Explanation of data formats: table and diary

Introduction

The environmental measurements are specific: the data are brought into order according to time. This important property essentially simplifies data processing but it is not sufficiently exploited in common database systems. Another specialty of environmental measurements is the internal structure of the set of variables. A specific physical quantity can be measured at different stations using different sensors and following different time schedules. Every combination is to be formally considered as an independent variable. Example: air temperature at a certain station is recorded every 5 minutes at the heights of 2 m and 10 m, and additionally as hourly averages at the height of 2 m. These three variables have no formal relations when the data format does not contain a description of the structure of the variable set. In a large environmental dataset with unstructured variable set, the relations between variables are not visible and the data are hard to survey.

Diverse data models are used in the management of the environmental data. The pertinence of a specific format depends on the nature and applications of the data. Descriptions of numerous data formats are available on the web. A good introduction can be found on the web site of the British Atmospheric Data Centre

http://badc.nerc.ac.uk/data/dataset_index/,

which includes hundreds of links to different datasets.

DataDiurna (abbreviation DD) is a data format or a data model, which is optimized for saving, arranging, and managing of routine environmental measurements in scientific research. The format is explained in the document *DataDiurna_manual.pdf*. The manual contains a lot of information that is necessary only when the user is going to develop existing or create the new DD-datasets. Many of users may be satisfied with a reduced instruction how to extract tables from a completed dataset. Therefore the knowledge required for extracting tables from an existing dataset is available as a brief separate document *How_to_extract_tables.pdf*.

Table

Environmental measurements may yield values of different quantities like air temperature, pressure, size fraction concentrations of aerosol particles etc. We expect that all values are expressed by numbers. Typically the measurements are made according to a regular schedule and arranged as a rectangular table, which can be inspected and analyzed using MS Excel. A sample table below consists of four columns, one header row and 5 data rows.

Time	T:C	RH:%	p:mb
2012-09-21 14:30	14.7	47.8	1002.2
2012-09-21 14:45	14.1	50.5	1002.4
2012-09-21 15:00	13.2	54.1	1003
2012-09-21 15:15	11.9	62.5	1003.6
2012-09-21 15:30	10.8	66.3	1004.2

The table is a good format in case of uniform structure of measurements when every row contains values of the same set of simultaneously measured quantities. Rectangular tables can be browsed as spreadsheets using the programs like MS Excel, which additionally allow carrying out simple calculations. The tables are accepted as input in all packages of statistical analysis. The tables are used in DD as the input when importing the data and as the output when extracting the data from the repository for mathematical and statistical analysis.

Traditionally, a table is solely a data matrix, which has no space for additional information about the encoding of the time stamp and the nature of the variables. Hence necessary information is included into a description file, which is attached to every data table in DD.

Large datasets often consist of measurements made by different institutions during different time periods and are presented with a set of numerous tables, which have different structures. The time periods are typically partially overlapping, but the time grids of different tables may be different. The collections of measured quantities are partially overlapping, but the same quantity may be expressed in different tables according to different units. There are used many different methods of writing the time stamps.

A collection of tables of different style can be formally managed using database management software. However, the analysis of such a data will be very troublesome. Popular database software, e.g. MS Access, is optimized for the processing of the business data and do not consider the peculiarities of regular time-arranged environmental measurements.

In principle, a collection of tables could be merged into a joint universal table, but in practice this procedure is often accompanied with crucial difficulties. The duration of the measurement series in a large dataset is long and the dataset often contains some short-period campaign measurements of many variables, which are not presented in long data series. Every new variable inserts a new column into the joint table. Many columns are necessary only for episodic measurements and will be empty during most of the time period covered by the table. Some variables may be measured only few times per day and some variables every minute. A minute-scale variable forces including a new row into the table at every minute, but typically, only few cells in this row will be filled. As a result, more than 99% of cells in a common rectangular table may be left empty; the table is hard to survey and occupies a large space in the computer memory and on the disk.

Diary

An alternative is the traditional diary format of data records. A diary is a plain text, which consists of independent diurnal records. A diurnal record includes the date, the variable name, and full diurnal data series for one variable. The values of a variable inside one day are written with uniform time step without need to waste space for time stamps. Different records can contain different numbers of measurements. If a variable is measured every minute then a diurnal record consists of 1440 values of this variable. The next record may contain measurements of hourly averages and consist of only 24 values. If a variable is not measured during a day then the record is skipped. This excludes the accumulation of empty cells, which is a trouble of large tables in case of non-uniform measurements.

A large dataset is stored in DD as a unique master diary and saved as a visually readable comma-delimited plain ASCII text. The diary is accompanied with relatively small attached files, which include the descriptions necessary for the interpretation of the data. The automated data processing presumes that all details of the table and diary format are well defined. Thus the diary format is rigorous in DD and the user must strictly follow the formal rules.

Different from rectangular tables or spreadsheets, a DD diary does not waste the disk space for saving a large number of missing value codes. The simple uniform structure and economical association of the data with the time makes the management of time series in DD more convenient when compared with a set of tables accumulated in a database. A DD dataset can be easily processed using a simple data processing program called the DD data manager. The main tasks of the data manager are importing the measurements from tables into a master diary, and exporting the selected data from the diary to the tables or spreadsheets. The table format is flexible in DD, but the user must correctly describe all details of the format in an attached file. The requirement to describe the data does not allow leaving any data

undocumented. The tables can be composed considering the requirements of the specific research.

A freeware data manager DD2007T32 is described in the third part of the present manual. The data manager enables to decode and encode diverse presentations of time stamps and variable values, shift the time scale, average or interpolate the data in time, make simple statistical overviews of the data, and carry out some other procedures of data processing. The interpolation and averaging functions make possible importing the data, which has originally an irregular time structure.

Some complications can become evident only when working with very large datasets. The diary format is efficient when all data is rewritten after inserting new data into the dataset. In the era of the first computers, the full rewriting of the data was strongly limiting the size of the dataset and this was one of the main reasons why the diary format was almost forgotten in computer data processing. Today, the time of copying of a gigabyte file is measured in seconds and does not create problems. The diary format may turn out to be technically inconvenient only when the amount of data exceeds many gigabytes. In this case, the database formats could be preferred.

On the other hand, the regular full rewriting of the diary results in a favorable side effect: keeping of full backup copies of the data is automatically supported.

DD is opportune for data management when the number of variables is large and the measurements are carried out at different stations following different time regimes. The dataset can simultaneously contain long-term measurements as well as results of episodic measurement campaigns.

Another subject of special attention in DD is the technique of describing the time stamps while paying simultaneous attention to the flexibility and the simplicity of form.

The aspirations of DD are:

- comprehensible user interface,
- visual readability of the data and file fragments as a plain text,
- compactness of the data repository,
- convenient management of the multidimensional data,
- convenience for preparing the data for scientific analysis.

The listed targets are largely contradictory and DD pursues a compromise which, to some extent, attends to all these targets.