Journal of Statistical Software

September 2007, Volume 22, Issue 8. http://www.jstatsoft.org/

Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R Package qualV)

Stefanie Jachner
Technische Universität Dresden

K. Gerald van den Boogaart
Ernst-Moritz-Arndt-Universität Greifswald

Thomas Petzoldt
Technische Universität Dresden

Abstract

Results of ecological models differ, to some extent, more from measured data than from empirical knowledge. Existing techniques for validation based on quantitative assessments sometimes cause an underestimation of the performance of models due to time shifts, accelerations and delays or systematic differences between measurement and simulation. However, for the application of such models it is often more important to reproduce essential patterns instead of seemingly exact numerical values.

This paper presents techniques to identify patterns and numerical methods to measure the consistency of patterns between observations and model results. An orthogonal set of deviance measures for absolute, relative and ordinal scale was compiled to provide informations about the type of difference. Furthermore, two different approaches accounting for time shifts were presented. The first one transforms the time to take time delays and speed differences into account. The second one describes known qualitative criteria dividing time series into interval units in accordance to their main features. The methods differ in their basic concepts and in the form of the resulting criteria. Both approaches and the deviance measures discussed are implemented in an R package. All methods are demonstrated by means of water quality measurements and simulation data.

The proposed quality criteria allow to recognize systematic differences and time shifts between time series and to conclude about the quantitative and qualitative similarity of patterns.

Keywords: ecological modeling, qualitative validation criteria, time shift, R.

1. Introduction

Dynamic simulation models are important tools in environmental science and environmental management. For both, scientists and decision makers, the validity of the model and its results are of high importance, in particular if private persons and/or economy are affected by decisions based on the model results. A commonly used definition of model validation is the substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy, consistent with the intended application of the model (Schlesinger *et al.* 1979). However, the question what is accurate enough might depend on very different features for different applications.

The general framework of this paper is the problem to compare a (more or less deterministic) computer simulation model of a nonstationary ecological process with a real world realization of the same process. Typically, we have measurements y_t at times τ_t of the real world process and corresponding simulated values $\hat{y}_t$, which are the output of the simulation model. Thus we assume that the process is dominated by its deterministic laws and not by its negligible stochastic variation. On one hand we assume that the deterministic simulation model (among other simulation models) is one of our best understandings of the process, but on the other hand we are well aware of the fact that it is just a model and can not describe the process completely or even completely up to measurement errors. Although the y_t and $\epsilon_t := y_t - \hat{y}_t$ are random, we can not model them as a realization of a stationary or otherwise describable stochastic process, because this would imply a model on the difference of our best model and the reality.

Commonly used validation methods include visualization techniques, where experts, if available, evaluate the quality of the results and quantitative assessment criteria, where the distance between measured and simulated values is analyzed. However, visual and quantitative methods can deliver contrasting results, as soon as processes have different velocities, shifts in time, or when systematic differences between measurement and simulation exist.

1.1. An introductory example

A simple example with three time series (Figure 1) will illustrate the problem (see also Höppner 2002b). In an ecological context, let's assume these curves were, for example, time series of an observed algae development over time (y_t) and the outputs ($\hat{y}_{t,A}$ and $\hat{y}_{t,B}$) of two alternative models A and B.

Asking humans to decide which model is more appropriate to describe the observed pattern, they would probably prefer model A over model B. Using a common deviance measure, e.g. mean absolute error (MAE, described below), yields B closer to the observation than A. The resulting assessment based on numerical differences (B is better than A), does not match with the human decision (A is better than B). The reason is that people do not decide about similarity by numerical values, but instead identify similar patterns. While both time series, y_t and $\hat{y}_{t,A}$ consist of two segments, an increasing and a decreasing part, the time series $\hat{y}_{t,B}$ contains only one single linearly increasing segment. Quantitative validation methods do not consider such patterns. The high numerical difference between time series y_t and $\hat{y}_{t,A}$ is a result of the shift of the maximum of the series in time. When y_t reaches its maximum, $\hat{y}_{t,A}$ is still small, and when y_t is at maximum $\hat{y}_{t,A}$ is small again.

If a shift in time leads quantitative methods to underestimate the quality of one model can-

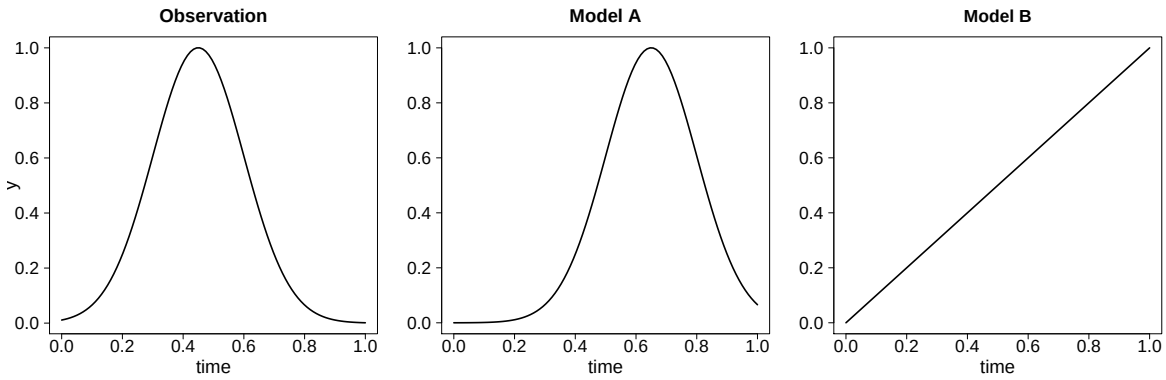

Figure 1: Similarity of time series: Using the mean absolute error (MAE) model B is closer to the observation than model A.

didate, but the other is preferred only by subjective decision, which of the models would you trust for the predictions of the water quality in next autumn?

In full accordance with this, Grimm *et al.* (1996) proposed that the main focus of ecological models should not be to deliver each measured detail, but instead to reproduce the observed patterns (see also Levin 1992; Wiegand *et al.* 2003). In this context, it is necessary to develop qualitative methods for the validation of the simulated patterns in addition to existing assessment techniques.

1.2. Approach

The qualitative comparison methods proposed are based on comparing the information remaining, when several quantitative aspects of the actual series y_t and $\hat{y}_t$ are ignored. In a first step we discuss ignoring location, scale and distance of values, in a second step we discuss ignoring exact time and speed, and in last step we discuss ignoring even inequality relations and time continuity. In this way we build up a system of descriptive index values, called deviance measures and similarity measures, which allow to describe, quantify, and investigate various types of qualitative and semiquantitative similarity.

These index values can be used in various ways: We can select the deviance measure describing best the type of accuracy needed for our application to select the model performing best with respect to this deviance measure. We could calculate and compare different similarity measures to identify the qualities in which model and reality differ. We can also take a random sample of real systems, calculate a deviance measure with respect to several models, and use inferential statistics to compare the different models. However, the aim of this paper is not to give brewing recipes what to do with the index values but to give the systematics and to provide the software.

Section 2 introduces the example data used to demonstrate the proposed methods. A brief overview of common validation methods and a support to identify essential differences between measurement and simulation are given in Section 3. Sections 4 and 5 present adequate deviance measures for different cases and qualitative validation criteria accounting for time shifts. Finally, inferential considerations and a discussion are offered in Sections 6 and 7.

2. Data set and data preparation

2.1. Example data set

The new R (R Development Core Team 2007) package for qualitative validation (**qualV**) provides time series with measured and predicted lake phytoplankton concentrations (`phyto`) in Bautzen Reservoir 1994 (Saxony, Germany).

```
R> library("qualV")
R> data("phyto")
```

The data set `phyto` contains two corresponding data frames `obs` and `sim`. The data frames include a time code `t`, and corresponding observed and simulated phytoplankton biovolume `y` (mg/L), respectively (Figure 2). The observed data are vertical mean values of phytoplankton biovolume, derived from microscopic cell counting (see Benndorf *et al.* 2001, for details about motivation and data). The respective simulated data are generated using a recent version of the dynamic lake model SALMO (Simulation of an Analytical Lake MOdel, Benndorf and Recknagel 1982).

This model belongs to a class of ecological models which predict time series of state variables (e.g. concentrations of nutrients, phytoplankton, zooplankton and oxygen) by means of ordinary differential equations, see the mass balance equation of phytoplankton (Y) as an example (with P photosynthesis, R respiration, S sedimentation, G zooplankton grazing, I import and export):

$$\frac{dY}{dt} = P - R - S - G + I$$

The applied model version consists of 13 state variables and an extensive set of nonlinear algebraic equations. The latter are used to describe functional relationships between state variables and external forcings (meteorology, matter import, physical structure of the water body). A short description of the model and required inputs can be found in Petzoldt and Uhlmann (2006).

2.2. Smoothing

Interpolation is required whenever the time steps of measurements and simulation differ. Furthermore, smoothing may be required in order to reduce the noise of the measured data. The problem is, however, that smoothing depends critically on bandwidth and the selected smoothing kernel. In order to avoid subjectivity we strongly suggest to apply automatic bandwidth selection methods.

