Title

Hunting data rogues at scale: Data quality control for observational data in research infrastructures

ISBN

9781538626863

Authors

Pastorello, G
Gunter, D
Chu, H
et al.

Publication Date

2017-11-14

DOI

10.1109/eScience.2017.64

Peer reviewed

Hunting data rogues at scale: data quality control for observational data in research infrastructures

G.Z. Pastorello, D.K. Gunter, H. Chu, D.S. Christianson, B. Faybishenko, Y. Cheah, S.W. Chan, S. Dengel, T.F. Keenan, F.L. O'Brien, A. Elbashandy, and D.A. Agarwal
Lawrence Berkeley National Laboratory
contact: gzpastorello@lbl.gov

C. Trotta and D. Papale
University of Tuscia

N.F. Beekwilder and M. Humphrey
University of Virginia

E. Canfora
Euro-Mediterranean Center on Climate Change

C.M. Poindexter
California State University, Sacramento

***Abstract*—Data quality control is one of the most time consuming activities within Research Infrastructures (RIs), especially when involving observational data and multiple data providers. In this work we report on our ongoing development of data rogues, a scalable approach to manage data quality issues for observational data within RIs. The motivation for this work started with the creation of the FLUXNET2015 dataset, which includes carbon, water, and energy fluxes plus micrometeorological and ancillary data measured in over 200 sites around the world. To create an uniform dataset, including derived data products, extensive work on data quality control was needed. The unpredictable nature of observational data quality issues makes the automation of data quality control inherently difficult. Developed based on this experience, the data rogues methodology allows for increased automation of quality control activities by systematically identifying, cataloging, and documenting implementations of solutions to data issues. We believe this methodology can be extended and applied to others domains and types of data, making the automation of data quality control a more tractable problem.**

I. Research Infrastructures

As Research Infrastructures (RIs) become more widespread, several new data management challenges are surfacing. Many of these RIs have to support observational (or sensor) data, in particular data contributed by independent third parties. This trend is prevalent enough to be seen in many domains, e.g.,: genomics (GenBank [www.ncbi.nlm.nih.gov/genbank/]), clinical medicine research (CDISC [www.cdisc.org]). In environmental sciences, this is even more pervasive, with many efforts focusing on central data repositories fed by many sources and serving long-term and synthesis types of studies. Examples of these RIs are plentiful: LTER, the Long Term Ecological Research Network [lternet.edu] and its agriculture-focused counterpart LTAR, or Long-Term Agroecosystem Research [ltar.nal.usda.gov]; the International Soil Carbon (ISCN [iscn.fluxdata.org]) and the International Soil Moisture (ISMN [ismn.geo.tuwien.ac.at]) Networks; the National Ecological Observatory Network (NEON [www.neonscience.org]); the ecosystem phenology web camera network PhenoCam [phenocam.sr.unh.edu]; among many others. For this work we focus on eddy covariance data: measuring carbon, water, and energy fluxes between land and the atmosphere. These measurements are uniquely positioned in spatial and temporal scales to allow answering a wide range of questions about ecosystems, from soil microbiology to long-term effects of disturbances. Regional networks around the world support data contribution from individually operated sites measuring these fluxes, e.g.: AmeriFlux [ameriflux.lbl.gov] in the Americas, ICOS [www.icos-ri.eu] and the EU DB cluster [www.europe-fluxdata.eu] in Europe, and OzFlux [www.ozflux.org.au] in Australia and New Zealand.

FLUXNET [fluxnet.fluxdata.org] is a network of networks bringing together regional networks, with the Fluxdata.org project serving as its RI. Synthesis datasets are a signature product of FLUXNET, with standardized data from many sites around the world allowing comparisons among sites, large-scale and long-term studies of land-use change, calibration and validation of remote sensing, and constraining and evaluation of ecosystem and Earth system models, and many other applications. The creation of these datasets involve extensive work to coordinate data sent from sites and mainly make the data uniform in terms of formatting, application of correction processing steps, generation of derived data products, and especially in terms of data quality uniformity.

II. Data Quality Assurance and Quality Control

For the creation the FLUXNET2015 dataset [1] – the most recent FLUXNET global fluxes dataset – we estimate that about 90% of the human effort was dedicated to data quality assurance and quality control (QA/QC) activities. This is not unusual for observational data: malfunctions or misconfigurations can affect sensors; external agents can interfere with measurements; and incorrect data processing choices might be used while handling data.

Most current QA/QC procedures involve human input and interaction to interpret and check results, making them hard to scale. While standardization (e.g., in data collection protocols) helps prevent some types of issues, sensor data will still be subject to problems. Automation of QA/QC tasks is the only way to scale such procedures along with the data deluge currently being created by sensor data.

Even if bad analyses can be performed with big data as much as small data, it is much harder to tell the difference in the former case. Data QA/QC scalability must be addressed or

978-1-5386-2686-3/17 $31.00 © 2017 IEEE
DOI 10.1109/eScience.2017.64

many science domains will not be able to take advantage of new developments in instrumentation and big data techniques.

