

NETLAKE toolbox for the analysis of high-frequency data from lakes



Factsheet #10

Pattern detection using Dynamic Factor Analysis (DFA)

Rosana Aguilera and Rafael Marcé

Objective

One of the main applications of time-series analysis is the identification of trends and cyclic patterns in the data. Many trend detection and frequency decomposition analyses already exist for those purposes, particularly to address single time-series. However, classical methodologies are not particularly well suited to cope with multivariate problems. Dynamic Factor Analysis (DFA) decomposes a collection of time-series into common patterns and associated error terms (Zuur *et al.* 2003a). Broadly speaking, this method resembles performing Principal Component Analysis (PCA) but it is specifically designed for time-series. The end-product is a collection of patterns shared by all time-series, the relative relevance of each pattern across time-series, and error terms.

DFA is a dimension-reduction method that estimates underlying common patterns in a set of time-series (Zuur *et al.* 2003a). An attractive feature of DFA is its ability to treat time-series that have been recorded irregularly over time, or have short duration. Moreover, DFA allows time-series to be short and thus the lack of sufficiently long records does not represent a problem (Zuur and Pierce 2004). The extracted patterns (e.g., cycles and/or trends) are associated to factor loadings, which indicate the weight that each pattern has for each monitoring point. These two end products, i.e., patterns and factor loadings, can be then analyzed in order to characterize the temporal and spatial variability of the extracted water quality signals. The resulting description of the extracted patterns thus facilitates the interpretation and the identification of potential drivers of change in the system.

Specific application

The main application is the **detection of hidden patterns** that are shared by sets of time-series. If the time-series belong to a network of monitoring points, the spatial dimension can also be considered by looking at the relevance of each extracted pattern at each particular point, based on the magnitude and sign of the associated so-called factor loadings.

The greatest advantage of this method is its **ability to cope with missing observations and uneven sampling resolution** in time-series.

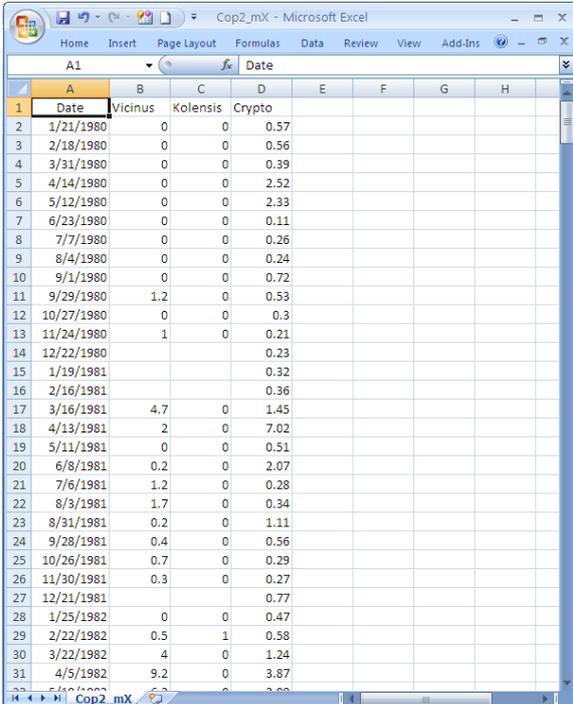
Background

The main tool is the **MARSS** (Multivariate Auto-regressive Space-State Model; Holmes *et al.* 2012) **R-Package**. A basic knowledge of 'R' would therefore be beneficial. Nevertheless, the MARSS manual and associated documents at the CRAN repository (<http://cran.r-project.org/web/packages/MARSS/index.html>) provide detailed information about setting up the DFA model.

It is also recommended to acquire some basic knowledge about time-series analysis (time-frequency domains, aliasing, autocorrelation function, etc.) before using DFA.

Type of data and requirements

The input files for DFA using the MARSS R-Package are .csv (Comma Separated Value) files with columns for each variable (an example is presented below). The data need not be standardized since the DFA script includes a previous data standardization step.



The screenshot shows a Microsoft Excel spreadsheet titled 'Cop2_mX'. The data is organized in columns: 'Date', 'Vicinus', 'Kolensis', and 'Crypto'. The rows represent time points from 1/21/1980 to 4/5/1982. The 'Date' column contains dates, while the other three columns contain numerical values representing the abundance of each variable.