In the following, we propose a pure exemplary preprocessing step for (ecological) time series, which does not rely on the R package **qualV**. A Gaussian kernel and an automatic bandwidth selection method based on the plug-in methodology (`dpill`) after Ruppert *et al.* (1995) are used (package **KernSmooth**, see Wand and Ripley 2006) to smooth `obs`:

```
R> bobs <- dpill(obs$t, obs$y)
R> n <- tail(obs$t, n = 1) - obs$t[1] + 1
```

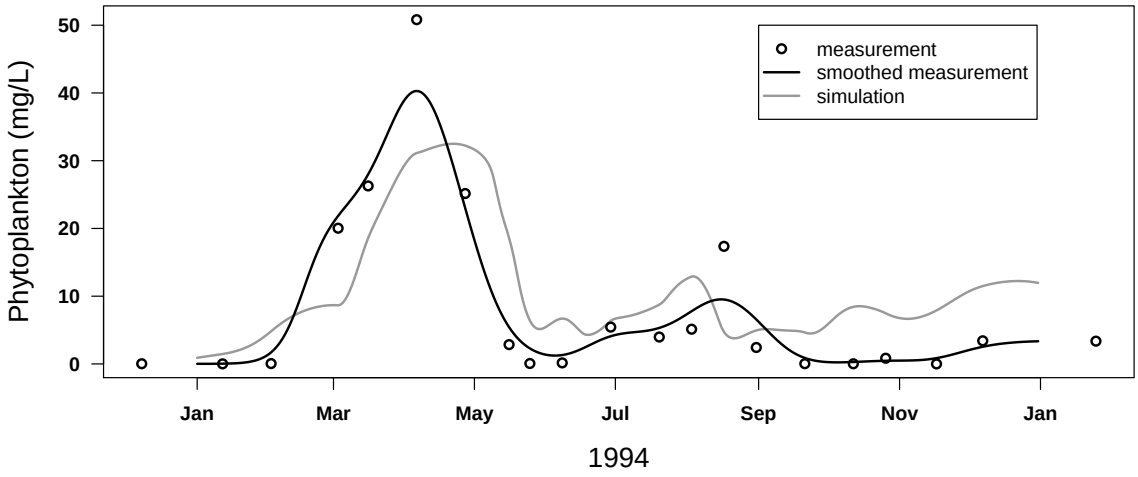

Figure 2: Measured (circles), smoothed measurement (black line), and simulated (gray line) phytoplankton data for Bautzen Reservoir (Data: TU Dresden, Workgroup Limnology)

```
R> obss <- ksmooth(obs$t, obs$y, kernel = "normal", bandwidth = bobs,
+     n.points = n)
R> obss <- as.data.frame(obss)
R> names(obss) <- c("t", "y")
R> obss <- na.omit(obss[match(sim$t, obss$t), ])
```

The result is a new data frame `obss` containing smoothed data `y` at a daily time step `t` aligned to `sim$t` (Figure 2).

In the following, these example data are used to demonstrate the proposed qualitative model criteria implemented in the package **qualV** and to show their performance in practice.

3. Model validation methods

Numerous methods for the validation of models already exist. They are described and discussed by authors of different disciplines, for example Balci and Sargent (1982), Rykiel (1996) or Sargent (2003). It is probably impossible to list all the methods, which have been proposed.

3.1. Classification of methods

Following Mayer and Butler (1993), we can divide the methods into the four groups "subjective assessment", "visual techniques", "deviance measures" and "statistical tests".

Examples for **subjective assessment** are *face validity* and *Turing tests*. Here experts are asked whether a model and its behavior are reasonable and whether it is possible to distinguish between real data and modeled output. These tests show if a model is feasible and how close measured and simulated data match in graphical display (Rykiel 1996).

Visual techniques are associated with subjective assessment. They are used to present values as time series, cumulative sums or two- or three-dimensional diagrams (e.g. predicted-observed plots). Here simulation $\hat{y}$ and measurement y are plotted against each other, and an identity of the lines $y = \hat{y}$ illustrates the best fit.

Deviance measures can be regarded as numerical validation techniques, which provide a measure of difference (or similarity) between measured and simulated values and are typically applicable if the values can be paired in space and/or time.

Specific **statistical tests** can also be used for model validation, if the design and the values satisfy the conditions of these tests. Most of them require independent and evenly distributed values. This assumption does often not hold for real-world data and, in particular, is not adequate in ecological models. In cases where the conditions can be met, the power of the test will be low due to the small sample size.

Since subjective assessment and visual techniques are subjective and since the applicability of tests is limited as discussed in Section 6, this paper will focus on deviance measures.

3.2. Classification of measures

To compare a simulated and a real time series semiquantitatively or qualitatively, it is important to define what type of difference is considered essential. Many values, like concentrations, amounts, energies, speeds or flows can only be positive. In this case relative errors are often more important than absolute differences. Similar consideration can be found in more detail in Pawlowsky-Glahn *et al.* (2003), van den Boogaart and Tolosana-Delgado (2007) or Tolosana-Delgado and Pawlosky-Glahn (2007). Examples for absolute and relative scale are given in Figures 3 and 4. Adequate deviance measures for different situations are proposed in Chapter 4 and summarized in Table 1. The following questions should help to select appropriate measures:

1. *Is the actual level of values of the simulation important or do we consider a scaled or shifted simulation qualitatively equivalent?* Typically, such simple scaling or shifting can be realized by changing parameters of the simulation. However, for a simulation based on many different externally validated parameters, one would not intend to change parameters just to increase the fit for one single dataset. Such differences can be ignored by centering or scaling the datasets before being compared as discussed in Section 4.1.

2. *Are absolute (arithmetic) errors or relative (geometric) errors important?* Is the difference from 1g/L to 1.01 g/L 100 times more or 10 times less important than the difference of 1mg/L to 1.1mg/L? Relative errors can be considered by using log-transformed data instead of the data itself and by downweighting differences of higher values as discussed in 4.2.

3. *Are the actual values of the simulation important or only their relation?* It might be a qualitative criterion that the second maximum is smaller than the first, but the precise value of the two maxima might be considered as a quantitative artifact. If the values are not important, but only their order relations, we might use a rank transform first, as discussed in Section 4.3.

4. *Should deviance be measured absolutely or relatively to the inherent variation of the process?* In the first case we can interpret the outcome in terms of the original units

but not relative to other examples. In the second case we can compare the goodness of fit for very different systems. The corresponding similarity measures are discussed in Section 4.5.

5. *Is the precise timing and speed of the process important?* Qualitatively similar processes can be quantitatively very different, when the speeds of different subprocesses are not modeled precisely enough. Related deviance measures ignoring precise speeds of processes are discussed in Section 5.1.

6. *Are the relations and curvature of the process only locally of interest?* We may think that it would be a qualitative feature to have two clear maxima at specific points in time, but want to disregard which maximum is higher because we think they are not comparable. In this case we propose the use of local features as discussed in Section 5.2.

4. Deviance measures

In this section we will systematically list and introduce deviance measures for all possible combinations of answers to the first 4 questions and answering the penultimate with yes and the last with no (Table 1). Methods for different answers in the last two questions will be discussed in later sections.

Since each measure produces different values for quite similar or very different time series we need a comparative value with a simple interpretation. Thus we will report for each measure the value of the deviance measure for a best fitting constant process, which is a reference value for "totally unrelated" processes, using the phrase "comparable to".

4.1. Classical deviance measures for absolute scale

We assume y_t to be the measurement of a nonstationary typically time continuous real world process and $\hat{y}_t$ simulated values of the same process at the same times $\tau_1, \ldots, \tau_n$. We only consider real valued y_t and not vectors. Concerning the methods of this section multiple outcomes can be studied individually.

In this situation some major deviance measures are defined by:

- *mean absolute error*: $\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|$
  is comparable to the *median absolute deviation*: $\mathrm{MAD}(y_t) := \frac{1}{n}\sum_{t=1}^{n} |y_t - median(y_t)|$

- *mean squared error*: $\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2$
  is comparable to the *variance*: $var(y_t)$

- *root mean squared error*: $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}$
  is comparable to the *standard derivation*: $sd(y_t)$

MAE, MSE and RMSE deliver absolute, but scale dependent measures of model performance, i.e. they can only be used for a relative comparison between different models. These deviance measures are zero if and only if the values are identical (Figure 3). A useful interpretation of the scale can be achieved when the deviance of a model is compared to the deviance of

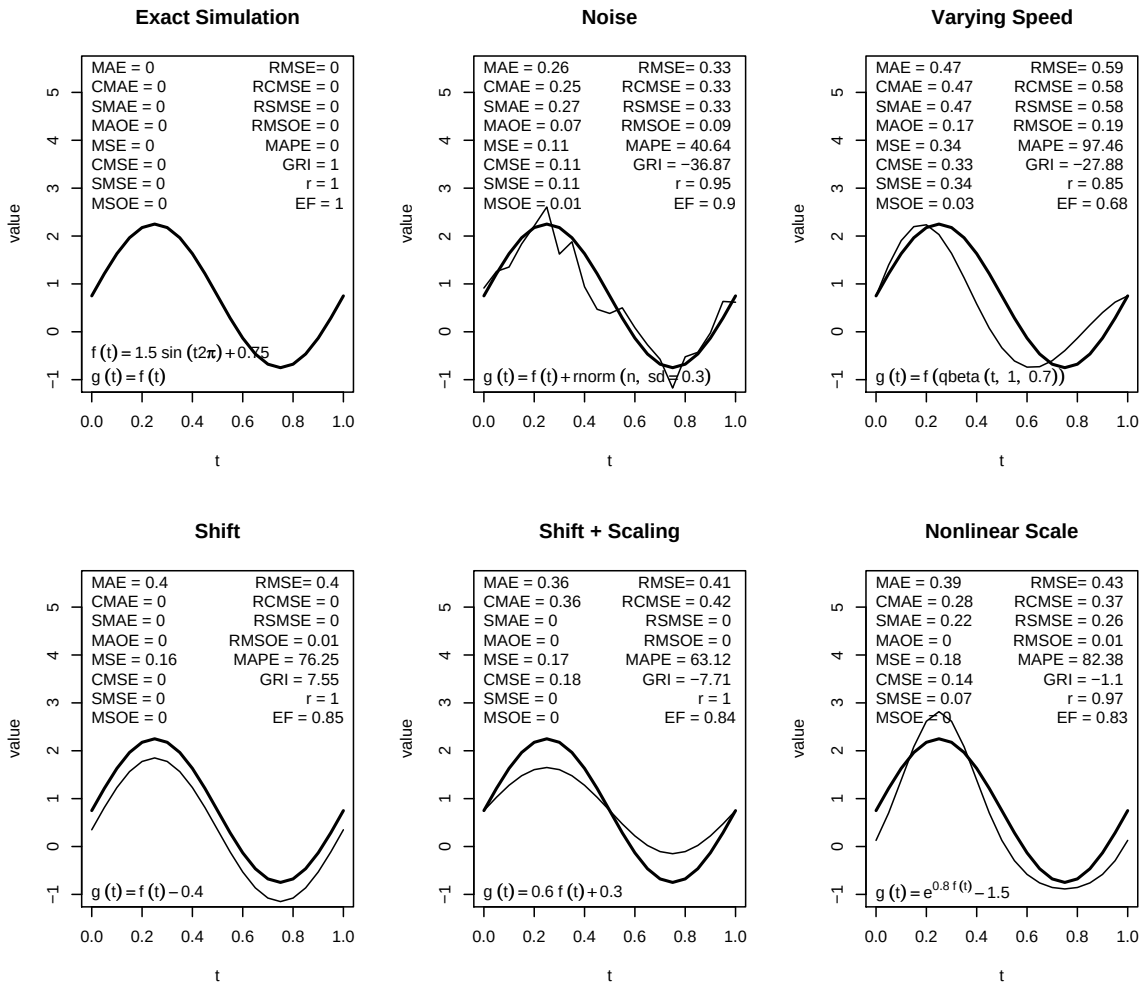

Figure 3: Deviance measures for absolute scale for special setups. $f(t)$ represents the underlying function (thick line) and $g(t)$ its modifications (thin line) for $t \in [0, 1]$ of length n.

a simpler model, e.g. to the simple model of a constant process. The deviance to a best fitting constant process is often computable by a simple statistical property of the dataset and reported with each deviance measure.

