SoftwareX 21 (2023) 101309

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

trec: An R package for trend estimation and classification to support integrated ecosystem assessment of the marine ecosystem and environmental factors

Hiroko Kato Solvang [a,*], Mineaki Ohishi [b]

[a] Institute of Marine Research, Nordnes, N-5817, Bergen, Norway
[b] Tohoku University, Aoba-ku, Sendai, Japan

A R T I C L E I N F O

Article history:
Received 12 October 2022
Received in revised form 9 January 2023
Accepted 9 January 2023

Keywords:
Integrated trend assessment
Common trend
Stakeholders
Multinomial logistic discriminator

A B S T R A C T

Solvang and Planque [ICES Journal of Marine Science, 77, pp.2529–2540, (2020)] provided a trend estimation and classification (TREC) approach to estimating dominant common trends among multivariate time series observations. This approach was developed to improve communication among stakeholders like marine managers, industry representatives, non-governmental organizations, and governmental agencies as they investigate the common tendencies between a biological community in a marine ecosystem and the local environmental factors. The tasks of trend estimation and classification in the original computational procedure have been revised, and new features include an automatic *icon* assignment algorithm using a multinomial logistic discriminator. In this paper, we present R package trec. Implementation of this package involves three partitions corresponding to TREC1) estimating trends from observed time series data; TREC2) classifying two/three rough patterns; and TREC3) generating a table summarizing categories of common configurations (trends) and the automatic *icon* assignments to them. The proposed **trec** focuses on investigating mean non-stationary long-term trends of data, and it works for any length of time steps. It is not necessary to apply a stationary Gaussian assumption to the estimated trends to investigate the common trends, which are interpreted as common variations of biological and environmental data.

Code metadata

Current code version	0.1.0
Permanent link to code/repository used for this code version	https://github.com/ohishim/trec/
Permanent link to reproducible capsule	https://github.com/ohishim/trec/
Legal code license	MIT License
Code versioning system used	
Software code languages, tools and services used	R(>=2.10)
Compilation requirements, operating environments and dependencies	
If available, link to developer documentation/manual	https://github.com/ohishim/trec/ https://arxiv.org/abs/2209.06619
Support email for questions	mineaki.ohishi.a4@tohoku.ac.jp

1. Motivation and significance

Integrated ecosystem assessment (IEA) is one approach to organizing scientific information at multiple scales and across sectors to support ecosystem-based fisheries management (EBFM) [1]. IEA results can reflect various aspects of an ecosystem beyond dynamics, status, and future risks. In particular, identifying common trends can be useful as a diagnostic tool to reveal past changes and to explore the relationships among biological communities, as well as between these communities and environmental conditions. For such investigation, trend estimation and classification (TREC) was proposed in [2]. The entire calculation procedure was originally implemented using MATLAB [3]. The method has been applied for annual trend analyses in ICES

* Corresponding author.
E-mail addresses: hirokos@hi.no (Hiroko Kato Solvang), mineaki.ohishi.a4@tohoku.ac.jp (Mineaki Ohishi).

