Proceedings of the 17th International Symposium on Mathematical
Theory of Networks and Systems, Kyoto, Japan, July 24-28, 2006

Comparison of identification algorithms on the database DAISY

Ivan Markovsky, Jan C. Willems, and Bart De Moor

K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium,
{ivan.markovsky,sabine.vanhuffel}@esat.kuleuven.ac.be

Abstract— **Five subspace and two optimization based identification methods are applied on seventeen datasets from DAISY. The validation criterion measures how accurate the model can fit a part of the data that is not used for identification. The average result over all data sets shows that the global total least squares method achieves the best fit for all partitionings of the data into identification and validation parts. The prediction error method, which is also optimization based and minimizes a closely related cost function, achieves the second best fit on the identification part of the data but not always on the validation part of the data. The difference in performance between the global total least squares and prediction error methods is likely to be due to the imposed stability constraint in the prediction error method. Among the subspace methods the best fit on the identification part of the data is achieved by the MOESP method. Fastest and probably most efficient is a method based on the shift-and-cut operator.**

Keywords: system identification, global total least squares, prediction error, subspace methods, DAISY.

I. Introduction

One of the main goals of the classical system identification theory is to find conditions for consistency and asymptotic efficiency of identification methods. Such conditions give the following certificate to the methods:

> asymptotically as more data is observed the methods work well (consistency) and are optimal (efficiency) *if the data generating mechanism is of a certain specified type*.

It seems, however, that there is little judgment as to what extent the commonly used stochastic assumptions are satisfied in applications. Three reasons for a theory–practice gap in system identification are:

1) the existing theory is asymptotic in nature and little is known about the practically relevant finite sample size case,
2) the assumption that there is a (deterministic) “true” model in a specified model class is almost never satisfied,
3) the difference between the measured output and a simulated output of the identified model is not a stationary stochastic process.

It is common practice to take into account the cumulative effect of model errors, disturbances, and measurement errors by adding an “error” signal to the model output. In the mainstream literature, the error signal is modeled as a stationary stochastic process. While the disturbances and the measurement errors are sometimes well modeled by a stationary stochastic process, the model errors are certainly not random and might not be approximated well by a stationary process. Moreover, the model–data mismatch in practice is often due to the infinite dimensional, nonlinear, time-varying nature of the object or phenomenon that produces the data and the finite dimensional, linear, time-invariant model class that is used, and not due to measurement errors and disturbances. This suggests that the approximation properties of identification algorithms are often more important than the stochastic ones.

In practice the assumptions for the theoretical certificate stated above are not fulfilled and the model is searched by an ad hock trail and error. Models obtained by trying different methods, tuning some parameters, and pre-processing the data are compared according to various validation criteria and the one that is believed to be the most “suitable” for the purpose at hand is selected. This common identification practice requires active participation of a specially trained human and is more of an art than a science. Certainly the selected model is no longer obtained by the identification method but by the human. It is strange that the mainstream identification theory invariably concentrates on the theoretical certificate and leaves the actual identification process to be carried out manually guided by intuition and ad hock rules.

For the present comparison we choose a validation criterion that reflects the predictive power of the model: how accurate the model can fit a part of the data that is not used for identification. Values for the identification methods’ parameters that correspond to this validation criterion are chosen and fixed for all data sets. The data sets are benchmark problems from the data base for system identification DAISY [DM05] and come from a number of applications: process industry, electrical, mechanical, and environmental. The data is not pre-processed because different methods might benefit from different pre-processing steps, which makes the comparison on the same data impossible. Our purpose is to apply the methods choosing only the model class (which reflects an a priori bound on the model complexity) and the identification/validation criterion (which reflects a desired notion of approximation).

II. Compared identification methods

The compared identification methods are listed in Table I. The first five—`subid`, `uy2ssbal`, `w2x2ss`, `cva`, and `moesp`—are subspace-type methods. They are based on standard numerical linear algebra operations like eigenvalue and singular value decompositions and do not involve nonconvex numerical optimization. The last two—`pem` and `gtls`—optimize certain nonlinear least squares cost functions by local optimization methods. In general, subspace methods are cheaper to compute than the optimization based methods but provide suboptimal results in terms of the particular criteria being used by the optimization based methods.

TABLE I

Compared methods.

Name	Description	Reference
`subid`	robust combined subspace algorithm	[VD96, Figure 4.8]
`uy2ssbal`	deterministic balanced subspace identification	[MWRM05, Algorithm 7]
`w2x2ss`	deterministic subspace identification using the shift-and-cut operator	[MWD05, Algorithm 2]
`cva`	subspace method `n4sid` with option `'N4Weight'` set to `'CVA'`	[Lju04]
`moesp`	subspace method `n4sid` with option `'N4Weight'` set to `'MOESP'`	[Lju04]
`pem`	output error identification in the prediction error setting	[Lju04]
`gtls`	output error identification using structured total least squares	[MWV$^+$05]

The methods `uy2ssbal` and `w2x2ss` are derived for solving exact identification problems, i.e., the data is assumed to be produced by a model in a model class considered. Here the methods are applied on data that does not necessarily satisfy this condition. The software can handle the non-exact case because the exact operations rank computation and solution of a compatible system of equations are automatically replaced by corresponding approximate operations numerical rank computation (up to a given tolerance) and solution of an overdetermined linear system of equations in a least squares sense. With this adaptation on the level of the numerical implementation, the exact identification algorithms `uy2ssbal` and `w2x2ss` become heuristic approximate identification algorithms.

The optimization based methods `pem` (Prediction Error Method) and `gtls` (Global Total Least Squares [RH95]) are similar in structure: they both minimize a nonlinear least squares cost function and the cost function evaluation involves solving a Kalman filtering problem (`pem`) or a deterministic smoothing problem (`gtls`). The motivation for the two methods, however, is rather different. The prediction error method is derived for ARMAX system identification, where statistical assumptions like stationarity, whiteness, and normal distribution play an important role. The global total least squares method is motived for deterministic approximation of the observed data. Similar dichotomy exists for the subspace methods as well: the `subid`, `moesp`, and `cva` methods are motivated in the ARMAX setting, while `uy2ssbal` and `w2x2ss` are derived for exact identification.