However, for qualitative comparison a simulation on a different base level or a different scaling might be considered as a qualitatively equivalent shape (Figure 3 Shift, Shift+Scaling). In order to remove the difference in mean, we could center the data beforehand by subtracting the mean for the datasets to be compared. We will encode this by the letter "C" for "centering" and get a set of deviance measures ignoring shifts:

- *centered mean absolute error*: $\mathrm{CMAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t - \mathrm{median}(y_t - \hat{y}_t)|$ is comparable to $\mathrm{MAD}(y_t)$ and to differences of y_t-values.
- *centered mean squared error*: $\mathrm{CMSE} = \frac{1}{n-1}\sum_{t=1}^{n} (y_t - \hat{y}_t - \mathrm{mean}(y_t - \hat{y}_t))^2$

is comparable to $var(y_t)$.

- *root centered mean squared error*: $\text{RCMSE} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \hat{y}_t - \text{mean}(y_t - \hat{y}_t))^2}$
  is comparable to $sd(y_t)$ and to differences of y_t-values.

The centering of MAE is done with the median, to ensure that the resulting measure is minimized among all possible centerings. For squared error based measures the usual degrees of freedom for the residuals are used.

Should scaling be ignored, we can proceed to standardize the datasets beforehand. To ensure comparability of the deviance of different models we need to guarantee that scaled or shifted model values are always giving the same deviance. We thus need to find parameters a, b, a', b' for $z_t = ay_t + b$ and $\hat{z}_t = a'\hat{y}_t + b'$, which minimize the deviance measure for z_t and $\hat{z}_t$ under such constraint. It is design decision to select the scaling of such a quantity, since we could select every scaling. We decided to keep the scaling of the reference part (e.g. the observed data) and to find an optimal fitting of the models to be compared, i.e. we fix $a = 1$. Thus if r_t denote the residuals of regression of y_t depending on $\hat{y}_t$, we would define using the letter "S" for "scaled":

- *scaled mean absolute error*: $\text{SMAE} = \frac{1}{n}\sum_{t=1}^{n}|r_t|$
  is comparable to $\text{MAD}(y_t)$ and to differences of y_t-values.

- *scaled mean squared error*: $\text{SMSE} = \frac{1}{n-2}\sum_{t=1}^{n} r_t^2$
  is comparable to $var(y_t)$.

- *root scaled mean squared error*: $\text{RSMSE} = \sqrt{\frac{1}{n-2}\sum_{t=1}^{n} r_t^2}$
  is comparable to $sd(y_t)$ and to differences of y_t-values.

Again the degrees of freedom need to be adjusted such that we get $n-2$ rather than $n-1$ as denominator.

If comparability between different datasets for the same model is desired one could exchange the role of data and model for these measures.

4.2. Deviance measures for relative scale

We consider data to be in a relative scale if they are strictly positive and the importance of the difference is given by the ratio and not by the arithmetic difference (see Figure 4 for an example). For positive data with a relative scale different measures have been defined in literature. The **qualV** package supports two of them:

- *mean absolute percentage error*: $\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{|y_t|}$
  is comparable to a percentage.
  The major problem is that the influence of total underestimation (e.g. saying 0) is limited, while radical overestimation can have unbounded influence on this deviance measure.

- *geometric reliability index*:

$$\mathrm{GRI} = \frac{1 + \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\hat{y}_t - y_t}{\hat{y}_t + y_t}\right)^2}}{1 - \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\hat{y}_t - y_t}{\hat{y}_t + y_t}\right)^2}}.$$

  according to Leggett and Williams (1981). GRI is a statistical method to determine the reliability of a model. The index is a number GRI ≥ 1, e.g.:

```
R> GRI(obss$y, sim$y)

[1] 3.483178
```

  One possible interpretation of GRI is that the simulation is accurate within a multiplicative factor, i.e. in our example the observed values fall between 1/3.48 and 3.48 times of the corresponding predicted values.

MAPE cannot be determined if measured values y_t are equal to zero and it tends to infinity if measurements are small or near to zero. This is a typical behavior, when relative errors are considered. The analysis of positive data with log-transforms has a long tradition and got a solid theoretical justification by the Euclidean space approach developed by Pawlowsky-Glahn and coauthors for compositional data (see Pawlowsky-Glahn and Egozcue 2001; Pawlowsky-Glahn 2003), and has recently been extended to positive data with a relative scale (Pawlowsky-Glahn *et al.* 2003; Pawlowsky-Glahn and Mateu-Figueras 2005; Egozcue 2005; van den Boogaart and Tolosana-Delgado 2007). A basic consequence of this work is that the relative character of the data is perfectly honored by the application of classical methods for real values to the log of positive observations. Eventually results need to be back-transformed after the analysis for a better interpretation. We can thus use the following quantities as deviance measures for the relative scale (Figure 4). The names are generated by adding an "L" for "logarithmic" in front of the "E" for "error":

- *mean absolute logarithmic error*: $\mathrm{MALE} = \frac{1}{n}\sum_{t=1}^{n} |\log(y_t/\hat{y}_t)|$
  is comparable to $\mathrm{MAD}(\log(y_t))$ and to log-ratios of values of y_t.
- *mean squared logarithmic error*: $\mathrm{MSLE} = \frac{1}{n}\sum_{t=1}^{n} \log(y_t/\hat{y}_t)^2$
  is comparable to $\mathrm{var}(\log(y_t))$.
- *root mean squared logarithmic error*: $\mathrm{RMSLE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \log(y_t/\hat{y}_t)^2}$
  is comparable to $\mathrm{sd}(\log(y_t))$ and to log-ratios of values of y_t.

The interpretation of the values is difficult because we have no feeling for expectations of absolute logs. However, the exponential of a mean log-ratio can be interpreted as a geometric mean of scaling factors. We therefore define "Geometric" rather than "Logarithmic" measures as deviance measures that can be interpreted as multiplicative factors:

- *mean absolute geometric error*: $\mathrm{MAGE} = \exp\left(\frac{1}{n}\sum_{t=1}^{n} |\log(y_t/\hat{y}_t)|\right)$
  is comparable to ratios of values of y_t.

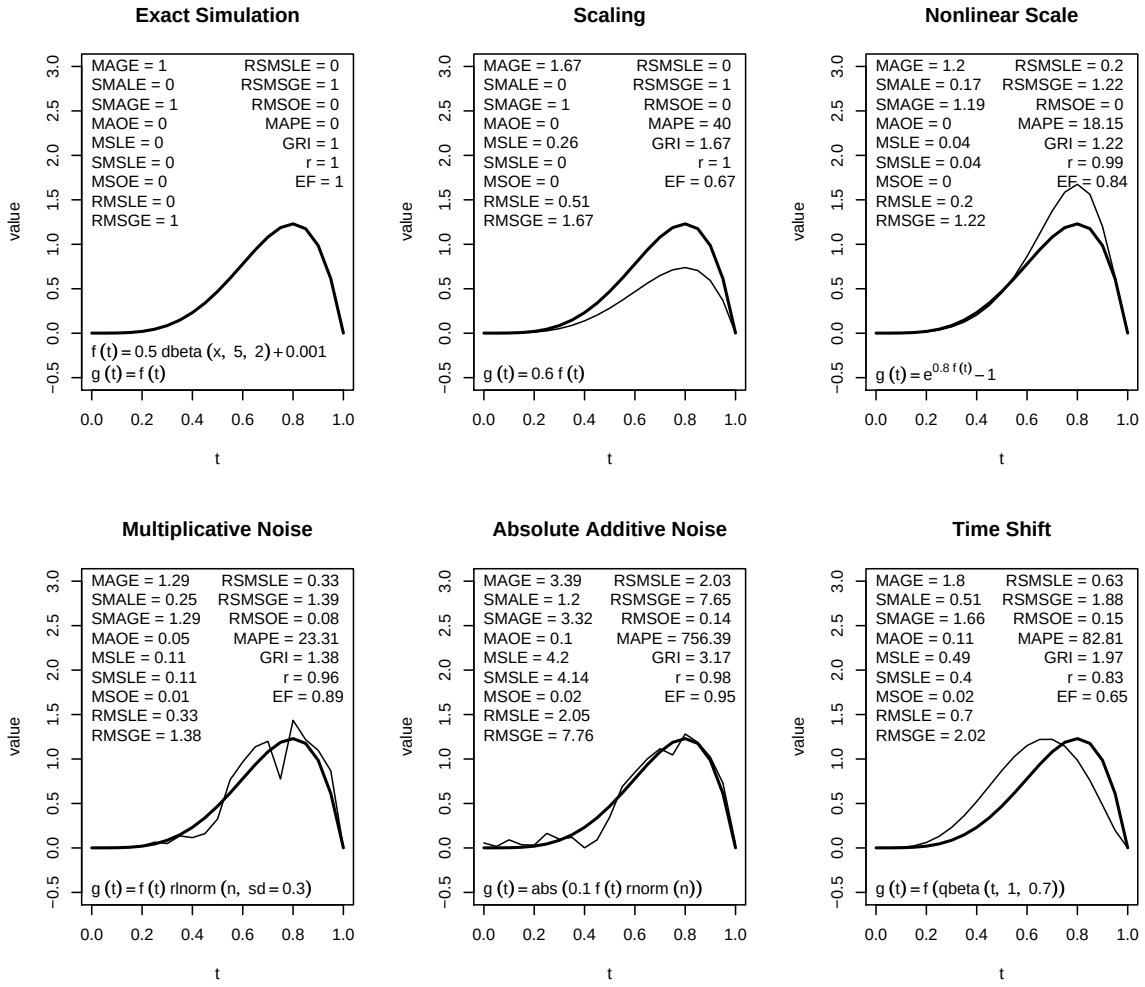

Figure 4: Deviance measures for relative scale for special setups. $f(t)$ represents the underlying function (thick line) and $g(t)$ its modifications (thin line) for $t \in [0, 1]$ of length n.