https://doi.org/10.1016/j.softx.2023.101309
2352-7110/

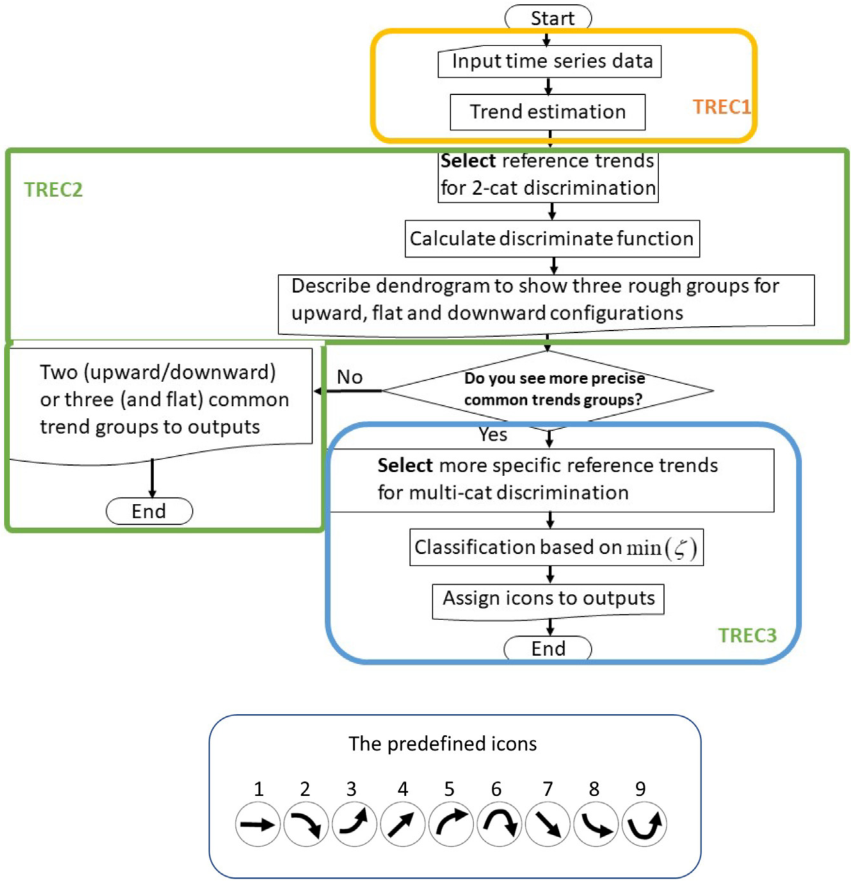

Fig. 1. The entire numerical procedure and the *icons* we predefined for communication with stakeholders.

working groups, and several ICES experts have asked us to package TREC in R language. In this article we present **trec** implemented in R software motivated by their request.

TREC involves tasks for trend estimation of the observed time series data and for classification of the estimated trends into common trend groups. In the trend estimation task, the original TREC used two different kinds of parametric trend models: a polynomial regression model and a stochastic difference equation model. The trend estimated by the polynomial regression model presents a simpler trend configuration than that estimated by the stochastic trend model using a difference equation model. [2] reported that the simpler patterns could be more robustly classified to common trend pattern groups than when using variable trend patterns by the stochastic difference equation model. Therefore, **trec** in R adopts the usage of the simple polynomial regression model, which aims to achieve common trend classification.

In the trend classification task, two-category discrimination is first applied to roughly divide trends into three groups representing configurations for upward, flat, and downward. The comparison made to find the dis-similarity between trends is conducted for time steps of the same length, not targeting trends of different lengths as considered in Dynamic time warping [4]. In the original procedure, a two-categorical discriminant function is calculated, and hierarchical clustering is applied by the discriminant function according to [5]. The **trec** in R includes a procedure to check the outputs by the discriminant function with/without hierarchical clustering.

If it is necessary to classify trends into groups for more concrete common patterns, multiple-category discrimination is secondarily applied to the target trends, which the users can define in the estimated trends. These common patterns may be interpreted as latent processes, which correspond to latent factors identified by smoothed dynamic factor analysis [6]. Our assumption of the common trend pattern is basically very simple, such as upward, flat, or downward, and is more general, not only a Gaussian stationary process but also a mean-nonstationary process. Our approach does not consider the functional relationship between predictor and response variables or a combination of basis trend patterns with an estimated trend for a single observation, which were provided in [7]. Finally, each target trend is assigned as a predefined reference, called an icon, which is set as an easily accessible form that can be used to serve the needs among stakeholders. In the original MATLAB code provided by [2], the assignment to the multiple-category groups was done manually by the user. The **trec** in R has been developed to automatically assign icons with a multinomial logistic regression model, which is an extended model for handling a response variable having more than two categories. For a random variable taking one of multiple categories, a multinomial logistic regression is used to estimate probabilities occurring in each category based on a multinomial distribution and outputs the category with the highest probability. Such a multinomial logistic regression model can be applied to multiple-discriminant analysis (e.g. [8–10]). In the **trec**, a multinomial logistic discriminator assigns icons from the estimated trends by estimating the probabilities that assign each icon.

