

Improving Quality of Water Related Data in a Cyberinfrastructure

Sorin N. Ciolofan, Mariana Mocanu, Florin Pop, Valentin Cristea
(sorin.ciolofan@cs.pub.ro, mariana.mocanu@cs.pub.ro, florin.pop@cs.pub.ro,
valentin.cristea@cs.pub.ro)

Computer Science Department, University Politehnica of Bucharest, Romania

Abstract— Data Quality is an important concern for any category of data, regardless of size. Acquiring BigData is of limited usefulness if that data is not qualitative. Although in this paper we refer to the specific case of water related data measured through sensor networks and persisted as time series, the identified challenges and the solutions proposed are generally applicable to other environmental resources observations (e.g. soil, air). Existing data is verified to assess the plausibility of the measured values. In this respect, a suite of tests to be applied on the sensor data are proposed. Where bad data or missing data is identified, it is replaced based on numerical methods.

Keywords – data quality, cyberinfrastructure, water pollution, sensor networks

I. INTRODUCTION

Big volumes of raw sensor data (>100 TB/ year) are expected to be gathered by emergent networked observatories (NEON project with 14000 sensors deployed on 60 sites over USA by 2017 [1]). Acquiring a vast amount of data from a sensor network is of limited usefulness if that data lacks quality. GSI defines “data quality” as being “complete, consistent, accurate, time-stamped and industry standards-based” [1a]. “Complete” means that missing data should be minimized (ideally reduced to zero). “Consistency” refers to the characteristic of data to be logically valid across multiple views. “Accuracy” describes the degree of

closeness of results of observations to the true values. “Time-stamped data” means that data is appended the real moment (yyyy-mm-dd-hh-ss.ms) when it was generated, so it can be later ordered on a timescale. As specified in [1b], “Data Standards are documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data”.

Wrong values that are made publically to stakeholders have as result, in the best case, loss of credibility in the project, and in the worst case could lead to tragic consequences, especially when the data should be used to support decisions in critical situations (floods, water pollution, etc.). For this reason global ongoing research efforts try to design the frame for the future Quality Assurance (QA) protocols and standards. In this article we present relevant scenarios which lead to data corruption and propose solutions in order to mitigate the associated risks.

Beside natural inherent processes that affects water quality (hydrological, physical, chemical or biological) the most significant impact factor results from human activities (urban sewage, agriculture, industrial and urban waste disposal, dredging, navigations and harbors) that dispose bacteria, nutrients, pesticides and herbicides, metals, oils and greases, industrial organic micro-pollutants[2].

Cyberinfrastructure is a relatively new holistic approach in environmental

research and generally designates a mix of advanced data acquisition through real-time sensor networks, scalable databases, high performance computational platforms, visualization tools, analytics, and data integration techniques [3]. This paradigm involves collaboration between various science, engineering and social disciplines. Laboratory sampling refers to the process of analyzing in a laboratory a sample collected in the field. Spot sampling is about using a field sensor to take measurements of interest in the field. This can further involve telemetry (measurement is made in the field and collected data is sent, usually by wireless transfer, to remote monitoring equipment). In this paper we will address the second case, i.e. *spot sampling*. A field sensor can give wrong values in case of poor calibration, changes suffered by the sensor in the course of transportation to the deployment site, vandalism, accumulation of algae, plants or other microorganisms on the surface of the sensor, extreme natural phenomena (such as extreme cold conditions, high flows), bad circuit boards, or just by aging. The quality of data depends also on the location of the equipment, effective deployment procedure, on the spot service interval, methods of protecting the sensors (such as anti-fouling materials) – but these aspects are not subject of this study.

Within a sensor network is essential to establish and enforce the same standard operation procedures, in order to gather consistent data and allow for comparison of data measured at different sites.

Further we make the assumption that sensor measured data is provided in the form of time series, where these values are accessible from a sort of historical database and that metadata describing sensors and measurement procedures is available.

Section II presents a high-level view of the Cyberwater project, the types of

data which are managed and what are the challenges raised when trying to maintain quality data.

Following section consists of a critical overview of currently adopted automated methods for quality control of sensor data.

Section IV focuses on proposing the procedure of testing and correcting data for Cyberwater project while the last Section is dedicated to conclusions and future research topics.