- *root mean squared geometric error*: $\text{RMSGE} = \exp\left(\sqrt{\frac{1}{n}\sum_{t=1}^{n}\log(y_t/\hat{y}_t)^2}\right)$ is comparable to ratios of values of y_t.

The quantitative interpretation of RMSGE^2 and MAGE^2 is quite similar to the interpretation of GRI. However, all three measures are differently sensitive to large relative errors. MAGE and RMSGE are sensitive to outliers just like MAD and variance. For GRI a difference of several orders of magnitude in 50% of the data can still produce a GRI-index of 6. The latter is similar to a mean simulation error by a multiplicative factor of 2.5. This is adequate when considering a qualitative simulation. It might be seen as "extreme robustness against" or as "pure ignorance of" extreme differences, when doing a qualitative comparison.

For data with a relative scale, zero is a special value. Centering is thus not an option when comparing such datasets. Thus there are no centered deviance measures for relative scale. However, scaling is similar to a centering on the log scale and thus the scaled deviance

measures for relative data are technically centered deviance measures on log-scale:

- *scaled mean absolute logarithmic error*:
  $\text{SMALE} = \frac{1}{n}\sum_{t=1}^{n} |\log(y_t/\hat{y}_t) - \text{median}(\log(y_t/\hat{y}_t))|$
  is comparable to $\text{MAD}(\log(y_t))$.
- *scaled mean squared logarithmic error*:
  $\text{SMSLE} = \frac{1}{n-1}\sum_{t=1}^{n}(\log(y_t/\hat{y}_t) - \text{mean}(\log(y_t/\hat{y}_t)))^2$
  is comparable to $var(\log(y_t))$.
- *root scaled mean squared logarithmic error*: $\text{RSMSLE} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}\log(y_t/\hat{y}_t)^2}$
  is comparable to $sd(\log(y_t))$.
- *scaled mean absolute geometric error*:
  $\text{SMAGE} = \exp\left(\frac{1}{n}\sum_{t=1}^{n} |\log(y_t/\hat{y}_t) - \text{median}(\log(y_t/\hat{y}_t))|\right)$
  is comparable to ratios of y_t-values.
- *root scaled mean squared geometric error*: $\text{RSMSGE} = \exp\left(\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}\log(y_t/\hat{y}_t)^2}\right)$
  is comparable to ratios of y_t-values.

4.3. Deviance measures for ordinal scale

Two time series can be compared as ordinal sequences of ranks, when we want to ignore the precise values and real geometry in the comparison. We propose transforming ranks to portions:

$$p_t := \frac{\text{rank}(y_t) - 1}{n - 1}$$

$$\hat{p}_t := \frac{\text{rank}(\hat{y}_t) - 1}{n - 1}$$

The measures for absolute scale, applied to portions, will then result in new interpretable deviance measures for ordinal scale. We use the letter "O" for "ordinal" to mark this type of deviance measures:

- *mean absolute ordinal error*: $\text{MAOE} = \frac{1}{n}\sum_{t=1}^{n} |p_t - \hat{p}_t|$
  is comparable to $\text{MAD}(p_t) := \frac{1}{4}$. The value can be interpreted as the mean portion of values between the modeled and the observed rank.
- *mean squared ordinal error*: $\text{MSOE} = \frac{1}{n}\sum_{t=1}^{n}(p_t - \hat{p}_t)^2$
  is comparable to $var(p_t) = \frac{1}{4}$.
- *root mean squared ordinal error*: $\text{RMSOE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(p_t - \hat{p}_t)^2}$
  is comparable to $sd(p_t) = \frac{1}{2}$ and to a portion of values.

4.4. Example

For the example data `phyto` described above a semiquantitative comparison with respect to several criteria can be produced by:

	absolute scale				relative scale		
	raw	**centered**	**scaled**	**ordered**	**raw**	**scaled**	**ordered**
mad-type	MAE	CMAE	SMAE	MAOE	MALE	SMALE	MAOE
as ratio					MAGE	SMAGE	
var-type	MSE	CMSE	SMSE	MSOE	MSLE	SMSLE	MSOE
sd-type	RMSE	RCMSE	RSMSE	RMSOE	RMSLE	RSMSLE	RMSOE
as ratio					RMSGE	RSMSGE	

Table 1: Adequate deviance measures for different situations: the columns specify the aspects to be ignored (raw: original values are compared, centered: differences in mean are removed, scaled: scaling is ignored, ordered: ordinal geometry is used), the rows indicate how distances are been measured (mad-type: mean absolute distances, var-type: squared distance, sd-type: root of mean squared distances).

```
R> sqc <- compareME(obs$y, sim$y, obs$t, sim$t, type = c("normalized"))
```

The table of normalized deviance measures calculated for various grid cells is given by:

```
R> print(sqc, digits = 3)

$normalized
                    time   fixed
                    ignore   raw centered scaled ordered
geometry    measure
real        mad             0.855    0.703  0.723   0.731
            var             0.389    0.356  0.342   0.219
            sd              0.624    0.597  0.585   0.468
logarithmic mad             0.240    0.815  0.815   0.731
            var             1.087    0.644  0.644   0.219
            sd              1.043    0.803  0.803   0.468
geometric   mad             0.945    0.702  0.702   0.731
            var             1.108    0.620  0.620   0.219
            sd              1.108    0.620  0.620   0.468
ordinal     mad             0.731    0.731  0.731   0.731
            var             0.219    0.219  0.219   0.219
            sd              0.468    0.468  0.468   0.468

R> compareME(type = "name")

$name
                    time    fixed
                    ignore    raw centered scaled ordered
geometry    measure
real        mad             MAE      CMAE   SMAE    MAOE
            var             MSE      CMSE   SMSE    MSOE
            sd              RMSE     RCMSE  RSMSE   RMSOE
```

```
logarithmic mad            MALE     SMALE  SMALE   MAOE
            var            MSLE     SMSLE  SMSLE   MSOE
            sd             RMSLE    RSMSLE RSMSLE  RMSOE
geometric   mad            MAGE     SMAGE  SMAGE   MAOE
            var            RMSGE    RSMSGE RSMSGE  MSOE
            sd             RMSGE    RSMSGE RSMSGE  RMSOE
ordinal     mad            MAOE     MAOE   MAOE    MAOE
            var            MSOE     MSOE   MSOE    MSOE
            sd             RMSOE    RMSOE  RMSOE   RMSOE
```

The latter command shows the names of deviance measures. Here `time fixed` indicates that actually the time is not transformed. The row `ignore` specifies the aspects of data to be ignored, where `raw` compares original values, `centered` removes differences in mean, `scaled` ignores scaling, `ordered` uses ordinal geometry. The column `geometry` indicates the geometry to be used for the data and the output. Here `real` corresponds to arithmetic differences and means, `logarithmic` handles relative data on a logarithmic scale, `geometric` determines geometric differences and means, and `ordinal` compares values as ordinal sequence of ranks. The column `measure` specifies how distances are been measured, i.e. as mean absolute distances (`mad`), as squared distances (`var`), or as root of mean squared distances (`sd`).

Reasonably small values with less than 50% deviance can be found for the variance measures on the ordinal and the real scale. We can thus assume that the simulated data have quite high relative errors, but fit reasonably well in general shape and in terms of mean squared differences.

4.5. Similarity measures

Often similarity measures, such as the **coefficient of correlation** according to Pearson, measuring in a dimensionless way the linear relationship between two variables, are preferred to deviance measures, because a direct interpretation of the value exist. A value of 1 always means complete similarity and 0 always means unrelated dissimilarity.

Classical similarity measures are:

- The Pearson correlation coefficient:

$$r = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}},$$

  where x_t and y_t are continuous random variables for $t = 1, \ldots, n$, and $\bar{x}$ as well as $\bar{y}$ are the corresponding means. It is assumed that x_t and y_t are independent of each other as well as normally distributed. The correlation coefficient ranges between -1 and 1, where -1 is a perfect negative correlation, 0 denotes no correlation, and 1 is a perfect positive correlation. The correlation coefficient r is a statistical standard function in R (see function `cor` for more informations).

- The *rank correlation coefficient* according to Spearman is similar to Pearson's correlation. However, instead of the data itself, the ranks are used. It is a non-parametric

measure of correlation. In general, it is capable to assess monotonic and also nonlinear relationships without any assumptions about the frequency distribution of the data.

- The *efficiency factor* of Nash and Sutcliffe (1970) is also a dimensionless statistical measure, which directly relates model predictions to observed data:

$$\begin{aligned} \text{EF} &= 1 - \frac{(SS \text{ over } y = \hat{y})}{(\text{improved } SS \text{ over } y)} \\ &= 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2}, \end{aligned}$$

  where SS is the sum of squares. The efficiency factor takes values between minus infinity and 1, whereas negative values do not recommend for the model and values near 1 indicate good model performance. EF is implemented in **qualV** as function `EF(o, p)` with two vectors `o` and `p` as data to be compared.

The advantage of similarity measures are the easy interpretation of 1 as the maximum possible similarity and 0 as the absence of similarity. Possible negative values indicate opposite behavior. Such a measure of similarity can be defined for each of the deviance measures given above by a simple formula: Let D be a deviance measure and C a comparative value for the similarity to a constant as given for each of the deviance measures above. We could get a normalized deviance measure N according to:

$$N = \frac{D}{C},$$

which is 0 for perfect similarity and 1 for absence of similarity (i.e. the similarity to a well chosen constant). To get a corresponding similarity measure S one could easily change the sign:

$$S = 1 - \frac{D}{C}.$$

The squared Pearson correlation coefficient is conceptually related to a similarity measure of SMSE, as Spearman correlation is to MSOE and the efficiency factor with MSE.

For multivariate time series similarity measures for different dimensions could be combined by taking the mean similarity.

5. Qualitative validation criteria

5.1. Time transformation method

Commonly used quantitative deviance measures often show a large dissimilarity of patterns, which are just shifts or different speeds in time between observation and simulation. To define deviance and to measure model performance independent of time shifts and time speed changes, we could transform the time of the simulation, i.e. to run the time faster or slower, in a way that the deviance of the measured time series and the transformed predicted time series gets minimal.

