Automated Assessment of Data Quality in Marine Sensor Networks

G. P. Timms
Tasmanian ICT Centre, CSIRO
GPO Box 1538
Hobart, TAS 7001 Australia

P. A. de Souza Jr.
Tasmanian ICT Centre, CSIRO
GPO Box 1538
Hobart, TAS 7001 Australia

L. Reznik
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623-5608 USA

***Abstract*-Inference on aspects influencing data quality should be based on or derived from experience. In this work we implement a Fuzzy Set Theory [1] approach to provide an assessment of data quality from a near-real time marine sensor network. The Fuzzy System firstly captures tacit knowledge using membership functions, and then uses these membership functions along with statistics derived from the data themselves to assess the quality of each measurement from the sensor network.**

I. Introduction

Quality assurance and quality control of data from Marine Sensor Networks (MSNs) are essential, but are neglected in many MSNs. This may be a result of the difficulty of effective implementation, a higher priority given to sensor deployment and data streaming compared with data quality, or a lack of critical applications of the network requiring high quality data. Any system aiming at the assessment of data quality in MSNs must consider, for each sensor, aspects such as time since deployment, time since last calibration, the quality of hardware supporting the sensor operation and data communication (e.g., electronics, wiring, welding, housing), the potential influence of human activity where the sensor is located (e.g., ports, storm water drains in urban areas), quality of the sensors (e.g., anti-biofouling capability), etc. Once the MSN reaches hundreds (or even tens) of sensors in size, quality assurance and quality control of data becomes a serious issue which needs to be addressed. Data streaming in a near-real time, large-scale sensor network needs to be followed by automated quality assessment. Clear presentation (visualisation) of the quality assessment for users is also critical.

Uncertainty classification

Over the years, many different approaches to evaluating and expressing the uncertainty of measurement results have been developed and adopted. Because of this lack of international agreement on the expression of uncertainty in measurement, in 1977 the International Bureau of Weights and Measures (BIPM), in collaboration with the various national metrology institutes, proposed a specific recommendation to solve this problem, the 100-page *Guide to the Expression of Uncertainty in Measurement* [2] (or GUM as it is now often called). The GUM has been adopted by the US National Institute of Standards and Technology (NIST) and most of NIST's sister national metrology institutes throughout the world, such as the National Research Council in Canada, the National Physical Laboratory in the United Kingdom, and the Physikalisch-Technische Bundesanstalt in Germany. The GUM standard approach includes the following points:

1. The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:

    - Type A. Those which are evaluated by statistical methods
    - Type B. Those which are evaluated by other means

2. The components in category A are characterized by the estimated variances s_i^2 (or the estimated "standard deviations" s_i) and the number of degrees of freedom v_i. Where appropriate the covariances should be given.

3. A Type B evaluation of standard uncertainty is usually based on scientific judgment using all of the relevant information available, which may include:
    - previous measurement data,
    - experience with, or general knowledge of, the behavior and property of relevant materials and instruments,
    - manufacturer's specifications,
    - data provided in calibration and other reports, and
    - uncertainties assigned to reference data taken from handbooks.

This recommended approach to uncertainty estimation therefore takes into account both statistics derived from the data themselves as well as expert knowledge derived from previous experience of the sensors and the systems under study.

The application of fuzzy logic to uncertainty estimation

Fuzzy set theory is particularly well-suited to applications where the solution is highly dependent on human experience;

978-1-4244-5222-4/10/$26.00 ©2010 IEEE

because of either imprecise information available or the empirical nature of the problem. A fuzzy system incorporates fuzzy heuristics and knowledge that defines the terms being used in the former level. Indeed, it is possible to encode linguistic rules and heuristics directly reducing the solution time since the used expert's knowledge can be built in directly. In addition, its qualitative representation form makes fuzzy interpretations of data natural and an intuitively plausible way to formulate and solve several problems. Qualitative aspects can be implemented and can also be updated making this system useful to solve problems of which an analytical solution is very difficult or not possible.

In this paper we develop a fuzzy-logic based approach to automated data quality assessment and apply it to data collected by the TasMAN sensor network [3].

II. THE TASMANIAN MARINE ANALYSIS NETWORK (TASMAN)

TasMAN has been setup to provide real-time and near real-time data from the waterways of south-east Tasmania. This data is crucial in managing the multiple uses of these estuaries and coasts, including heavy industry, aquaculture, shipping, tourism, commercial fishing and recreational uses, while ensuring environmental sustainability.

