Technical report 97-70

DAISY: A database for identification of systems*

B. De Moor, P. De Gersem, B. De Schutter, and W. Favoreel

[*] This report can also be downloaded via https://pub.bartdeschutter.org/abs/97_70.html

DAISY: A Database for Identification of Systems [*]

Bart De Moor[†] Peter De Gersem[‡] Bart De Schutter[§] Wouter Favoreel[¶]

ESAT/SISTA, Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium,
tel.: (+32)-(0)16-32.17.09, fax: (+32)-(0)16-32.19.70
{bart.demoor,peter.degersem,bart.deschutter,wouter.favoreel}@esat.kuleuven.ac.be
web: http://www.esat.kuleuven.ac.be/sista

Abstract

We point out the existence of a disturbing deficiency in the field of system identification, namely the fact that many results, published in papers, are *not reproducible*. In many cases, datasets and time series, that are used to illustrate identification methods and algorithms in these publications, are **not** freely available. We propose to remedy this serious deficiency by setting up a publically accessible website, called **DAISY**, to which authors can submit datasets that are used to illustrate certain claims and algorithms in their papers. Several additional benefits are discussed as well.

Keywords: System identification, signal processing, time series analysis, data analysis, modeling, datasets.

1 To measure is to know...

Reproducibility is one of the most basic characteristics of scientific research. Yet, in the fields of *data analysis*, *system identification* and *signal processing*, this very aspect is often neglected or even completely ignored. By this we mean the following: Too often, papers under review or papers published in conference proceedings and journals, contain an illustration of a certain algorithm, applied to a given data set. A typical statement is then that '*such and such method works well on such and such dataset*'. The problem is that this specific dataset is almost always unavailable and inaccessible. The critical reader is confronted with the paradox that the theoretical derivation of the algorithm in the paper seems to be right but that the verification of its behavior, when applied to the real dataset, is merely impossible. Therefore, the value of such an illustration of a method when applied to a real dataset, is scientifically

[*] Work supported by the Flemish Government (BOF (GOA-MIPS), AWI (Bil. Int. Coll.), FWO (projects, grants, res. comm. (ICCoS)), IWT (IWT-VCST (CVT), ITA (ISIS), EUREKA (Sinopsys))), the Belgian Federal Government (IUAP IV-02, IUAP IMechS), the European Commission (HCM (Simonet), TMR (Alapedes), ACTS (Aspect), SCIENCE (ERNSI)), NATO (CRG) and industry (Electrabel). For more details on these projects, see web-coordinates.

[†] Research Associate with the FWO (Fund for Scientific Research-Flanders), Associate Professor K.U.Leuven;

[‡] Research Assistant with the FWO;

[§] Senior Research Assistant with the FWO;

[¶] Research Assistant supported by the IWT.

void and may be only aesthetic. However, in many cases, the author of the paper had to go through a lot of trouble to obtain the given dataset. Everyone who has been active in experimental work, knows how difficult and time-consuming it is to set up an experiment, obtain measurements, decide on filters, sampling frequencies, sensors, data-acquisition and logging, etc.... When all of this is done, there remains the confrontation of reality with the theoretical framework, which is always based on assumptions and hypotheses, that never seem to be satisfied in practice. The central challenge in system identification and signal processing is precisely this confrontation between experimentally obtained measurements and mathematically derived algorithms. Yet, what we see in most papers is an emphasis on the mathematics and the algorithmic derivations, for which (at least for good papers), all necessary details are provided, so that the algorithm can be understood and reproduced without much difficulty. When it comes to the data or time series to which these algorithms are applied, only nice pictures or generic statistics on the performance of the algorithm are provided, which are barely reproducable.

2 Turn an art into science...

Of course, this lack of reproducibility basically originates in practical considerations, as one could not expect that papers would contain the complete dataset, especially when it is huge. A scientifically acceptable solution would be to make datasets publically available on floppys or CD-roms. Of course, while rather expensive, this solution would also have its practical limitations of compatibility of data formats between different measurement and computing environments.

