computation of point estimates, the derivation of interval estimates, and the testing of hypotheses. However, we can identify several critical assumptions. These assumptions include the following [1]:

- The dependent variable is normally distributed and the variance is constant.
- The relationships between a dependent variable and each independent variable are linear.

Monika Chuchro is with the Geoinformatics and Applied Computer Science Department, Geology, Geophysics and Environmental Protection Faculty, AGH- University of Science and Technology, Krakow, Poland (e-mail:chuchro@geol.agh.edu.pl)

This work was supported in part by the AGH- University of Science and Technology, Faculty of Geology, Geophysics and Environmental Protection, as a part of statutory project number No.11.11.140.032and by the AGH-University of Science and Technology, Faculty of Geology, Geophysics Environmental Protection as dean grant number 15.11.140.212

- The error term is a random variable that has a mean equal to zero and constant variance.
- The independent variables are linearly independent of each other

Those assumptions are difficult to prove in case of modeling environmental data. Firstly, environmental data are affected by noise in two different ways: as measurement error and as process noise. Secondly, the variability of the data depends on many factors, also affected by noise and not fully recognized, so knowledge about modeled phenomenon is incomplete and fraught with errors and noise. Also, the nature of environmental phenomena may causes the time series not to have normal distribution, or else, variance of data is changing in time.

There are ways to approach such data. The first of them is the use of data transforms to obtain a normal distribution and linear relationships between variables [2]-[4]. In order to obtain a normal distribution, the min-max or Box-Cox transforms can be used, as well as logarithms.Transformations to achieve linearity are: exponential, quadratic, reciprocal, logarithmic, and power [4].

Also, methods without assumptions, such as Artificial Neural Networks described above, may be used [5].

Another method which could be used in the case when a dependent variable can be split into few subgroups with normal distribution and constant variance will be presented in this paper. This method is based on the use of clustering and regression execution for each cluster of data. Due to the fact that clustering is an unsupervised method, an algorithm was used in order to increase the reliability of division into cluster meta-clustering.

II. Methodology

A. Clustering K-Mean Algorithm

Data clustering is a group of methods in which subgroups (clusters) are made from instances that are somehow similar in characteristics. The instances are thereby organized into representations that characterized the population being sampled. Each instance belongs to exactly one cluster [6].

The simplest algorithm, which was used in this work, is a K-mean algorithm. This algorithm splits the data into K clusters represented by their centers of mean. The center of each cluster is calculated as the mean of all the instances belonging to that cluster. This algorithm also employsa squared error criterion. Thealgorithm partitioned N data points into K disjoint subset S_j containing N_j data point so as to minimize the function (1) seen below [7]:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \left| x_n - \mu_j \right|^2 \quad (1)$$

where: x_n is a vector representing the *n*th data point, andμ_j is the centroid of data points in S_j

The algorithm starts a predetermined number of cluster with centers randomly chosen. In every iteration, each instance is assigned to its nearest cluster center according to the distance measure function. Next, the centroid is computed. The algorithm stops when there is no further change in the assignment of the data points [8].

B. Clustering EM algorithm

Expectation Maximization algorithm is one of the simplest methods of clustering. Just like K – mean, algorithm works iteratively. But instead of assigning observations into clusters order to the differences between the means of the groups which were the highest, the EM algorithm calculates the probability of belonging to each cluster, assuming one or more probability distributions. The aim of the algorithm is to maximize the overall probability for the distribution of the clusters. The method consists of two steps performed interchangeably. In the first step, the estimated parameters of the probability distribution assign instances to the clusters. In the second step, current distribution parameters are converted in such a way that causes the model to be more compatible with the data [9].

The result of the EM algorithm is dependent on the choice of initial parameter values. It can also be interpreted as an alternate, setting a lower bound for the logarithm of the likelihood function and maximizing the performance of this limitation. Detailed information on the algorithm can be found in [9].

C. Meta-Clusters

The clustering methods are regarded as unsupervised learning for a lack of a class label or a quantitative response variable. Therefore some questions could be generated: whether this division is the most appropriate or if the centroids have been appropriately selected?

