# MEASURED DATA QUALITY ASSESSMENT TOOL FOR HYDROINFORMATIC SYSTEMS

Nemanja Branisavljević*, Dušan Prodanović*, Dragana Obradović **, Ivan Kovač **
* University of Belgrade/Faculty of Civil Engineering, Belgrade, Serbia
** Jaroslav Černi Institute for the Development of Water Resources, Belgrade, Serbia
nbranisavljevic@grf.bg.ac.rs
dprodanovic@grf.bg.ac.rs
dragana.obradovic989@gmail.com
ivankovac1984@gmail.com

***Abstract*—the data quality assessment method specially customized to environmental measurement systems is presented in this paper. The methodology is based on mathematical data relations and may be used in any acquisition system where sufficient data redundancy or functional correlations between measured quantities exist. It relies on quantitative comparison of the measured data and predicted data with the result that is expressed in the form of continuous data quality grade. It may be customized to be used on-line, with real-time data streams, for model data assimilation purposes or off-line for off-line applications and analysis. The system is tested on both testing data, where the capabilities of the framework were defined and on real world real time data, obtained from the system of cascading dams. Although the system provided fairly good results in the real time use, some pitfalls were identified that are mostly focused on the data quality interpretation once that the data quality grade is obtained.**

## I. INTRODUCTION

One of the main tasks in the hydroinformatic analysis is the proper representation and interpretation of the state of the system. One of the most common ways to quantitatively represent the state of the system is to use the measured data obtained from the gauging stations. Nevertheless, sometimes measured data quality may be in question (figure 1). Therefore, poor data quality leads to poor information obtained, which further leads to unreliable decisions and various side effects.

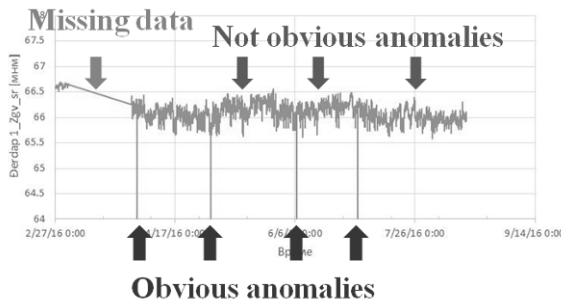


Figure 1. Anomalies in data

Data quality aspects are numerous. Data quality may reflect its completeness, accuracy, precision, adequacy, etc. The selection of data quality aspects that should be considered is usually based on case to case bases and may not be generalized in any way. Nevertheless, for hydroinformatic problem focused on hydraulic and hydrologic modeling or at data assimilation, the data precision and accuracy play the crucial role. The aim of data quality analysis presented in this paper is to detect and classify the data value due to presence of numeric errors.

Nevertheless, sometimes some subtle errors in data may not be that obvious, even if the data may be represented graphically, as may be seen in the figure 1. In those cases, some kind of more precise and preferably automatic system has to be used.

Since the quality of acquired data varies constantly because of its dynamic nature, data acquisition systems themselves must perform the data quality check, assess the data quality and, if possible, work on data quality improvement.

From the point of data analysis procedures and applications, its implementation may be suitable for offline analysis, real-time analysis or near real-time analysis that comprises the batch data analysis as the data arrives in the database.

The methodology and implementation strategy for efficient automatic data quality evaluation, suitable for real-time data acquisition systems is presented in this paper. The system is developed specifically as a part of the hydroinformatic system used for managing the cascaded hydropower plants (HIS Djerdap) [5], but can be customized for any acquisition system where sufficient data redundancy or functional correlations between measured quantities exist.

Various authors addressed the issue of measured data quality [3]. Most of them performed custom off-line data processing before running the simulation models [2] and just few of them addressed the possibilities for simple on-line data validation [1]. There are also some developed methodologies for data quality assessment where, due to complex nature of data acquisition and processing, several validation tests have to be performed on single measured data value [3], but robustness of applied method prevented

the full-scale implementation, suitable for on-line, real-time use. Some other authors [4] defined some special preprocessing and post-processing procedures to improve data quality assessment.

In this paper, we introduce the data quality assessment framework especially suitable for real-time environmental data. The methodology is presently customized to be suitable to hydroinformatic systems, but may be implemented and adjusted for any kind of measurement system, as long the data values may be computationally related to some environmental parameters, other measured values or some general rules and constrains.

