

ANOVATS: A SUBSAMPLING-BASED TEST TO DETECT DIFFERENCES AMONG SHORT TIME SERIES IN MARINE STUDIES

BY YUICHI GOTO^{1,a}, HIROKO KATO SOLVANG^{2,b},
MASANOBU TANIGUCHI^{3,c}, AND TONE FALKENHAUG^{4,d}

¹*Faculty of Mathematics, Kyushu University, yuichi.goto@math.kyushu-u.ac.jp*

²*Marine Mammals Research Group, Institute of Marine Research, hiroko.solvang@hi.no*

³*Faculty of Science and Engineering, Waseda University, taniguchi@waseda.jp*

⁴*Plankton Research Group, Institute of Marine Research, Tone.Falkenhaus@hi.no*

Assessing marine ecosystems is important for understanding the impacts of climate change and human activity, as well as for maintaining healthy oceans and ecosystems. In marine science, it is common for biologists and geologists to identify regional differences based on expert knowledge, frequently through data visualization. However, time series data collected through surveys in marine studies typically span only a few decades, limiting the applicability of classical time series methods. Additionally, without expert knowledge, detecting significant differences becomes challenging. To address these issues, we introduce ANOVATS (ANOVA for small-sample time series data), a subsampling-based method to detect regional differences in small-sample time series data with a fixed number of groups. This method bypasses the need for spectral density estimation, which requires a large number of time points in the data. Furthermore, after detecting differences in homogeneity across all areas using the ANOVATS procedure, we devised a simple ANOVATS post hoc procedure to group the areas. Finally, we demonstrate the effectiveness of our method by analyzing zooplankton biomass data collected in different strata of the North Sea, showing its ability to quantify differences in species between geographical areas without relying on prior biological or geographical knowledge.

1. Introduction. Recent increases in the effects of climate change, together with other human factors and environmental pressures, are accelerating changes in marine ecosystems and threatening limited natural resources. For example, changes in commercial fish stocks in the oceans are no longer attributable solely to catch levels; it has become necessary to consider various biological interactions (e.g. food-web) within the marine ecosystem and the environmental factors that alter them. To address these issues, the Integrated Ecosystem Assessment (IEA) (Levin, Murawski and Fluharty 2009) was proposed as a scientific basis for ecosystem-based fishery management by National Oceanic and Atmospheric Administration (NOAA), The United Nations Educational, Scientific and Cultural Organization (UNESCO), and the International Council for the Exploration of the Sea (ICES), which consider marine resources on an international scale. The IEA approach analyzes and synthesizes information on a wide range of ecosystem components and pressures, including natural and social scientists, stakeholders, and resource managers (Clay et al. 2023). Based on the results analyzed, the IEA identifies the status, changes, relationships, and processes at the ecosystem level (ICES 2018), assesses risks, and ultimately evaluates strategies to implement ecosystem-based management measures (Clay et al. 2023). In particular, a working group established by scientific experts from the International Council for the Exploration of

Keywords and phrases: clustering, climate change, homogeneity test, marine ecosystem, time series analysis.

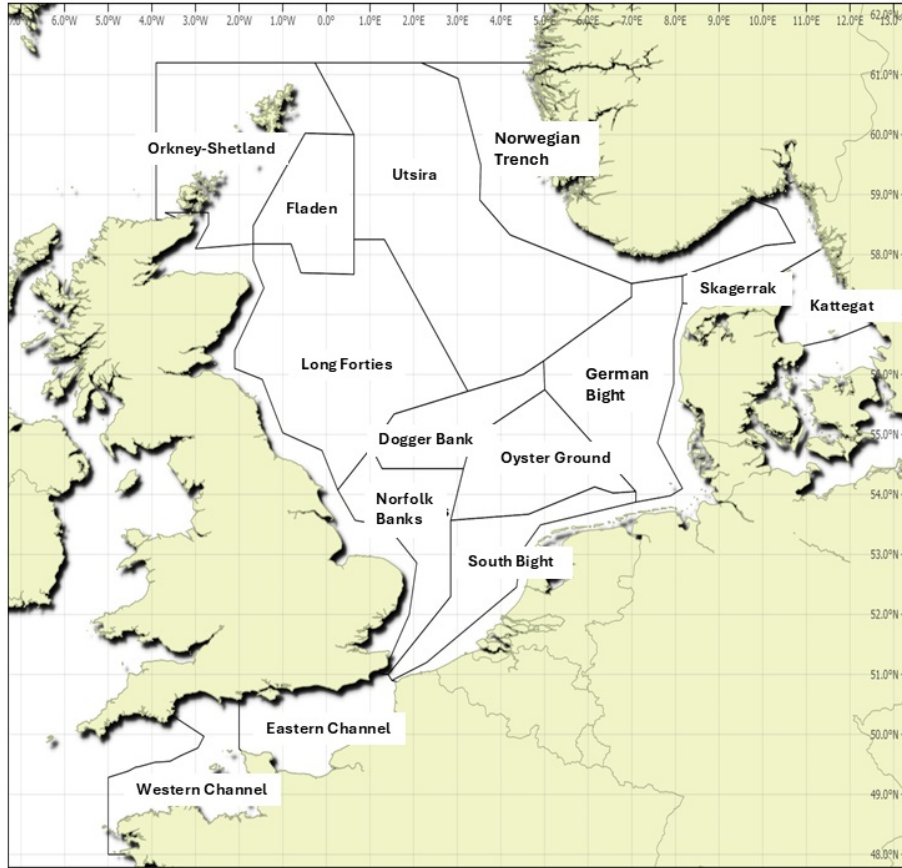


FIG 1. Fourteen strata in the North Sea considered in [Belgrano et al. 2023](#). This classification represents a typical approach to dividing the sea into several strata in marine studies. In the real data analysis in Section 4, we analyze data obtained from these areas.

the Sea (ICES) is conducting IEAs in European waters (ICES, <https://www.ices.dk>). The IEA regional group within ICES focuses on collecting time series data for key physical and biological variables and conducts integrated trend analyses to understand dynamics and ecosystem changes (ICES 2018, ICES 2022a). The time series datasets considered by the IEA are collected through various scientific surveys and have been selected as indicators that form the basis for assessing the status and trends of ecosystems (ICES, <https://www.ices.dk>, OSPAR, <https://odims.ospar.org>).

However, there are potential challenges when analyzing changes over time. Data have been collected for only a few decades since awareness of climate change emerged, resulting in insufficient information to apply conventional time series models that account for temporal correlations and relationships between multivariate time series ([Hardison et al. 2019](#), ICES 2018, ICES 2022a, [Solvang and Subbey 2019](#)). Additionally, the IEA is investigating differences in environmental factors across various subregions (referred to as strata ([Belgrano et al. 2023](#)) or polygons ([Ohlsson and Eklöf 2020](#))), considering the distribution of biological communities. This is important because the assessment areas are adjacent to different oceans and possess ecologically distinct marine geological environments, which are likely to result in changes in the ecosystems. For example, there are fourteen strata containing diverse ecosystems, ranging from the shallow waters south of the North Sea to the deep waters of the Norwegian Trench, as illustrated in Figure 1 ([Belgrano et al. 2023](#), numerical examples in this region are presented in Section 4). The IEA regional groups working in Norwegian

waters (including Iceland, Denmark, Svalbard, the Faroe Island, and the UK, in addition to Norway) have provided trend analyses for short time series data. This includes a method for classifying biotic (such as zooplankton or biological community) and abiotic (environmental factors) data into common trend patterns (Solvang and Ohishi 2023, Belgrano et al. 2023). Additionally, they have developed a method for identifying observations marked with a flag, indicating the latest observation value that is deemed to deviate from recent trends (Belgrano et al. 2023, Solvang and Arneberg 2024). Furthermore, the regional groups need consider changes in the state of the ecosystem across multiple strata (or polygons). For this reason, analysis of variance (ANOVA) is a possible statistical method for examining the homogeneity within the strata based on the time series data. Within ICES, as all variations and tendencies derived from time series are analysed as part of trend analysis, examining differences between strata constitutes one important task within this trend analyses.