The function `pem` is implemented in the System Identification Toolbox of MATLAB. We call it with the options

- `'dist','none'`, which chooses output error model structure,
- `'nk',0`, which requires a feedthrough term to be estimated, and
- `'LimitError',0` which disables the default robustification of the cost function.

The output error model structure is chosen because it is compatible with the selected “simulation fit” validation criterion, see Section IV.

The function `gtls` is called with an option that specifies the inputs as exact. This again corresponds to an output error identification problem. In the multi-output case, however, the cost function minimized by `gtls` is the trace of the output error sample covariance matrix while the cost function minimized by `pem` is the determinant of the same matrix. Therefore in the multi-output case the two cost functions are not necessarily equivalent. In addiiton, the `gtls` function does not constrain the identified model to be stable while the `pem` does so. The initial approximation for `gtls` is the model computed by the function `n4sid` from the System Identification Toolbox with the default value of the option `'N4Weight'`.

The function `uy2ssbal` computes a finite time balanced model. The finite time balancing parameter is selected to be $5l$, where l is the lag of the identified system, i.e., a degree of a difference equation representation, or equivalently the observability index.

III. Datasets of DAISY

The database for system identification DAISY [DM05] is used for verification and comparison of identification algorithms. The considered data sets are listed in Table II.

Next we give references and some details about the meaning and origin of the data:

1) *Lake Erie [GLM80]:* data of a simulation related to the identification of the western basin of Lake Erie. The inputs are the water temperature, water conductivity, water alkalinity, NO_3, and total hardness. The outputs are the dissolved oxigen and algae.

TABLE II

EXAMPLES FROM DAISY. T—NUMBER OF DATA POINTS, m—NUMBER OF INPUTS, p—NUMBER OF OUTPUTS, l—LAG OF THE IDENTIFIED MODEL.

#	Data set name	T	m	p	l
1	Data of a simulation of the western basin of Lake Erie	57	5	2	1
2	Data of ethane-ethylene distillation column	90	5	3	1
3	Heating system	801	1	1	2
4	Data from an industrial dryer (Cambridge Control Ltd)	867	3	3	1
5	Data of a laboratory setup acting like a hair dryer	1000	1	1	5
6	Data of the ball-and-beam setup in SISTA	1000	1	1	2
7	Wing flutter data	1024	1	1	5
8	Data from a flexible robot arm	1024	1	1	4
9	Data of a glass furnace (Philips)	1247	3	6	1
10	Heat flow density through a two layer wall	1680	2	1	2
11	Simulation data of a pH neutralization process	2001	2	1	6
12	Data of a CD-player arm	2048	2	2	1
13	Data from a test setup of an industrial winding process	2500	5	2	2
14	Liquid-saturated steam heat exchanger	4000	1	1	2
15	Data from an industrial evaporator	6305	3	3	1
16	Continuous stirred tank reactor	7500	1	2	1
17	Model of a steam generator at Abbott Power Plant	9600	4	4	1

2) *Distillation column [GLM82]:* simulated data of an ethane-ethylene distillation column. The inputs are the ratio between the reboiler duty and the feed flow, ratio between the reflux rate and the feed flow, ratio between the distillate and the feed flow, input ethane composition, and top pressure. The outputs are top ethane composition, bottom ethylene composition, and top-bottom differential pressure.
3) *Heating system:* the experiment is a simple SISO heating system. The input drives a 300 Watt Halogen lamp, suspended several inches above a thin steel plate. The output is a thermocouple measurement taken from the back of the plate.
4) *Industrial dryer:* data from an industrial dryer (by Cambridge Control Ltd). The inputs are fuel flow rate, hot gas exhaust fan speed, and rate of flow of raw material. The outputs are dry bulb temperature, wet bulb temperature, and moisture content of raw material.
5) *Hair dryer [Lju99], [Lju04]:* laboratory setup acting like a hair dryer. Air is fanned through a tube and heated at the inlet. The air temperature is measured by a thermocouple at the output. The input is the voltage over the heating device (a mesh of resistor wires).
6) *Ball-beam [Ove95, pages 200–206]:* data of a the ball and beam practicum at ESAT-SISTA. The input is the angle of the beam. The output is the position of the ball.
7) *Flutter [FBPT98]:* wing flutter data. Due to industrial secrecy agreements, details are not revealed. The input is highly colored.
8) *Robot arm:* data from a flexible robot arm. The arm is installed on an electrical motor. The transfer function from the measured reaction torque of the structure on the ground to the acceleration of the flexible arm is modeled. The applied input is a periodic sine sweep. The input is reaction torque of the structure. The output is acceleration of the flexible arm.
9) *Glass furnace [VD94]:* The inputs are the heating input and cooling input. The outputs are produced by 6 temperature sensors in a cross section of the furnace.
10) *Two layer wall:* heat flow density through a two layer wall (brick and insulation layer). The inputs are internal wall temperature and external wall temperature. The output is heat flow density through the wall.
11) *pH neutralization process:* simulation data of a pH neutralization process in a constant volume stirring tank. The inputs are the acid solution flow in liters and base solution flow in liters. The output is the pH of the solution in the tank. This process is a highly non-linear system.
12) *Data of a CD-player arm [HS93]:* data from the mechanical construction of a CD player arm. The inputs are the forces of the mechanical actuators while the outputs are related to the tracking accuracy of the arm. The data is measured in closed loop, and then through a two-step procedure converted to open loop equivalent data. The inputs are highly colored.
13) *Winding:* the process is a test setup of an industrial winding process. The main part of the plant is composed of a plastic web that is unwinded from first reel (unwinding reel), goes over the traction reel and is finally rewinded on the rewinding reel. Reel 1 and 3 are coupled with a DC-motor that is controlled with input set point currents I1 and I3. The angular speed of each reel (S1, S2 and S3) and the tensions in the web between reel 1 and 2 (T1) and between reel 2 and 3 (T3) are measured by dynamo tachometers and tension meters. The inputs are the angular speed of reel 1 (S1), angular speed of reel 2 (S2), angular speed of reel 3 (S3), set point current at motor 1 (I1),

and set point current at motor 2 (I3). The outputs are tension in the web between reel 1 and 2 (T1) and tension in the web between reel 2 and 3 (T3).