The network consists of both fixed sensors (shown in Fig. 1) and a mobile node, a small AUV [4].

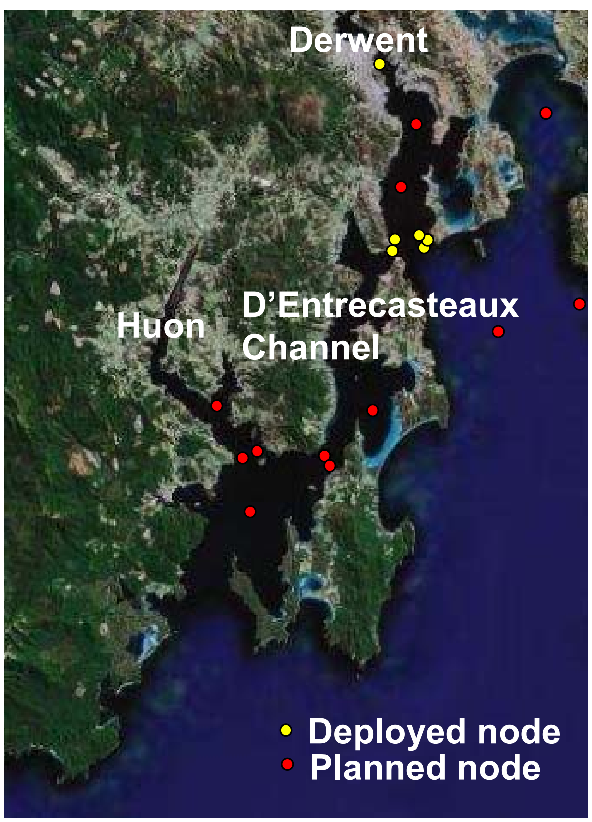

Figure 1. Location of fixed nodes in the TasMAN sensor network.

Each fixed node has a number of conductivity, temperature and depth sensors at different depths in the water column, with dissolved oxygen sensors to also be deployed in the near future.

The fixed nodes generally communicate data via a 3G mobile phone network although, where a cluster of nodes are close together, they communicate with each other via a proprietary radio protocol, with a "sink" node transmitting data via the mobile telephone network. The real-time data are then integrated with models to produce nowcasts and forecasts of river conditions, and are made available via web and mobile applications for use by our partners.

III. AUTOMATED DATA QUALITY ASSESSMENT

The structure of the TasMAN sensor network is hierarchical in nature. Each node in the network contains one or more sensors. The nodes themselves are arranged in groups, called clusters and the network is made up of a number of clusters. The network covers a large area (roughly 40km x 40km) and the cost and practicalities of distributing sensors evenly throughout that area, as well as the power requirements of long-range communication, make it desirable to concentrate the nodes in a number of clusters which communicate their data back using existing 3G mobile phone infrastructure. Above the level of the network itself, a number of networks and data sources in the region can also be combined as part of the Sensor Web [5].

As a consequence, the QC checks performed on data also follow this hierarchy. The purpose of QC is to flag data which may be faulty or which may require further follow-up (*e.g.*, data which may correspond to a pollutant release or other event in the waterways).

A. Sensor QC

Relevant aspects potentially influencing the quality of MSN data were captured in membership functions (i.e., the fuzzification process). The aspects considered in this implementation were: quality of the sensor, time deployed in water, time since last calibration, and rate of change of measured parameter. Fig. 2 depicts the membership function for the time since deployment for a number of sensors. The information for each measurement from the MSN was assessed against the membership function and the membership functions combined in a defuzzification process that produced an overall assessment of the quality of a measurement.

In this initial implementation, data quality flags were derived from the "defuzzification" process, that were a subset of the IODE flags used by both the Argo floats [6] and the Australian Integrated Marine Observing System (IMOS) [7]:

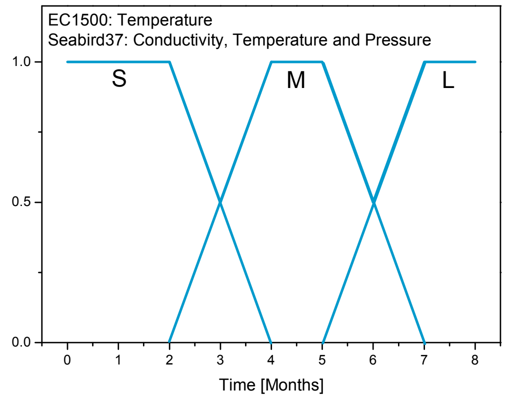

Figure 2. Location Membership function for the parameter, time deployed in water, for a number of sensors used in the TasMAN network.