## II. METHODOLOGY

The methodology used in presented data quality assessment framework is based on predicting the measured data values using predefined relations. The set of relations are specifically determined for the specific measured variable data stream. The selection of relations that are used for one data stream data quality assessment is based on several criteria. The nature of measured variable, availability of relation parameters, availability of other measured data and the nature of expected errors in data are some of them.

In the process of data quality estimation, every relation is used for measured data value prediction and to map the measured data value to the data quality grade between 0 and 1.

Usually, more than one relation can be used, so more than one data quality grades are obtained. Therefore, some aggregation function (average, max, min, etc.) may be applied to provide the final data quality grade. After that, quality grade may be used to accept the data, reject it or to use the data only for certain types of analysis, what should be determined on case to case basis.

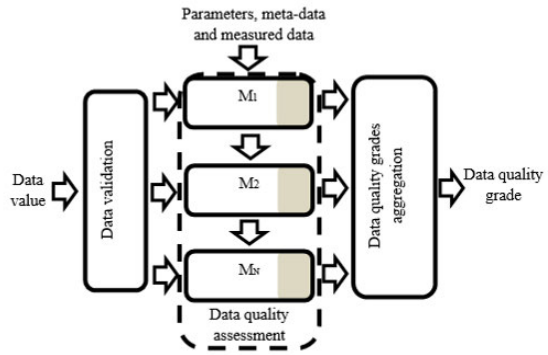


Figure 2. Data validation and data quality assessment

The methodology is divided into two successive stages: data validation and data quality assessment (figure 2). Data validation stage is comprised of two basic steps - check of expected data format and check of the global value boundaries for specific measured variable in the specific measuring micro-location under specific environmental circumstances.

If the measured data doesn't pass the data validation stage, due to extremely low data quality, the data is rejected at that level and it doesn't pass to the next, data quality assessment stage. This stage is therefore, used as a kind of filter, that filters the data with the extreme low quality (usually due to measurement equipment malfunction) and therefore this stage, in some cases, saves some amount of CPU time.

The second stage (figure 2 and 3), data quality assessment, is used for more granular data quality check and is consisted of set of computational methods, based on predefined data relations ($M_1$, $M_2$, …, $M_N$), that are used for measured data prediction and the data quality grade assessment.

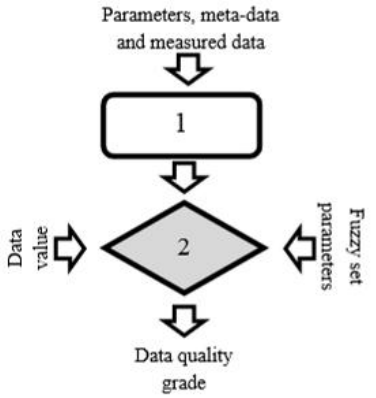


Figure 3. Data quality method procedure: 1 – prediction, 2 – mapping to data quality grade

In the process of data quality estimation, every relation is used for measured data value prediction and to map the measured data value to the data quality grade between 0 and 1.

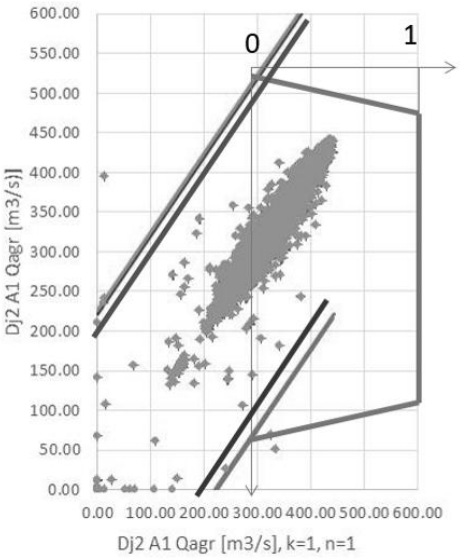


Figure 4. Data quality method calibration diagram

Data quality computational method consisted of two stages is presented on figure 4. First, the data value under consideration is predicted using predefined relation. Second, the fuzzy set is determined to map the measured data value to the data quality grade. The parameters of fuzzy set are determined using four boundaries that are determined in the process of the method calibration (figure 4):

1. Two inner boundaries that bounds the data that should be considered as accurate

2. Two outer boundaries, that bounds the data that have the quality grades between 1 and 0

Proper determination of those boundaries is crucial for the data quality assessment precision and accuracy, and therefore during that process the special attention should be payed. It should be emphasized that data values uncertainty may be used for more reliably determine the boundaries and the fuzzy set parameters. Usually historical data are used in the process of calibration, but experience of the expert may be of great value in the absence of the historical data.