II. DATA QUALITY CHALLENGES IN CYBERWATER PROJECT

The Cyberwater research project [4] goal is to monitor the water pollution on a specific section of Cotmeana watershed, and respectively a 50 km length sector on Dâmbovița River, both localized on the Romanian river network. In this context the chemical and physical indicators of interest are: pH, Alkalinity, Conductivity, Total Phenols, dissolved Oxygen, N-NH₄, N-NO₂, N-NO₃, Total N, P-PO₄, and Magnesium.

The web based e-platform will allow visualizing, both monitored and modeled data in a custom built application based on ArcGIS, help managers to make informed decision and send alerts. The architecture of the cyber-system was presented in detail in [5], but from a data perspective, the data quality is a concern for three main categories of data, as depicted in Fig. 1:

- Monitored data (data acquired through sensor networks)
- Predicted data (that is data which is forecast based on statistical models)
- Propagated data (data which is mainly the output of the propagation of pollutant simulation algorithm, based on Mike11 software). The QA module has as input the type of data to be checked (individual measurement or a set of measurements)

and has access to historical relevant data (in case of monitored data, the time series recorded till that moment) and outputs a label which is appended to the data and which states the degree of confidence that the data is good and what was the specific problem identified (e.g. “Probably false spike”). The QA module and the database form a closed loop, the first is feed with historical data from the second and when the data is labeled with the QA tag it is persisted in the appropriate tables in the database.

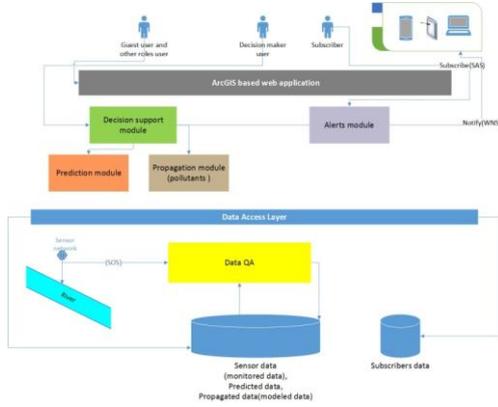


Fig.1 Data QA module within Cyberwater high level architecture

The “industry based standard data” criteria enumerated in the definition of “data quality” presented in Section I is achieved in Cyberwater using OGC (Open Geospatial Consortium) standards such as SOS (Sensor Observation Service) [1c] and XML implementation of Observations and Measurements (O&M) conceptual model [1d].

The tests described further are intended to address the other remaining criteria for quality and also to propose methods to replace bad or missing data.

III. CRITICAL REVIEW OF STATE OF THE ART ON AUTOMATED QUALITY CONTROL OF SENSOR DATA

1) False spikes

In a time series, a *spike or peak* is defined as being a point having

$$f(xi) > \theta$$

where $f(xi)$ is a function which associates a positive score to the i -th element of the time series (i.e x_i), θ is a user defined threshold value. The challenge further is to give a definition of the peak function $f(x_i)$ and compute θ . In [6] there were proposed five possible definitions of function f . One of them uses the local context of $2k$ points around x_i , k left neighbors and k right neighbors of x_i (first k points and last k points are not analyzed) to compute a mean of the differences between the considered point and its neighbors.

$$f(x_i) = \frac{(x_i - x_{i-1} + x_i - x_{i-2} + \dots + x_i - x_{i-k}) + (x_i - x_{i+1} + x_i - x_{i+2} + \dots + x_i - x_{i+k})}{2k}$$

Experiments show that choosing $k=5$ give good results. It was shown that a sampling distribution constructed on a statistical quantity of the measured data reach a Gaussian distribution when the number of samples tend to infinite regardless the population distribution which may be or not normal distribution (Central Limit Theorem) [7]. As new measurements are conducted and the valid values are added to time series thus leveraging the number of samples needed for statistics. If we consider as a particular case of statistic, the f function defined above, then this will closely approximate the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean and σ is the standard deviation.

The 68-95-99.7 rule (Fig.2) states that in a Gaussian distribution

almost all data (99.7%) is included within 3 standard deviation of the mean. If we chose $\theta = \mu + 3\sigma$ that means only 0.15% of data have such a big value for the peak function $f(x_i)$, so we can consider x_i as a candidate for being a spike. If a time series has too many spikes (more than N% from points presents spikes) in a given interval of time Δt , then this is a sign that the time series is candidate to be rejected (for example, spikes generated by communication interference).