As mentioned, the time series datasets handled by IEA are generally short, but time dependence is unavoidable. ANOVA was originally conceived as a method of multivariate analysis, which basically assumed identically independent distribution for the data (Searle, Casella and McCulloch 1992, Sahai and Ojeda 2004, Clarke 2008), and does not take time dependence into account. If a conventional analysis method that ignores time dependence is used despite the presence of actual time dependence, serious problems may arise (Brillinger 1980). For time series data, smoothing spline ANOVA for time-dependent spectral analysis (Guo et al. 2003), tests for the equality of several spectra (Dette and Paparoditis 2009, Jentsch and Pauly 2015), ANOVA for time series with independent and correlated groups, studied by Nagahata and Taniguchi (2018) and Goto et al. (2023a), involved estimation of spectral density. These approaches require time series datasets containing more than 100 time points, as seen in several textbooks on time series (e.g., Brillinger 1981, Hamilton 1994, Shumway and Stoffer 2006).

To face this challenge, we propose an ANOVA test for short Time Series (ANOVATS) using a subsampling method, which makes it possible to circumvent the need for estimating the spectral density. Estimation of the long-run variance typically requires a large sample size, and when the sample size is insufficient, tests based on its estimation often suffer from severe size distortion. To address this problem, one may consider applying resampling methods such as the bootstrap to approximate critical values. In the frequency domain, subsampling methods are known to perform effectively in several contexts (Meyer, Paparoditis and Kreiss 2020; Yu, Kaiser and Nordman 2023, 2024). For lattice data, Heagerty and Lumley (2000) applied the subsampling technique to estimate the variance of parameters of interest, thereby overcoming the difficulties arising from replacing the unknown mean with the sample mean. Building on these insights, we apply a subsampling method to time series data. The idea is to divide the observed time series into shorter time series with overlap and then construct the empirical distribution of the test statistic based on the shorter time series (see Politis, Romano and Wolf 1999). Subsampling-based methods have been investigated in several studies (e.g., Carlstein 1986; Politis and Romano 1994; Fukuchi 1999; Berg, McMurry and Politis 2010; McMurry, Politis and Romano 2012; Dudek and Lenart 2023), but not in the context of ANOVA for dependent or short time series.

After rejecting the homogeneity hypothesis, Fisher's least significant difference procedure has been widely used to identify which groups exhibit significant differences. This method applies the t-test to each pair of groups, leading to a severe multiple comparisons problem (Hayter, 1986). We propose a method that efficiently detects significant differences among groups while reducing the number of tests.

This paper is organized as follows: In Section 2, we introduce the one-way model with regional differences, time-dependent errors, and correlated groups, and we define a test statistic and a subsample-based p -value. The asymptotic behavior of the proposed test is investigated.

The analysis method is introduced after rejecting the null hypothesis, which helps in dividing or grouping the data into clusters. In Section 3, we present a numerical example to demonstrate our proposed method. In Section 4, we demonstrate our approach applied to two real data sets from the North Sea. In Appendix, we provide proofs of our theoretical and numerical results from a simulation study to validate the proposed approach.

2. Statistical methodology.

2.1. Proposed ANOVATS method. We consider the one-way model with regional differences, time-dependent errors, and correlated groups defined as

$$(1) \quad \mathbf{z}_{it} = \boldsymbol{\mu} + \boldsymbol{\psi}_i + \mathbf{e}_{it}, \quad i = 1, \dots, a; t = 1, \dots, n,$$

where $\mathbf{z}_{it} = (z_{it1}, \dots, z_{itp})^\top$ is a p -dimensional observation from the i -th group at time t , $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$ is a general mean, $\boldsymbol{\psi}_i = (\psi_{i1}, \dots, \psi_{ip})^\top$ is a nonrandom regional effect such that $\sum_{i=1}^a \boldsymbol{\psi}_i = \mathbf{0}$, and $\mathbf{e}_{it} = (e_{it1}, \dots, e_{itp})^\top$ is a disturbance process such that $\mathbf{e}_t = (e_{1t}, \dots, e_{at})^\top$ is a centered strictly stationary sequence with an ap -by- ap spectral density matrix $\mathbf{f}(\lambda) = (\mathbf{f}_{ij}(\lambda))_{i,j=1,\dots,a}$ for $\lambda \in [-\pi, \pi]$, where $\mathbf{f}_{ij}(\lambda)$ is a spectral density matrix of the processes $\{\mathbf{e}_{it}\}$ and $\{\mathbf{e}_{jt}\}$.

We are interested in the existence of regional differences, that is, the following hypothesis testing problem:

$$(2) \quad H_0 : \boldsymbol{\psi}_1 = \dots = \boldsymbol{\psi}_a \quad \text{v.s.} \quad H_1 : H_0 \text{ does not hold.}$$

For this hypothesis, we propose the test statistic defined as

$$T_n = n \sum_{i=1}^a (\bar{\mathbf{z}}_{i.} - \bar{\mathbf{z}}_{..})^\top (\bar{\mathbf{z}}_{i.} - \bar{\mathbf{z}}_{..}).$$

The estimation of spectral density from short time series poses a significant obstacle to achieving proper size control. To address this issue, the proposed test statistic is designed to exclude spectral estimators. As a consequence, the asymptotic variance includes information from the underlying process, making the statistic not asymptotically distribution-free. Therefore, tailored approaches are required to compute the critical values. In this paper, we overcome this limitation by employing a subsampling method. Here, we note that in classification problems under high-dimensional and low sample size settings, the inverse of the sample variance matrix in the classifier is replaced by an identity matrix or other invertible matrices, as the sample variance matrix is always singular (see, e.g., [Aoshima and Yata 2014](#)). From this perspective, our statistic is comprehensible and natural.

We make the following assumption to establish the theoretical results:

ASSUMPTION 2.1. (i) *The disturbance process $\{\mathbf{e}_t\}$ is geometrically α -mixing, that is, the α -mixing coefficient $\alpha(\cdot)$ defined as*

$$\alpha(n) := \sup_{k \in \mathbb{Z}, A \in \mathcal{F}_{-\infty}^k, B \in \mathcal{F}_{k+n}^\infty} |\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)|,$$

where, for $a \leq b$, $\mathcal{F}_a^b$ is the σ -field generated by $\{\mathbf{e}_t : a \leq t \leq b\}$, satisfies $\alpha(n) \leq C_\alpha \rho^n$ for some constants $C_\alpha \in (1, \infty)$ and $\rho \in (0, 1)$.

(ii) *The moments of all orders exist in the sense that*

$$\sup_{s_2, s_3, \dots, s_\ell \in \mathbb{Z}} |\mathbb{E} e_{i_1 0 d_1} e_{i_2 s_2 d_2} \dots e_{i_\ell s_\ell d_\ell}| < \infty$$

for any $\ell \in \mathbb{N}$, any $(i_1, \dots, i_\ell) \in \{1, \dots, K\}^\ell$, and any $(d_1, \dots, d_\ell) \in \{1, \dots, p\}^\ell$.

REMARK 2.1. Under Assumption 2.1, it holds, for any $d \in \mathbb{N}$, any $\ell \in \mathbb{N}$, any $(i_1, \dots, i_\ell) \in \{1, \dots, K\}^\ell$, and any $(d_1, \dots, d_\ell) \in \{1, \dots, p\}^\ell$, that

$$(3) \quad \sum_{s_2, \dots, s_\ell = -\infty}^{\infty} \left(1 + \sum_{j=2}^{\ell} |s_j|^d \right) |\text{cum}\{e_{i_1 0 d_1}, e_{i_2 s_2 d_2}, \dots, e_{i_\ell s_\ell d_\ell}\}| < \infty.$$

See Lemma 2.1. of Goto et al. (2023b) for this result.