After the data quality grades are obtained, they are then aggregated to single representative value (figure 4) in the range 0-1, using the aggregation module. Seven data quality assessment methods were developed with particular aim to cover most of the existing relations of environmental data:

1. Constant relation,
2. Periodical continuous relation,
3. Managed relation,
4. Seasonal constant relation,
5. AR relation,
6. Linear (regressive) relation,
7. Nonlinear relation.

The aim of first relation (constant relation) is to predict the measured data value with single constant value. That value is usually determined as average of the historical dataset. The inner and outer boundaries are determined using the historical extreme values supported by data uncertainty. This relation is the most basic one and it may be used for any measured variable.

Periodical continuous relation is usually used for the meteorological periodic variables, like air temperature. Parameters of this relation have to be estimated using some kind of optimization algorithm - manual, semi automatic or automatic. If the temporal periodicity is present in data, this method may be used as the basic one.

Managed relation is reserved for managed variables, like discharge through turbines that is managed by the valves. For this kind of relation, external variable that can be used as management indicator have to be available. In the case of discharge through turbine, an indicator of valve opening may be used.

In some cases measured variables may have seasonal periodicity. In that case, seasonal constant relation may be used. Therefore, the measured variable is represented by a constant value during different seasons (e.g. spring, summer, autumn, winter).

Autoregressive (AR) relation is used if the measured variable has strong autoregressive features. That is the case of most the environmental variables. In general, AR method is the special case of linear method and for calibration purposes the historical data is needed.

Linear relation represents the method based on two variables that are linearly related. For this method, additional variable is needed.

The last one, nonlinear relation represents the most general relation that may be formed between two or more variables.

To ensure that the data value has got the data quality grade before it is being entered into the database system, a fallback system is provided. The fallback system encompasses several data quality sequential schemas that are used in situations when the envisaged data quality methods can't be used due to the various reasons like missing data. The fallback system is usually consisted of several data quality assessment schemas as presented on figure 5.

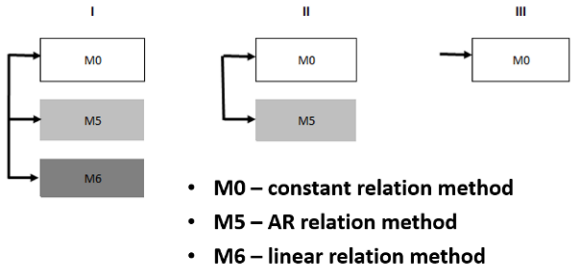


Figure 5. Data quality system schema with fallback

Proposed methodology can be used for both continuous and discrete real-time data streams as well as for off-line collected data. The system may be used as fully automatic, where the expert is involved only in the system preparation stage since he has to calibrate the data relations and estimate parameters based on historical data and his/her experience. The proposed methodology is also considered to be extremely flexible, since the relations and its parameters may be adjusted or updated later at any time.

## II. CASE STUDY

To reduce the number of different constellation of methods in data quality schema, several classes of variables are defined. One of the classes defined is class of measured water level in large rivers. One member of that class is measured water level of river Danube in the Ram gauging station that is presented as case study (figure 6).

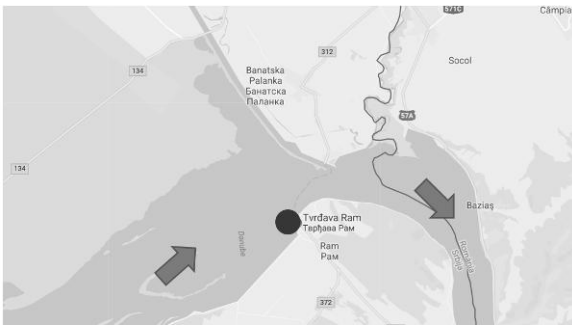


Figure 6. Position of gauging station near the town Ram on the bank of the Danube river

The schema for water levels on big rivers is consisted of three stages (figure 5). First stage is consisted of three prediction and data quality assessment methods, the second one is based on two methods (when the additional measured data value for the linear relation is missing) and the third stage is based only on the base constant method.

Historical data are divided into two series. First data series is used for data quality methods calibration, and the second one is used for the testing purposes (figure 7).