The majority of studies dedicated to sensor data quality focuses only on the automated aspect of quality control. At this point we want to emphasize that not any spike is a false one and there is the risk to automatically discard good data for the sake of quality control. For example, an algal bloom event can determine a real spike in chlorophyll values. For this reason, the automated identification of spikes is not enough to draw conclusions, extra domain specific knowledge from scientists being required. We suggest that this knowledge can be formally expressed using ontologies [8] (such as polluter's ontology, regulations ontology, measurements ontology) and the expert *if-then* rules can be expressed with Semantic Web Rule Language (SWRL). The proposed method will comprise two phases:

-The first step is to automatically detect the spikes for the physical/chemical parameter measured. This can be achieved using numerical algorithms described in [6].

-Second step is to correlate the detected peak for a given parameter with the values for other parameters using domain specific knowledge. This can be also intuitively visualized if all time series values are plotted in the same coordinate system. For example, when it rains then this event usually will cause a

decrease in conductivity and an increase in turbidity. If a peak for turbidity is detected and there is no decline for conductivity this suggests that the spike for turbidity is a false one and there could be a problem with corresponding sensor. A visit to the site can identify and remediate the problem (for example, a mass of algae grown inside isolated the sensor from the natural environment). A false spike is removed and replaced with an interpolated value based on the previous and next P points (e.g. $P=4$).

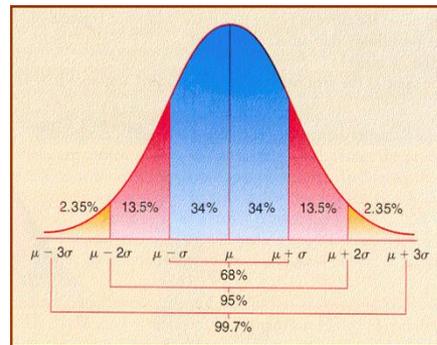


Fig.2 68-95-99.7 rule (source: <http://www.oxfordmathcenter.com/drupal7/node/290>)

2) Data blocked at a fixed value

If sensor reported the same measured value for a number of consecutive measurements this may be a sign of a defect. Using the same metrics described at III.1) this time we can look for points in the time series that minimize $f(x)$.

$f(x) < \mu - 3\sigma$ indicates with a probability of 99.85% that data are possible frozen at a given constant value in the time window defined by the k points. An alternative to the usage of statistical indicators is to choose for θ the resolution of the measurement device.

3) Data outliers

This is data which lies outside the admissible interval. As an example, a bug in the firmware can determine the omis-

sion of a decimal point, the value measured by the sensor being correct, but the wrong value being written in the data logger.

The admissible interval $[\min_value, \max_value]$ can be:

a) *User defined*

This can be flagged as a soft error and can be addressed by interpolating N values. In other cases, when a constant bias through time is observed, an offset adjustment can be applied.

b) *Defined in the specification of the sensor.*

This is a hard error and the value is rejected.

For example, one of the sensors used in the Cyberwater project has the associated metadata described in Fig.3

Manufacturer	Honeywell
Model	AH657E00
Measured parameter	pH
Min value	6.5
Max value	8.5
Measurement unit	Standard pH units

Fig.3 Metadata from sensor specification

The expected measured values must be within $[6.5, 8.5]$ interval, meaning that any value not in this range (e.g 8.7) is rejected.

4) *Missing data*

a) *Data gaps in sensor data*

Generally this can be defined as missing data in time series which can affect the project goals and can be checked based on timestamps. As an example, a defect connection between a sensor and the data logger can result in an increased

number of dropped data points. We propose the use of following metrics:

-Gap width, $w(g)$ – the number of consecutive points that are missing and which together form a gap

-Distance between two gaps, $d(g_i, g_{i+1})$ – the number of existing points between two consecutive gaps, g_i and g_j

-Number of gaps NG – total number of gaps in a time series.

The time series depicted in Fig.4 has a time step between observations of 3 seconds, $NG=3$, $w(g_1)=w(g_3)=3$, $w(g_2)=4$, $d(g_1, g_2)=5$, $d(g_2, g_3)=1$.