REMARK 2.2. Note that $\sqrt{n}(\bar{\mathbf{z}}_{1\cdot}^\top - \bar{\mathbf{z}}_{\cdot\cdot}^\top, \dots, \bar{\mathbf{z}}_{a\cdot}^\top - \bar{\mathbf{z}}_{\cdot\cdot}^\top)^\top$ converges in distribution to the centered normal distribution with variance $\mathbf{H}$, where $\mathbf{H} = (\mathbf{H}_{ij})_{i,j=1,\dots,a}$ with

$$\mathbf{H}_{ij} = 2\pi \mathbf{f}_{ij}(0) - \frac{2\pi}{a} \sum_{s=1}^a \{\mathbf{f}_{sj}(0) + \mathbf{f}_{is}(0)\} + \frac{2\pi}{a^2} \sum_{s,k=1}^a \mathbf{f}_{sk}(0),$$

under the cumulant summability condition (3). under the cumulant summability condition (3). Then, under H_0 , T_n converges in distribution to $\sum_{j=1}^r \lambda_j \chi_{1,j}^2$, where $\lambda_1 \geq \dots \geq \lambda_r > 0$ are the nonzero eigenvalues of $\mathbf{H}$, and $\chi_{1,j}^2$ are independent χ_1^2 random variables.

The subsampling test statistic and the subsample-based p -value with block length b are defined, for $t = 1, \dots, n - b + 1$, as

$$(4) \quad T_{n,b,t} = \frac{b}{1 - \frac{b}{n}} \sum_{i=1}^a (\bar{\mathbf{z}}_{i\cdot,b,t} - \bar{\mathbf{z}}_{\cdot\cdot,b,t})^\top (\bar{\mathbf{z}}_{i\cdot,b,t} - \bar{\mathbf{z}}_{\cdot\cdot,b,t})$$

$$\text{and } p_n = \frac{1}{n - b + 1} \sum_{t=1}^{n-b+1} \mathbb{I}\{T_{n,b,t} > T_n\},$$

where $\bar{\mathbf{z}}_{i\cdot,b,t} = \sum_{j=t}^{t+b-1} \mathbf{z}_{ij}/b$, $\bar{\mathbf{z}}_{\cdot\cdot,b,t} = \sum_{i=1}^a \sum_{j=t}^{t+b-1} \mathbf{z}_{ij}/(ab)$, and $\mathbb{I}\{\cdot\}$ is an indicator function. The coefficient $(1 - b/n)^{-1}$, which is asymptotically negligible, corresponds to the finite population correction and provides improvement of the empirical size of the test. We propose a subsampling base test (ANOVATS) that rejects H_0 whenever $p_n < \varphi$.

This test can equivalently be expressed in the quantile-based form as

$$p_n = 1 - \hat{F}_{n,b}(T_n) < \phi \quad \Leftrightarrow \quad T_n \geq \inf\{x \in \mathbb{R} : \hat{F}_{n,b}(x) > 1 - \phi\},$$

where $\hat{F}_{n,b}(x) = (n - b + 1)^{-1} \sum_{t=1}^{n-b+1} \mathbb{I}\{T_{n,b,t} \leq x\}$ denotes the empirical distribution function of the subsampling statistics. Hence, rejecting H_0 when $p_n < \phi$ is equivalent to rejecting it when T_n exceeds the empirical $(1 - \phi)$ -quantile of $\hat{F}_{n,b}$.

Note that in this paper, we set the block length as $b = \lfloor 2.5n^{1/3} \rfloor$ because the order $n^{1/3}$ corresponds to the optimal rate in the sense of bias-variance tradeoff for subsampling-based variance estimation of the sample mean in mixing stationary time series (Politis, Romano and Wolf, 1999, Section 9.2.1). The constant factor 2.5 is chosen empirically as it provides good finite-sample performance (see Section A.2 in Appendix). For the block length $b = \lfloor 2.5n^{1/3} \rfloor$ and significance level $\varphi = 0.05$, if there is at least one (or two) t such that $T_{n,b,t} > T_n$, then $p_n \geq \varphi$ for $n = 3, \dots, 26$ (or 48). Thus, for small-sample time series (those with $n \leq 26$), if there is at least one t such that $T_{n,b,t} > T_n$, this immediately implies H_0 is not rejected. The following theorem shows that our test has fundamental properties:

THEOREM 2.1. Suppose Assumption 2.1 and that the subsampling block of length b satisfies $b \rightarrow \infty$ and $b/n \rightarrow 0$ as $n \rightarrow \infty$. For significance level φ , the test that rejects H_0 whenever $p_n < \varphi$ has asymptotic size φ and is consistent, that is,

$$\mathbb{P}(p_n < \varphi \mid H_0) \rightarrow \varphi \quad \text{and} \quad \mathbb{P}(p_n < \varphi \mid H_1) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

2.2. Post-ANOVATS procedures. Let $p = 1$ and let Area_i represent the area we observed $\{z_{it}\}_{t=1,\dots,n}$. In Section 2.1, we considered the testing method based on subsampling for hypothesis (2). If H_0 is not rejected, there is no evidence for the mean-difference among areas. If H_0 is rejected, we conclude the area-means are not homogeneous. Then when the means are arranged in ascending order (say $\mu + \psi_{[1]}, \dots, \mu + \psi_{[a]}$), there exists at least one index $i \in \{1, \dots, a-1\}$ such that the difference $\psi_{[i+1]} - \psi_{[i]}$ is strictly positive. It is reasonable to divide the areas into two groups $(\text{Area}_{[1]}, \dots, \text{Area}_{[i']})$ and $(\text{Area}_{[i'+1]}, \dots, \text{Area}_{[a]})$ for the index i' corresponding to one of the largest differences, that is,

$$i' := \arg \max_{i=1,\dots,a-1} (\psi_{[i+1]} - \psi_{[i]}).$$

In practice, $\mu + \psi_i$ is replaced with the sample mean $\bar{z}_i$ for Area_i . Note that hierarchical splitting methods based on the sum of squares were considered by, e.g., [Edwards and Cavalli-Sforza \(1965\)](#), [Scott and Knott \(1974\)](#), [Caliński and Corsten \(1985\)](#), and [Wang and Xu \(2014\)](#).

REMARK 2.3. [Wang and Xu \(2014\)](#) proposed a procedure for clustering data by repeatedly applying a test procedure and verified its consistency. Their idea is to take the significance level φ_n converging to zero as $n \rightarrow \infty$ to eliminate type 1 errors. However, φ_n must converge slow enough to ensure the power of the test tends to one.

In our setting, for φ_n such that $\varphi_n \rightarrow 0$ and $\varphi_n / (n^{\frac{a}{2}-1} \exp(-\frac{n}{2})) \rightarrow \infty$ as $n \rightarrow \infty$, the procedure for rejecting H_0 in favor of H_1 whenever $p_n < \varphi_n$ satisfies

$$(5) \quad P(p_n < \varphi_n | H_0) \rightarrow 0 \quad \text{and} \quad P(p_n < \varphi_n | H_1) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

The proof of the convergences in (5) is deferred to Section B.2 in Appendix. The derivation of the order of φ_n is based on the upper bound of the tail of a chi-squared distribution ([Inglot and Ledwina, 2006](#), Lemma 1).

2.3. Analyzing procedure. Based on Sections 2.2 and 2.3, the analyzing procedure based on ANOVATS is summarized as follows:

Step 1 Hypothesis testing for all areas: Let $p = 1$ and let Area_i represent the areas $\{z_{it}\}_{t=1,\dots,n}$ that were observed. Consider the hypothesis test for the absence of a regional difference against the presence of a regional difference, defined in (2). Applying our test to the data provides the p -value p_n .

If H_0 is not rejected ($p_n \geq 0.05$), terminate the procedure.

If H_0 is rejected ($p_n < 0.05$), proceed to the next step.