14) *Exchanger [BP97]:* the process is a liquid-satured steam heat exchanger, where water is heated by pressurized saturated steam through a copper tube. The output variable is the outlet liquid temperature. The input variables are the liquid flow rate, the steam temperature, and the inlet liquid temperature. In this experiment the steam temperature and the inlet liquid temperature are kept constant to their nominal values. The heat exchanger process is a significant benchmark for nonlinear control design purposes, since it is characterized by a non minimum phase behavior. The input is the liquid flow rate. The output is the outlet liquid temperature.
15) *Industrial evaporator [ZVDL94]:* a four-stage evaporator to reduce the water content of a product, for example milk. The inputs are feed flow to the first evaporator stage, vapor flow to the first evaporator stage, and cooling water flow. The outputs are the dry matter content, flow of the outcome product, and temperature of the outcome product.
16) *Tank reactor [LI99]:* The process is a model of a continuous stirring tank reactor, where the reaction is exothermic and the concentration is controlled by regulating the coolant flow. The input is coolant flow l/min. The outputs are concentration mol/l, and temperature Kelvin degrees.
17) *Steam generator [PB96]:* the data comes from a model of a Steam Generator at Abbott Power Plant in Champaign, IL. The inputs are fuel scaled 0–1, air scaled 0–1, reference level inches, and disturbance defined by the load level. The outputs are drum pressure PSI, excess oxygen in exhaust gases %, level of water in the drum, and steam flow kg/s. To make possible the open loop identification the water level was stabilized by applying to the water flow input a feed-forward action proportional to the steam flow and a PI action. The reference of this controller is the third input.

Note 1 (Excluded data sets). Five data sets from DAISY are not included in the comparison. Three of them come from autonomous systems, one (“Step response of a fractional distillation column”) comes from a step response experiment, and one (“Data of a 120 MW power plant”) has inputs that are not persistently exiting of sufficient order (ramp signals). These data sets are excluded from the comparison because they can not be treated by all methods.

IV. Validation criterion and results

The data $w=(u,y)$ in all examples is split into identification and validation parts. For a chosen $x \in [0,100]$, the first or last x% of the data, denoted w_{idt}, are used for identification, and the remaining $(100-x)$% of the data, denoted w_{val}, are used for validation. A model $\hat{\mathscr{B}}$ is identified from w_{idt} by an identification method and is validated on w_{val} by the validation criterion defined next. The model class is linear time-invariant systems with a bound l on the lag (degree of a difference equation representation or equivalently observability index). Bounding the lag by l corresponds to bounding the order by lp, where p is the number of outputs.

The validation criterion corresponds to the “simulation fit” computed by the function `compare` of the System Identification Toolbox, see Note 2. Given a time series $w=(u,y)$ and a model $\mathscr{B}$, define the approximation $\hat{y}$ of y in $\mathscr{B}$ as follows:

$$\hat{y}\big((u,y),\mathscr{B}\big) := \min_{\hat{y}} \|y-\hat{y}\| \quad \text{subject to} \quad \operatorname{col}(u,\hat{y}) \in \mathscr{B}.$$

(The optimization is carried over the initial conditions that generate $\hat{y}$ from the given input u.) Let $\bar{y}$ be the mean of y, i.e., $\bar{y} := \sum_{t=1}^{T} y(t)/T$. With this notation, the fit of w by $\mathscr{B}$ is defined as

$$F(w,\mathscr{B}) := 100 \max\big(0, 1-\|y-\hat{y}(w,\mathscr{B})\|/\|y-\bar{y}\|\big).$$

We compare the fitting criterion $F(w_{\text{val}},\hat{\mathscr{B}})$ for the models produced by the compared identification methods.

The average results for all data sets are given in Table III. “70i/30v” is a short notation for “first 70% of the data is used for identification and the remaining 30% for validation”. Similarly “30v/70i” stands for “first 30% of the data is used for validation and the remaining 70% for identification”. Bar plots for all methods, all data sets, and all experiments (splitting of the data into identification and validation parts) are presented on Figures 2 to 7.

Note 2 (Unstable models). Some identification methods do not impose a stability constraint on the model, so that depending on the data, they can computed unstable model. Contrary to a common misconception unstable model is not necessarily a “bad” model. It could still achieve a good fit of the validation data, however, one should pay attention in computing the fitting trajectory. Simulated forward in time, the unstable part of the model is sensitive to the initial conditions, which makes the computation numerically ill conditioned. Simulation of the unstable part backward in time (from a final condition) avoids this difficulty.

TABLE III

AVERAGE FIT IN % ON ALL DATASETS. (THE BEST FITS AND SMALLEST EXECUTION TIMES OBTAINED BY SUBSPACE AND OPTIMIZATION METHODS ARE MARKED WITH **bold face**.)

Experiment		subid	uy2ssbal	w2x2ss	moesp	cva	pem	gtls
70i/30v	identification	51.18	49.27	46.39	**55.52**	49.79	57.43	**68.46**
	validation	32.14	31.57	32.34	**38.97**	33.38	37.77	**48.40**
30v/70i	identification	46.34	47.46	48.83	**53.86**	50.78	59.13	**68.87**
	validation	36.96	37.69	38.15	**40.43**	37.10	45.17	**53.72**
80i/20v	identification	49.14	46.82	45.56	**55.13**	50.88	56.84	**68.36**
	validation	30.01	28.20	29.75	**33.01**	31.75	36.17	**44.14**
20v/80i	identification	49.47	48.20	48.07	**54.48**	51.90	58.93	**68.48**
	validation	**46.09**	37.30	40.81	39.79	39.81	45.28	**56.88**
90i/10v	identification	50.92	47.61	48.59	**54.79**	51.25	58.39	**68.95**
	validation	**40.47**	32.89	31.46	37.06	35.07	39.48	**48.55**
10v/90i	identification	48.16	48.46	47.34	**53.93**	50.71	58.78	**69.06**
	validation	**45.58**	43.71	45.13	44.12	39.71	43.62	**56.28**
Typical execution times		0.11	0.95	**0.05**	4.45	5.03	**14.79**	25.14

V. CONCLUSIONS

The conclusions are based on the average results over all data sets, see Table III.

For all partitionings of the data into identification and validation parts, the best fit on both identification and validation parts of the data is obtain by `gtls`. The second best fit on the identification part of the data is obtained by `pem`. The `gtls` and `pem` methods are the optimization based; they explicitly minimize the validation criterion so it is not surprising that they outperform the subspace-based methods on the identification part of the data. A good fit on the validation part of the data, however, is not guaranteed by a good fit on the identification part of the data. Indeed in four out of the six partitionings of the data, a subspace method achieves the second best fit on the validation part of the data. With respect to the execution time `pem` is about twice faster than `gtls`.