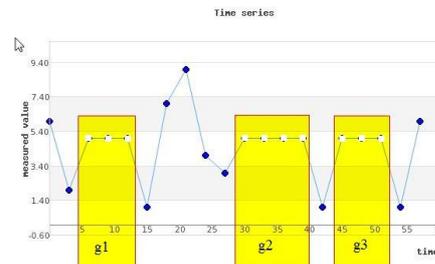


Fig.4 Data gaps in a time series

Based on predefined threshold, the series can be rejected in if some conditions (or logical OR/AND combinations) evaluates to *true*:

- There exist g and $w(g) > Tw$ (if gap has more than Tw missing points)
- $NG > N$ (there are too many gaps)
- $d(g_i, g_{i+1}) < d$ (frequent consecutive gaps)

Values for the thresholds Tw , N and d can be computed in the same manner as described at 1) and 2), by creating a sampling distribution of the corresponding statistic of interest (e.g. number of points in a gap) and then applying the 68-95-99.7 rule.

When gaps are considered reasonable enough to not reject the series (above logical conditions evaluate to *false*) the gap can be replaced by a segment of a

cubic best fit spline constructed to guess the curvature of time series.

b) Null data points

Null data points are measurement values reported with 0.0 values and can flag a problem of the instrument. In other cases these can represent real natural phenomena so, again, as for spikes, they cannot be rejected automatically but can be flagged as *suspect values* and further information is needed to decide if there was a problem with sensor or not. If the test described at III.2 also passes indicating that there is not an isolated case, but rather data frozen on zero over a time window, that may suggest that the sensor was calibrated during that interval.

c) Surrogates

Sometimes chemical/physical parameters of interest, rather than being directly measured, can be deduced from other measured parameters (surrogates) using regressive equations or neural networks ([9], [10]). A couple of examples are presented below ([10]):

$$\begin{aligned}
 Cl &= 1.74 \log SC - 3.14 \\
 SSC &= 3.29 NTU - 6.54 \\
 TN &= 0.00317 NTU + 0.0234 T - 0.0000655 SC + 0.469
 \end{aligned}$$

where Cl is chlorides (mg/l), SC is specific conductance (microSiemens/cm at 25°C), SSC is suspended sediment concentration (mg/l), NTU is Turbidity (measured in NTU), TN is total azote (mg/l) and T is water temperature (°C).

The estimation of parameters using surrogates can prove very useful for a couple of use cases:

- Missing data points in a time series

- Missing an entire time series for a specific parameter because an adequate sensor is not available (e.g. there are not sensors to measure phosphor's compounds or sediments [10]). This is especially of high importance for decision support systems where fast actions are needed to avoid catastrophic events.
- For concordance and validation purposes, check the numerical methods based estimated values (e.g. interpolation) vs. surrogate's estimation.

5) *Mean shifts*

Mean shifts corresponds to changes of the mean on some intervals (segments).

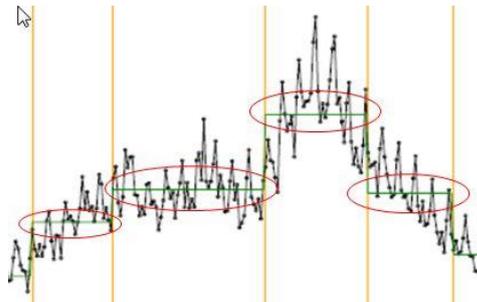


Fig.5 Changes in means of a time series

In literature this problem was studied for signal processing and time series in general and for environmental regime shift in particular [11]. Significant changes of the mean (Fig.5) can give important indications in the context of water quality monitoring:

- An increasing/decreasing of pollution
- A possible problem with the sensor

If the difference between means of two consecutive means is greater than a three-

shold value then it can be decided that is a sensor problem and both segments can be rejected (hard error). To practically asses the means shift we will be using the R implementation available on the open collaborative project CPA [12].

IV. QUALITY CONTROL PROCEDURE FOR CYBERWATER DATA

Given the quality criteria described in Section I we designed a Quality Control (QC) procedure consisting in a suite of seven tests (listed in Fig 5). All tests,

except T4, have as input an individual measurement, while T4 applies to an entire time series. All data, including the one that is identified and labeled as having problems (and thus rejected) is persisted for archiving. The rejection process takes place at logical level, the value being removed from the time series but not physically removed from database.