Step 2 Division of areas into two groups After rejecting H_0 , sort the sample means for all areas in ascending order, denoted as $\bar{z}_{[1]}, \dots, \bar{z}_{[a]}$, corresponding to the sorted areas $\text{Area}_{[1]}, \dots, \text{Area}_{[a]}$. Next, compute the differences between the sample means of adjacent areas: $\bar{z}_{[i+1]} - \bar{z}_{[i]}$ for all $i = 1, \dots, a-1$, and find the index i' that maximizes these differences:

$$i' := \arg \max_{i=1,\dots,a-1} (\bar{z}_{[i+1]} - \bar{z}_{[i]}).$$

The areas are then divided into two groups:

$$\text{Group 1: } \text{Area}_{[1]}, \dots, \text{Area}_{[i']}, \quad \text{Group 2: } \text{Area}_{[i'+1]}, \dots, \text{Area}_{[a]}$$

Step 3 Further division For each of the two groups, hypothesis tests are conducted to determine if further division is necessary. The following hypotheses are considered for each group:

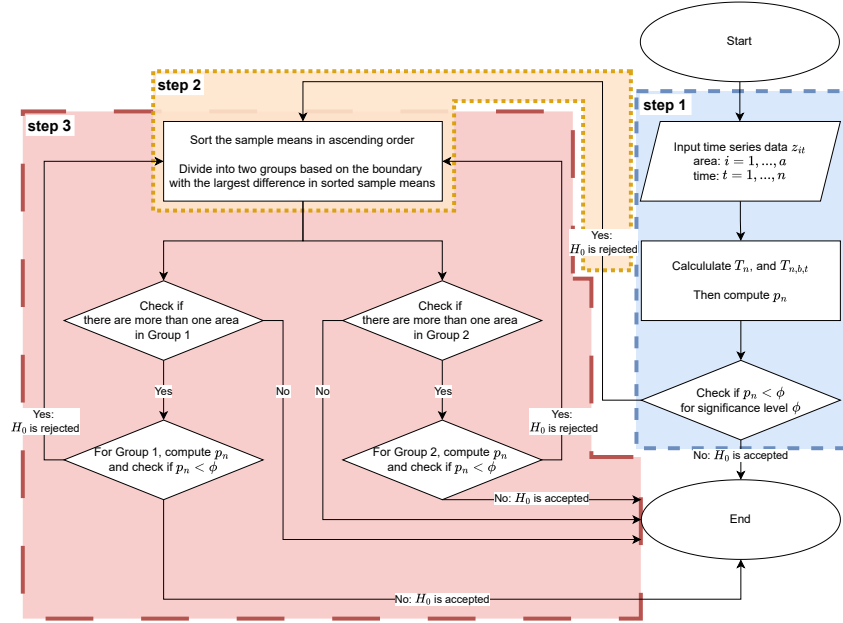


FIG 2. The flowchart of the ANOVATS-based proposed procedure in Section 2.3. Step 1: Apply ANOVATS to the full dataset. Step 2: If the homogeneity hypothesis is rejected in Step 1, divide the dataset into two subgroups. Step 3: Repeat Steps 1 and 2 for each subgroup until the hypothesis is not rejected or each subgroup contains only one area.

For Group 1, $H_0 : \psi_{[1]} = \dots = \psi_{[i']}$ v.s. $H_1 : H_0$ does not hold.

For Group 2, $H_0 : \psi_{[i'+1]} = \dots = \psi_{[a]}$ v.s. $H_1 : H_0$ does not hold.

The above steps are repeated for each group, testing and subdividing until the hypothesis is not rejected or the number of areas in the group is one.

By following this procedure, the areas can be grouped into statistically significant clusters.

A flowchart of this procedure is shown in Figure 2. The numerical procedure is implemented using R code (Team 2023) given in Appendix.

3. Demonstration example.

1. **Model setup:** Consider the moving average model of order 1 $y_{it} = \psi_i + e_{it}$, where $(\psi_1, \dots, \psi_4) = (0, 2, 2, 4)$, $e_{it} = \nu_{it} + 0.5\nu_{i(t-1)}$, and $\{\nu_{it}\}$ follows i.i.d. standard normal distribution with $n = 20$ and $a = 4$. Let Area_i denote the area name where $\{z_{it}\}$ was observed. Clearly, we know $\psi_1 < \psi_2 = \psi_3 < \psi_4$. Figures 3 and 4 show the plots and boxplots of these realizations.
2. **Initial hypothesis test:** We performed our proposed test for the hypothesis $H_0 : \psi_1 = \dots = \psi_4$ vs. $H_1 : H_0$ does not hold. The test provided a p -value of 0, leading to the rejection of H_0 .
3. **Sorting areas by sample mean:** The sample means of $y_{1t}, \dots, y_{4t}$ are $-0.064, 1.88, 1.75$, and 3.87 , respectively. We sorted the areas in ascending order according to their sample means, resulting in the order: $\text{Area}_1, \text{Area}_3, \text{Area}_2, \text{Area}_4$.
4. **Calculating differences between adjacent areas:** We computed the differences between the sample means for each pair of adjacent rearranged areas:

$$1.75 - (-0.064) = 1.81, \quad 1.88 - 1.75 = 0.13, \quad 3.87 - 1.88 = 1.99.$$

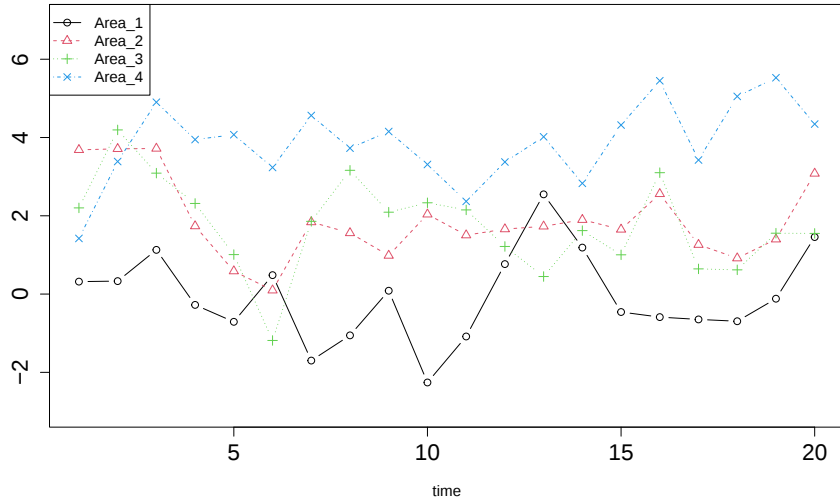


FIG 3. Plots of realizations of moving average models of order one, with means 0, 2, 2, and 4, corresponding to Area₁–Area₄, respectively, to demonstrate our analysis methods proposed in Section 2.3.

5. **First division of areas:** Based on the largest difference in sample means (1.99) between Area₂ and Area₄, we divided the areas into two groups: the first group is (Area₁, Area₃, Area₂), and the second group is (Area₄). We stopped the procedure for the second group because it contains only one area.
6. **Hypothesis testing for the first group:** For the first group (Area₁, Area₃, Area₂), we consider the hypothesis $H_0 : \psi_1 = \psi_3 = \psi_2$ vs. $H_1 : H_0$ does not hold. Our test provided a p -value of 0, leading to the rejection of H_0 .
7. **Further division of the first group:** We then divide the first group into two subgroups based on the largest difference in sample means: the first subgroup is (Area₁) and the second subgroup is (Area₃, Area₂). We stopped the procedure for the first subgroup because it contains only one area.
8. **Final hypothesis test for the remaining areas:** For the second subgroup (Area₃, Area₂), we consider the hypothesis $H_0 : \psi_3 = \psi_2$ vs. $H_1 : H_0$ does not hold. Applying our test, we obtained a p -value of 0.73, indicating that $H_0 : \psi_3 = \psi_2$ cannot be rejected.
9. **Final grouping:** In the end, we obtained three statistically significantly different groups (Area₁), (Area₃, Area₂), and (Area₄).

4. Empirical studies. Two numerical examples, related to zooplankton biomass data in the North Sea, applying our proposed ANOVA for time series data are summarized in this section. As mentioned in Introduction, the North Sea includes different subregions (strata) of oceanography, as shown in Figure 1. Zooplankton are at the bottom of the food chain and have an impact on the marine ecosystem of fish. However, owing to recent climate change, there is debate about changes in the currents and biological factors of the North Sea (Hátún et al. 2020), and existing knowledge alone is insufficient to interpret changes in biomass. We demonstrate the following examples using ANOVATS, which conduct systematically and statistically evaluating differences in plankton biomass in strata of the North Sea without biological or geological knowledge as prior information.