A time transformation is thus an increasing, bijective mapping of an interval of time (e.g. $[0, 1]$) to itself. Our package provides three types of time transformations (compare Figure 5):

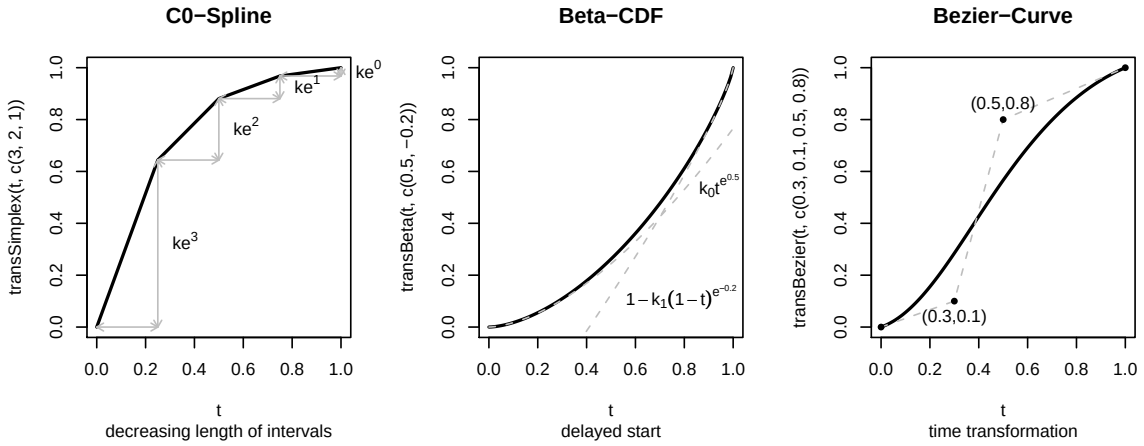

Figure 5: Transformation functions and the influence of parameters on the time transformation models. The C_0-Spline is parameterized by (3,2,1) and increases proportional to e^{p_i}. The Beta-CDF with parameters $p_1 = 0.5$ and $p_2 = -0.2$ is on both sides asymptotically proportional to a power. The Bezier-Curve is parameterized by the vectors (0.3,0.1) and (0.5,0.8) and asymptotic to corresponding segments. Proportionality constants are denoted by k, k_0, and k_2.

- *Increasing C0-Splines (*`transSimplex`*)*:
  $C0$-Splines are a very simple method to approximate a function. The Spline is parameterized by $(p_i)_{i=1,\dots,d-1} \in \mathbb{R}^{d-1}$. The time of the simulation is split into d intervals of the same length. The duration of the i-th interval in observation time is than proportional to $\exp(p_i)$, where p_d is set to 0.

- The *Beta Cumulative Distribution Function (*`transBeta`*)*:

$$\mathrm{CDF}(x;\alpha,\beta) := \frac{\int_0^x u^{\alpha-1}(1-u)^{\beta-1}du}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$$

  is a simple two parametric family of strictly increasing functions on the interval $[0,1]$, defined for $\alpha > 0$, $\beta > 0$, and $u \in \mathbb{R}$. We therefore parameterize the function as $\alpha = e^{p_1}$ and $\beta = e^{p_2}$ for $p \in \mathbb{R}^2$.

- *Bezier-Curves (*`transBezier`*)*:
  Bezier-Curves (e.g. Aumüller and Spitzmüller 1993) are general purpose free form curves from computer graphics, which handle the x and y coordinates interchangeably. Bezier-Curves of order n are parameterized by a starting point $c_0 = (0,0)$, an end point $c_n := (1,1)$, and a sequence of $n-1$ control points $c_i := (p_{2i-1}, p_{2i})$. Roughly speaking the Bezier-Curve is a smoothed version of the segmented curve going near $c_0, \dots, c_n$ in sequence. We thus parameterize a Bezier-Curve by $p \in [0,1]^{2(n-1)}$. However, there is no guarantee that this corresponds to a function or is increasing for $n \geq 3$. This transformation method is nevertheless useful for a fine tuned curve with small variation from the identity function.

This idea would lead to a definition of a time shifted version of some mean error ME (e.g.

MSE, SMAGE, ...) given by formula like:

$$T\mathsf{ME}_1((y(t_i))_i, (\hat{y}(t_i))_i) := \inf_{T \in \text{Transformations}} \mathsf{ME}((y(t_i))_i, (\hat{y}(T(t_i)))_i),$$

where T is the time transformation, $y(t_i)$ the observed and $\hat{y}(t_i)$ the predicted time series for $i = 1, \ldots, n$. Here ME is any of the deviance measures defined in Chapter 4 and chosen depending on our perception of the space of values. Adding the possibility of a time shift is an orthogonal concept and can be applied to each of the measures defined in the preceding chapter.

However, such a naive definition leads to two practical problems:

- *$\hat{y}(T(t_i))$ is typically difficult to compute for a general time transformation T.*
  Say, we would like to compare a time series $(y(t_i))_{i=1,\ldots,n}$ of observations with a time series $(\hat{y}(s_j))_{j=1,\ldots,m}$ of predictions calculated at different times. To resolve this problem, we propose to interpolate both time series at all data points available in one or the other time series, and compare each pair. Thus let $z(t)$ denote an interpolation of the observed time series, $\hat{z}(t)$ an interpolation of the predicted time series, and $x_k = t_k, k = 1, \ldots, n$ and $T(x_k) = s_{k-n+1}$ or $x_k = T^{-1}(s_{k-n+1})$ respectively for $k = n+1, \ldots, n+m$ to extend the definition to:

$$T\mathsf{ME}_2((y(t_i))_i, (\hat{y}(t_i))_i) := \inf_{T \in \text{Transformations}} \mathsf{ME}((z(x_k))_k, (\hat{z}(T(x_i)))_k).$$

- *Extreme transformations are sometimes quantitatively optimal*
  Sometimes, especially with MSE or very flexible time transformations, a minimal deviance is produced by compressing one of the time series into a very short range of the other. However, these "best fits" solutions are pure artifacts and do not reveal anything about the qualitative true similarity. For avoiding such extreme deformations it is sometimes useful to penalize the fitting by a deviance measure for the time:

$$T\mathsf{ME}((y(t_i))_i, (\hat{y}(t_i))_i) := \inf_{T \in \text{Transformations}} (\mathsf{ME}((z(x_k))_k, (\hat{z}(T(x_i)))_k) + \alpha \mathsf{ME}_t((x_k)_k, (T(x_i))_k)),$$

  which is the final definition for a time shifted version of some ME. The previous definition $T\mathsf{ME}_2$ can be found as a special case for $\alpha = 0$. The choice of the second deviance measure ME_t and of the weighting α is a tricky task. The choice of α tackles the objectivity of the measure. We therefore use $\alpha = 0$ as default option.

Regardless of α there is no need to do any time transformation for a constant time series. The reference value for a TME is thus the same as for ME itself.

For multivariate time models the individual outcomes could either be studied individually or using a joint criterion and a common time transformation. However, this second approach is currently not implemented in our package.

Example

Using the smoothed data `obss` we can generate a comprehensive table of deviance measures by:

```
R> sqc <- compareME(obss$y, sim$y, obss$t, sim$t, time = c("fixed",
+     "transform"), measure = "var", geometry = c("real",
+     "logarithmic"), type = c("normalized"), trials = 5,
+     col.vars = "ignore")
```

The command allows you to specify each independent parameter (`time` for the optional time transformation, `measure` for the measure of dissimilarity being used, `geometry` for the assumed geometry of the data, `type` for the type measure being used, and `col.var` to specify the display in the columns of the resulting table) of the deviance measures separately. In case of multiple choices the measures for all possible combinations are given.

Here the comparison is done simultaneously for non-transformed (`"fixed"`) and for transformed time (`"transform"`). Note, that in `compareME` the beta distribution is applied as default time transformation function. For arithmetic differences and means (`"real"`) and on logarithmic scale (`"logarithmic"`) `"normalized"` squared distances (`"var"`) are measured. `trials` gives the number of random starting values that should be used during the optimization of the time transformation. The argument `col.vars = "ignore"` just specifies the column variables of the table output and is used here for formatting purpose only.

```
R> print(sqc, digits = 3)
```

```
$normalized
                              ignore    raw centered scaled ordered
geometry    measure time
real        var     fixed          0.4029   0.3549 0.3545  0.2353
                    transform      0.1852   0.0886 0.0687  0.1022
logarithmic var     fixed          0.8794   0.5264 0.5264  0.2353
                    transform      0.8103   0.4703 0.4703  0.1021
```

See the example in Section 4.4 for the declaration of column and row names. The result shows that a very good performance (0.069) of the model can be achieved in scaled real geometry, if a time shift is allowed (Figure 6, dashed line). The comparable bad performance (0.47) on relative logarithmic scale suggests that the model has high relative errors for small values.

The time transformation of simulated values using the beta distribution and the deviance measure SMSE is determined by:

```
R> tt <- timeTransME(obss$y, sim$y, obss$t, sim$t, ME = SMSE,
+     time = "transform", type = "normalized", trials = 5,
+     timeME = MAE, timeMEtype = "dissimilarity")
```

result in an error `totalME` of:

```
R> print(tt, digits = 2)
```

```
totalME  timeME
  0.069  17.407
```

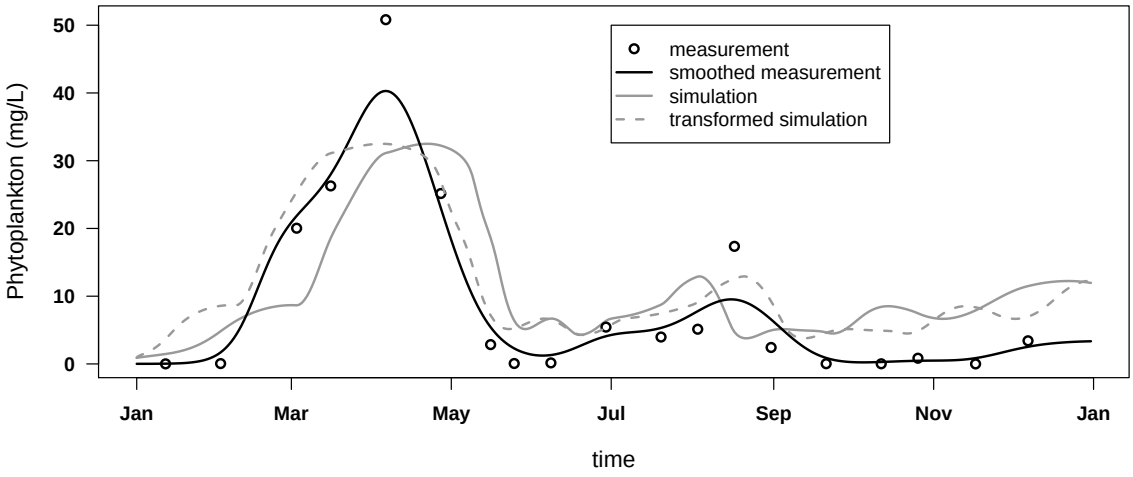

Figure 6: Illustration of time transformation for phytoplankton data (circles: measured data, solid black line: smoothed measurements, solid gray line: simulated data, dashed gray line: time transformed simulation data)

The second output of the function `timeTransME` is a dissimilarity measure of the temporal deformation determined as MAE in time between observation and simulation. For our example the value can be interpreted that the similarity of the pattern of the phytoplankton concentration increases with a difference in time of around 17 days.