Perhaps the most important reason for the superior performance of `gtls` over `pem` is the fact that `gtls` does not impose a stability constraint on the identified model. (Minimization over a larger set guarantees smaller value of the cost function.) Other possible reasons for the difference in performance of `gtls` and `pem` are that the cost functions in the multi-output case differ (trace vs determinant), the optimization algorithms differ (Levenberg–Marquardt vs Gauss–Newton), the initial approximations and the convergence criteria (most probably) also differ.

Among the subspace methods the best fit on the identification part of the data is achieved by `moesp`. In three partitionings of the data, `moesp` achieves also the best fit on the validation part of the data. In the other three cases, the best fit is achieved by `subid`. Fastest and perhaps most efficient of all methods is `w2x2ss`, the method based on the shift-and-cut operator. The subspace methods `uy2ssbal` and `w2x2ss` that are designed for exact system identification and are applied as heuristics for approximate identification, do not give as accurate results as the stochastic subspace identification methods.

VI. ACKNOWLEDGMENTS

Research supported by: **Research Council KUL:** GOA-Mefisto 666, GOA-Ambiorics, IDO/99/003 and IDO/02/009 (Predictive computer models for medical classification problems using patient data and expert knowledge), several PhD/postdoc & fellow grants; **Flemish Government: FWO:** PhD/postdoc grants, projects, G.0078.01 (structured matrices), G.0269.02 (magnetic resonance spectroscopic imaging), G.0270.02 (nonlinear Lp approximation), G.0240.99 (multilinear algebra), G.0407.02 (support vector machines), G.0197.02 (power islands), G.0141.03 (Identification and cryptography), G.0491.03 (control for intensive care glycemia), G.0120.03 (QIT), G.0452.04 (QC), G.0499.04 (robust SVM), research communities (ICCoS, ANMMM, MLDM); AWI: Bil. Int. Collaboration Hungary/ Poland; **IWT:** PhD Grants; GBOU (McKnow) **Belgian Federal Government:** DWTC (IUAP IV-02 (1996-2001) and Belgian Federal Science Policy Office IUAP V-22 (2002-2006) (Dynamical Systems and Control: Computation, Identification & Modelling)); PODO-II (CP/01/40: TMS and Sustainibility); **EU:** PDT-COIL, BIOPATTERN, eTUMOUR, FP5-Quprodis; ERNSI; Eureka 2063-IMPACT; Eureka 2419-FliTE; Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard.

REFERENCES

[BP97] S. Bittanti and L. Piroddi. Nonlinear identification and control of a heat exchanger: a neural network approach. *Journal of the Franklin Institute*, 334(1):135–153, 1997.

[DM05] B. De Moor. DaISy: Database for the identification of systems. Dept. EE, K.U.Leuven, `www.esat.kuleuven.be/sista/daisy/`, 2005.

[FBPT98] E. Feron, M. Brenner, J. Paduano, and A. Turevskiy. Time-frequency analysis for transfer function estimation and application to flutter clearance. *AIAA J. on Guidance, Control & Dynamics*, 21(3):375–382, 1998.

[GLM80] R. Guidorzi, M. Losito, and T. Muratori. On the last eigenvalue test in the structural identification of linear multivariable systems. In *Proceedings of the V European meeting on cybernetics and systems research*, Vienna, 1980.

[GLM82] R. Guidorzi, M. Losito, and T. Muratori. The range error test in the structural identification of linear multivariable systems. *IEEE Trans. on Aut. Control*, 27:1044–1054, 1982.

[HS93] P. Van Den Hof and P. Schrama. Function estimation from closed loop data. *Automatica*, 29(6):1523–1527, 1993.

[LI99] G. Lightbody and G. Irwin. Nonlinear control structures based on embedded neural system models. *IEEE Tran. on Neural Networks*, 8(3):553–567, 1999.

[Lju99] L. Ljung. *System Identification: Theory for the user*. Prentice Hall, 1999.

[Lju04] L. Ljung. *System Identification Toolbox: User's guide*. The MathWorks, 2004.

[MWD05] I. Markovsky, J. C. Willems, and B. De Moor. State representations from finite time series. In *Proc. of the Conf. on Decision and Control*, pages ??–??, Seville, Spain, 2005.

[MWRM05] I. Markovsky, J. C. Willems, P. Rapisarda, and B. De Moor. Algorithms for deterministic balanced subspace identification. *Automatica*, 41(5):755–766, 2005.

[MWV^{+}05] I. Markovsky, J. C. Willems, S. Van Huffel, B. De Moor, and R. Pintelon. Application of structured total least squares for system identification and model reduction. *IEEE Trans. on Aut. Control*, 50(10):1490–1500, 2005.

[Ove95] P. Van Overschee. *Subspace identification: Theory, Implementation, Application*. PhD thesis, K.U.Leuven, 1995.

[PB96] G. Pellegrinetti and J. Benstman. Nonlinear control oriented boiler modeling: A benchamrk problem for controller design. *IEEE Tran. Control Systems Tech.*, 40(1), 1996.

[RH95] B. Roorda and C. Heij. Global total least squares modeling of multivariate time series. *IEEE Trans. on Aut. Control*, 40(1):50–63, 1995.

[VD94] P. Van Overschee and B. De Moor. *N4SID:* subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30:75–93, 1994.

[VD96] P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer, 1996.

[ZVDL94] Y. Zhu, P. Van Overschee, B. De Moor, and L. Ljung. Comparison of three classes of identification methods. In *Proc. of SYSID*, volume 1, pages 175–180, Copenhagen, Denmark, 1994.