The **trec** in R is interactively implemented by the user in three procedures: TREC1, TREC2, and TREC3. In this article, we first introduce the methodological background in the next section and then give practical instructions for the **trec** package in Section 3.

Table 1
Functions that users can implement in **trec**.

Function	Objective	Implementation by users
TREC1	Trend estimation	Plot of original data Plot of each estimated trend Plot of all estimated trends
TREC2	Common trend classification	Set two target trends (default automatically sets two artificial target trends) Plot of two/three grouped trends based on output by clustering Plot of dendrogram obtained by clustering Plot of two/three grouped trends based on discriminant function
TREC3	Common trend classification	Define multi-group classification Show summary table for common trends and *icon* assignment

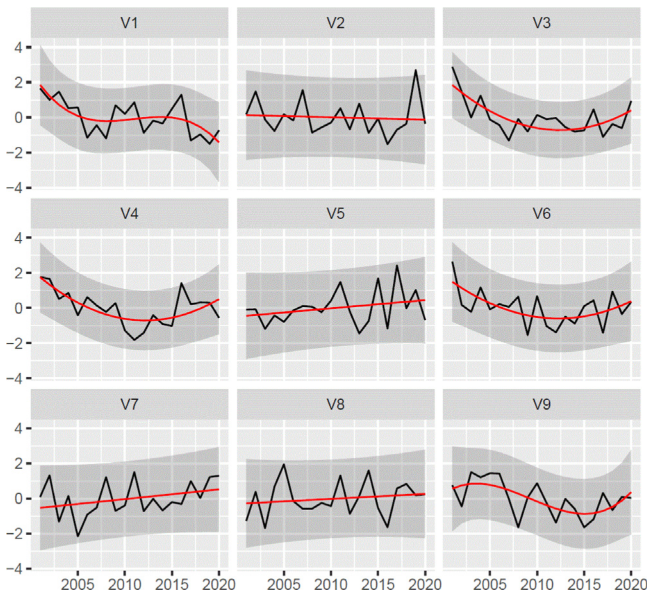

Fig. 2. The estimated trend with 95% prediction bands.

2. Software description

2.1. Software architecture

TREC involves two statistical procedures, trend estimation by polynomial trend model and discriminant analysis for classifying common trend classes. The methodological details are summarized in Supplementary file. The entire numerical procedure and the *icons* we predefined for communication with stakeholders are summarized in Fig. 1. The **trec** package is implemented in R. It is available in a GitHub repository and can be installed using the following command:

```
install.packages("devtools")
devtools::install_github("ohishim/trec")
```

2.2. Software functionalities

Time series datasets containing m variables and N time steps should be prepared. As mentioned above, **trec** involves three estimation procedures, which are also conducted interactively for users. The relevant functions and objectives are summarized in Table 1.

3. Illustrative examples

TREC1 We apply an example dataset called *exData* in the package, including $m = 9$ variables with $N = 20$ time steps. Here, *exData* takes the form of a "data.frame" object with 20 rows and 10 columns. The first column indicates 'year' corresponding to time points, and it is used for the axis label when describing plots for the original data or trend estimation. An argument of TREC1 includes a data.frame object with N rows and m columns, like *exData*. TREC1 estimates trends for each variable, where each variable is standardized and estimated. TREC1 can be executed as follows:

```
res1 <- TREC1(exData)
```

In TREC1, the variable names of *exData* are automatically represented by V1, $\cdots$, V9. The relationship between the original and the represented names is automatically output on an R console. If variables are removed for estimation, e.g. there are many missing values at certain time points, they are also displayed under the output:

```
The following variable(s) is/are removed:
```

The missing values between time points (e.g., years) are handled using the following rules:

1. The following variables (corresponding to columns of input data) are removed from the TREC process:

- All time points are missing.
- Either or both of the start and end points are missing.
- The number of missing values is greater than *remove.num*, where *remove.num* is a value given by the user.

2. Missing values for the remaining variables are filled by a linear interpolation via the R function **approx** of the R package **stats** [11] for each variable.

Table 2
Output of TREC1.

