QA/QC of Measurements collected on a very large scale: Rainfall and Streamflow datasets

Sparks, Ross
CSIRO Mathematics, Informatics and Statistics
Locked Bag 17, North Ryde NSW2113
Sydney, Australia
E-mail: Ross.Sparks@csiro.au

Okugami,Chris
CSIRO Mathematics, Informatics and Statistics
E-mail: Chris.Okugami@csiro.au

Abstract

The paper offers simple robust algorithms for checking spatio-temporal multivariate consistency of large volumes of measured data. The checks involve temporal data collected on a spatial grid, as well as several related measurements collected on the same spatial grid over time. The checking process involves computationally efficient methods of estimating expected values and variances used to judge measurement consistency. CUSUM and EWMA plans are advocated for flagging consistently small biased measures.

1 Introduction

Data quality is defined as data being fit for purpose (Borowski and Lenz, 2008). This paper focuses on the narrower aspect of measurement consistency checking. This checking process helps isolate poor measurements in large datasets. This is meant to make the task of ensuring quality data easier by focussing the manual data checking effort on inconsistent measurements only. Measurement inconsistency is assessed in terms of a combination of:

- **Spatial consistency:** The measurement is consistent with several measures at several geographical close locations.
- **Temporal consistency:** Measurements are consistent with measures at the same geographical location at earlier or later times.
- **Multivariate consistency:** Measurements are consistent with related measures made at the same geographical location and same times.

Often we are interested in the joint spatial, temporal and multivariate consistency checking rather than marginal consistency checks. An example of this is the checking of river flows at a specific location given several upstream flows and rainfall measurements. If all stream flow measurements are unusually large in the catchment at a specified time, but all are spatially consistent with each other in the whole catchment and consistent with the local rainfall values, then these measurements will not be classified as unusual even when they are unusual in the marginal sense. Thus, avoiding a paradox very similar to Simpsons paradox (Blyth, 1972). However, if one of the stream flows is unusual relative to local flow measurements made further upstream or downstream, then this will be flagged as inconsistent. This joint multivariate, spatio-temporal consistency checking is where we depart from other recommended univariate data quality checking.

The process of checking for unusualness is based on measurement departures from their conditional expected values. The measures we condition on are called explanatory variables (e.g. for river flows, rainfall is an explanatory variable). These explanatory variables are usually measured at the same time, before or after the measured value being considered.

The explanatory variables are selected to provide the interpolation of y with the least absolute error. Find the conditional expected value for the measured value (y) given the set of explanatory variables (x) (denoted $\mathrm{E}(y|x)$). Unusualness is measured in terms of how much y departs from

$\mathrm{E}(y|x)$. Let the variance for this departure be denoted by σ^2, then assign unusualness codes on the scale of extremely unusual, very unusual, unusual, within the expected range, and close to expected using the following rules:

1. $||y - \mathrm{E}(y|x)|| > 5\sigma$ **extremely unusual**
2. $4\sigma < ||y - \mathrm{E}(y|x)|| < 5\sigma$ **very unusual**
3. $3\sigma < ||y - \mathrm{E}(y|x)|| < 4\sigma$ **unusual**
4. $2\sigma < ||y - \mathrm{E}(y|x)|| < 3\sigma$ **somewhat unusual**
5. $\sigma < ||y - \mathrm{E}(y|x)|| < 2\sigma$ **within expected range**
6. $||y - \mathrm{E}(y|x)|| < \sigma$ **close to expected**

We consider the most common more complicated situation where no duplicate measurements are available. Section 2 discusses the consistency checking process. Section 3 looks at detecting persistent errors commonly caused by measurement device failure. Section 4 discusses how our quality consistency check fit into a larger QAQC framework. We end the paper with some remarks regarding the future challenges and dealing with missing values.