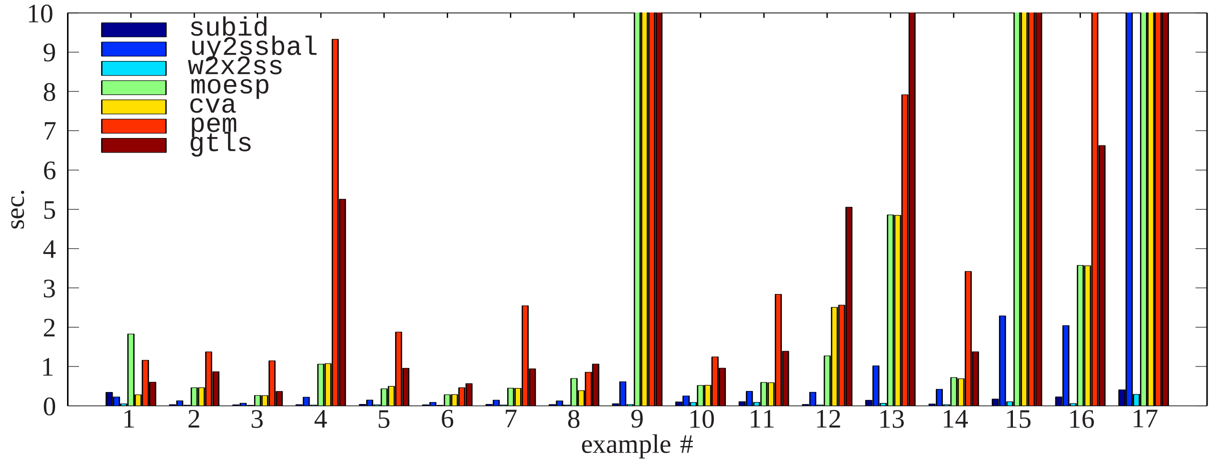

Fig. 1. Typical execution times.

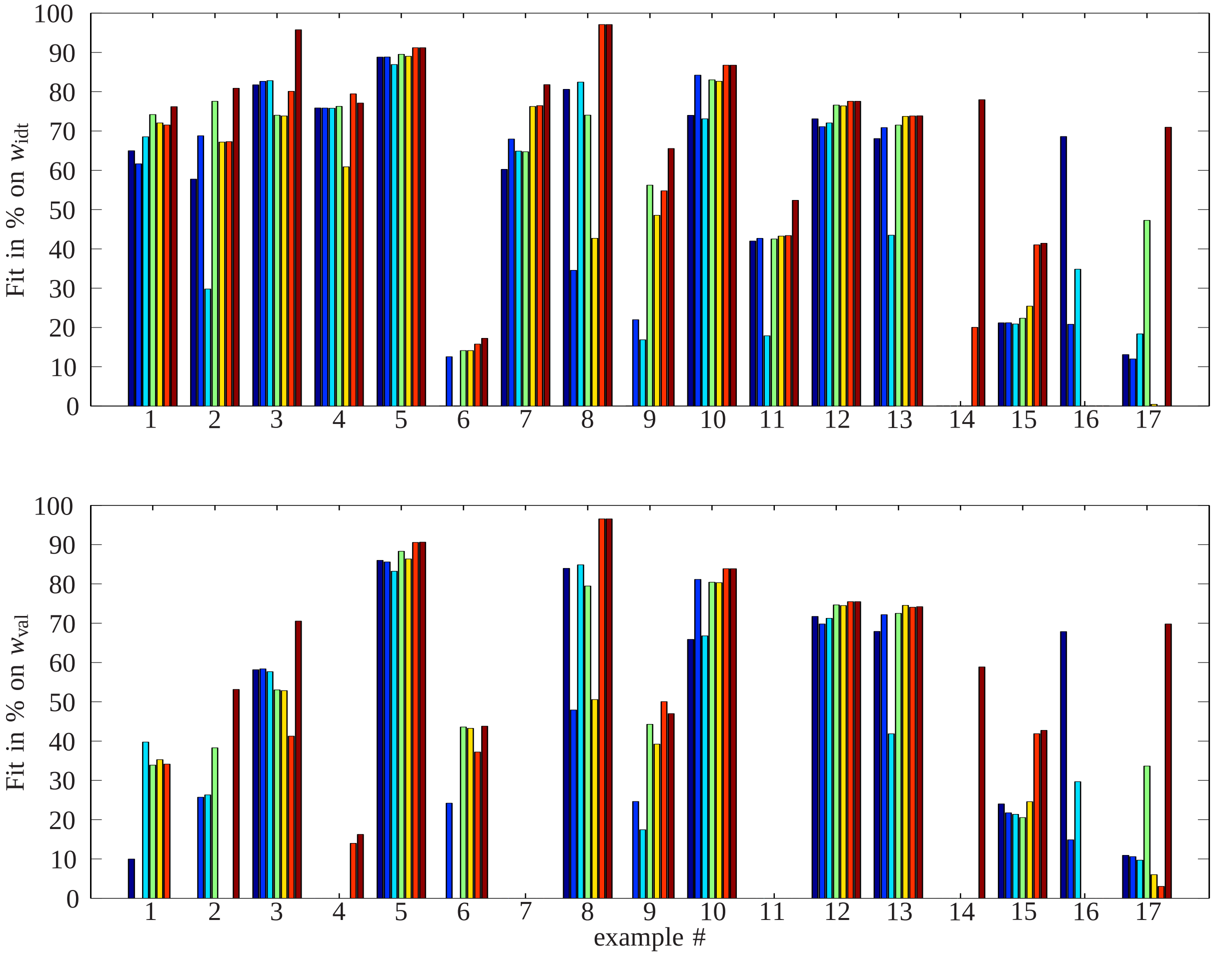

Fig. 2. Splitting of the data into 70% identification 30% validation.