In Table 2 a set of model evaluation criteria is selected to show the performance of the time transformation.

5.2. Interval sequences

Note that the procedures of combining deviance measures and accepting a time shift still keep one relation constant: They assume that the values are comparable in each time series of the whole time sequence. However, sometimes we would assume that values more than some time apart can not be compared directly any more. In this case the qualitative behavior of the time series is defined by the sequence of its local shapes.

Interval sequences have been proven suitable to analyze and to compare the qualitative behavior of time series (Agrawal *et al.* 1995a; Höppner and Klawonn 2001; Höppner 2001, 2002b; Cuberos *et al.* 2002). Similar methods are also used in genetics to search for DNA patterns in large databases, and to localize frequently recurring patterns. Here the method of interval sequences is used to compare two time series. Both time series are disaggregated into equal segments. The segments are described by qualitative features. The features can be named with symbols (like "A", "B", "C", ...) resulting in one string for each time series, which represents the interval sequences. The comparison occurs on the basis of the two interval sequences. Differently from the proposed methods in Sections 4 and 5.1 not the distance between the values of two time series decides about the similarity, but their qualitative behavior.

In some cases it is useful to standardize the time series before dividing them into segments.

		Time transformation	
Criterion	Symbol	before	after
mean absolute error	MAE	0.83	0.63
mean absolute percentage error	MAPE	2.81	4.76
mean square error	MSE	0.4	0.19
centered mean square error	CMSE	0.35	0.09
scaled mean square error	SMSE	0.35	0.07
mean square log error	MSLE	0.88	0.93
scaled mean square log error	SMSLE	0.53	0.49
mean square ordinal error	MSOE	0.24	0.12
correlation coefficient	r	0.8	0.97
model efficiency	EF	0.6	0.81
geometric reliability index	GRI	3.48	3.47
time MAE		0	0.17

Table 2: Selected model evaluation criteria for the example in Figure 6 before and after time transformation using the beta distribution. All measures are given in their normalized version.

One common method is scaling into the interval $[0, 1]$. Thereby quantitative differences between time series are removed.

Possible features for defining interval sequences

Which features are used to define interval sequences depends on the application. No generally accepted criteria exist. Here some possible features of time series are presented, which are implemented within the package **qualV**.

One obvious possibility is the description of the time series using the first derivative (`f.slope`, Höppner 2002b; Cuberos *et al.* 2002). The time series are separated into sequences which increase, decrease or stay constant. This property provides a first impression of the rough development of the values and distinguishes between the shapes "increase" and "decrease".

The second derivative (`f.curve`) is another possible feature for interval sequences. This property provides more detailed information about the shape of the time series - convex, concave or linear. In combination with the first derivative, the time series can be qualitatively described with terms like "convex decreasing" or "concave increasing" (see Höppner 2002b).

The characterization with various degrees of steepness (`f.steep`) is an alternative to the second derivation, e.g. in terms like "very steep", "steep", and "not steep" (see Höppner 2001) or low, moderate and high increasing or decreasing, and constant segments (see Cuberos *et al.* 2002).

Another possible feature arises from the categorization into high, middle and low values (`f.level`, compare Höppner 2001). The thresholds for this property depend on the values and the considered question. For example, are the values scaled into the interval $[0, 1]$, all values greater 0.8 can be considered as high and values below 0.2 as low. The purpose of this feature is to distinguish quantitative differences independent of scaling.

For a multivariate time series features from several dimensions can be combined to joint features. There is no generally applicable universal set of features. Using more features

provides a more detailed comparison, but requires a higher degree of similarity of the time series to give meaningful results.

Determination of similarity

After converting each time series into an interval sequence represented as string, a method is required to determine the similarity. Cuberos *et al.* (2002) proposed the "quality similarity index" (`QSI`) by dividing the length of a longest common subsequence (LCS) over the maximum length of both interval sequences. A subsequence of a string is defined as a subset of characters of the string, which maintain their original order, i.e. a subsequence does not need to consist of neighboring characters of the string. A longest common subsequence of two finite strings is defined as the longest accumulation of characters, which have the same order in both sequences (Gusfield 1997). Thereby this method allows insertions and deletions, i.e. characters can be ignored or missed. This behavior is complementary for the LCS of two strings, i.e. when characters are skipped in one string, the characters are missing in the other string and vice versa.

The LCS problem is that of finding a LCS of two given strings and to measure the length of it. The LCS problem can be applied in many areas, e.g. in molecular biology to compare DNA-sequences, or in computer science to compare two versions of one text. For solving the LCS problem various dynamic programming algorithms are possible, as described by several authors (e.g. Wagner and Fischer 1974; Hirschberg 1977; Nakatsu *et al.* 1982; Apostolico *et al.* 1992; Gusfield 1997). Paterson and Dančík (1994) give a survey of algorithms and their time and space complexity. One common algorithm is provided by the package **qualV**.

`QSI` takes values between 0 and 1. An index equal to 0 indicates no match, i.e. the qualitative behavior of two time series is completely different concerning the considered feature. An index near to 1 characterizes a similar qualitative behavior. Note that an index equal to 1 does not necessarily indicate identical values of both time series. Only the identity of the features is considered.

Example

For the interval sequence comparison, we use the smoothed data `obss` like in the time transformation method. Choosing from several possible features we use the most basic feature (`f.slope`), which is essentially a measure of the first derivative:

```
R> obs_f1 <- f.slope(obss$t, obss$y)
R> sim_f1 <- f.slope(sim$t, sim$y)
```

The result is a character vector with, in this case, the symbols "A" for increasing, "B" for decreasing and "C" for constant. The resulting sequences of the observed and simulated time series are printed to give an impression:

```
R> cat(obs_f1, sep = "", fill = 76)

CCCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
BBCCCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAC
```

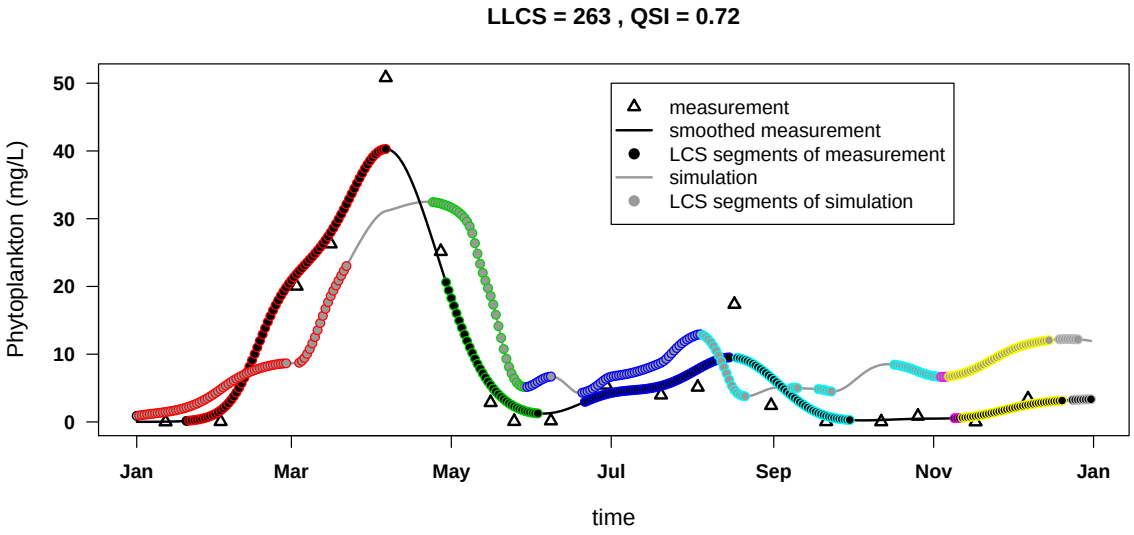

Figure 7: Illustration of interval sequence comparison for phytoplankton data (triangles: measured data, solid black line: smoothed measurements, solid gray line: simulated data, color filled circles: matching intervals)

```
CBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACCCCCCCCCCC
```

```
R> cat(sim_f1, sep = "", fill = 76)
```

```
CAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACCCCAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACCCBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAA
AAAAAAACBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBB
BBBBBCAAAAAAAAAAAACCCCBBBCCCCCCCBBBBBBCAAAAAAAAAAAAAAAAAAAACCBBBBBBBBBBBBBB
BBBCCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACCCCCCCCCBBBBB
```

Function `LCS(a, b)` is used to find the longest common subsequence using a dynamic programming algorithm. The result of this function is a list with vectors `a` and `b` holding the original interval sequences, `LCS` and `LLCS` are the longest common subsequence and its length, respectively, `QSI` being the LCS based similarity index, `va` and `vb` containing a possible LCS solution (visualized in Figure 7) for the input vectors `a` and `b`, respectively:

```
R> lcs <- LCS(obs_f1, sim_f1)
```

which results in `lcs$QSI`=0.72. The `QSI` is not affected by systematic deviations or shifts in time, but the duration of the considered feature is of importance. In the example the decreasing period of the first maximum is longer for the measurement than for the simulation (compare the green parts in Figure 7). Thus, the observation contains additional segments, which are not reproduced by the simulation. These interval sequences remain unconsidered in the `QSI`.

6. Inferential considerations

The deviance measures provided are descriptive and exploratory tools. With a clear idea of the relevant difference we can select the appropriate deviance measure and describe the difference between data and model numerically. This is similar to the use of a mean or standard deviation to describe location and spread of a specific sample. Without this clear idea one could calculate a whole set of deviance measures. This provides informations in which aspects data and model are similar and in what quality they differ. We can also use different data sets or different models to compare which of them are more similar to their corresponding opposite. However, this is not an inferential comparison by a statistical test but an exploratory comparison like the comparison of two sample means or two values in a dataset. This is saying something in an exploratory context, but it needs confirmation by a test to transform from an observation to a law.