Output	Content
fig. RawData	Figure of each variable
fig. StdData fig.ctrend	Figure of each standardized variable Figure of each standardized variable with estimated trend
fig.trend	Figure of all estimated trends
argTREC	List required in later steps TR: trend estimates ggD3: a form to draw figures Y: standardized data dim: indicator for dimension of polynomial equation coef: coefficients of polynomial equation
remove	Removed variables
Vnames	Original and represented variable names

The output of TREC1 takes the form of a list object and is summarized in Table 2.

TREC1 can output the plots for original data, standardized data, estimated trend (one figure by gathering all trends), and estimated trend with 95% prediction bands (Fig. 2) as follows:

```
res1$fig. RawData
res1$fig. StdData
res1$fig.trend
```

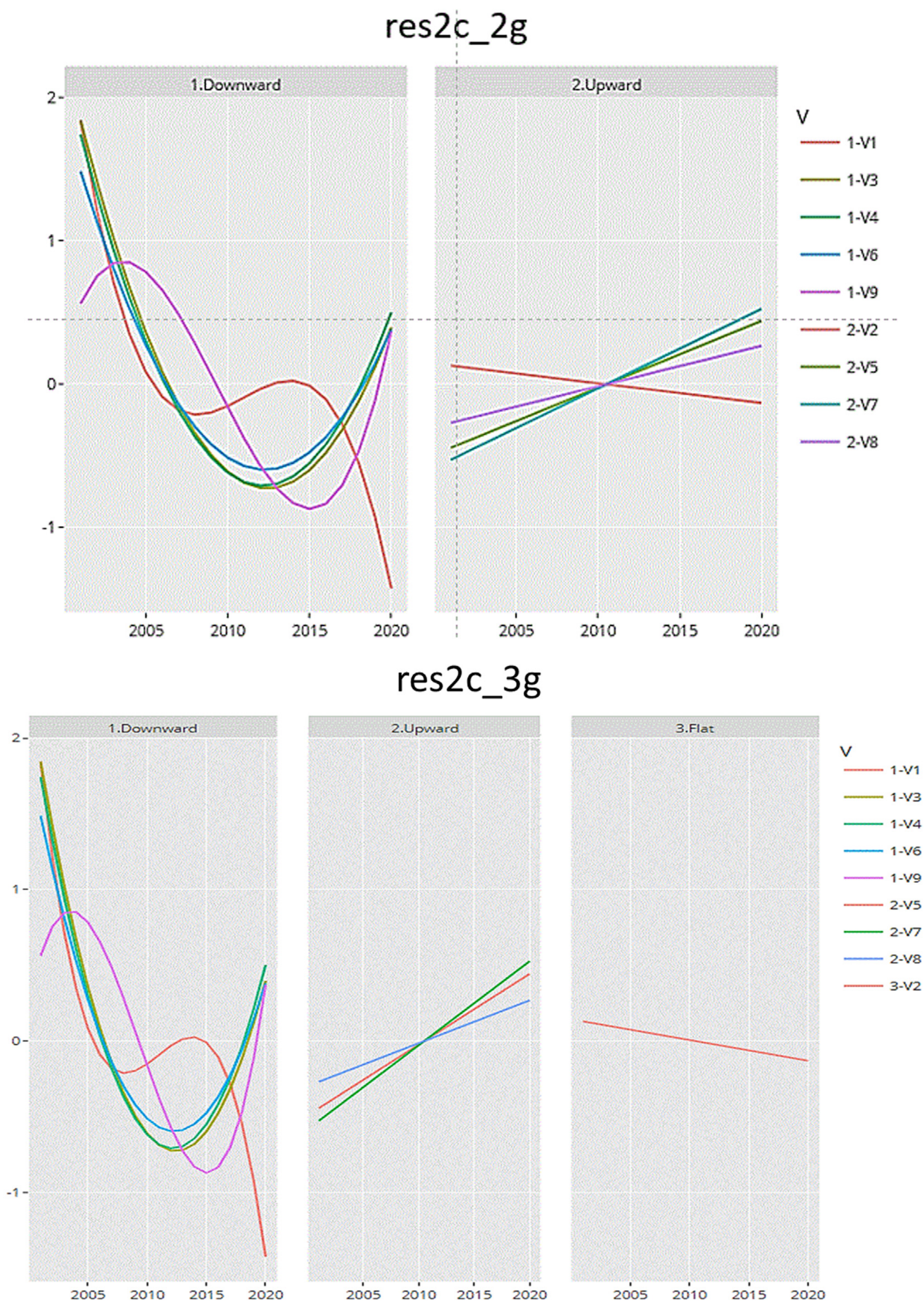

Fig. 3. Trend groups obtained using two- or three-group detection by clustering based on two-category discrimination.

and

`res1$fig.ctrend` .

If $m_0 > 16$, *fig.ctrend* takes the form of a list object, where $m_0 (\leq m)$ is the number of variables subjected to estimation. Hence, the above code is replaced by