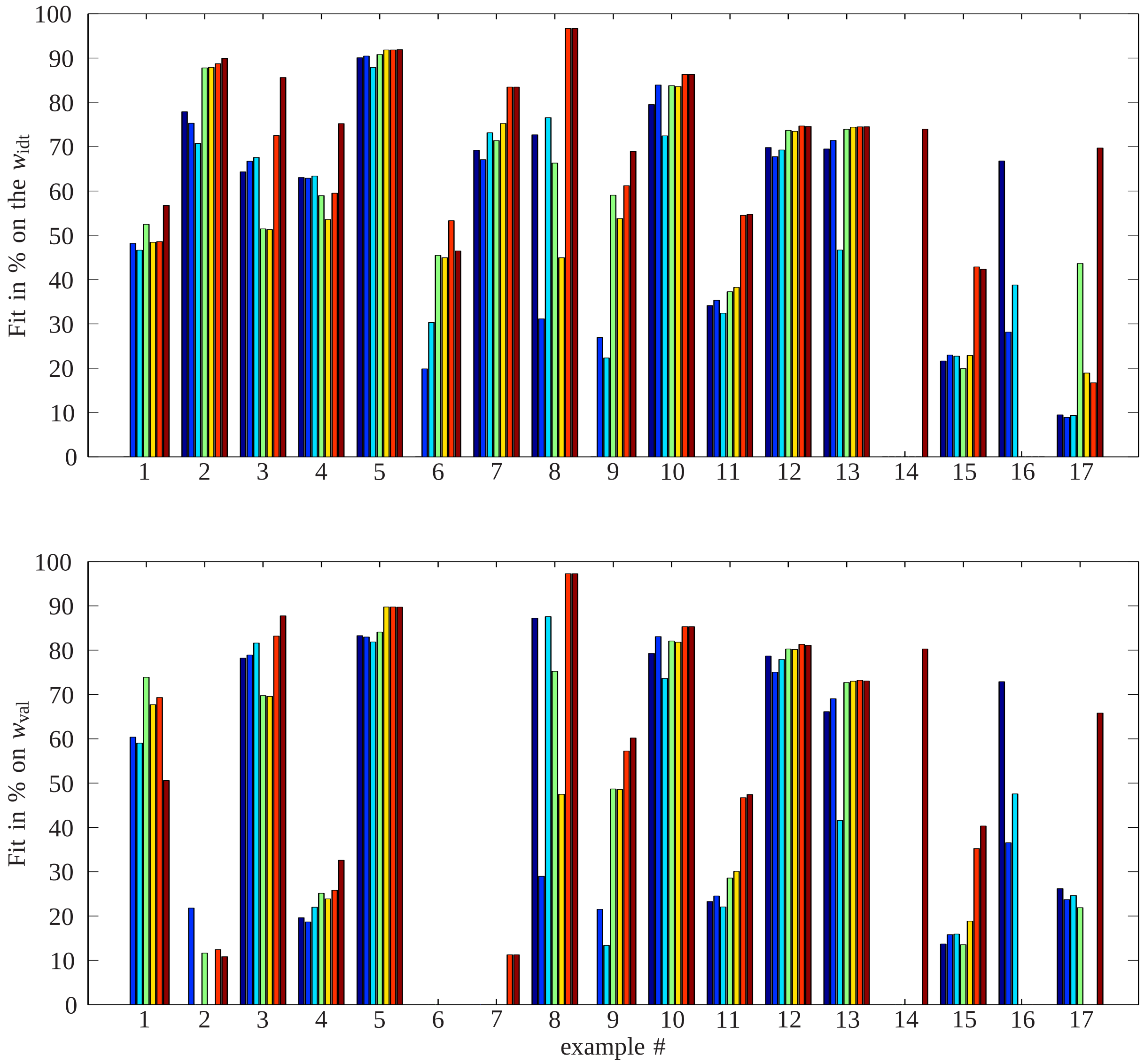

Fig. 3. Splitting of the data into 30% validation 70% identification.

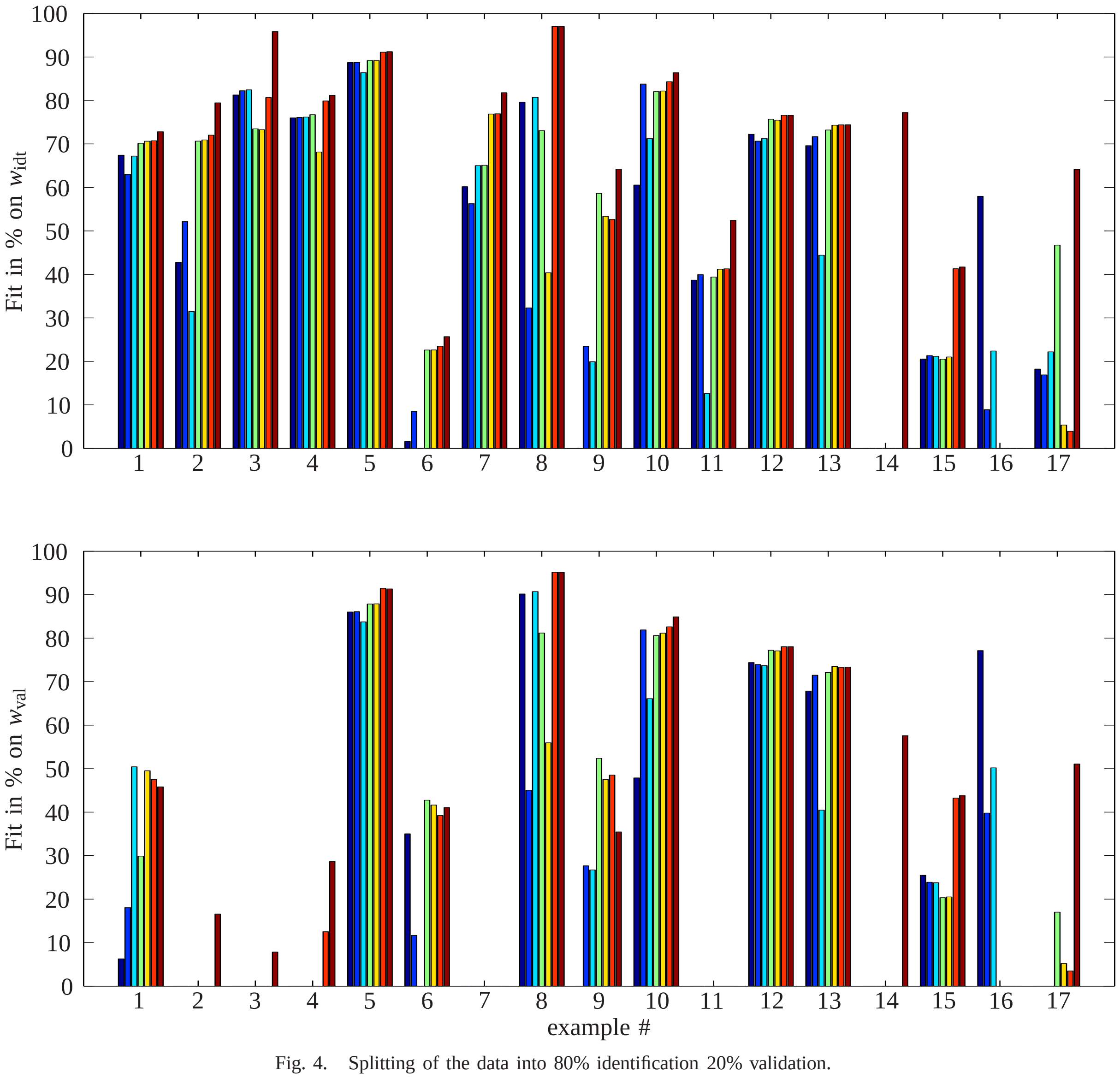

Fig. 4. Splitting of the data into 80% identification 20% validation.