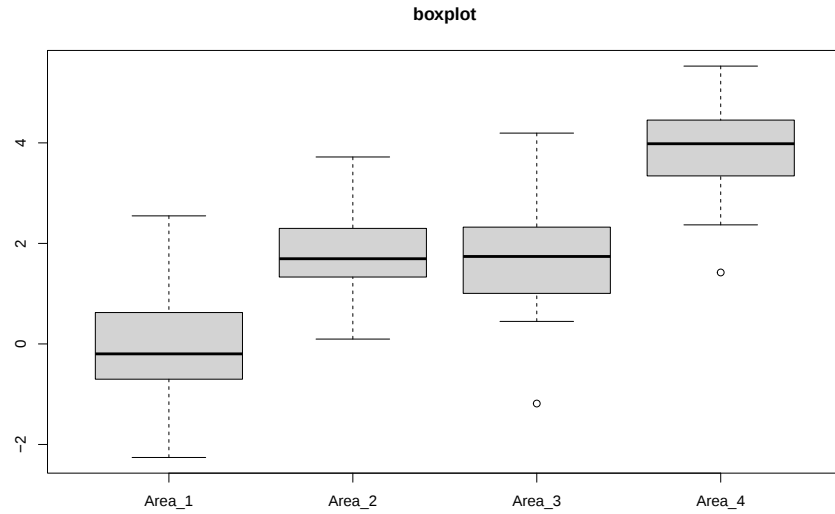


FIG 4. Boxplots of realizations of moving average models of order one, with means 0, 2, 2, and 4, corresponding to Area₁–Area₄, respectively, to illustrate the distribution and variability of the data.

4.1. *Zooplankton data in the North Sea.* In this subsection, we analyze the total biomass data (dry weight) of zooplankton between the years 2006 and 2023 in the North Sea (Figure 5). This dataset was collected by the Institute of Marine Research (IMR), Norway, as

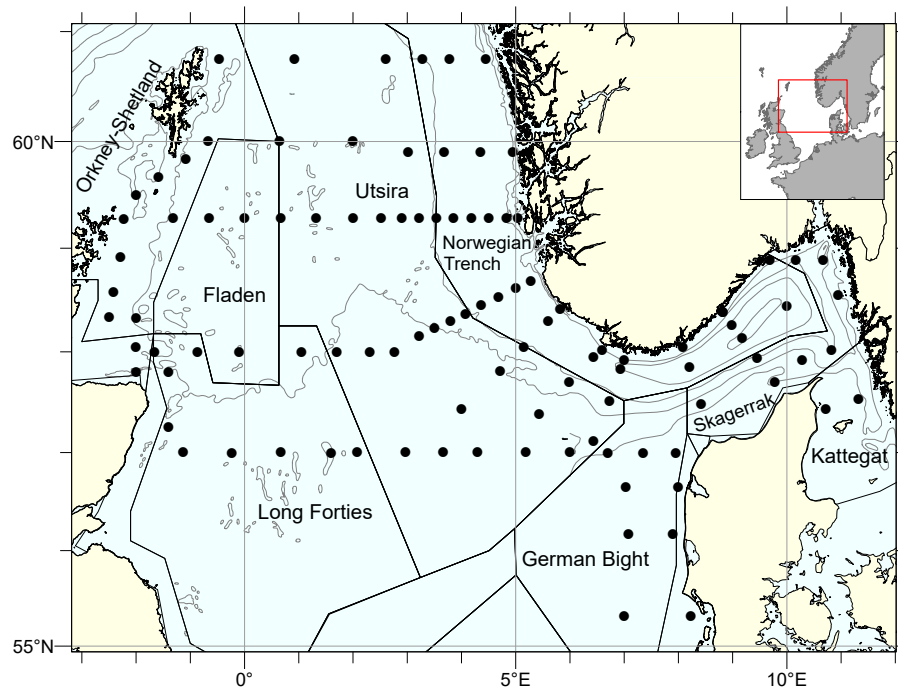


FIG 5. Zooplankton sampling areas in the North Sea, including Orkney-Shetland, Fladen, Utsira, Long Forties, German Bight, Skagerrak, Kattegat, and Norwegian Trench, surveyed as part of the climate and plankton monitoring program conducted by the Institute of Marine Research.

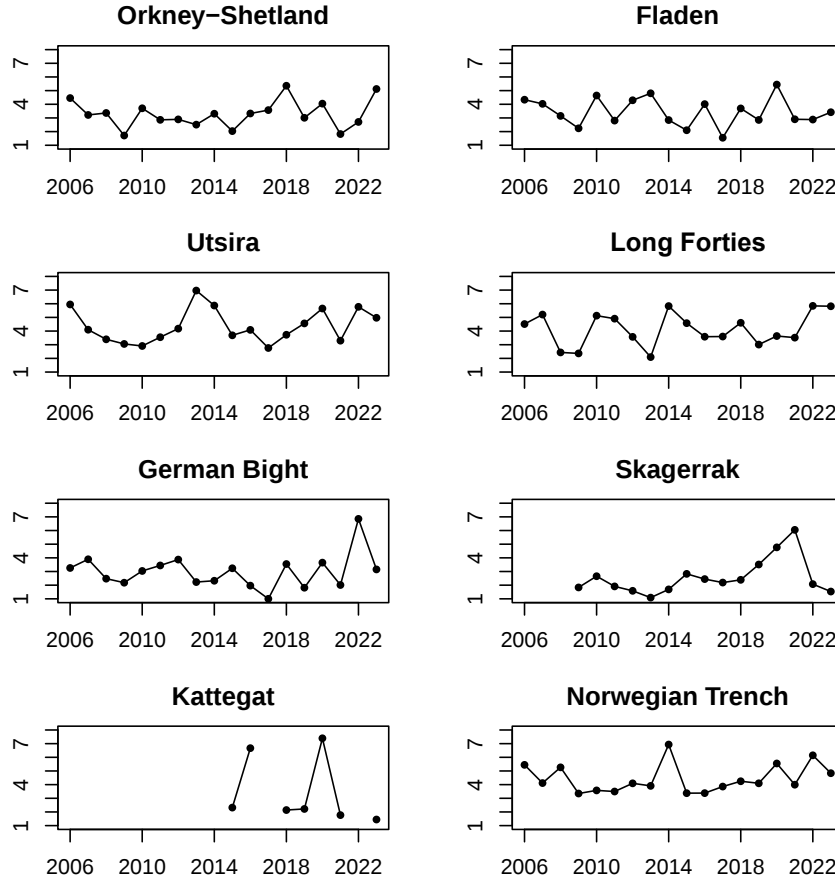


FIG 6. Plots of the time series data of zooplankton total biomass for eight sub-areas in the North Sea.

part of the monitoring program of climate and plankton in the North Sea-Skagerrak. This dataset includes zooplankton biomass from three main transects (Utsira-Orkney, Hanstholm-Aberdeen, and Torungen-Hirtshals) covered several times per year, as well as the large-scale North Sea Ecosystem cruise, conducted in spring from April to May (Franze et al. 2024). The dataset includes all seasons and is geographically restricted to the area between latitudes 55 N and 61 N and between longitudes 2.5 W and 11.5 E.

Zooplankton sampling and sample treatment were made according to the IMR standard procedure (Hassel et al. 2013), which includes vertical tows with a WP2 plankton net (0.25 m^2 opening, 180 μm mesh size (D.J. Tranter and J.H. Fraser 1968) from near the bottom to the surface. Each sample was split into two equal parts with a Motoda splitter (Motoda 1959). One half of the sample was preserved in 4% formaldehyde and stored for later analysis (data not presented here). The other half was used to determine the dry weight biomass of three size fractions by successive sieving the sample through three meshes (2000 μm , 1000 μm , and 180 μm). Samples were transferred to preweighed aluminum trays and dried at 65°C for > 24 h until at a constant weight. The data analyzed here are the total zooplankton biomass for the entire water column, derived as the sum of the three fractions, and expressed as gram dry weight per square meter surface area.