III. Data Rogues

Data rogues seek to systematically capture tacit knowledge from domain experts into automated tests and corrections for data QA/QC. The clear procedures to build a data rogue reduce the dependence on the semantics of a particular dataset, building all common tasks into set of steps and supporting tool, and leaving only domain- and dataset-specific details to be documented and coded into the data rogue.

Fig. 1 depicts the creation of a data rogue, beginning with the identification of a data issue – usually a manual process that depends on visual and analytical inspection of one or more data variables. Next, the effects of a data issue are characterized into an observational data pattern [2], which will define what to look for when evaluating data. Whatever is known about causes and possible corrections to the issue are cataloged next. Although highly interdependent, it is important to clearly distinguish between causes for a data issue and its effects in the data. If these are mixed, the requirements for implementing detection methods might become too broad, making such methods hard to code and prone to problems like large numbers of false positives.

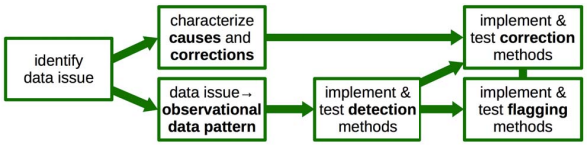

Fig. 1. Sequence of steps in the creation of a data rogue.

The implementation of detection methods can use a wide variety of techniques, from simple range tests (like physically plausible values) to machine learning-based approaches (depending on availability of sufficiently large number of instances of the data issues for training sets). In some cases, it is possible to apply correction methods (e.g., forcing agreement between variables using linear regression relationships). Finally, flagging methods are an important component of a data rogue, not only documenting the occurrence of an issue and its scope within the data, but also documenting the application of correction methods when applied. Flagging can be done with additional variables to mark individual records, or proving both original and filtered data as separate variables.

The issue illustrated in Fig. 2 motivates our example of a data rogue. Shortwave solar radiation measured using two different sensors is plotted along with theoretical maximum radiation, calculated for the same site. The diurnal cycle of the measurements is shown for two days of the year, and the measurements are aggregated in 20-day windows to compensate for cloud coverage. It is clear that the diurnal cycle of the measured radiation is offset (by about one hour) with respect to the theoretical curve (**data issue**). This is commonly caused by data logger miscofiguration (**cause**) and can be corrected by shifting the timestamps (**known correction**). This

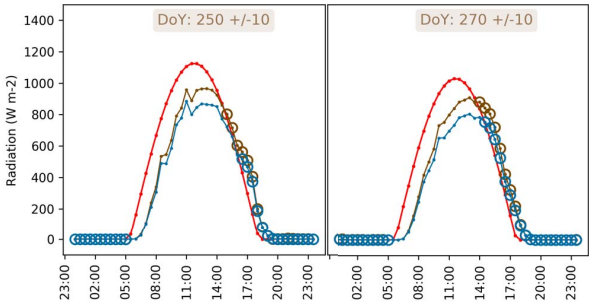

Fig. 2. Incoming shortwave solar radiation measurements (blue and brown lines) are temporally shifted with respect to theoretical maximum radiation (red line) for a site in North America.

timestamps shift (**data pattern**) can be characterized by a misalignment between the measured and theoretical curves or different times for the peak of each curve. The circles around data points for the measured data show when these points go over the theoretical curve, and counting them in morning and afternoon periods allow the identification of the shift (**detection method**). Applying a corrective shift to the timestamps of the affected records in the measured data (**correction method**) and creating an additional variable in the dataset showing the amount by which each timestamp was shifted (**flagging method**) fixes and documents the data issue.

IV. Concluding Remarks

The identification of data issues is highly dependent on domain knowledge and difficult to predict: new and exciting types of problems always seem to come up in observational data. So it is likely that data quality control for observational data can never be fully automated. But once an issue is first detected, checking new data for the same issue should be made simple and systematic. In a parallel with software testing, tools like integrated frameworks for unit testing, code versioning, and continuous integration allow testing of large and complex codebases, providing systematic and reproducible safeguards against errors and mistakes. Data rogues seek to offer similar functionality for data quality control.

Acknowledgments. This work was supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (Deduce Project) and the Office of Biological and Environmental Research (AmeriFlux Management Project and FLUXNET Partnership Project), under Contract No. DE-AC02-05CH11231. It was also partially supported by the European H2020 project COOP+ (GA 654131) and ENVRIPLUS (GA 654182).

References

[1] G. Pastorello, D. Papale, H. Chu, C. Trotta, D. Agarwal, E. Canfora, D. Baldocchi, and M. Torn, "A new data set to keep a sharper eye on land-air exchanges," *Eos*, vol. 98, 2017, https://doi.org/10.1029/2017EO071597.

[2] G. Pastorello, D. Agarwal, D. Papale, T. Samak, C. Trotta, A. Ribeca, C. Poindexter, B. Faybishenko, D. Gunter, R. Hollowgrass, and E. Canfora, "Observational data patterns for time series data quality assessment," in *Proc. IEEE 10th International Conference on e-Science*, 2014, pp. 271–278, https://doi.org/10.1109/eScience.2014.45.