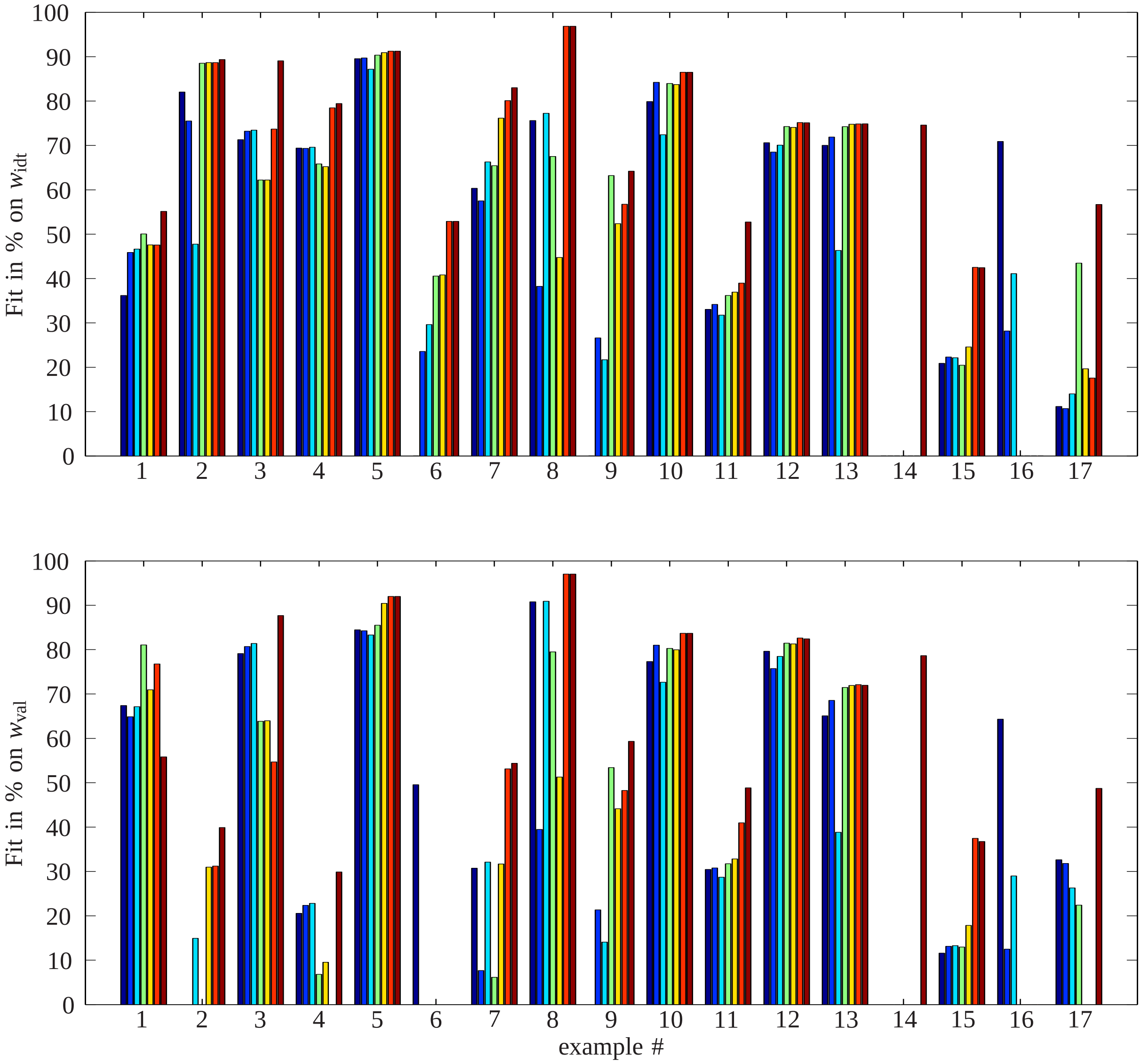

Fig. 5. Splitting of the data into 20% validation 80% identification.