2 Algorithms for measurement consistency checking

As mentioned earlier, consistency checks can be broken down into temporal, spatial, and multivariate checks. Multivariate spatio-temporal consistency checks are typical in environmental or hydrological applications. Algorithms appropriate for each of these situations are described in the subsections to follow. Note that the emphasis here is on automatic consistency checking in large scale data collection applications such as sensor networks.

Assume an example of consistency checking stream flow measures in a catchment (Figure 1). We develop algorithms that will check consistency of these jointly with:

- identical measurements made earlier in time at the same site for real-time checking, or made earlier or later in time at the same site for batch checking, i.e. temporal consistency;
- identical measurements made at neighbouring sites (at the same time or earlier), i.e. spatial consistency;
- related measurements made at the same site or neighbouring sites, i.e. multivariate consistency (e.g. checking consistency of flow considering local rainfall as a related measure).

It is important to emphasis that we build interpolating time series models as opposed to forecasting models. In other words, we allow explanatory variables collected at the same time as the response. Therefore, the models used are generally more accurate than the usual time series models.

Regression models for predicting the measurement is used to check for consistency. The model predicted value is taken as the conditional expected value. The spatial variables are considered as explanatory variables. Upstream flow measurements in a river used to explain downstream flows are time adjusted for the duration it takes for water to flow between the two sites, i.e. if y_t is the downstream flow measurement made at time t, x_t is the upstream flow measurement made at time t, and it takes q units of time for water to flow between the upstream site and the downstream site in the river, then the correlation between y_t and x_{t-q} should maximize the $\mathrm{cor}(y_t, x_{t-\tau})$ over all selections of τ. In most systems a constant delay of q time units is unrealiastic, but in some circumstances, for example, where velocity is measured the delay can be estimated. Alternatively, delays are likely to be relate to rainfall and therefore could be estimated as a function of rainfall.

Two data sources are used: training data (assumed clean) and test data. The training data should be representative of the time period to be assessed for consistency. The test data are to be checked for consistency. Training data provides starting values for estimated regression parameters and are used to estimate the best time lags between explanatory variables and response values. These parameters are used to find expected values and variances for the first measurements in the test data.

The algorithms supplied below have two phases:

1. The start up phase used to provide initial estimates of model parameters.
2. The recursive estimation phase used to update parameter estimates with each new observation in the test data.

The training data are used to establish the starting estimates for parameters. Cleaned training data directly prior to the time period being checked is the default. The training data are also used to establish the explanatory variable temporal lags which best explained the response variable. Data checks are carried out for each catchment separately as described below.

The process has the following steps (stream flows used as an example):

Step 1: Specify the sequence measures will be checked in the catchments.

Step 2: Specify the set of explanatory variables to be used.

Step 3: Specify the checking start and finish date.

Step 4: Specify the training dataset.

Step 5: Estimate the time delay (q_i) for each variable ($x_{i\,t-q_i}$) that best predict the response. Measures of explanatory usefulness consider here are correlation measures. That is, we examine correlation measures as a function of q_i; selecting the q_i that maximises the correlation.

Step 6: Fit the regression model, e.g., for flows fit:

$$y_t = \beta_0 + \beta_1\, x_{1\,t-q_1} + \ldots + \beta_k\, x_{k\,t-q_k} + \alpha_1 y_{t-1} + \alpha_2\, y_{t-2} + e_t$$

where:

- y_t is the flow being checked for consistency measured at time t;
- $x_{i\,t-q_i}$ is either an upstream flow or an upstream rainfall measurement measured at time $t-q_i$ to account for delays;
- β_i are the regression parameters associated with the influence the ith explanatory variable;
- α_i are the regression parameters associated with carryover influences of measurements made at the same site at earlier times.

Note the time series model has no moving average component only autoregressive components; this is to avoid the need for iterative estimation.

Step 7: Check that $t+1$ observation for unusualness by calculating the differences between measured and expected (predicted using the model).

Step 8: Repeat steps 6 and 7 for all measurements in the catchment made at time $t+1$

Step 9: Update estimates using observation $t+1$**.** Repeat steps for $t+2$ until all observations in the test sample have been exhausted,

3 Detection of errors that persist

Detecting persistent problems early is paramount for improving data quality.