Figure 6 presents the plots of the time series data for the eight areas (Orkney-Shetland, Fladen, Utsira, Long Forties, German Bight, Skagerrak, Kattegat, and Norwegian Trench) shown in Figure 5. The data in Kattegat are missing over half of the time points and the data

in Skagerrak are missing the first three years. Therefore, we exclude Kattegat and other areas are used for data from 2009 to 2023.

Our primary objective is the identification of statistically significant regional trends or regional differences concerning the total biomass dataset. Therefore, the number of areas and observations are $a = 7$ and $n = 15$, respectively, and the significance level φ is set as 0.05. To this end, we carry out the following procedure: for each distinct biomass dataset, we apply our test with $b = 6$ to the data from seven areas.

The abbreviations of area's names are defined as follows: OSN for Orkeny-Shetland, FG for Fladen, NCNS for Utsira, UKN2 for Long Forties, Ger3 for German Bight, Sk1 for Skagerrak, and NorC for Norwegian Trench.

We set the hypothesis for applying step 1 in subsection 2.3 as

H_0 : Mean levels for the overall time series are homogeneous for all areas.

The overall sample mean supporting homogeneity is 3.56 g/m^2 , and this hypothesis is rejected (p -value 0).

Next, step 2 identifies two groups: Group 1 includes Sk1, Ger3, OSN, and FG, and Group 2 includes UKN2, NorC, and NCNS.

In step 3, we set the hypothesis for each group as

H_0 : Mean levels for the overall time series in Group 1 are homogeneous for all areas.

H_0 : Mean levels for the overall time series in Group 2 are homogeneous for all areas.

The sample mean of the four areas in Group 1 is 3.02 g/m^2 , and the sample mean of the three areas in Group 2 is 4.27 g/m^2 . For these means, the null hypothesis for each group is not rejected (p -value 0.7 for each group).

The sample means in the groups are 2.57 g/m^2 for Sk1, 2.95 g/m^2 for Ger3, 3.20 g/m^2 for OSN, 3.37 g/m^2 for FG, 4.14 g/m^2 for UKN2, 4.32 g/m^2 for NorC, and 4.34 g/m^2 for NCNS. Figure 7 shows the boxplot of the time series data for each area. The boxes with solid and dashed borders indicate two different groups, based on the output from step 2.

The zooplankton biomass presented here represents the vertically integrated abundance for the entire water column, which is affected by the water depth. Consequently, the shallow areas OSN and FG contain less biomass compared with the deeper NorC, where the plankton net is towed through a larger volume of water. In addition, Ger3 and Sk1 are shallow areas, which suggests that both areas include less biomass. Furthermore, deeper areas contain zooplankton communities with larger-sized zooplankton, as seen in NorC, NCNS, and UKN2. From ANOVATS, it is verified that this pattern is consistent with previous studies where significant differences in abundance of zooplankton across various regions in the North Sea have been documented (Fransz et al. 1991, Krause et al. 2003). Physical conditions such as ocean currents, temperature, and bottom topography are major drivers for shaping zooplankton distributions and species composition in the North Sea. Areas in the central North Sea and NorC tend to have higher zooplankton concentrations owing to the continuous supply of Atlantic water transporting large copepods such as *Calanus finmarchicus* into the area from the north (Heath et al. 1999, Gao et al. 2021). In contrast, the shallow areas in Sk1 and Ger3 are dominated by small short-lived plankton forms (Krause et al. 2003).

4.2. Zooplankton at four sites in Skagerrak. In this subsection, we apply our method to zooplankton biomass data collected monthly by the IMR from 2013 to 2021 at four coastal sites (Arendal, Langesund, OF2, Risør) in Skagerrak (Figure 8) as a part of the IMR Coastal Monitoring program. While the data include a larger number of time points compared with the time series shown in Section 4.1, there are missing values in some months in some years.

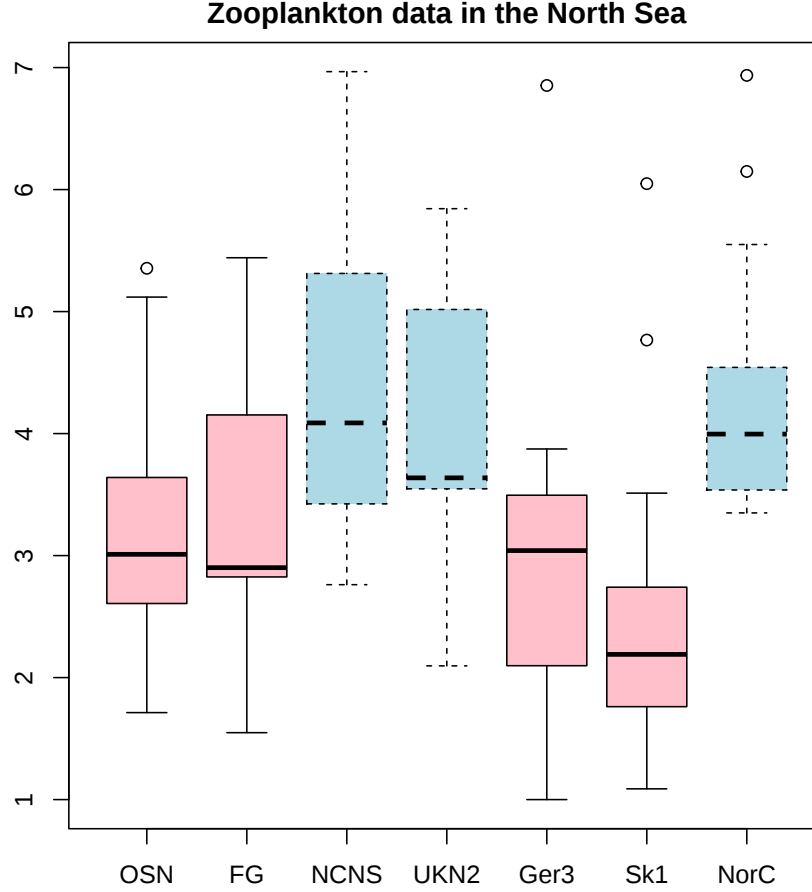


FIG 7. Boxplot of the time series data of total biomass from 2009 to 2023 for each sub-area in the North Sea. The abbreviations listed on the horizontal correspond to: OSN for Orkney-Shetland, FG for Fladen, NCNS for Utsira, UKN2 for Long Forties, Ger3 for German Bight, Sk1 for Skagerrak, and NorC for Norwegian Trench. The vertical axis indicates biomass [g/m^2]. The boxes with solid and dashed borders indicate two different groups based on the output of the method proposed in Section 2.3.

For this reason, we have divided the data into four seasons: spring (March, April, May), summer (June, July, August), autumn (September, October, November), and winter (January, February, and December of the previous year) and used quarterly data. Because there are still missing values in autumn 2019 for Risør and Langesund, and in spring 2014 and autumn 2019 for OF2, those data are applied by Box-Cox transformation and imputing the parts by an autoregressive model. This procedure is conducted by the functions *boxcox* and *arfit* in R package *TSSS* (Kitagawa 2023) and R code to impute the missing values, written based on the theory of state space modeling (Kitagawa 2021). The final time series includes 36 quarterly points.

Zooplankton sampling and sample treatment followed the IMR standard procedure (Hassel et al. 2013) as described in Section 4.1. The dataset is given as grams dry weight per square meter surface area (g/m^2) and grouped into three categories according to size fractions: $180 \mu m$ - $1000 \mu m$, $1000 \mu m$ - $2000 \mu m$, and $> 2000 \mu m$. The time series data for each size fraction are presented in Figure 9.

We apply our procedure explained in Section 2.2 with $b = 8$ and the result is given in Figure 10.

