Dafit – a new work flow oriented approach for time efficient data preparation, validation and flagging of time series data from environmental monitoring.

Ludwig Ries[1]

Abstract

Standardized quality assurance according to UN/WMO Global Atmosphere Watch (GAW) data quality objectives is essential for a homogeneous high level of data quality in GAW world data centers. However data processing and data preparation often times is done individually and in a non-standardized way. As an additional problem interactive data validation can be a very time consuming step on a GAW measurement station.

In front of this background a new approach for a software solution with a set of time efficient and standardized methods is proposed which also can be used as a tool for future standardization of data quality assurance on GAW measurement stations.

In general in order to prepare measurement data a set of methods is required. In the framework of this program the methods are ranked in two groups: first – simpler methods for managing structural changes and corrections in the time series data and second – higher developed methods for assuring a correct time structure, graphical control and flagging non representative data and for the calculation of differently or higher aggregated mean values and statistical values.

By fulfilling the auxiliary condition that data treatment at first has to be finished with application of methods of group one the user is practically free in finding and selecting its way for a solution. Once the user has found an ordered set of methods and parameters which is a sufficient solution for the preparation of the actual data, the solution can be stored as a set of parameters for the individual project. This enables the user for repeating the solution at another time to another set of time series data which are produced with the same data format from the same instrument. This structure is characteristic and useful for continuous environmental monitoring which produces a high amount of time series data with a constant data format. By saving all intermediate results in a hierarchical way with a fixed naming convention, in a later time the effect of all steps of the application of methods can be traced back in the workflow history.

Because the proposed methodology separates methods and its parameters, it can be a good substitution to many situations in data preparation whereas up to now normally for each little change the program source would have to be changed and compiled. This gives an essential facilitation to the whole process of data preparation of time series data in the daily monitoring. The set of methods works with ascii data files, in case each line has a time stamp with own date and time. The program works on files with arbitrary minute data or higher aggregated time data and provides precision up to the second. Dafit runs on Windows XP-Win 8.

1. General

This paper describes a work flow oriented methodology for efficient data preparation, validation and data flagging of monitoring time series data, following a more general approach.

[1] Federal Environment Agency, Global Atmosphere Watch Observabory Zugspitze/Hohenpeissenberg, Zugspitze, Germany

This software was written in order to ease the tedious and time consuming process of data preparation and to facilitate its standardization on Global Atmosphere Watch monitoring stations. It is relatively fast to install and to learn and applicable for all cases, when up to now elementary data preparation was required.
It helps in such situations, when practically the same program for data processing had to be altered only for a little change and then to be recompiled again and again.
It separates methods and parameters from each other and saves the parameters in an own project file. Because the program uses a format interpreter it can process a larger variety of ASCII files which are produced as an output of different data acquisition programs for different instruments or from larger data acquisition systems which are used at GAW Stations.

1.1 Concept

- Separate methods, parameters and processed data.
- Store parameters and data of a problem in an own project.
- Save each step of processed data, from the beginning to the final result in a hierarchical history (Workflow), which allows backtraceability.
- You receive an ordered set of methods and parameters, stored in the project (Solution).

2. How to work with (Workflow)

The program is directory oriented. At first a directory has to be created for data with a certain time interval, for example, one month. The maximum interval of data which can be processed at once is one year. Then a new project name for this data file is created and stored. In the next step the data structure of the file will be opened with an editor and has to be investigated by the user. At next the user will have to find what preparatory steps with the data file will have to be done in which order to receive a usable file for data validation. Usable file means: Each line contains a datum of one time interval. Each line (time interval) is readable. Missing gaps (sometimes e.g. during a power down, data acquisition does not work.) in the time series data time structure are substituted in the file with a missing data flag. The time stamps are in ascending order.
Finally a plan for the stepwise processing of the data will have to be made.
Besides this the following elementary requirements have to be regarded:
The data file has to be in ascii. Each line of the data file has to contain a date time header with fields of fixed width YYYY or YY, MM, DD, hh, mm (and ss optional). The order of time and date is arbitrary, as the time date fields are interpreted. Each line must contain a minimum of one column of data values and a maximum of 30 columns. At one time only one column can be processed. It is recommended to use list directed data format which means: numbers are separated by a comma or a semicolon, or a blank or tab. The columns after the time data header may consist of numbers and strings in arbitrary order.