Evidently it would be very nice to have some statistical tests for the qualitative comparison of qualitative behavior of inhomogeneous time series. One might come up with simple statistical tests such as a correlation test. However, the situation is more difficult. To establish inferential statistical procedures, we need at least two things:

- A relevant scientific question.
- A valid model of the inherent randomness.

A clear source of randomness are the measurement errors, which might be modeled as independent with known or unknown variances. Other sources of variation in transient random systems are the variability of starting conditions, imprecise process parameters, simplifications in the model description, incompletely known boundary conditions such as weather, and indeterministic or chaotic progression of subsystems. This sort of variability is stochastically dependent by nature and has a complex unmodeled structure. To cope with this second type of randomness, we either need a precise model of the inner structure of the system, which is an approach far beyond the scope of this paper, or multiple independent realizations of the system. A statistical modeling of the dependence by standard models like autoregressive or moving average processes seems inappropriate for nonstationary times series. We thus have no strong general model of error in the individual realizations of the system.

Several types of questions seem to be relevant:

- *"Is the model somehow similar to nature?"*
  This is a very shy question of "is my model predicting anything" and would need a test of the form:
  $$H_0 : S = 0 \text{ v.s. } H_0 : S > 0$$
  where S is an appropriate measure of similarity and 0 has a meaning. For a single sequence of data pairs this is e.g. done by correlation tests. However, these tests rely on stochastic independence of the observations and thus is inappropriate for our situation. A way out might be to read:
  $$H_0 : E[S] \leq 0 \text{ v.s. } H_0 : E[S] > 0$$
  or
  $$H_0 : P(S \leq 0) \leq 0.5 \text{ v.s. } H_0 : P(S \leq 0) > 0.5$$

instead and to draw a random sample from a statistical population of systems modeled by our model and to check this by a t-test or a binomial test.

However, often 0 similarity is an abstract concept and has no true meaning. Thus maybe the more appropriate question would be:

- *"Is the prediction better than pure guesswork?"*
  In this case we could replace the H_0 hypothesis by a model for guesswork. Since a canonical pure guesswork does not exist, we have to consider different definitions of guesswork: a) another simpler model, b) or the chosen model but in another time sequence, c) the outcomes from a set of different models for different situations. For a) the most simple model is that of a constant system, which is the inbuilt comparison model for our similarity measures. Thus the question is equivalent to the first possibility.

  We can get b) e.g. by reordering predicted outcomes by a random permutation. The similarity measure S would get the test statistics. Its distribution under the 0 hypothesis can be calculated by Monte-Carlo methods and the test would reject the 0 hypothesis for high values of S. However, still both time series have in common to be somehow correlated in time. This results in a higher probability of extreme values for S and thus a falsely significant test. In conclusion approach b) could deliver significant results simply based on the similarity of being both nonrandom.

  We can get c) by pairing every prediction with every dataset and ask whether the similarities of matching pairs are higher than the similarities of nonmatching pairs. This can be done by the Wilcoxon-Rank-Sum-Test of matching similarities against the nonmatching similarities. However, this is a relative comparison checking that the matching partner gives a better prediction than nonmatching. This answers the question: Is the adaption to different situations pure guesswork?

  In conclusion there is no generally satisfying concretization of the first problem.

- *"Does the model fit nature?"*
  The mathematical concretization of this is: Does the model perfectly describe nature up to acceptable sources of variation. This being the more sparse assumption it must serve as the hypothesis of the test. However, this hypothesis is in general false, especially in bio-related sciences, since a model is a model and not a perfect image of the truth. The whole approach of qualitative comparison is based on accepting some deviations as inevitable. The whole concept of accepting measurement errors and variation renders the model as incomplete. The approach of providing the model as time series is not able to represent such details. We thus consider this problem of "Does the model fit nature?" as irrelevant and not answerable in the given context. However, the whole approach focuses on comparison and thus the question of interest might be:

- *"Is model A describing the observation better than model B?"*
  This is exactly the answer provided by the deviance measures. However, since each realization of an observed time series is only one independent observation, we can not prove a general law of one model being better than the other from a single comparison. We thus need a whole independent sample of realizations of the system to choose which model performs better for the whole statistical population of possible realizations. However, if we have such a set of different realizations we could compare each of them to both models with our preferred deviance measure and get a dataset of paired deviances.

These can be compared by a paired test like the sign-test or Wilcoxen-Signed-Rank-Test depending on what the appropriate alternative of deviances seems to be.

We decided to put no tests into the package because the only appropriate tests are standard tests to be performed on the outcome of our deviance measures for whole datasets of datasets.

7. Discussion

Dynamic models are useful tools for process understanding, for consistency checking of existing knowledge, for deriving hypotheses, and for making predictions. For both, theory and applications, model validation is crucial. Validation is necessary for accepting a model and for identifying the range of its applicability, as there are often uncertainties about the scope of questions to be answered with one particular model (see Mayer and Butler 1993; Reckhow 1994; Elliott *et al.* 2000).

Numerous methods already exist for the evaluation of models, but for different reasons not all methods are suitable for a certain model. In many cases a set of different techniques are required in order to obtain an overall assessment of model performance, but there is no combination of methods applicable to all models, which can be suggested as universal solution (see Mayer and Butler 1993; Sargent 1998). The set of validation methods introduced in Sections 3.1 and 3.2 are without exception either subjective assessments or based on quantitative measurements with corresponding values in space and time. In practice methods for the qualitative assessment of models by means of observed values are seldom used (Konstantinov and Yoshida 1992; Agrawal *et al.* 1995a,b; Höppner and Klawonn 2001; Höppner 2002a; Cuberos *et al.* 2002), but important, when numerical validation techniques do not apply.

McCuen and Knight (2006) investigated the behavior of the commonly used similarity measure EF of Nash-Sutcliffe in more detail. They showed that outliers can significantly influence sample values of EF. Furthermore time delays and bias in magnitude can have an adverse effect on EF. They pointed out that EF can only be a reliable goodness-of-fit statistic if it is properly interpreted.

Elliott *et al.* (2000) used a total of ten validation statistics to test the quality of a new model, but these were all conventional methods based on a quantitative comparison (e.g. MAE, MAPE, RMSE, EF, ...). For the authors the results of the evaluation were disappointing and they refer to Grimm (1994) and Grimm *et al.* (1996) that it is not necessarily of interest to simulate data in high precision for concentrating on the pattern observed. They concluded that only some of the statistics they used were suitable and that they should be used in combination with visual techniques, i.e. a subjective assessment.

The weak point in using only quantitative methods for validation is that in the case of time delay the comparison fails in spite of similar patterns without identifying the reason (the time delay). Here the orthogonal set of deviance measures that we propose is not only an approach to complete commonly used deviance measures for model validation, but to identify the type of difference. The complicate question which difference should be regarded as important can not be answered by this method. That remains a user decision dependent on the application of the model.

In case of time delays Marron and Tsybakov (1995) suggested an error criteria using "visual" instead of vertical distances. They defined visual distance to be the shortest distance from a

given point of the estimation to any point of the observation. This approach is applicable in finding the "best" estimate close to visual impression.

Similarly the new proposed method of time transformation has a special focus on the problem of shifts in time. In addition to the method of Marron and Tsybakov (1995) a value of the goodness of fit can be determined as well as the size of the difference in time. For example the pattern of the measured values in Figure 2 points two maxima of phytoplankton growth, a large one in spring and a smaller one in the end of summer. The model also simulates two maxima, but with a small time delay for the first peak, and the second one appears earlier in comparison to the measurement. A statistical model evaluation with a correlation coefficient of 0.8 and an efficiency factor of 0.6 already delivers a quite "good" model performance, but may not satisfy everybody. Transforming the time of the simulation with the beta distribution improves the model performance considerably (Figure 6). Now a correlation coefficient of 0.97 proves an increasing linear relationship. The efficiency factor has improved to 0.81, but in relation to the high correlation coefficient it indicates systematic deviations, which is up to a factor of 3.47 (geometric reliability index).

Konstantinov and Yoshida (1992) considered that measured values, which are afflicted with measurement errors are volatile and never follow exactly the same pattern, and any attempt to compare and analyze time profiles in a strict quantitative fashion will be inadequate. They suggested to reduce the accuracy to an appropriate level of abstraction, preserving only the underlying shape feature. Therefore they used a qualitative analysis, removed quantitative details and determined the similarity by means of qualitative features. Analogously, Agrawal *et al.* (1995b) developed a "shape definition language" to define the qualitative behavior of time series. Similar to the method of interval sequence this approach translates the shape of the time series into symbols. The symbols describe classes of transitions. Dividing measurement and simulation into interval sequences by using the first derivative and determining the similarity of the pattern with decreasing, increasing and constant features delivers a `QSI` of 0.72 for the above example. Here systematic deviations and shifts in time are downweighted. The order of the features is kept and additional (insertions) or missing segments (deletions) are allowed. Solely the duration of a feature affects the `QSI`.

Qualitative comparison of a model and nature can be based on rather different aspects of similarity: Similar values (quantitative similarity), similar distribution (stochastic similarity), similar processes (analogy), similar behavior (phenomenological similarity). The approach of this paper is focused on phenomenological similarity of the qualitative behavior of the outcome. We justify this as a pragmatic approach, since quantitative similarity is often not within reach, stochastic similarity needs much data, and analogy of the process is typically the basis the model has been build on originally. On the other hand if a model is both analog to nature and phenomenological validated it might adequately present the qualitative behavior of the modeled system.

We propose a system of elementary deviance and similarity measures in theory and software for this special type of comparison. From the viewpoint of descriptive statistics the measures allow a comparison of the modeling fitness between different models for the same situation, between different situations simulated by one individual model, and with similarity measures even between different models applied to different situations. For each situation the adequate measure can be selected based on the answers to six simple questions. Important limitations for inferential statistics in the given context were discussed in detail. However, in case of representative samples of real world systems described by various models, common statistical

tests can be used to compare the resulting quality measures and to assess the phenomenological performance of the applied models. Moreover a combination of multiple deviance measures used in an exploratory context can help to identify both, deficiencies and capabilities of given and newly developed ecological models.