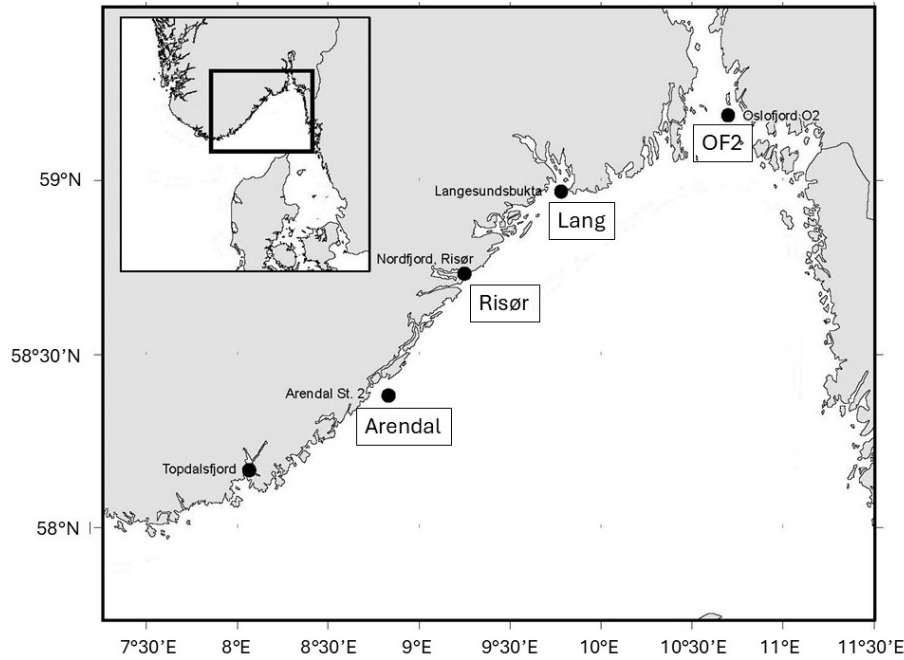


FIG 8. Zooplankton sampling areas in Skagerrak, part of the North Sea, including Oslofjord (OF2), Langesund (Lang), Risør (Risør), and Arendal (Arendal), surveyed as part of the Coastal Monitoring program conducted by the Institute of Marine Research. Zooplankton biomass data are believed to exhibit variations in size.

In step 1, the overall sample mean levels supporting the null hypothesis are 0.705 g/m^2 for size fractions $180 \mu\text{m} - 1000 \mu\text{m}$, 1.01 g/m^2 for size fractions $1000 \mu\text{m} - 2000 \mu\text{m}$, and 0.619 g/m^2 for size fractions $> 2000 \mu\text{m}$. The null hypothesis is rejected for all size fractions (p-value 0 for each). Furthermore, step 2 divides the sites into two groups as follows:

180–1000 μm : Group 1: Arendal, Risør; Group 2: Lang, OF2

1000–2000 μm : Group 1: Arendal, Risør, Lang; Group 2: OF2

>2000 μm : Group 1: Arendal, Risør; Group 2: Lang, OF2.

For $180 \mu\text{m} - 1000 \mu\text{m}$, step 3 does not further subdivide Group 1 (p-value 0.86, sample mean for null 0.488 g/m^2) or Group 2 (p-value 0.069, sample mean for null 0.921 g/m^2).

For the case of $1000 \mu\text{m} - 2000 \mu\text{m}$, step 3 subdivides Group 1 into two groups: Group 1-1 includes Arendal and Risør, and Group 1-2 includes Lang (p-value 0, sample mean for null 0.447 g/m^2). Step 3 does not further subdivide Group 1-1 (p-value 0.14, sample mean for null 0.322 g/m^2).

Finally, for $> 2000 \mu\text{m}$, step 3 subdivides Group 1 into two groups, one with Arendal and the other with Risør (p-value 0, sample mean for null 0.151 g/m^2). However, Group 2 is not subdivided into two groups by step 3 (p-value 0.24, sample mean for null 1.09 g/m^2).

Table 1 summarizes the sample mean for each site according to size fractions. The biomass data sampled at Langesund and OF2 are identified in the same group for each size fraction. The data from Risør is divided into subgroups only for size fraction $> 2000 \text{ g/m}^2$, and the data from Arendal remains undivided for all size fractions. These results can be explained from the point view of the sampling depth and the position of the sampling station in relation to the coast. The sampling depths and the positions of the sampling stations are provided in Table 2. Sampling sites with deep sampling depths (OF2 and Langesund) usually have larger total biomass values. This is explained by the plankton net being towed through a larger

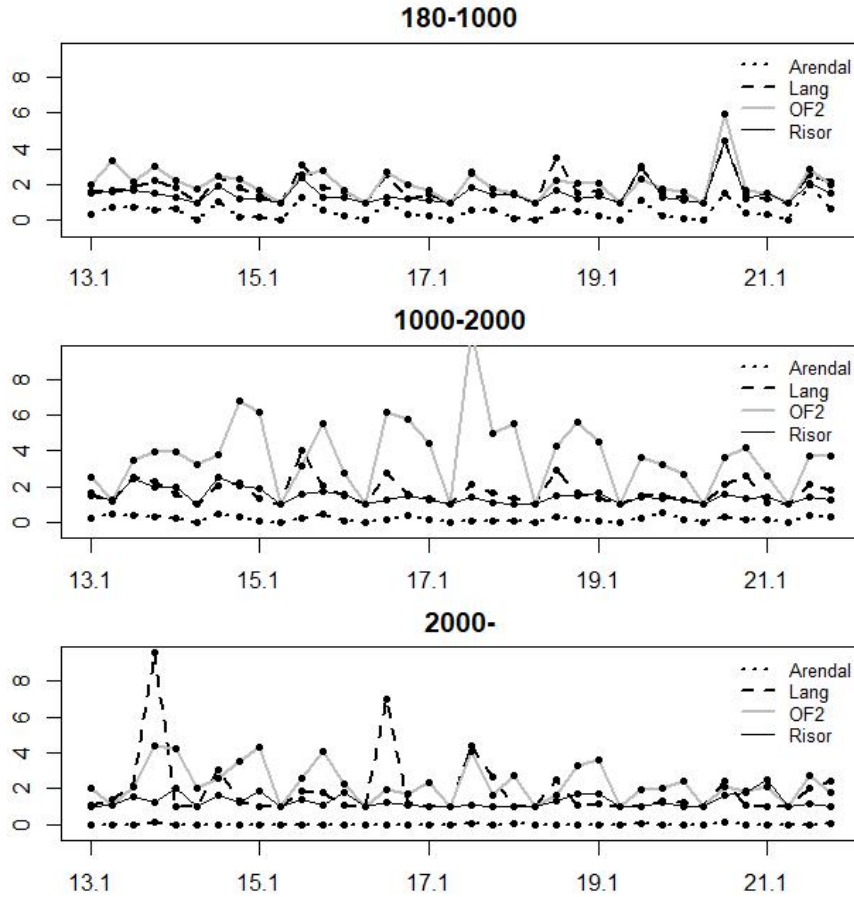


FIG 9. Plots for the quarterly time series data of zooplankton biomass for each size group ($180 \mu\text{m} - 1000 \mu\text{m}$, $1000 \mu\text{m} - 2000 \mu\text{m}$, and $> 2000 \mu\text{m}$), sampled at four sites in Skagerrak.

sampling volume from the bottom to the surface. Furthermore, deeper water layers often harbor larger zooplankton species (Aksnes et al. 1989), contributing to higher biomass values. This is confirmed by the high proportion of the size fraction $> 2000 \mu\text{m}$ at these sites. Risør is located inside the fjord sill and may have a somewhat different zooplankton community. The

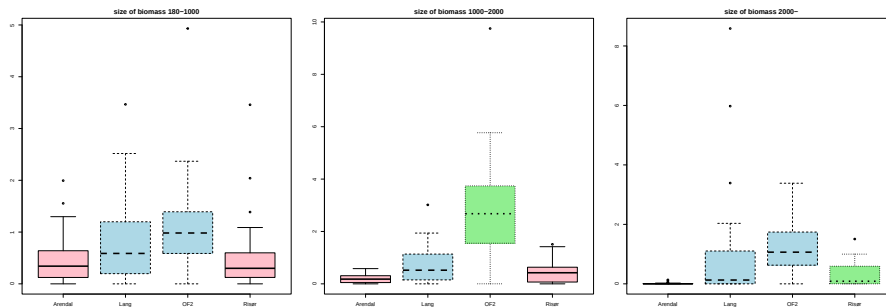


FIG 10. Boxplots for the biomass data. The horizontal corresponds to sites. The boxplots with solid borders are the subareas for which the null hypothesis was rejected by ANOVATS. The plots with dashed and dotted borders are areas that were divided into two subgroups within a group for which the null hypothesis was rejected. The boxplot for Arendal with biomass size $> 2000 \mu\text{m}$ has a box with a solid border.