Date	Vicinus	Kolensis	Crypto
1/21/1980	0	0	0.57
2/18/1980	0	0	0.56
3/31/1980	0	0	0.39
4/14/1980	0	0	2.52
5/12/1980	0	0	2.33
6/23/1980	0	0	0.11
7/7/1980	0	0	0.26
8/4/1980	0	0	0.24
9/1/1980	0	0	0.72
9/29/1980	1.2	0	0.53
10/27/1980	0	0	0.3
11/24/1980	1	0	0.21
12/22/1980			0.23
1/19/1981			0.32
2/16/1981			0.36
3/16/1981	4.7	0	1.45
4/13/1981	2	0	7.02
5/11/1981	0	0	0.51
6/8/1981	0.2	0	2.07
7/6/1981	1.2	0	0.28
8/3/1981	1.7	0	0.34
8/31/1981	0.2	0	1.11
9/28/1981	0.4	0	0.56
10/26/1981	0.7	0	0.29
11/30/1981	0.3	0	0.27
12/21/1981			0.77
1/25/1982	0	0	0.47
2/22/1982	0.5	1	0.58
3/22/1982	4	0	1.24
4/5/1982	9.2	0	3.87

Figure 1. Example of input data for DFA analysis

In this case (Figure 1), three time-series are being considered in the analysis: the abundances of two copepods species (*Cyclops vicinus* and *Cyclops kolensis*) as well as the abundance of cryptophytes. The Date column is included here for posterior reference but it is omitted in the analysis. Of course, the columns may represent other arrangements (a variable in different locations, systems, etc).

The data need to be **evenly spaced**; i.e., the user must decide on a time-step if the data are unevenly sampled and adjust the observations to a specific resolution (e.g., monthly, daily, etc.). However, the analysis accepts missing values.

Basic procedures

1. Preparation of input file as indicated above.
2. Read MARSS documentation to understand the basic procedures and the different options of the analysis. Key model parameters are:
 - Number of common patterns (m) to be tested. The analysis does not automatically find the most efficient model in terms of the number of patterns to be extracted. So a trial-and-error procedure may be useful at this point, using a model selection criteria like Akaike Information Criterion (AIC), already included in the MARSS package.
 - Structure of error variance-covariance matrix (R). This tries to account for measurement errors and their covariance structure. Trial-and-error may also be needed here.