These questions are often more troublesome than the actual process of analysis. For this reason, we used meta-clustering. In the case of clustering multi-dimensional data, finding the best distribution of the concentration is a very complex issue which can generate n-different and equally good solutions.

In order to improve the quality of clustering and reduce the time needed to analyze all potential clusters, the meta-clustering method is useful. The first step of meta-clustering is carrying out a number of alternative divisions. In the second step, similarity between the different divisions into clusters is measured. In the next step, given input clustering results are used for measured similarities between different clustering [10], [11].

A variety of methods could be used to measure the similarity; methods such as connectivity matrix, par counting, set matching, and variation of information [12], [13]. In this paper a measure of clustering similarity related to the Rand index will be used [14].

Given *N* points, X_1, X_2,…X_N, and two clustering of them $Y=\{Y_1,\ldots, Y_{K1}\}$ and $Y'=\{Y_1',\ldots Y_{K1}'\}$, we can define:

$$c(Y,Y') = \sum_{i<j}^{N} \gamma_{ij} \Big/ \binom{N}{2} \quad (2)$$

where: γ_{ij} is 1 if there exist *k* and *k'* such that both X_i and X_j are in both Y_k and Y_k' or 1 if there exist *k* and *k'* such that X_i is in both Y_k and Y_k' while X_j is in neither $Y_k$orY_k', otherwise 0.

The measured value of similarity is between 0 and 1. 0 when the clustering has no similarities and 1 when clustering are identical.

The measured distances between all pairs of clustering were a kind of clustering quality assessment, and also clustersthemselves clustered at the meta level. For comparing the results of simple meta-clustering via Rand Index, the clustering with agglomeration method were used whereby in agglomeration clustering objects, here clusters, are grouped in increasing concentration, using a measure of similarity or distance [15],[16].

In the next step, data points are classified into the final clusters. If the data point in each clustering was assigned to the same cluster, it is assigned to the same final cluster. If that data point changes clusters depending on methods or initial centers of clusters, data point is finally assigned to the cluster, which was assigned more often.

D. Regression

Final clusters were assessed for normality of distribution in each cluster, and correlated with the explanatory variables. Quantification of the relationship between the explanatory variables and the dependent variable can be described as a simple equation of multiple regression (3) presented below [9]:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + \varepsilon \quad (3)$$

where: $a_0$is the intercept parameter, a_1,a_2,…a_n are the slope (coefficient for each independent variables), ε is model deviations.

III. Analysis

A. Data

Meta-clustering and regression were evaluated on environmental time series data. Data set contains the dependent variable which was the inflow into wastewater treatment plant in unit m^3/day, parameters determined in raw and treatment wastewater and weather data. Weather data contain the information about temperature in °C, precipitation in mm of water column, humidity in %. Each time series consisted of almost three thousand daily observations. A characteristic feature of the data was the lack of normal distribution with a distinct skewness of the data. In Table I below, the basic characteristics of the dependent variable and

selected variables and the histogram of dependent variable on Fig. 1 can be seen.

TABLE I
BASIC STATISTICS OF SELECTED VARIABLES

Variables	Mean	Median	Min.	Max.	Variance	Std. dev.	Skewness
Inflow	52.3	50.6	23.8	166	77.0	8.78	3.29
Temperature	8.95	9.60	-22.6	27.9	73.9	8.60	-0.29
Precipitation	1.84	0.00	0.00	64.8	21.5	4.64	6.13

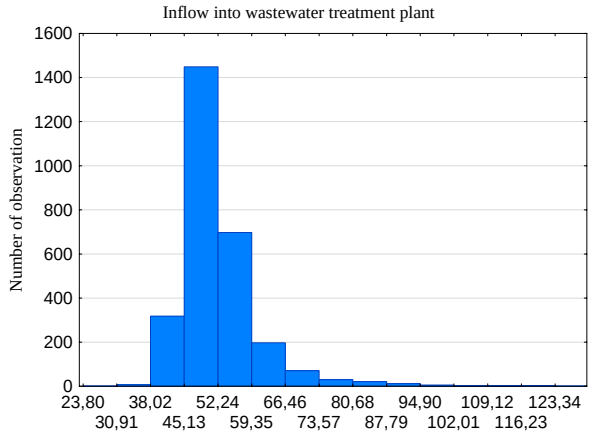

Fig. 1 Histogram of inflow (m^3/day)