`plot(res1$fig.ctrend[1]); plot(res1$fig.ctrend[2]); ...`

argTREC takes the form of a list object, which includes *TR*, *ggD3*, *Y*, *dim*, and *coef*. TR takes the form of a matrix object with N rows and m_0 columns, which includes prediction by the trend model for each variable. *ggD3* takes the form of a data.frame object with m_0N rows and 4 columns, which is used to draw figures in later steps. *Y* takes the form of a matrix object with N rows and m_0 columns, which has standardized

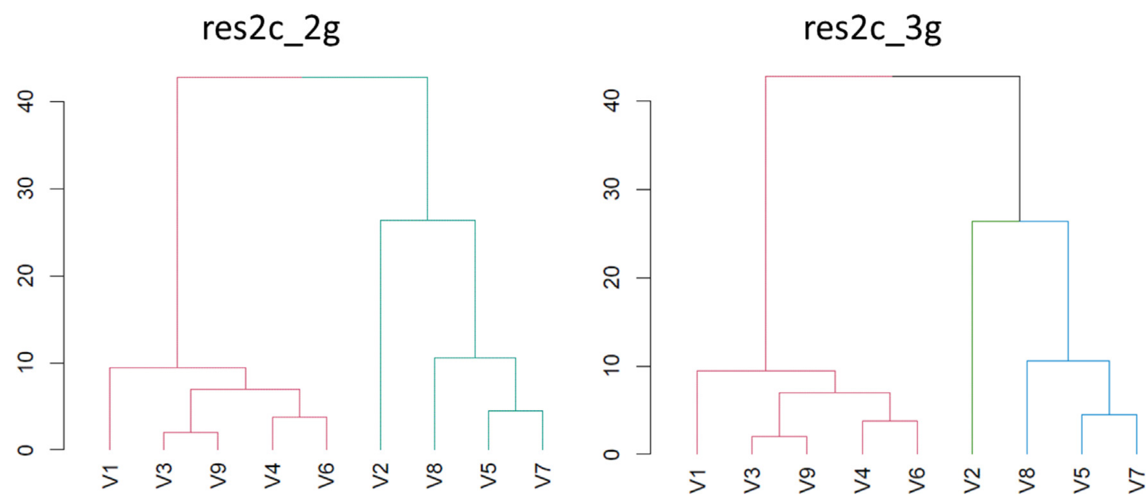

Fig. 4. Dendrograms obtained by clustering with centroid link and distance obtained by the discriminant function.

data for each variable. *dim* and *coef* take the form of matrix objects with m_0 rows and 2 and 4 columns, which are defined by $dim: \begin{pmatrix} d_{1,1} d_{1,2} \\ \vdots \quad \vdots \\ d_{m_0,1} d_{m_0,2} \end{pmatrix}$, $coef: \begin{pmatrix} \hat{\gamma}_{1,0} & \cdots & \hat{\gamma}_{1,3} \\ & \vdots & \\ \hat{\gamma}_{m_0,0} & \cdots & \hat{\gamma}_{m_0,3} \end{pmatrix}$.