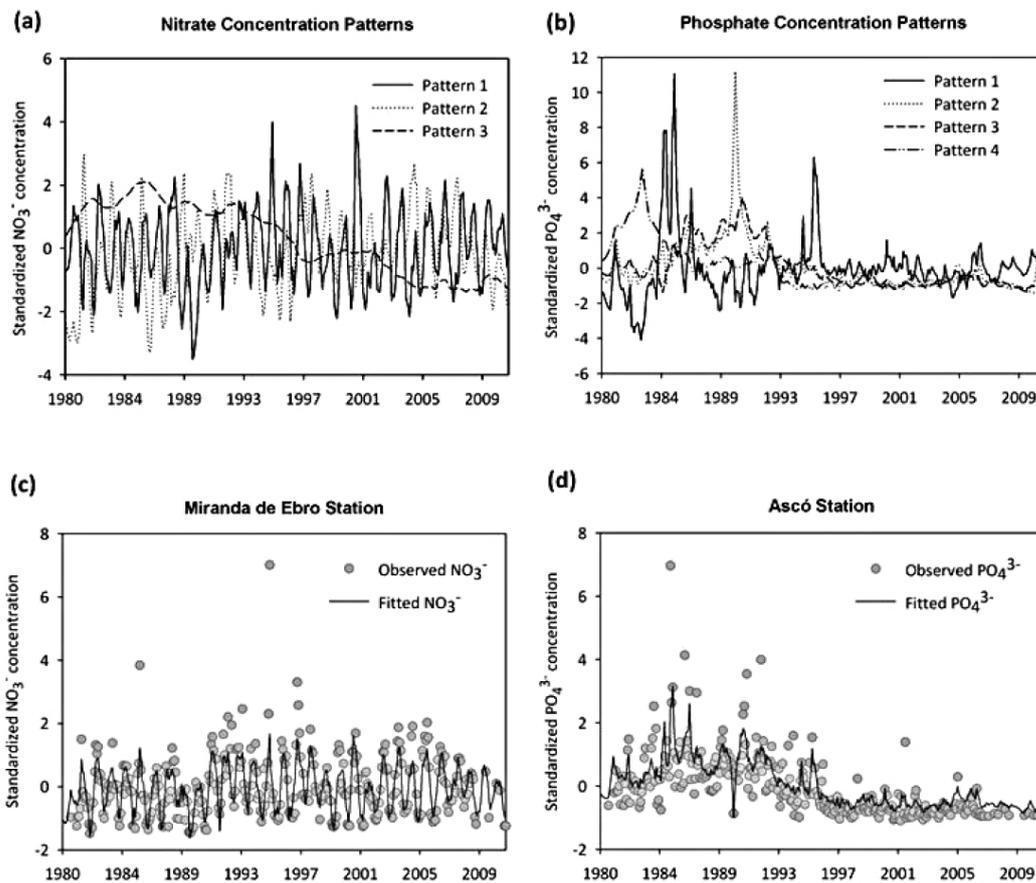


Figure 2. Patterns extracted from 50 time-series of nitrate (left) and phosphate (right) in different locations along the Ebro River basin, and the fit for two selected sampling points. The modelled lines in C and D are a linear combination of the patterns in A and B.

Pitfalls and tips

A visual assessment of the fit of the DFA model against the observed time-series may be useful to decide if the DFA analysis performed well (Figure 2). However, the absence of fit for a

variable or sampling location does not mean that the analysis is not working in that case, it may simply imply that there are no common patterns detected for that particular site or variable.

Depending on the number of time-series involved, as well as the length of data series, DFA can be time consuming and computationally demanding. Consider the use of a High Performance Computer if you are working with long series from many sites. At least you can run in parallel all trial-and-error runs related to parameter selection, which are independent.

Further reading

Key References:

Holmes, E.E., Ward, E., Wills, K. 2012. MARSS: Multivariate Autoregressive State-space Models for analyzing Time-series Data. *The R Journal* 4: 11-19.

Zuur, A.F., Fryer, R.J., Jolliffe, I.T., Dekker, R., Beukema, J.J. 2003a. Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics* 14: 665-685.

Other useful references:

Aguilera, R., Marcé, R., Sabater, S. 2015. Detection and attribution of global change effects on river nutrient dynamics in a large Mediterranean basin. *Biogeosciences* 12: 4085–4098.

Holmes, E.E. 2013. Derivation of the EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models. Technical Report. *arXiv preprint arXiv: 1302.3919*.

Zuur, A.F., Ieno, E.N., Smith, G.M. 2007. *Analysing ecological data*. Springer. New York.

Zuur, A.F., Pierce, G.J. 2004. Common trends in Northeast Atlantic Squid time series. *Journal of Sea Research* 52: 57-72.

Zuur, A.F., Tuck, I.D., Bailey, N. 2003b. Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences* 60: 542-552.

Code

The code for this technique was written in the 'R' language and is available in the MARSS package. For a complete application of the technique including several sampling locations you may use the 'R' code by Rosana Aguilera and included in her PhD project (contact her for a copy of the R-codes).

Contact details

Rosana Aguilera. Catalan Institute of Water Research, Girona, Spain.
r.aguilerabecker@gmail.com

Rafael Marcé. Catalan Institute of Water Research, Girona, Spain.
rmarce@icra.cat

Suggested citation

Aguilera, R. and Marcé, R. 2016. Pattern detection using Dynamic Factor Analysis (DFA). In Obrador, B., Jones, I.D. and Jennings, E. (Eds.) *NETLAKE toolbox for the analysis of high-frequency data from lakes* (Factsheet 10). Technical report. NETLAKE COST Action ES1201. pp. 52-56. <http://eprints.dkit.ie/id/eprint/541>.

Acknowledgement

This factsheet is based upon work from the NETLAKE COST Action ES1201, supported by COST (European Cooperation in Science and Technology).