Basic analysis also included correlation matrix, which examined the relationship between the dependent variable and selected explanatory variables. Correlation was measured by Pearson product-moment correlation coefficient and Spearman's rank correlation coefficient. Differences between the two coefficients were low, which indicates the existence of linear relationship between the dependent variable and the independent variables. Time series of inflow into wastewater treatment plant shows a significant positive correlation with precipitation (0.6), temperature (- 0.2), day of week (0.1). Dependent variable also shows a significant positive correlation with lagged data, precipitation lag of 1 (0.45), and dependent variable lag of 1 (0.4). With increasing lags, correlation with dependent variable decreases; only a lagged dependent variable of 7 and 14 observation show a significant positive correlation (0.3 and 0.25).

B. Analysis

The occurrence of the relationship between precipitation and dependent variable made us decide to use the clustering of data. Also, the correlation between the inflow data and precipitation were the cause of inflow data division into three clusters. Each clusters corresponding to days with no precipitation, days with small precipitation and days with high precipitation. Six divisions into clusters were made using EM and K-mean algorithms and choosing different initial parameters. K-mean algorithm for splitting the inflow time series data, used different measures of distances. In the first one, a Euclidean distance; in the second, a squared Euclidean distance; and in the third, a Manhattan distance. In EM algorithm different initial cluster centers were used. Table II present numbers of data classified to each cluster.

TABLE II
NUMBERS OF OBSERVATION IN EACH CLUSTER

Method	I	II	III
K-mean 1	2261	496	30
K-mean 2	2196	550	41
K-mean 3	2411	344	32
EM 1	2236	348	3
EM 2	2238	537	12
EM 3	2190	590	7

Computed value of the Rand index varies from 0.9 to 0.76. The values of Rand index show that every used method of clustering could be considered as corrected. For this reason, all computed division on clusters was used for the meta-clustering. The agglomeration method of clustering indicates strong similarities between k-means algorithms with Euclidean distance, and squared Euclidean distance, which was confirmed with value of Rand index (0.9).

In the final division on clusters, 2263 observations were classified into the first cluster, 478 into second, 46 observations into third. Mean value computed for the first cluster is equal to 46.11 m^3/day, for the second cluster it is 54.26 m^3/day, for the third cluster it is 58.35 m^3/day. The number of instances assigned to the third cluster is not sufficiently large for multiple regression model computation. For this reason data from second and third cluster were summed.

Two models of regression were computed. First of them was calculated for original time series of inflow into wastewater treatment plant. The second one was calculated for data of inflow after clustering. The second model consisted of two submodels - one for each cluster. Both models used the same set of explanatory variables: inflow lag of 1, 2, 7, 14 observation, precipitation, precipitation lag of 1 observation, temperature, and variable indicating the day of the week.

IV. RESULTS

For both models, the determination coefficient is comparable, which is shown in Table II. The values of standard error of estimation are also comparable.

The distribution of residues of the regression model has shown heteroscedasticity and autocorrelation. The distribution of residues of cluster regression model also has shown heteroscedasticity but autocorrelation is weaker. Comparison of models on a graph shows a better fit of the cluster regression model to a real inflow data (Fig. 2).

TABLE III
RESULTS OF REGRESSION MODELS

Results	Regression model	Cluster regression model
Determination coefficient R^2	0.69	0.75
p value	0.00	0.00
Std. error of estimation	6.52	5.89

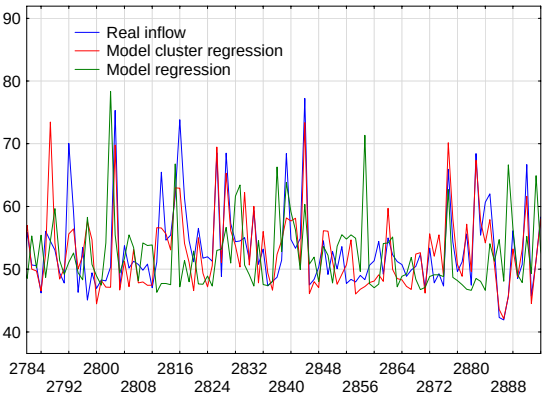

Fig. 2 Comparison of fitting models to the real data

The final step of analysis includes forecasting four observations and their evaluation by mean absolute percentage error MAPE. The value calculated by the formula shown below (4) should be multiplied by 100, to be presented in the percentage form, like in Table IV.