TREC2 performs rough classification of the estimated trends into two (upward and downward) or three (upward, flat, and downward) groups based on the two target trends. In the default setting, these two trends are fixed as linear trends. Using the output for *argTREC*, TREC2 can be executed as follows:

```
argTREC <- res1$argTREC
res2 <- TREC2(argTREC)
```

If users want to select target trends from the estimated trends, they can input two variables in pvar as

```
res2 <- TREC2(argTREC, pvar=c("V2", "V7"))
```

Moreover, two-category discrimination can be expanded to three-category discrimination by the following option:

```
res2 <- TREC2(argTREC, groups=3)
```

In [2], hierarchical clustering was applied using the two-category discriminant function as a distance measurement. Two or three groups were detected by the dendrogram with a centroid link of clusters. This flow was obtained according to the method in [5], and TREC2 calculated a two-category discriminant function and made a dendrogram based on the following function:

```
res2c_2 g <- TREC2(res1$argTREC) # two groups detection by clustering
# based on two-category discriminant function
res2c_2g$dend() # show the dendrogram
res2c_3 g <- TREC2(res1$argTREC, group = 3)# three groups detection by
# clustering based on two-category discriminant function
res2c_3g$dend() # show the dendrogram
```

The trend groups detected by clustering based on two-category discrimination are presented in Fig. 3. The dendrograms corresponding to these groups are shown in Fig. 4. The panel on the right-hand side of Fig. 4 shows that one more group is divided into upward and flat trends in the estimated linear trends (V2, 8, 5 and 7). The difference between two and three categories is whether the estimated trend for V2 belongs to upward or flat. As seen in Fig. 5, if it is meaningful for a user to detect the trend of V2 as flat compared with the other linear trends, it should belong to the flat group. Alternatively, we can also detect roughly common trend groups without clustering as follows:

```
res2d <- TREC2(res1$argTREC, clustering=FALSE)
```

The output for trend groups is shown in Fig. 5. This procedure indicates that variable 2 belonged to the downward group. Even if we perform the command

```
res2d_3 g <- TREC2(res1$argTREC, clustering=FALSE, group=3),
```

TREC2 returns the following message to us:

```
The following group(s) is/are not applicable:
[1] "Flat",
```

which means that three-category discrimination is not applicable in this case. As seen in Fig. 8, the trend of V2 belonging to the downward group may be useful when the intended classification is limited to either upward or downward discrimination. On the other hand, the trend of V2 could also lead to belonging to the flat group as shown above. The **trec** in R includes two options, detecting groups by clustering based on two-category discrimination and by discrimination alone. The final decision for the outputs obtained by those options would depend on the aim of the user. Comparing detected common trend groups with the estimated trend pattern summarized in Fig. 2 is an important step in interpreting the best grouping of common trends for the user.

TREC3 performs multi-category discrimination to classify more common trend groups based on target trends selected by the user, and it assigns *icons* to common trend groups. If we set the target trends for V1, 6, and 9 to downward, V8 to upward, and V2 to flat, the following setting is required:

```
tvard <- list(
        Downward = paste0("V", c(1,6,9)),
        Upward = paste0("V", c(8)),
        Flat = paste0("V", c(2))
        )
```

Using this, TREC3 can be executed as follows:

```
res3_2c3 g <- TREC3(tvard, res2c_3g$trn, res1$arg TREC)
```

TREC3 outputs a table of assigned *icons* and variable names for each group (Fig. 6). The output by TREC3 can show trend groups for downward, upward, and flat separately as summarized in Table 3.