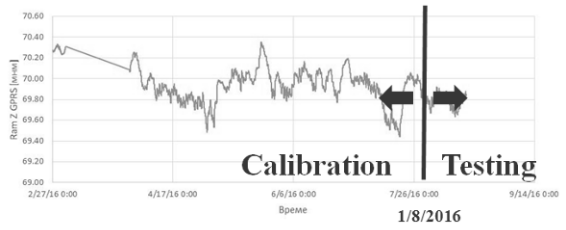


Figure 7. Calibration and testing period of historical data

The methods are calibrated and the calibration results are presented on the figure 8.

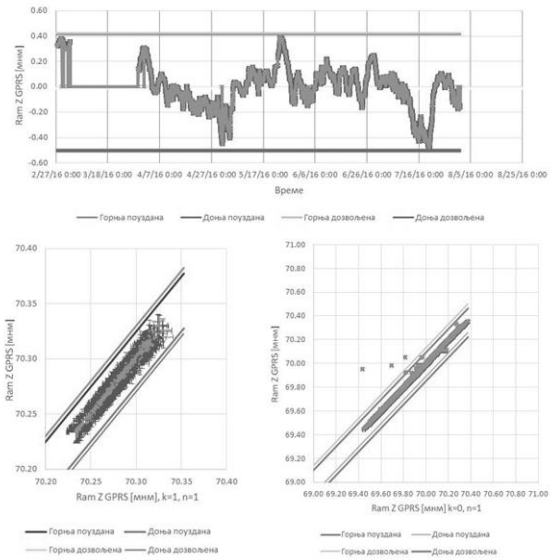


Figure 8. Calibration of methods M0, M5 and M6

Average value of calculated data quality grades is used as final data quality grade. Four data quality classes are defined according to final data quality grades: [0, 0-0.33, 0.33-0.66, 0.66-1, and 1].

After 1502 data from the testing period analyzed for data quality, most of the data had data quality grade equal to one (figure 9). Nevertheless, three subtle anomalies were detected and classified with data quality in the class 0.66-1.

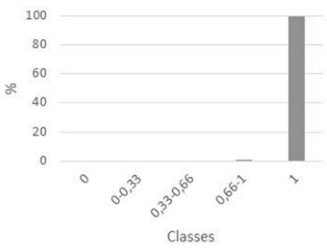


Figure 9. Calibration and testing period of historical data

## II. CONCLUSION

The methodology used in presented data quality assessment framework is based on predicting the measured data values using predefined relations. The set of relations are chosen for the specific measured variable data stream. The selection of relations that are used for one data stream data quality assessment is based on several criteria. The nature of measured variable and the availability of relation parameters are some of them.

The methodology is implemented in for two different applications. First one is for offline data and is implemented in MSExcel environment. That implementation is also used for manual data quality assessment and calibration of data relations. The second implementation is based on a server that takes the data from the database, compute the data quality and returns data quality grade to the database as meta-data value.

Nevertheless, this methodology has some pitfalls. First, the data quality grades in the continuous form may be ambiguous in the case of its interpretation. That's why its interpretation should be defined on case to case bases, with specific focus to data application and its further transformation. The major challenge in this field still remains the issue of rare events that may be considered as faults due to equipment malfunction. The fact is that the rare events are usually the most valuable to be acquired, recorded and analyzed. That's why the role of expert in such a system is crucial.

## ACKNOWLEDGMENT

The authors are grateful to the Serbian Ministry of Education, Science and Technological Development for its financial support (Projects No. TR37010 and TR37013).

## References

[1] Bertrand-Krajewski, J. L., Laplace, D., Joannis, C., & Chebbo, G. (2000). *Mesures en hydrologie urbaine et assainissement*. Tec & Doc., ISBN13: 978-2-7430-0380-7

[2] Branisavljević, N., Prodanović, D., Arsić, M., Simić, Z., and Borota, J. (2009). Hydro-Meteorological Data Quality Assurance and Improvement. *Journal of the Serbian Society for Computational Mechanics/Vol*, *3*(1), 228-249.

[3] Branisavljević, N. (2012). Methodology for quality assessment of environmental data, *PhD thesis in Serbian*, Faculty of Civil Engineering, University of Belgrade

[4] Branisavljević N., Kapelan Z., Prodanović D. (2011): *Improved real-time data anomaly detection using context classification*, Journal of Hydroinformatics Jul 2011, 13 (3) 307-323; DOI: 10.2166/hydro.2011.042.

[5] Branisavljević N., Prodanović D. (2016): *Development of the hydroinformatic system for Djerdap cascading dams – data quality*, Report for Jaroslav Černi Institute for the Development of Water Resources, Belgrade, Serbia