$$MAPE = \frac{1}{T-n}\sum_{t=n+1}^{T}\left|\frac{y_t - y_{tP}}{y_t}\right| \quad (4)$$

where y_t is the actual value, y_{tp} is a forecast value

TABLE IV
FORECASTING RESULT

Nr.	Real data	Regression model	Cluster regression model
1	87.5	82.9	85.2
2	83.0	87.8	87.3
3	84.7	88.2	87.8
4	71.5	91,7	90.1
MAPE		**10.9 %**	**9.34 %**

MAPE value for cluster regression model is acceptable because there is less than 10%.The difference between both values of MAPE is only 1.5%.

V. CONCLUSION

The clustering of data before performing a regression could be useful in the case when data do not have normal distribution and clusterscould easily separate the original data set. The result in such a simple example shows that clusteringmakes sense even for regression problems in time series data. In the case of assigning a higher number of clusters or choosing different algorithms of clustering, the obtained results could be better.

Planned work on this issue includes comparison of different algorithms of clustering and meta-clustering. Future work will also focus on the analysis of the influence of numbers of clusters on the accuracy of prediction after meta- clustering of noisy data.

REFERENCES

[1] E. W. Steyerberg, “Assumptions in regression models: Additivity and linearity”, in *Clinical Prediction Models*, New York: Springer, 2009, pp. 213-230.
[2] J. W. Osborne, “Improving your data transformations: Applying the Box-CoX transformation”, in *Practical Assessment Research and Evaluation*, Vol. 15, No. 12, 2010, pp. 1-9.
[3] R. H. Myers, Classical and Modern Regression with Applications, 2nd Edition. Duxbury Press. Belmont, California, 1990.
[4] A. M. Glenberg, Learning From Data, 2nd Edition. Lawrence Earlbaum Associates, Mahwah, New Jersey, 1996.
[5] N. Karunanithi, D. Whitley, Y. K. Maalaiya, “Using neural networks in reliability prediction” in IEEE Software, vol. 9, issue:4, pp. 53-59, Jul. 1992.
[6] A. K. Jain, M. N. Murty, P. J. Flynn, “Data Clustering: A review”, in ACM Computing Surveys, Vol. 31, No 3, 1999, pp. 264-323.
[7] J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, in: *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297.
[8] C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford, England: Oxford University Press, 1995.
[9] I. H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Elsevier, San Francisco, 2005.
[10] E. Lozano, E. Acuna, Comparing Clustering and Metaclustering Algorithms, in:*Machine Learning and Data Mining in Pattern Recognition*, Lecture Notes in Computer Science Vol. 6871, 2011, pp 306-319.
[11] Z. Yujing, T. Jianshan, J. Garcia-Frias, G.R. Gao, An adaptive meta-clustering approach: combining the information from different clustering results in: *Bioinformatics Conference*, 2002. Proceedings. IEEE Computer Society, pp. 276-287.
[12] M. Meila, Comparing Clusterings - An Axiomatic View, in: *Proceedings of the 22nd International Conference on Machine Learning*, 2005.
[13] L. Hubert and P. Arabie. Comparing partitions, 1985,Vol.2, pp.193–218
[14] W. Rand, Objective criteria for the evaluation of clustering methods. The American Statistical Association, Vol.6, 1971, pp.846–850.
[15] D. J. Divya, D. B. Gayathri, A Meta Clustering Approach For Ensemble Problem, in: *International Journal of Image Processing and Vision Sciences* (IJIPVS), Vol-1 Iss-3,4, 2012, pp. 98-102.
[16] R. Caruana, M. Elhawary, N. Nguyen, C. Smith, Meta Clustering http://www.cs.cornell.edu/~caruana/ICDM06.metaclust.caruana.pdf.