0 – No QC performed
1 – Good data
2 – Probably good data
3 – Bad data that are potentially correctable
4 – Bad data

Table 1 shows the characteristics of the membership functions for a subset of the sources of uncertainty in the TasMAN network.

For a given measurement, small, medium and large uncertainty values were assigned for each of the Type A and B parameters. Weightings were provided to the different parameters and they were then combined by simple addition. Once combined the overall small value was referred to as the QC green fraction, the overall medium value as the QC yellow fraction and the overall large value as the QC red fraction. The fractions were normalized to one. Finally the data was compared against thresholds based on the range of historical measurements in the region. Where data fell outside the expected range, the fuzzy output was overwritten and the data's red fraction set to 1.

TABLE I
MEMBERSHIP FUNCTIONS FOR THE PARAMETERS CONTRIBUTING TO DATA QUALITY

Parameter	Small	Small → Medium	Medium	Medium → Large	Large
Rate of change of temperature (°C/min)	< 0.03	0.03 to 0.05	0.05 to 0.07	0.07 to 0.10	> 0.10
Rate of change of conductivity (µS/cm/min)	< 50	50 to 100	100 to 150	150 to 250	> 250
Time deployed in water (days) – EC1500 conductivity sensor	< 7	7 to 28	28 to 42	42 to 56	> 56

B. *Node QC*

Node QA/QC compares data obtained from individual sensors (of the same type) within the node and identifies any large discrepancies. The output from this process is used to modify the overall data quality assessment at the sensor level. The modification could be either downwards or upwards. For example, if all sensors in the node have recorded a decrease of 2 °C in water temperature in the past 30 minutes, then it may indicate an event at this location and we can have more confidence in the data from the individual sensors. However, if only one sensor records this decrease, it is suggestive of a fault with that sensor.

In this initial implementation, where there are only two sensor locations on a node, the percentage difference between the sensor readings was used as the metric for Node QA/QC. While this works well in the case of our node in Sullivan's Cove, Hobart, it is recognised that this is an exception, as the water is well-mixed at this location In general, layering of the water column would make an alternative such as testing for hydrostatic stability of the water column by calculating water density at each depth more appropriate at the node level.

The "defuzzification" step was carried out by mapping the Green, Yellow and Red QC fractions to the IODE flags as follows:

IODE 1:	Green QC ≥ 0.9
IODE 2:	0.9 > Green QC ≥ 0.7 and Red QC < 0.2
IODE 3:	Not used
IODE 4:	All other data

In the present work we describe results obtained up to the node level. An extension of this work, which is in progress, is to bring the fuzzy inference to cluster, network and web level.

IV. RESULTS AND DISCUSSION

The system was initially applied to data collected at the CSIRO Wharf in Hobart between 20 February 2008 and 13 July 2008. This node was composed of two EC-1500 Temperature and Conductivity sensors fixed to the wharf, one at a depth of 1.0 meter below chart datum and the other at a depth of 9.5 meters below chart datum (chart datum was at the level of the lowest possible astronomical low tide). The sensors were calibrated in mid January 2008 and deployed on 19 February 2008.

Fig. 3 shows a subset of the temperature data for 22 February 2008. At around 7:40am there was a small spike in temperature. While the rate of change of the temperature fell into the medium and large sections of the membership function for this parameter, the fuzzy system maintained an overall higher QA/QC green fraction on data during this period. This was a consequence of the similar behavior of both temperature sensors on the node (making a faulty sensor less likely), and the small amount of time that had passed since both calibration and deployment of the sensors.