Test number	Problem verified	Label applied	Solution	Threshold	Data Quality criteria addressed	Input
T1	False spikes	“Probably false spike”, „False spike”	“False spike” is replaced with interpolated value	Statistic	Accuracy	Individual observation
T2	Data blocked at constant value	“False constant”	Data rejected*	Statistic/ Resolution of instrument	Accuracy	Individual observation
T3	Data outliers	“Not in range”	If user defined- interpolate If spec defined – reject data	Specification/User defined	Accuracy	Individual observation
T4	Data gaps	“Missing data”	Either reject or replace with best fit spline segment or using surrogates(if possible)	Statistics/ User defined	Completeness	Entire time series
T5	Null values	“Probably false null”, „False null”	If T2 failed then reject data	0.0	Accuracy	Individual observation
T6	Mean shift	“Segmentation”	Reject both segments	User defined	Consistency, Accuracy	Individual observation
T7	Date/Time values are not as expected	“Bad time-stamp”	Adjust date/time based on Counter and interval of sampling	N/A	Time-stamped, Consistency	Individual observation

*) When data is rejected and not replaced then automatically execute T4 to check the impact of removing the bad data

Fig.5 Tests included in the QC procedure for Cyberwater

V. CONCLUSIONS

In this paper we presented criteria for evaluating the quality of sensor acquired data (time series) and further discussed methods to verify if data is plausible or not. Bad data is replaced using numeri-

cal methods (spline segments, interpolation) or surrogates relations. While we were discussing about water pollution data, the solutions described can be applied also for other environmental

management systems (air, soil, etc). Future research topics and developments will consider: redundant data, uncertainty estimation, parallelization of tests, and assessment of sensor health based on identified wrong values.

ACKNOWLEDGMENT

This research is part of the CyberWater project supported by the UEFISCDI PN II, PCCA 1, nr. 47/2012.

REFERENCES

- [1] „Automated quality control methods for sensor data: a novel observatory approach”, J. R. Taylor and H. L. Loescher, *Biogeosciences Discussions*, Volume 9, Issue 12, 2012, pp.18175-18210
- [1a] <http://www.gs1.org/gdsn/dqf>
- [1b] http://ofmpub.epa.gov/sor_internet/registry/datastds/home/whatisadatastandard/#a2
- [1c] <http://www.opengeospatial.org/standards/sos>
- [1d] <http://www.opengeospatial.org/standards/om>
- [2] *Water Quality Monitoring - A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes*, Edited by Jamie Bartram and Richard Ballance, Published on behalf of United Nations Environment Programme and the World Health Organization, 1996 UNEP/WHO, ISBN 0 419 22320 7
- [3] *Revolutionizing Science and Engineering through Cyberinfrastructure*. Blue-Ribbon Advisory Panel on Cyberinfrastructure, NSF Report, (Jan, 2003). <http://www.nsf.gov/od/oci/reports/atkins.pdf>
- [4] PN-II-PT-PCCA-2011-3 joint research project Cyberwater <http://cyberwater.cs.pub.ro/>
- [5] Sorin Ciolofan, Mariana Mocanu, Anca Ionita, *Cyberinfrastructure architecture to support decision taking in natural resources management*, *Control Systems and Computer Science (CSCS)*, 2013 19th International Conference on, May 2013
- [6] G. Palshikar, *Simple Algorithms for Peak Detection in Time-Series*, In *Proceedings of 1st IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence*, Ahmedabad, India, Jun. 2009.
- [7] John A. Rice, „*Mathematical Statistics and Data Analysis*”, Cengage Learning; 3-rd edition (2006)
- [8] Ahmedi, Lule, Jajaga, Edmond and Ahmedi, Figene. "An Ontology Framework for Water Quality Management" Paper presented at the meeting of the SSN@ISWC, Sydney, 2013.
- [9] "Application of Artificial Neural Networks for the Prediction of Water Quality Variables in the Nile Delta", Bahaa Mohamed Khalil, Ayman Georges Awadallah, Hussein Karaman, Ashraf El-Sayed, *Journal of Water Resource and Protection*, 2012, 4, 388-394
- [10] J.S. Horsburgh, J.A. Spackman, D.K. Stevens, D.G. Tarboton, N.O. Mesner, "A sensor network for high frequency estimation of water quality constituent fluxes using surrogates", *Environmental Modelling and Software*, vol. 25 (9), pp. 1031-1044, 2010.
- [11] Sergei Rodionov, "A brief overview of the regime shift detection methods", In: *Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability*, V. Velikova and N. Chipchev(Eds.), UNESCO-ROSTE/BAS Workshop on Regime Shifts, 14-16 June 2005, Varna, Bulgaria, 17-24.
- [12] CPA project <https://sites.google.com/site/changepointanalysis/home>