2.1 Prepare and repair data structure

After the first preparatory steps which have been explained above, it has to be tested, whether methods from a first group of elementary methods to correct deficiencies in the data structure have to be applied. The application of these methods is optional. Following methods presently are given:

- **concatenate files** (e.g. daily files to a monthly or annual
  file or concatenation over an arbitrary interval)
- **inverse time order** (the data have to be in ascending order)

- **remove multiple empty lines and multiple separation marks** (for the list directed format each column has to be separated by one separating character)**.**
- **substitute separation marks** (e.g. TAB by semicolon, or comma or blank)
- **filter corrupted lines** (sometimes data acquisition produces corrupted lines. This filter removes such lines and produces a log file of removed information.)

After this the application of methods of the following second group of methods has to be planned.

2.2 Apply higher methods for time structure correction, validation, flagging and final aggregation

The following sequence of methods has shown practical usability for high quality atmospheric gas or areosol measurements for global atmosphere watch data.

1. **test and repair for a correct and consistent time structure**. This tests check for an equidistant time grid. Each single time step between start time and end time must be contained in the resulting output. In case there is a gap in the data file then it will be substituted with the appropriate time date fields and a missing data code for the missing data value.
2. **aggregate to a time series with higher time resolution in the minute scale** e.g. from 3 sec, 31sec, 146 sec data to 1 minute means. In case already 1 min data exist, then this step has to be omitted.
3. **first validation: check interactive plausibility of time series and flag with graphical editor**. Flag spikes, calibrations, outliers etc. (In one minute data files with higher time resolution, typically artifacts like spikes, outliers, short time deviations because of technical issues are found.)
4. **aggregate to a time series with lower time resolution, preferably in the hour scale.** E.g. from 1 min values to 30 min or 1hr mean values.
5. **second validation: interactive check with graphical editor**: Flag non representative episodes, local pollutions etc. (In lower time resolution data, process related information can be read and understood and validated. For example causes for short time sharp shifts in the level of measured values can be detected with a trajectory model like Hysplit or pollution events can be flagged. This step has development potential for inclusion of automated statistical and chemical or physical models for data validation.)

2.3 Methods of the second group for aggregation and validation are:

2.3.1 Time structure

This test is contained in the methods 2.3.2 and 2.3.3. It is performed each time when an aggregation of data is calculated.

2.3.2 Means of even intervals

This method works with a programmed chain of conditions. It works only for whole minute data. Data intervals with uneven seconds may not be treated. It is slower than method 2.3.3. (Both methods 2.3.2 and 2.3.3 calculate the arithmetic mean, standard deviation, and the percentage of available underlying data. A threshold value for this percentage can be selected. An actual existing percentage over or equal this boundary then allows calculation of a mean value.)