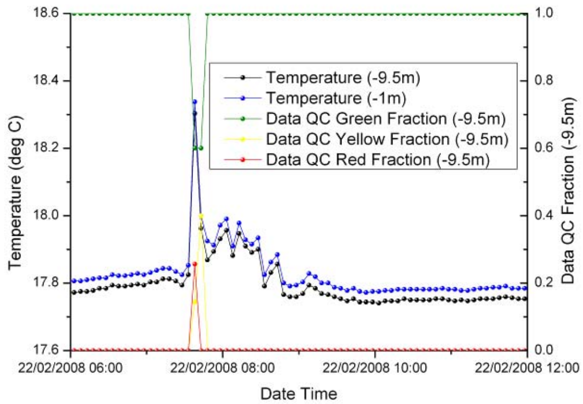

Figure 3. Measured water temperature at the CSIRO Wharf during 22 February 2008, along with the data QC fractions automatically generated for these data.

Fig. 4 shows the temperature data flags automatically generated for the two temperature sensors for the period shown in Fig. 3.

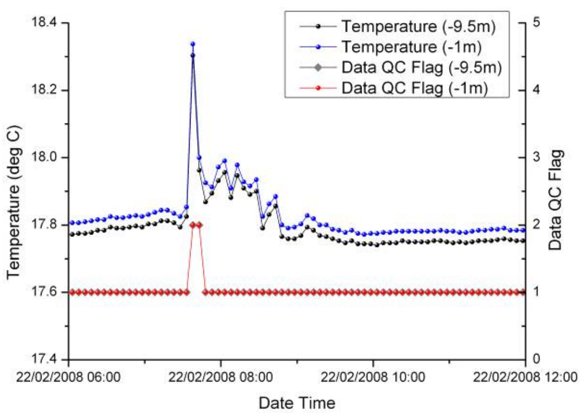

Figure 4. Measured water temperature at the CSIRO Wharf during 22 February 2008, along with the data quality flags automatically generated for these data.

Fig. 5 shows the data flags for conductivity data on 7 March 2008. At approximately 11:30 am there was a drop in the measured conductivity at a depth of one meter. In this case both the rate of change of conductivity and intranode membership functions are large; the latter because the conductivity at 9.5 meters remained unchanged. The data from this sensor for this period is flagged as 4 (bad).

The remaining data shown in Fig. 5 is flagged as 2 (probably good). It is not flagged as good because this conductivity sensor has been deployed for three weeks. Because of previous experience with this type of sensor in moored deployments, the "time deployed in water" membership function has been set up so that the quality of the data determined by the fuzzy approach begins to decrease after two weeks.

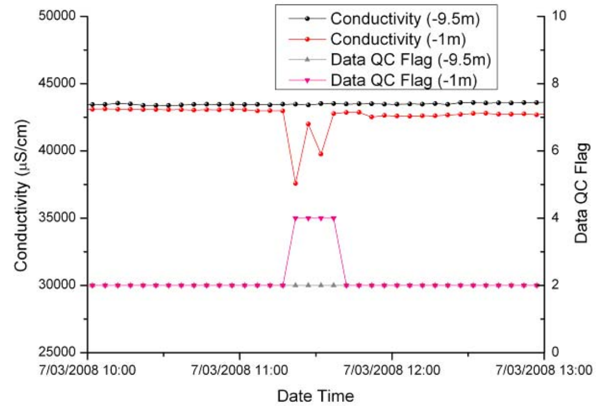

Figure 5. Measured conductivity at the CSIRO Wharf during 7 March 2008, along with the data QC flags automatically generated for these data.

Fig. 6 shows the system in operation over 4 days in late May, at which time the sensors were due for maintenance.

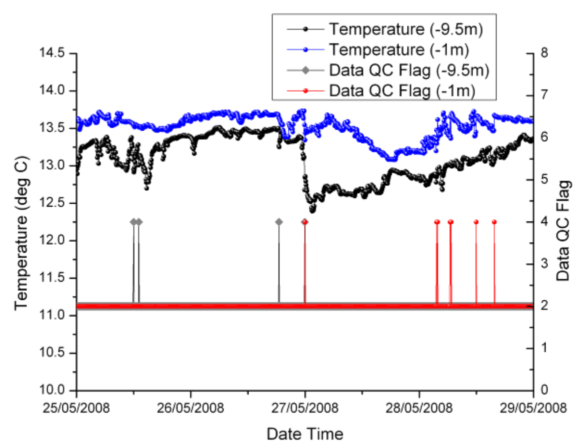

Figure 6. Measured water temperature at the CSIRO Wharf from 25 to 29 May 2008, along with the data quality flags automatically generated for these data.

