

Session I3: Data Quality Management Tools and Techniques

Room B117-119
10:00 – 11:30 am

0460
I3-1

Bayesian Surprise as a Tool for Validating Data from Sensor Networks

Wesley Brooks^{1,2}, Mike Fienen², John Walker² and Randy Hunt²

¹*Univ. of Wisconsin, Madison, Wis., USA*, ²*US Geological Survey Wisconsin Water Science Center, Middleton, Wis., USA*

Rapid growth in the quantity of monitoring data is accompanied by a need to assure the data quality. Here we demonstrate applying the method of Bayesian surprise as a tool for automating part of the quality-assurance process.

The method of Bayesian surprise is a statistical technique that was developed for computer vision applications, with the goal of mimicking the way that the human brain allocates attention. We use it to detect surprising or anomalous events in a stream of data.

Consider an automatic sensor embedded on a platform (*e.g.*, a dissolved oxygen sensor on a buoy). With each successive observation, we update our belief about the distribution of the measured quantity. The Bayesian surprise is measured by an f-divergence (such as the Kullback-Leibler divergence) between the prior and posterior distributions.

We demonstrate that the method of Bayesian surprise can be used in the quality-assurance process to quickly identify when a sensor is malfunctioning, or to detect real but surprising events. We begin with the simplest case, in which a stream of data is modeled by a distribution that is defined by just a few parameters. The parameters are chosen so that when we use each observation to update our belief about the quantity being measured, the parameters may change but the functional form of the distribution will not. This setup is convenient because the entire data model can be summarized by the distribution of the parameters and because the Bayesian surprise has a simple closed-form expression. Thanks to its simplicity, this setup is appropriate for calculating surprise in real time onboard an embedded sensor platform.

From this starting point, we consider extensions of the method of Bayesian surprise. One extension is to a multivariate data stream from a sensor platform where there is some unknown relationship between the various quantities (*e.g.*, turbidity and dissolved oxygen) being measured. Another is to a network of sensors where the spatial relationship between sensors may allow the method of Bayesian surprise to detect and locate anomalous events, and to distinguish them from sensor faults.

0058
I3-2

Quality Control of Observational Datasets Collected in a Real-Time Monitoring Network

Mohammad Islam^{1,2}, James Bonner^{1,2}, William Kirkey¹, Chris Fuller¹ and Temitope Ojo¹

¹*Clarkson Univ., Potsdam, N.Y., USA*, ²*Beacon Institute for Rivers and Estuaries, Beacon, N.Y., USA*

Real-time continuous environmental observatories are essential to characterize the intensity, frequency, and effects of episodic events, both natural and anthropogenic, in dynamic aquatic systems. Assuring the data quality created by these systems can be a challenge, due to the large volumes of data generated in an unattended fashion. Both automated and manual methods are employed in a four tiered process to maintain data quality from monitoring nodes in the River and Estuary Observatory Network (REON). REON is a joint venture of the Beacon Institute, Clarkson University, General Electric Inc. and IBM Inc. to monitor New York's Hudson and Mohawk Rivers using various sensor platform types at multiple nodes within the network. At the first level, pre-deployment laboratory calibrations are performed on each instrument. In the automated second level, post-processed results are filtered to remove outliers and improperly formatted data resulting from sensor malfunctions and/or data acquisition software glitches. At the third tier, filtered data is visually inspected over the web in near real-time on a daily basis. Any observed data anomalies are subjected to detailed analysis (*e.g.*, spatial-temporal analysis, sensor synergy) and if necessary, any questionable sensors are inspected, cleaned, and recalibrated or replaced as necessary. At the fourth level, data quality is verified by comparing data across two different systems. This is accomplished using a mobile sensor package that is essentially collocated with an autonomous monitoring system, and the two systems are operated simultaneously. The results from this fourth level evaluation at an autonomous profiling station located in a Superfund site, with active dredging to remove legacy polychlorinated biphenyl (PCB) contamination, are presented. These results indicate that data quality degrades over time for optical sensors where the optics are exposed to the water column, while other sensors

can function for extended durations with no loss of data quality. They also show that field cleaning and recalibration procedures can generally restore the optical sensors to normal working order.

0385
I3-3

Streamlining and Automating Water-quality Time-series Records Processing

Patrick Rasmussen

US Geological Survey, Lawrence, Kans., USA

The United States Geological Survey (USGS) has recently developed three new software tools to streamline and automate water-quality (WQ) time-series records processing in order to improve record quality and consistency and decrease costs. A handheld software package called CHIMP (Continuous Hydrologic Instrumentation Measurement Program) is used to collect field data during WQ monitor site visits. The data is then automatically uploaded to the USGS database where the Automated Correction Loader (ACL) retrieves the data for computing and uploading data corrections to the database. Water quality monitor review (WQMReview) compiles data tables and plots for hydrographers to review and provide quality assurance. The presentation will include examples of how USGS uses CHIMP, ACL and WQMReview to streamline and automate WQ time-series records processing.

0473
I3-4

Visualization and Exploration of Time-Dense Monitoring Data with the USGS Data Grapher

Stewart Rounds

US Geological Survey, Portland, Oreg., USA

Time-dense, high-frequency datasets collected from continuous water-quality monitors and stream gages present challenges for data visualization and exploration. Exploration of patterns and trends in the data, as well as comparisons to other datasets at the same or different sites can be greatly facilitated by software tools written specifically for such tasks. The Data Grapher system is a set of Internet-based tools developed by USGS that allows users to access high-frequency time-series data-including water-quality, meteorological, streamflow, and other data- and quickly create customized graphs, color maps, and tables.

Time-series graphs, XY plots, color maps, wind-rose diagrams, and customized data tables can be created through the Data Grapher. Daily statistics and running averages can be applied to the high-frequency data, thereby allowing the creation of datasets to match a criterion used by a water-quality standard. Time-series data can be compared from two or three sites on the same graph or compared for specific months over multiple years. Color maps can be created to visualize daily or seasonal patterns. Percent saturation data for dissolved oxygen and total dissolved gas can be computed and graphed. Specialized XY plots of dissolved oxygen versus temperature, with contours of percent saturation, are easily constructed. Wind rose diagrams and polar plots of wind speed and wind-direction can be created. Many XY and polar plots offer the option to highlight data from a user-selected time period. Graphs are displayed on-screen and can be downloaded in many common formats such as EPS, PDF, PNG, JPG, and WMF.

The Data Grapher is available for use with USGS datasets from many areas of the United States (Oregon, Tennessee, Idaho, Colorado, California, others) but is not yet available nationwide. In Oregon, the system serves more than 1000 time-series records from close to 300 sites. To use Oregon's Data Grapher, go to <http://or.water.usgs.gov/grapher/>. Example graphs and video tutorials are available on the website to serve as an introduction to the use of the system.