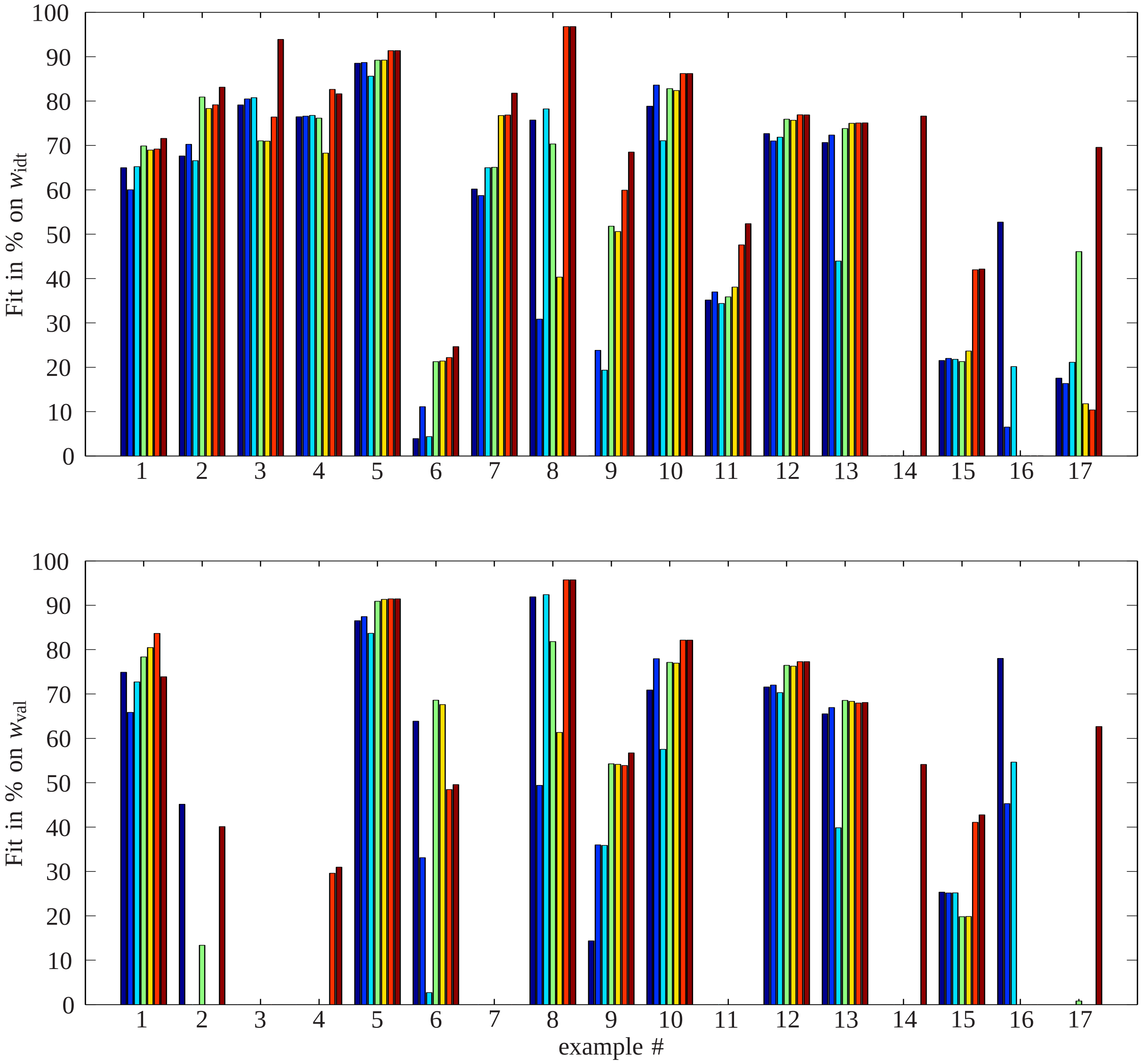

Fig. 6. Splitting of the data into 90% identification 10% validation.

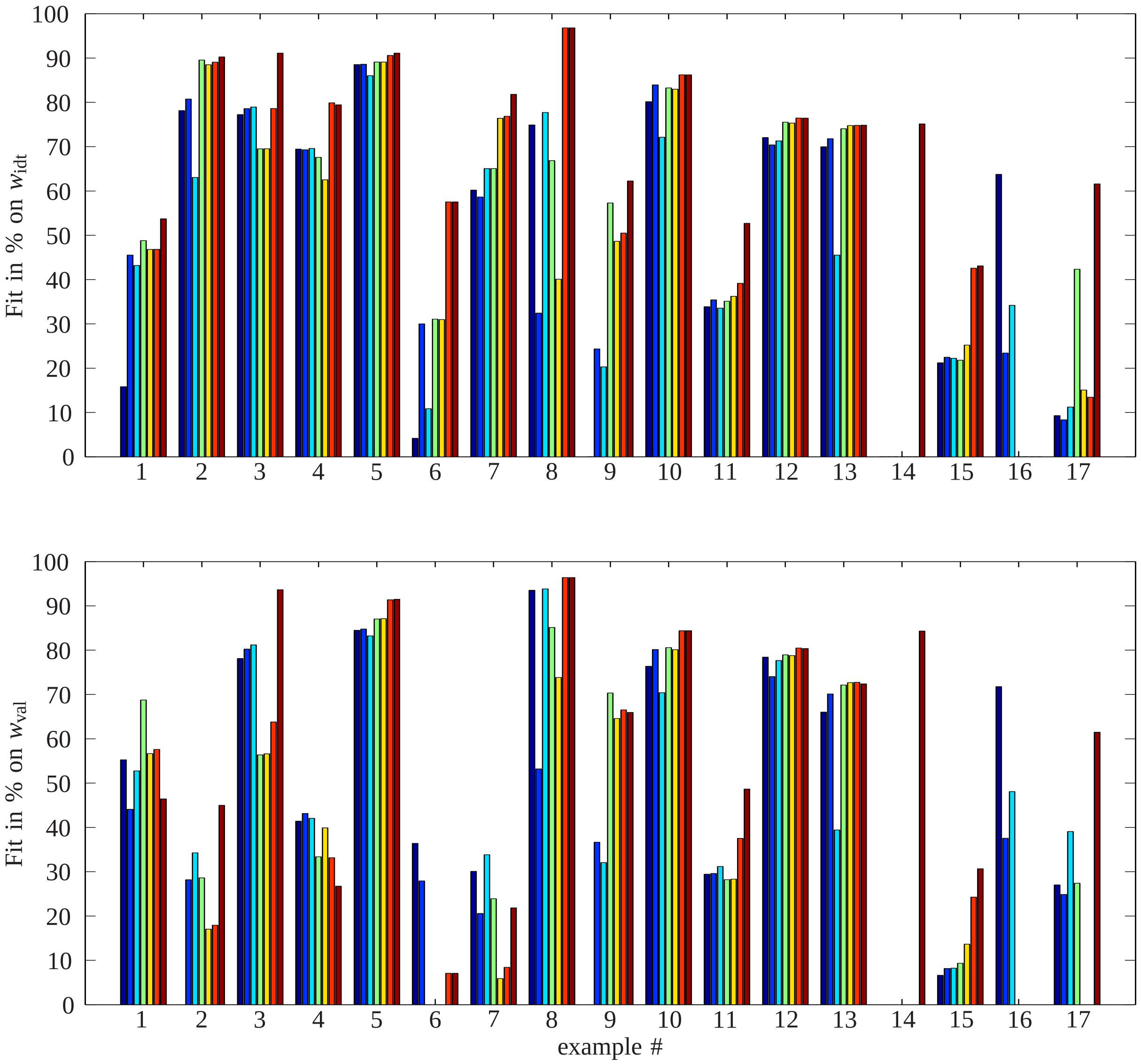

Fig. 7. Splitting of the data into 10% validation 90% identification.