While the system is able to successfully detect spikes and jumps in the data, it has difficulties with determining the quality of the data following a sudden jump. The jump may have been caused by a sensor fault and this would suggest that future data at this location is also unreliable. A further membership function will be added in future to differentiate between jumps caused by sensor faults and those likely to be caused by changes in the measured parameter.

One key advantage of the fuzzy set approach to automated data QC is that as well as the output data quality flag, fractional information about data quality is also available through the QC fractions. This creates a few possibilities for furthering our understanding of data quality from individual sensors and from the network as a whole. For example, by plotting the green, yellow and red QC fractions from a particular sensor on a ternary graph (Fig. 7) and then seeing how they evolve with time, we can quickly visualize whether data from this sensor is slowly decreasing in quality and whether sensor maintenance is likely to be needed in the near future.

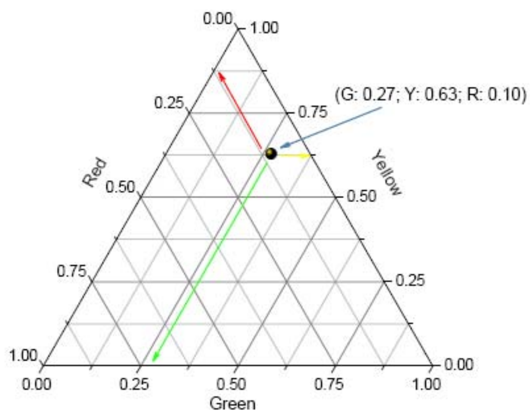

Figure 7. Assessment of overall data quality using a ternary graph. The point is positioned (quality assessed) based on the combination of all membership functions (e.g. Figure 1).

Another option is to plot the most recent QC fractions from all sensors in the network; this provides a snapshot of network health and will assist in prioritizing resources for network maintenance.

V. CONCLUSIONS

The automated collection of data (e.g. through sensor networks) has led to a large increase in the quantity of environmental and other data available. The sheer quantity of data means that automated QA/QC is necessary to ensure that the data collected is fit for purpose.

We have demonstrated a framework for automated data quality assessment which incorporates not only data statistics but also the expert knowledge of domain specialists and also takes advantage of the hierarchical nature of many real-world sensor networks.

The approach outlined here is able to quickly assess data quality of datasets containing millions of records and it is planned to adapt the system so that it can be applied to real-time data streaming from the network.

Further work is underway to develop visualisation tools which will enable users to both quickly assess network health and, once the system has been extended to cluster and network level, to detect events within the network.

ACKNOWLEDGMENT

The Tasmanian ICT Centre is jointly funded by the Australian Government through the Intelligent Island Program and CSIRO. The Intelligent Island Program is administered by the Tasmanian Department of Economic Development, Tourism and the Arts.

This research was conducted as part of the CSIRO Wealth from Oceans National Research Flagship.

REFERENCES

[1] M. K. Urbanski and J. Wasowski "Fuzzy Approach to the Theory of Measurement Inexactness", *Measurement*, **34**, 67-74 (2003).

[2] "Guide to the Expression of Uncertainty in Measurement", International Organization for Standardization, Geneva, 1995.

[3] G. P. Timms, J. W. McCulloch, P. McCarthy, B. Howell, P. A. de Souza, M. D. Dunbabin, and K. Hartmann, "The Tasmanian Marine Analysis Network (TasMAN)", presented at IEEE Oceans 2009, Bremen, Germany, 11-14 May 2009.

[4] M. Dunbabin, J. Roberts, K. Usher, G. Winstanley, and P. Corke, "A Hybrid AUV Design for Shallow Water Reef Navigation," in *Proc. IEEE International Conference on Robotics and Automation ICRA 2005*, 18–22 April 2005, pp. 2105–2110.

[5] K. Delin, "The Sensor Web: A Macro-Instrument for Coordinated Sensing", *Sensors,* **2**, pp. 270-285 (2002).

[6] A. Wong, R. Keeley, T. Carval, et al., "Argo Quality Control Manual", Version 2.5, October 2009, p32.

[7] S. Mancini, K. Tattersall and R. Proctor, "IMOS NetCDF User's Manual", Version 1.2, April 2009, p46.