Workshop to define quality management standards for data completeness in derived data products

Wade Sheldon (GCE), Don Henshaw (AND) and Ken Ramsey (JRN)

Introduction

Integration of data sets from multiple sources is an important prerequisite for network-level and cross-site synthesis studies in LTER, and scaling or re-sampling of primary data is often performed as part of this process. In the "Trends" project, for example, fine-scale (e.g. hourly) measurements collected by LTER sites are rescaled to monthly or yearly time steps for analysis and display of long-term trends. However, recent discussions within the LTER IM community and in ASM workshops have elucidated problems with both handling and reporting of missing and flagged measurements in source data when derived data sets are produced by and for the Trends project and also by the LTER ClimDB/HydroDB database. **We believe it is essential that LTER establish standards for re-sampling and temporally scaling incomplete and flagged measurement data used in synthesis projects, and for documenting this process to provide quality control information for the derived data products**. In addition to benefiting LTER synthesis projects, these standards will inform development of data integration tools by LTER sites, LNO, and other projects and organizations such as NCEAS, SEEK, the Canopy Databank and CUAHSI.

Background

Missing values are ubiquitous in long-term environmental data sets, due to unexpected equipment failure, maintenance downtime, data loss, extreme events, and many other factors. Statistically, missing values can be random with respect to time and parameter magnitude, or they can be correlated with value ranges, events or specific conditions and represent significant measurement bias. Both the number and distribution of missing values in primary data can affect statistical analyses (e.g. sum, mean, and distribution) and therefore influence results when data are temporally rescaled. For example, yearly mean temperature estimates would be biased towards higher values if a temperature sensor tended to fail during prolonged freezing events. Similarly, failure of rain gauges during hurricanes or monsoons would bias monthly or annual precipitation totals downward. The presence of biased monthly or annual statistics in rescaled data could then produce false trends or signals in the derived data set, potentially leading to incorrect scientific conclusions.

In addition to missing values, many data sets contain values that have been "flagged" with various qualifiers by the data contributor or information manager during quality control and quality assurance analysis. These flags may indicate values are outside the expected range for conditions (i.e. questionable), are invalid due to recording or processing errors, or are estimated to replace missing values. The statistical consequences of including flagged values in aggregated data are often unclear; however, information about the numbers and types of qualified values in source data is important for proper interpretation and use of derived data products.

A general approach is clearly needed for managing quality of derived data products and retaining critical information about the completeness and quality of source data in metadata for derived data sets and integrated databases. This represents a key part of the infrastructure required in data analysis workflow environments and automated data integration tools for tracking data lineage and understanding data provenance of derived products.

Workshop Goals

We propose to conduct an initial two day workshop to discuss the issues described above, particularly as applied to data integration and synthesis currently being performed in LTER by the Trends project and ClimDB/HydroDB. Workshop topics will include:

- LTER site practices for reporting missing values and QC/QA flags in data sets, and potential for standardization
- Statistical algorithms currently used for aggregating data in Trends, ClimDB
- Strategies for temporally rescaling incomplete data series (i.e. literature review)
- Strategies for reporting on data completeness and quality in metadata for derived data products and for flagging or omitting derived values based on this information
- Potential extension of the EML metadata standard to support QC/QA information for values in data columns (e.g. defining column dependencies, semantics)

During this workshop we will identify a core working group that will carry on the discussion beyond the workshop and formulate specific recommendations and standards.

Participation

Informal discussions at the 2006 LTER ASM meeting indicated that there is very broad interest in this subject across the ecoinformatics community. We will therefore try to include representatives from multiple LTER sites, LNO, the Canopy Databank Project, NCEAS, SEEK and CUAHSI in the workshop, particularly individuals directly involved with large-scale data integration projects (e.g. Trends, ClimDB, Grasslands Long-term NPP Databank, CUAHSI Hydrologic Information System) and synthesis tools (e.g. Kepler). We anticipate inviting 16 people to the workshop, with additional participation via PolyCom VTC.

We will conduct pre-meeting planning and post-meeting discussions via PolyCom VTC and email, and establish a general email list that also includes interested parties who could not attend the workshop.

Workshop Products

The initial product of this workshop will be a formal set of recommendations (i.e. white paper with potential for publication) for handling and reporting missing and flagged measurements present in source data in all derived data sets and associated metadata. Draft recommendations will be vetted through the IM Committee, NISAC and informatics partner groups (NCEAS, SEEK, Canopy Databank Project and CUAHSI) for comment, then final recommendations will be submitted to NISAC as a proposed standard for derived data sets produced as part of LTER synthesis projects and NIS modules.

We will publicly post all workshop documents on the LTER web site and also pursue publication of our findings in the scientific literature. We anticipate that this effort will lead to NSF proposals for additional workshops and development of necessary cyber-infrastructure and standards for quality management of derived data.

Workshop Budget

We are requesting a total of $8500 to conduct the workshop described above. In order to maximize participation while minimizing costs, we plan to hold the workshop in Las Cruces, New Mexico in conjunction with the upcoming Trends Editorial Committee meeting (1-3 Feb 2007 at JRN). This will eliminate travel costs for participants from JRN (Ramsey, Laney, Peters) and minimize travel costs for invited participants from LNO (Brunt, Servilla, San Gil) and those already attending the Trends meeting (Sheldon, Schildhauer, Kratz). Consequently, we are not including costs for three participants from JRN and are budgeting reduced amounts for six participants from LNO and Trends (i.e. hotel and meals only, using cost estimates provided by JRN staff).

Because the remainder of the participant list is not yet finalized, we are estimating $1000 per person for seven additional participants to cover air fare, ground transportation, lodging, meals and miscellaneous meeting expenses, as suggested by LNO staff.

Budget Breakdown

Lodging (2 nights) and meals (3 days) for 3 LNO participants (Brunt, Servilla, San Gil)	$750
Lodging (2 nights) and meals (3 days) for 3 Trends meeting participants (Sheldon, Kratz, Schildhauer)	$750
Air travel plus lodging (2 nights), ground transportation and meals (3 days) for 7 additional participants (Henshaw, 3 LTER reps, Canopy Databank rep, CUAHSI rep, SEEK rep), including miscellaneous meeting expenses	$7000
Total Request	**$8500**