It goes without saying that the World Wide Web can contribute significantly to solve the reproducibility problem hinted at in the introductory section. We propose to construct a website, which we have called **DAISY**, which stands for **Da***tabase for* **I***dentification* of **Sy***stems*. The key idea is that authors, having published a paper on system identification or signal processing, submit the dataset that was used as an illustration, to **DAISY**, hence making it publically available[1].

The best way to get acquainted with **DAISY** is to consult it at its World Wide Web URL:

http://www.esat.kuleuven.ac.be/sista/daisy

The central objects in **DAISY** are datasets, which, once submitted, undergo a (moderate) review procedure (to filter out 'impossible' or low quality datasets) and, when accepted, are publically available on the Web[2]. Datasets are grouped according to *data categories*, which at the time of writing consist of process industry systems (e.g. ethane-ethylene destillation column, glass furnace, ...), electrical systems, mechanical systems (e.g. wing flutter data, CD-player arm data, ...), biomedical systems (e.g. Fetal ECG measurements, ...), biochemical systems, econometric data, environmental systems, 'classical' datasets, thermal datasets.

[1] The issue of reproducibility requires that we agree on how to refer to **DAISY** in papers that will use some of its datasets. We propose the following reference: De Moor B. (ed.). **DAISY: Database for the Identification of Systems**, Dept. of Electrical Engineering, ESAT/SISTA, K.U.Leuven, Belgium, URL: http://www.esat.kuleuven.ac.be/sista/daisy, + date of visit, name of dataset, name of section and code number.

[2] We take it for granted that all submitted datasets have been cleared from any confidentiality agreement between the owner of the system on which the data were obtained and the person and/or organization that submitted the dataset to **DAISY**.

There is an automatic submission procedure in which some characteristic parameters of the dataset need to be described (sampling frequency, number of data, number of inputs and outputs and their units, references, etc. . .)(see the website for details). Also available are an extended bibliography of more than 100 books on system identification and signal processing, a survey with World Wide Web hyperlinks to existing software packages and existing databases of datasets on the Web.

3 Making it work: *L'appétit vient en mangeant !*

While providing a basic solution to the problem of reproducibility in system identification, we achieve other benefits as well: Some of the datasets in **DAISY** will evolve in due time into real *benchmarks* [3], that will facilitate a comparison of the performance of algorithms. More generally, **DAISY** can become instrumental in establishing *comparisons* of concepts, methods and algorithms. One and the same dataset could be used to assess the quality of several variations of the same method, or, more generally, to compare the performance of different methods derived in different frameworks (e.g. 'classical' system identification, prediction error methods, subspace methods, maximum likelihood methods, structured total least squares, time versus frequency domain approaches, linear versus nonlinear, neural, fuzzy, etc...) or different software environments (like Matlab, Xmath, etc. . .). **DAISY** will also stimulate *collaboration* and *interaction* between researchers, organizations and companies active in system identification. In particular, such a collaboration might enhance the *cost-effectiveness* of experiments, since measurement set-ups will not have to be repeated. **DAISY** will also be instrumental in *providing inspiration* to people in industry, when they *see* how certain datasets are reminiscent to the application they have in mind. And why not use **DAISY** as a *didactical tool*, by inviting students to apply to *real datasets*, in their homeworks, the methods taught in system identification courses. Last but not least, a dataset submitted to **DAISY**, will, on the average, be longer available in time than it would be with the author, who might have decided to move to another address or started a career in another domain than system identification. While experimental set-ups cease to exist at a certain moment in time, the datasets that were obtained from it, will remain available under **DAISY**.

4 Conclusions

We have been describing how an important deficiency in the field of system identification, namely the reproducibility of datasets and time series, can be cured via **DAISY**, a Database for Identification of Systems. We would like to invite all researchers active in system identification, data and time series analysis to submit datasets and provide us with feedback, suggestions for improvement etc. . . . **DAISY** can be consulted at

http://www.esat.kuleuven.ac.be/sista/daisy

[3] All accesses to datasets in **DAISY** are logged. A dataset will evolve into a benchmark when it becomes a leader in the hitting statistics. These can be consulted in **DAISY** by just hitting a push button.