2.3.3 Means of broken intervals

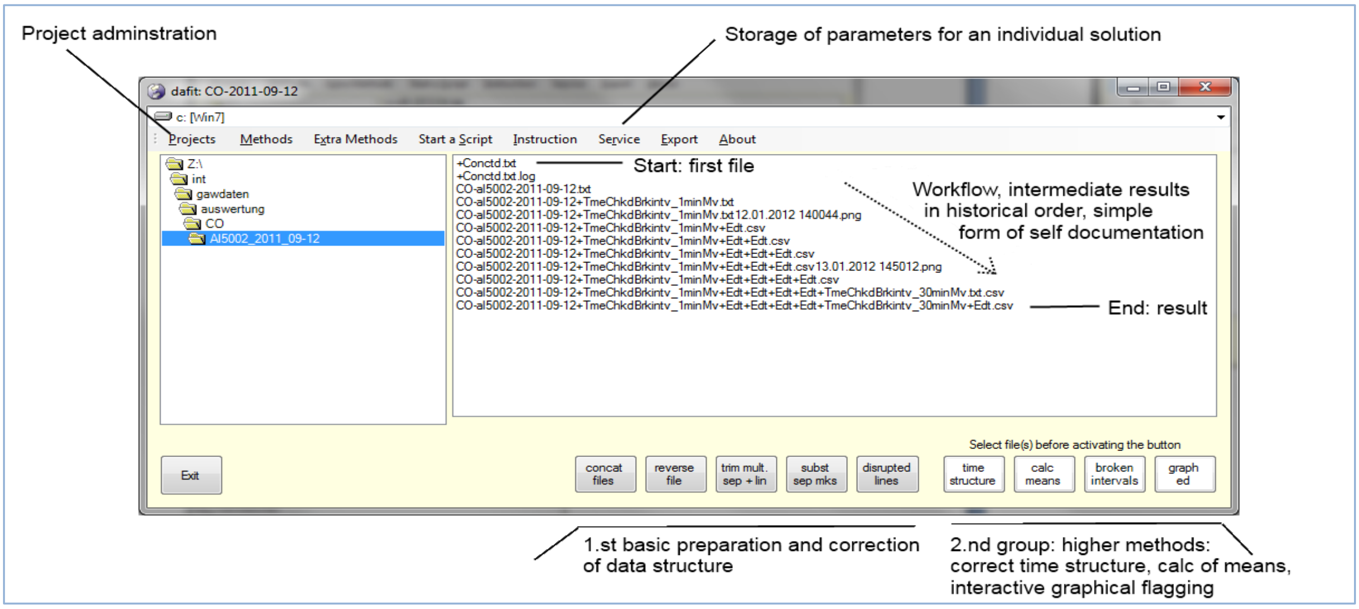
This method works with a large array which contains every second of one whole year. According to the start and end time of each mean value for input it's value is stored correctly. In a second step, new mean values will be calculated according to a selected equidistant time grid with arbitrary grid length in the minute range. With this method a 1min data file can be calculated on the same 1min data file. In that case input equals output, but the time structure has been tested for gaps and gaps have been filled in the output with missing data code.

2.3.4 Graphical editor

The graphical editor has various methods for navigation through large data files. It allows interactive graphical flagging with line editors. This means everything above (or under) an interactively drawn line will be flagged. Additional functions are given for scaling, zoom in and out, selecting a single point and it's flags and coordinates, for determination of n, mean and median from selected points in a predefined graphical rectangle. Flagging results and comments can be retrieved and analyzed at any time after the processing of the data. On a modern laptop with an I5 processor one whole year with about 500 thousand data is read in about 25 seconds. It allows simultaneously two hand operation: One hand works with predefined keys and the other operates the computer mouse. Against becoming tired, the function of both hands can be exchanged.

3. Program user interface

3.1 Start window and graphical editor



Project adminstration

Storage of parameters for an individual solution

```
+Conctd.txt
+Conctd.txt.log
CO-al5002-2011-09-12.txt
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv.txt
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv.txt12.01.2012 140044.png
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv+Edt.csv
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv+Edt+Edt.csv
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv+Edt+Edt+Edt.csv
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv+Edt+Edt+Edt.csv13.01.2012 145012.png
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv+Edt+Edt+Edt+Edt.csv
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv+Edt+Edt+Edt+Edt+TmeChkdBrkintv_30minMv.txt.csv
CO-al5002-2011-09-12+TmeChkdBrkintv_1minMv+Edt+Edt+Edt+Edt+TmeChkdBrkintv_30minMv+Edt.csv
```

Start: first file

Workflow, intermediate results in historical order, simple form of self documentation