3.1 Detecting persistent biases

The CUSUM plan of Page (1954) is useful for accumulating enough memory of small persistent one-sided departures from expected for detecting them early. The CUSUM for high sided departures is given by ($S_t^U = 0 = S_t^L$)

$$S_\tau^U = \max(0, S_{\tau-1}^U + \widehat{e}_{\tau|\tau-1}/\widehat{\sigma}_{n\,\tau} - k)$$

and CUSUM for low sided departures is given by

$$S_\tau^L = \min(0, S_{\tau-1}^L + \widehat{e}_{\tau|\tau-1}/\widehat{\sigma}_{n\,\tau} + k)$$

for $\tau = t+1, t+2, \ldots$, where $k(>0)$ is some suitable offset. Measurements are classified as persistently biased at time τ and before if either

$$S_\tau^L < -h_c \quad \text{or} \quad S_\tau^U > h_c$$

where h_c is a suitable threshold designed to deliver an acceptable false alarm rate. After each bias signal, the CUSUM statistic is restarted at zero for the calculation of the next CUSUM value. This allows users to assess whether the bias persists after a signal. Other work on CUSUM of recursive residuals (Ploberger and Kramer, 1992) differs slightly from our approach outlined above. We use standardised interpolation errors whereas they use recursive residuals, the standardisation we use involves only past data for real-time consistency checking, whereas their approach used all the data, and unlike our CUSUM statistic, their CUSUM is not based on Page (1954) with an offset k.

3.2 Detecting increased uncertainty in measurements early

Assume that the expected variance is given by σ_T^2. We recommend the following statistic

$$V_\tau = max(\sigma_T^2, (1-\lambda_4)V_{\tau-1} + \lambda_4 \widehat{e}^2_{\tau|\tau-1}).$$

This statistic tracks local changes in variance in the same way as the EWMA statistic tracks the moving average, and therefore has all the advantages of the EWMA in terms of being simple and easy to update with large volumes of data. The V_τ statistic flags a significant increase in uncertainty from the target whenever $V_\tau > \sigma_T^2 h_v$ where $h_v(>1)$ is a suitable threshold designed to deliver an acceptable low false alarm rate. Not allowing the V_τ statistic to drift below the expected value helps the fast reponse to flagging any sudden increases in its value. If V_τ is allowed to drift very much lower than expected, then, in its worst case, the monitoring process can suffer a severely delayed response in detecting the change.

4 Concluding remarks

The paper discusses the development of simple QAQC algorithms for automatically detecting and flagging unusual measurements. When data are missing, a first pass interpolated value is used to patch the missing data. An alternative is to use regression trees or random forests to fit the models, because these models can be fitted to datsets with partially missing values without the need for patching.

Future data quality challenges include:

1. Optimal design for multivariate spatio-temporal consistency checking.
2. Scalable real-time systems.
3. Integrating the consistency checking process into a data quality assurance plan.
4. Periodic reviews to improve the measurement process.
5. Monitoring the number of errors and the percentage of missing data over time.

Acknowledgement: We acknowledge: Hydro Tasmania for their applied hydrological knowledge that helped facilitate the technology emphasis and for supplying the datasets, Bureau of Meteorology for their funding of this work. We also would like to thanks Dr Peter Toscas for his comments on an earlier version of the paper resulting in improvements.

References

[1] Blyth, C.B. (1972). "On Simpson's Paradox and the Sure-Thing Principle". Journal of the American Statistical Association. 67: 364366.

[2] Borowski, E. and Lenz, H-J. (2008). Design of a workflow system to improve data quality using oracle warehouse builder. JAQM. 3: 198-206.

[3] Page, E. S. (1954). Continuous Inspection Scheme. Biometrika, 41: 100115.

[4] Polberger, W. and Kramer, W. (1992). The CUSUM Test for OLS Residuals. Econometrica, 60(2), 271-285.