Acknowledgments

The authors appreciated very much the helpful and detailed comments of the two reviewers, which significantly helped to improve the manuscript substantially. We would also like to thank Ben Poulter for his helpful comments.

References

Agrawal R, Lin KI, Sawhney HS, Shim K (1995a). "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases." In "VLDB '95: Proceedings of the 21. International Conference on Very Large Data Bases," pp. 490–501. Morgan Kaufmann Publishers Inc. ISBN 1-55860-379-4.

Agrawal R, Psaila G, Wimmers EL, Zait M (1995b). "Querying Shapes of Histories." In "VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases," pp. 502–514. Morgan Kaufmann Publishers Inc. ISBN 1-55860-379-4.

Apostolico A, Browne S, Guerra C (1992). "Fast Linear-Space Computations of Longest Common Subsequences." *Theoretical Computer Science*, **92**(1), 3–17. ISSN 0304-3975. doi:10.1016/0304-3975(92)90132-Y.

Aumüller G, Spitzmüller K (1993). *Computerorientierte Geometrie*. BI Wissenschaftsverlag, Mannheim u.a.O. ISBN 3-411-16021-7.

Balci O, Sargent RG (1982). "Some Examples of Simulation Model Validation Using Hypothesis Testing." In "WSC '82: Proceedings of the 14th conference on Winter Simulation," pp. 621–629. edited by Highland, Chao, and Madrigal. ISBN 0-123-454678-0.

Benndorf J, Kranich J, Mehner T, Wagner A (2001). "Temperature Impact on the Midsummer Decline of *Daphnia galeata* from the Biomanipulated Bautzen Reservoir (Germany)." *Freshwater Biology*, **46**, 199–211.

Benndorf J, Recknagel F (1982). "Problems of Application of the Ecological Model SALMO to Lakes and Reservoirs Having Various Trophic States." *Ecological Modelling*, **17**, 129 – 145.

Cuberos FJ, Ortega JA, Gasca RM, Toro M, Torres J (2002). "Qualitative Comparison of Temporal Series - QSI." *Topics in Artificial Intelligence. Lecture Notes in Artificial Intelligence*, **2504**, 75–87.

Egozcue JJ (2005). "Applications of Statistical Analysis on Coordiantes to Compositional and Positive Data." In "International Statistical Institute. Abstract book: 55th session of the International Statistical Institute (ISI), 5-12 April 2005, Sydney Convention & Exhibition Centre, Sydney, Australia.", ISBN 1-877040-28-2.

Elliott JA, Irish AE, Reynolds CS, Tett P (2000). "Modelling Freshwater Phytoplankton Communities: an Exercise in Validation." *Ecological Modelling*, **128**(1), 19–26.

Grimm V (1994). *Stabilitätskonzepte in der Ökologie: Terminologie, Anwendbarkeit und Bedeutung für die Ökologische Modellierung*. Dissertation, Philipps-Universität Marburg.

Grimm V, Frank K, Jeltsch F, Brandl R, Uchmanski J, Wissel C (1996). "Pattern-Oriented Modelling in Population Ecology." *Science of the Total Environment*, **183**(1-2), 151–166.

Gusfield D (1997). *Algorithms on Strings, Trees and Sequences*. Cambridge University Press. ISBN 0-521-58519-8.

Hirschberg DS (1977). "Algorithms for the Longest Common Subsequence Problem." *Journal of the Association for Computing Machinery*, **24**(4), 664–675. ISSN 0004-5411. doi:10.1145/322033.322044.

Höppner F (2001). "Discovery of Temporal Patterns – Learning Rules About the Qualitative Behaviour of Time Series." In "PKDD: Proceedings of the 5. European Conference on Principles and Practice of Knowledge Discovery in Databases," pp. 192–203. Springer-Verlag.

Höppner F (2002a). "Learning Dependencies in Multivariate Time Series." In "Proceedings of the ECAI'02 Workshop on Knowledge Discovery in (Spatio-)Temporal Data (Lyon, France)," pp. 25–31.

Höppner F (2002b). "Lernen lokaler Zusammenhänge in multivariaten Zeitreihen." In "5. Göttinger Symposium Soft Computing," pp. 113–125.

Höppner F, Klawonn F (2001). "Finding Informative Rules in Interval Sequences." In "Advances in Intelligent Data Analysis. Proceedings of the 4. International Symposium," volume 2189, pp. 123–132. Lecture Notes in Computer Sciences.

Konstantinov KB, Yoshida T (1992). "Real-Time Qualitative Analysis of the Temporal Shapes of (Bio)process Variables." *American Institue of Chemical Engineers Journal*, **38**(11), 1703–1715.

Leggett R, Williams LR (1981). "A Reliability Index for Models." *Ecological Modelling*, **13**, 303 – 312.

Levin SA (1992). "The Problem of Pattern and Scale in Ecology." *Ecology*, **73**(6), 1943–1967.

Marron JS, Tsybakov AB (1995). "Visual Error Criteria for Qualitative Smoothing." *Journal of the American Statistical Association*, **90**(430), 499–507.

Mayer DG, Butler DG (1993). "Statistical Validation." *Ecological Modelling*, **68**, 21–32.

McCuen RH, Knight Z (2006). "Evaluation of the Nash-Sutcliffe Efficiency Index." *Journal of Hydrologic Engineering*, **11**(6), 597–602.

Nakatsu N, Kambayashi Y, Yajima S (1982). "A Longest Common Subsequence Algorithm Suitable for Similar Text Strings." *Acta Informatica*, **V18**(2), 171–179. doi:10.1007/BF00264437.

Nash JE, Sutcliffe V (1970). "River Flow Forecasting Through Conceptual Models, I. A Discussion of Principles." *Journal of Hydrology*, **10**, 282–290.

Paterson M, Dančík V (1994). "Longest Common Subsequences." *Mathematical Foundations of Computer Science*, **841**, 127–142.

Pawlowsky-Glahn V (2003). "Statistical Modelling in Coordinates." In S Thio-Henestrosa, JA Martin-Fernandez (eds.), "Compositional Data Analysis Workshop," Universitat de Girona. ISBN 84-8458-111-X.

Pawlowsky-Glahn V, Egozcue JJ (2001). "Geometric Approach to Statistical Analysis on the Simplex." *Stochastic Environmental Research and Risk Assessment (SERRA)*, **15**(5), 384–398.

Pawlowsky-Glahn V, Egozcue JJ, Burger H (2003). "An Alternative Model for the Statistical Analysis of Bivariate Positive Measurements." In J Cubitt (ed.), "CD Proceedings of IAMG'03," University of Portsmouth, Portsmouth (UK). ISBN 0-9734220-0-9.

Pawlowsky-Glahn V, Mateu-Figueras G (2005). "The Statistical Analysis on Coordinates in Constraint Sample Spaces." In "Abstract book: 55th session of the International Statistical Institute (ISI)," International Statistical Institute, Sydney Convention & Exhibition Centre, Sydney, Australia. ISBN 1-877040-28-2.

Petzoldt T, Uhlmann D (2006). "Nitrogen Emissions Into Freshwater Ecosystems: is There a Need for Nitrate Elimination in All Wastewater Treatment Plants?" *Acta Hydrochimica et Hydrobiologica*, **34**, 305–324. doi:10.1002/aheh.200500638.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reckhow KH (1994). "Water Quality Simulation Modeling and Uncertainty Analysis for Risk Assessment and Decision Making." *Ecological Modelling*, **72**(1/2), 1–20.

Ruppert D, Sheather SJ, Wand MP (1995). "An Effective Bandwidth Selector for Local Least Squares Regression." *Journal of the American Statistical Association*, **90**, 1257–1270.

Rykiel EJ (1996). "Testing Ecological Models: the Meaning of Validation." *Ecological Modelling*, **90**(3), 229–244.

Sargent RG (1998). "Verification and Validation of Simulation Models." In "WSC '98: Proceedings of the 30. conference on Winter Simulation (Washington, D.C., United States)," pp. 121–130. IEEE Computer Society Press. ISBN 0-7803-5134-7.

Sargent RG (2003). "Verification and Validation: Verification and Validation of Simulation Models." In "WSC '03: Proceedings of the 35.Conference on Winter Simulation (New Orleans, Louisiana)," pp. 37–48. ISBN 0-7803-8132-7.

Schlesinger S, Crosbie RE, Gagné RE, Innis GS, Lalwani CS, Loch J, Sylvester RJ, Wright RD, Kheir N, Bartos D (1979). "Terminology for Model Credibility." *Simulation*, **34**(3), 103 – 104.

Tolosana-Delgado R, Pawlosky-Glahn V (2007). "Kriging Regionalized Positive Variables Revisited: Sample Space and Scale Considerations." *Mathematical Geology*. (in press).

van den Boogaart KG, Tolosana-Delgado R (2007). "**compositions**: A Unified R Package to Analyze Compositional Data." *Computers and Geosciences*. In press.

Wagner RA, Fischer MJ (1974). "The String-to-String Correction Problem." *Journal of the ACM*, **21**, 168–173.

Wand M, Ripley BD (2006). "**KernSmooth**: Functions for Kernel Smoothing for Wand & Jones (1995)." S original by Matt Wand. R port by Brian D. Ripley, R package version 2.22-19, URL http://CRAN.R-project.org/.

Wiegand T, Jeltsch F, Hanski I, Grimm V (2003). "Using Pattern-Oriented Modeling for Revealing Hidden Information: A Key for Reconciling Ecological Theory and Application." *Oikos*, **100**(2), 209–222.

Affiliation:

Stefanie Jachner
Dept. Global Change and Natural Systems
Potsdam Institute for Climate Impact Research
P.O. Box 601203
D-14412 Potsdam, Germany
E-mail: jachner@pik-potsdam.de
URL: http://www.pik-potsdam.de/members/jachner/

K. Gerald van den Boogaart
Ernst-Moritz-Arndt-Universität Greifswald
Institut für Mathematik und Informatik
Jahnstr. 15a
D-17487 Greifswald, Germany
E-mail: boogaart@uni-greifswald.de
URL: http://www.math-inf.uni-greifswald.de/statistik/boogaart.html

Thomas Petzoldt
Institut für Hydrobiologie
Technische Universität Dresden
D-01062 Dresden, Germany
E-mail: thomas.petzoldt@tu-dresden.de
URL: http://tu-dresden.de/Members/thomas.petzoldt/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 22, Issue 8 *Submitted:* 2007-02-26
September 2007 *Accepted:* 2007-06-25