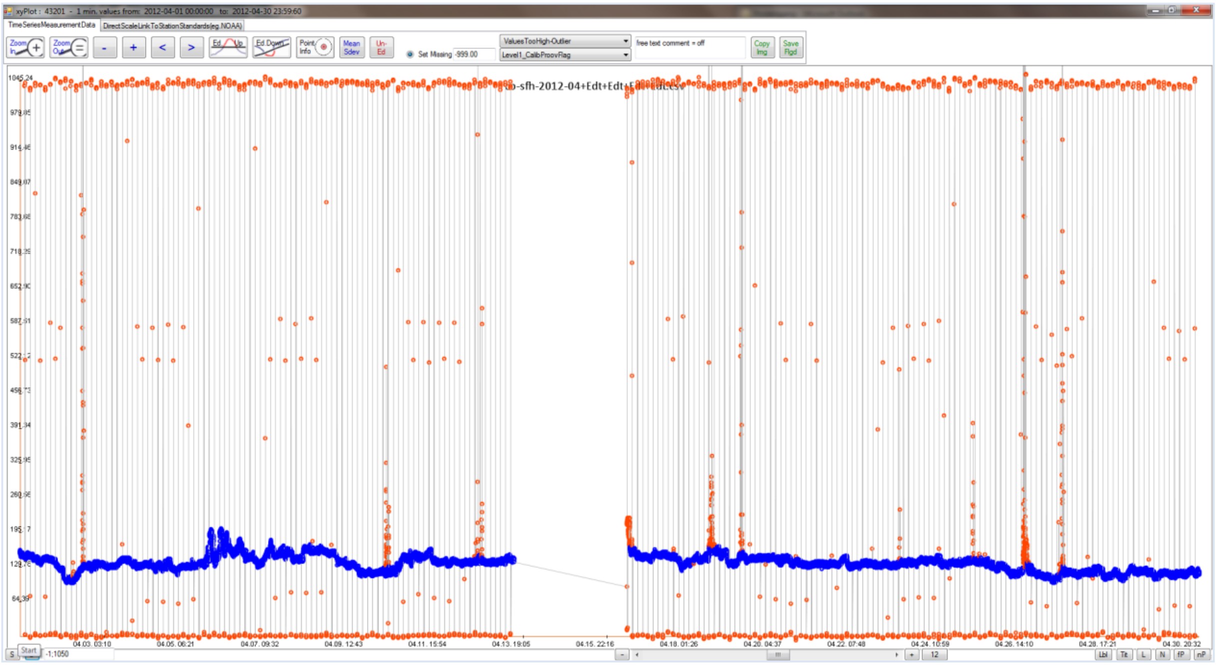
End: result

Select file(s) before activating the button

1.st basic preparation and correction of data structure

2.nd group: higher methods: correct time structure, calc of means, interactive graphical flagging

Figure 1
Start window of the program and its essential properties. 1. Create directory and copy data into it. 2. Create project and save. 3. Apply methods for basic preparation and correction 4. Apply higher developed methods for time structure correction, flagging and validation and final aggregation of data



xyPlot : 43201 - 1 min. values from: 2012-04-01 00:00:00 to: 2012-04-30 23:59:60

Set Missing -999.00

ValuesTooHigh-Outlier

Level1_CalibProovFlag

free text comment = off

Figure 2
Graphical editor for interactive control, flagging and plausibility check, with several functions for navigating in the data enables for a productive screening, validation and flagging of the processed time series data

Proposed methodology and software can facilitate essentially the process of data preparation of monitoring time series data. Dafit runs on Windows and is free for station personnel and institutions which work for the UN/WMO Global Atmosphere Watch program.

Bibliography

L.Ries, How to Keep the Quality Chain, Work Flow-Oriented Data Processing for Global Atmosphere Watch Measurement Stations. pp.109-112. In: Environmental Informatics and Systems Research. Vol. 2 Workshop and application papers. Eds. Olgierd Hryniewicz, Jan Studzinski, Anna Szediw. Shaker Verlag, 2007, Aachen, 292 p.

L.Ries. Standardized and automated data quality assurance at GAW Stations: concept, methods and tools, presentation at GGMT-2013 Conference, Peking.

GAW Technical Report No.206. 16th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases and Related Measurement Techniques (GGMT 2011) Wellington, New Zealand, 25-28 October 2011.