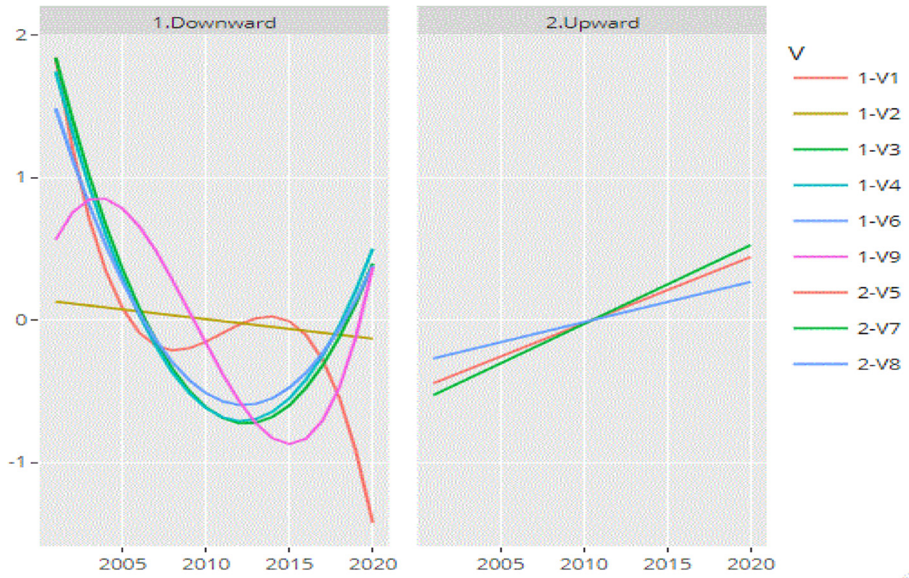

Fig. 5. Trend groups obtained by two-category discriminant analysis without clustering.

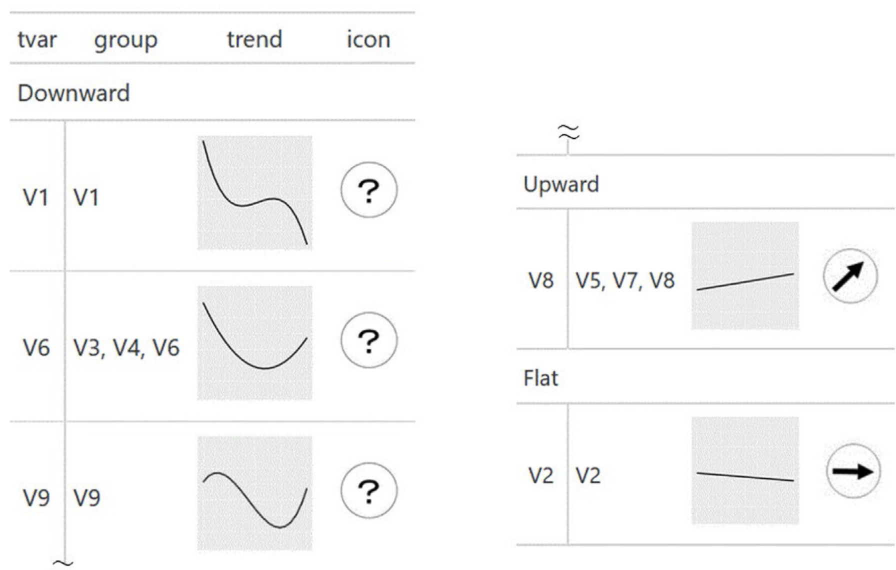

Fig. 6. Outputs by TREC3.

Table 3
Contents of the output obtained by TREC3.

Output	Content
fig.down	Trend groups based on selected target trends in downward
fig.up	Trend groups based on selected target trends in upward
fig.flat	Trend groups based on selected target trends in flat
fig.icon	Figure of assigned *icons*

The *icons* are automatically assigned based on the information by learning many trend patterns. For the predefined *icons* in Fig. 1, the **trec** in R was set with the rules that *icons* 2, 7, and 8 be assigned to the downward group, *icons* 3, 4, and 5 be assigned to the upward group, and *icons* 1, 6, and 9 be assigned to the flat group. For example, the target trend of V6 was interpreted by the **trec** in R as any *icon* being available for assignment. However, if we review the trends of V3, 4, and 6 in Fig. 2, the patterns may be recognized as *icon* 8. Therefore, the output for *icon* assignment should be confirmed by comparison with the estimated trend according to the aim of the user.

4. Impact

The original idea in **trec** has been applied for annual trend analyses in ICES working groups in the Barents Sea, the Norwegian Sea, and the North Sea, including abiotic, biotic, and human impact data. The ICES (International Council for the Exploration) Scientific Report [12] that validated several trend analyses through benchmarking practice recommended to apply the approach as a practical method, which assumption of Gaussian or identical distribution is not necessary for the data. The **trec** focuses on the analysis for long-term trends of data, and it works for any length of time steps in high-dimensional variables. Based on classified trend groups by **trec**, communication among stakeholders can be enhanced by showing the common tendency between a biological community in a marine ecosystem and the environmental factors as well as the icons allocated. The performance could be also expected in any scientific field where needs long-term trend estimation and common trend classification for their multivariate time series data set. This package could become a useful tool for IEA and studies conducted to investigate precise ecosystem functions using this package are expected as further extensions of it.

5. Conclusions

We presented the R package **trec** for analyzing common trends in a marine ecosystem by using statistical trend estimation and classification. The original idea of TREC has been applied annually to the ICES integrated assessment working groups since 2019. Developing the R package **trec** is expected to not only contribute to integrated ecosystem assessment [13,14] but also to help research professionals to more widely and practically conduct trend analysis for any scientific and social field.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by JSPS Bilateral Program, Japan Grant Number JPJSBP120219927 and the second author's research was partially supported by JSPS, Japan Grants-in-Aid for Scientific Program Number JP20H04151 and JP21K13834.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.softx.2023.101309.

References

[1] Levin PS, Fogarty MK, Murawski SA, Fluharty D. Integrated ecosystem assessments: developing the scientific basis for ecosystem-based management of the ocean. PLoS Biol 2009;7(e14):299–314.
[2] Solvang HK, Planque B. Estimation and classification of temporal trends to support integrated ecosystem assessment. ICES JMS 2020;77(7–8):2529–40. http://dx.doi.org/10.1093/icesjms/fsaa111.
[3] Matlab. version.9.5 (R2018b). Natick, Massachusetts: The MathWorks Inc.
[4] Giorgino T. Computing and visualizing dynamic time warping alignments in R: The dtw package. J Stat Soft 2009;31(7). http://dx.doi.org/10.18637/jss.v031.i07.
[5] Solvang HK, Taniguchi M, Nakatani T, Amano S. Classification and similarity analysis of fundamental frequency patterns in infant spoken language acquisition. Stat L Meth 2008;5(5):187–208. http://dx.doi.org/10.1016/j.stamet.2007.08.005.
[6] Ward EJ, Anderson SC, Hunsicker ME, Litzow MA. Smoothed dynamic factor analysis for identifying trends in multivariate time series. Meth Ecol Evol 13:908–18. http://dx.doi.org/10.1111/2041-210X.13788.
[7] Pedersen E, Miller DL, Simpson GL, Ross N. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7:e6876. http://dx.doi.org/10.7717/peerj.6876.
[8] Jeune W, Francelino M, de Souza E, Filho E, Rocha C. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. Rev Bras de Ciên Do Solo 2018;42:e0170133. http://dx.doi.org/10.1590/18069657rbcs20170133.
[9] Mirzaei S. Hyperspectral image classification using non-negative tensor factorization and multinomial logistic regression. J Appl Rem Sen 2019;13(2):1–18. http://dx.doi.org/10.1117/1.JRS.13.026501.
[10] Amri M, Sumertajaya I, Syafitri U. Comparison of multinomial logistic discriminant analysis (mlgda) and classification and regression tree (cart) performance in classifying the impact of working children. J Phys Conf Ser 2020;1490(1):012033. http://dx.doi.org/10.1088/1742-6596/1490/1/012033.
[11] Becker RA, Chambers JM, Wilks AR. The new S language: A programming environment for data analysis and graphics. In: Wadsworth brooks/cole computer science series. 1988.
[12] WKINTRA-3. Third workshop on integrated trend analysis to support integrated ecosystem assessment. ICES Sci Rep 2022;4(32).
[13] ICES. Report of the workshop on integrated trend analyses in support to integrated ecosystem assessment (WKINTRA). ICES Doc CM 2018;15:23.
[14] Otto S. Integrated Ecosystem Assessment Tools. Hamburg, Germany: ICES WKINTRA; 2022, URL https://gitlab.rrz.uni-hamburg.de/saskiaotto/IEAtools.