TABLE 1
Sample means for each site according to different size fractions

size fractions [μm]	sites	sample mean [g/m^2]
180 – 1000	Arendal	0.480
	Langesund	0.800
	OF2	1.04
	Risør	0.497
1000 – 2000	Arendal	0.198
	Langesund	0.698
	OF2	2.69
	Risør	0.446
> 2000	Arendal	0.011
	Langesund	0.886
	OF2	1.289
	Risør	0.289

TABLE 2
Sampling depths and positions of sampling stations

sites	sampling depth	position
Arendal	50 m	in the coastal current
Langesund	200 m	outside the fjord sill
OF2	350 m	outside the fjord in the coastal current
Risør	160 m	inside the fjord sill

presence of a sill at the mouth of the fjord serves as a barrier, limiting water exchange between the fjord and the adjacent coastal current, and generally the zooplankton populations of the fjord differ significantly from those of open ocean waters, with a decrease in the abundance of zooplankton towards the inner regions of the fjord (Salvanes et al. 1995). The boxplots of the data are shown in Figure 10. The boxplots with solid borders correspond to the sites for which the null hypothesis was rejected. The boxplots clearly show the grouped areas have larger means than those of the rejected areas from ANOVATS. The results are supported by the following facts: Langesund and OF2 differ in depth and species composition. OF2 has a higher proportion of large zooplankton, and it is probably affected more by advection of water masses from NorC. OF2 is deeper (with a basin) and thus a better habitat for large plankton like euphausiids and copepods.

These results estimated by ANOVATS are supported by multiple sources in the literature and are consistent with biological/ecological facts. In the field of marine ecology, data analysis using ANOVATS represents the first attempt to statistically prove these facts. Variations in marine resource abundance based on regional differences are presented within a simple visualization and descriptive statistical framework not only for the North Sea but also for the Barents Sea (Eriksen et al. 2025) and the Norwegian Sea (ICES 2025). Data analysis using ANOVATS must be pursued to establish the statistical credibility of these variations.

5. Conclusion. We provided a subsampling-based test, ANOVATS (ANOVA for small-sample time series data), for exploring regional differences. While the existing ANOVA method for time series data requires estimation of spectral density, which is not appropriate for small-sample time series data, our proposed method avoids spectral estimation and achieves better size control, even for small-sample time series data. Time series data often contains trend and/or seasonality. Our approach is extend to take these properties into account.

There are only a few decades of annual data available for studying the impact of climate change on the diverse species of marine resource and environmental changes caused by human activity in IEA. The proposed method is practical for comparing the mean levels of these changes based on short time series data. This study shows an example of the application of zooplankton data. The differences among strata are attributable to topographical factors, and although they are shown qualitatively, it has become possible to show statistically significant differences without prior information using the proposed method. The fact that the output from our proposed method is supported by previous studies also indicates that this method can be used to verify whether the data contain any artifacts. The North Sea is considered to be particularly vulnerable to oil activities in the Northern Sea area (Lundberg, Harris and Trout 2024), but there are many fish groups in each strata that should be investigated for their biological production, diversity, and ecosystems. This information is stored in multiple databases, such as the trawl surveys database by ICES (Belgrano et al. 2023) and an indicator assessment based on the Oslo and Paris Commissions (Commission 2022). Investigations on ecosystem status based on subregions are also conducted in several sea areas (e.g., ICES 2022b), and it is possible to apply our approach to the differences between ecological regions to investigate the biodiversity for each species. The approach we proposed has potential to be a standard method within the IEA. By adding quantitative comparisons of time series data obtained from ANOVATS, the outcomes are incorporated into risk assessments in IEA, enabling more reliable implementation of ecosystem-based management.

Acknowledgments. We would like to thank Prof. Li Wang and Prof. Xingzhong Xu for kindly sharing a copy of their paper with us. We also thank the captains and crews of the various research vessels of the Institute of Marine Research, as well as the many people involved in collecting and processing the data here analyzed. We are grateful to Edanz (<https://jp.edanz.com/ac>) for editing the English text of a draft of this manuscript. We used ChatGPT (based on GPT-3.5, developed by OpenAI) for the purposes of language improvement and coding assistance.

Funding. The first author gratefully acknowledge JSPS Grant-in-Aid for Early-Career Scientists JP23K16851 and Research Fellowship Promoting International Collaboration of the Mathematical Society of Japan.

REFERENCES

- AKSNES, D. L., AURE, J., KAARTVEDT, S., MAGNESEN, T. and RICHARD, J. (1989). Significance of advection for the carrying capacities of fjord populations. *Marine Ecology Progress Series* **50** 263-274.
- AOSHIMA, M. and YATA, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Ann. Inst. Statist. Math.* **66** 983-1010.
- BELGRANO, A., FALKENHAUG, T., FRANZE, G., GLYKI, E., HAMON, K., HASSELLOV, I.-M., JONGBLOED, R., JUDD, A., KENNY, A., LYNAM, C., MARTINEZ, I., MASON, C., MOTOVA, A., OLSEN, E., PARMENTIER, K., PIET, G., SCHARFF-OLSEN, C. H., SKOEN, M., SOLVANG, H., STAGE, J., TAMIS, J., TOMCZAK, M., VANAVERBEKE, J., VON HELLFELD, R. and WEHDE, H. (2023). Working group on integrated assessments of the North Sea (WGINOSE). *ICES Scientific Reports* **5**.
- BERG, A., McMURRY, T. L. and POLITIS, D. N. (2010). Subsampling p-values. *Stat. Probab. Lett.* **80** 1358-1364.
- BRILLINGER, D. R. (1980). *Handbook of Statistics, Analysis of variance and problems under time series models*. Elsevier.
- BRILLINGER, D. R. (1981). *Time series: Data Analysis and Theory*. San Francisco: Holden-Day.
- CALIŃSKI, T. and CORSTEN, L. (1985). Clustering means in ANOVA by simultaneous testing. *Biometrics* 39-48.
- CARLSTEIN, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* 1171-1179.

- CLARKE, B. R. (2008). *Linear Models: The Theory and Application of Analysis of Variance*. John Wiley & Sons.
- CLAY, P. M., FERRETTI, J., BAILEY, J. L., GOTI, L., DANKEL, D. J., SANTURTUN, M., FULLER, J., SEBASTIAN, L., SCHMIDT, J., NIELSEN, K. N., GOLDSBOROUGH, D., GROENEVELD, R., FRAGA, A. R., ELEGBEDE, L. and RÖCKMANN, C. (2023). Implementation of integrated ecosystem assessments in the International Council for the Exploration of the Sea - conceptualizations, practice, and progress. *ICES Journal of Marine Science* **80** 1516–1528.
- COMMISSION, O. (2022). CEMP Guideline - FW3 Combined guideline for processing of survey data for fish and food webs common indicators FC2, FC3, FW3 and FW7. *OSPAR Agreement 2018-05*.
- DETTE, H. and PAPARODITIS, E. (2009). Bootstrapping frequency domain tests in multivariate time series with an application to comparing spectral densities. *J. Roy. Stat. Soc. Ser. B* **71** 831–857.
- TRANTER, E. D. J. and FRASER, J. H. (1968). *Zooplankton sampling. Monographs on oceanographic methodology*. the United Nations Educational Scientific and Cultural Organization, Paris.
- DUDEK, A. E. and LENART, Ł. (2023). Spectral density estimation for nonstationary data with nonzero mean function. *J. Amer. Statist. Assoc.* **118** 1900–1910.
- EDWARDS, A. W. and CAVALLI-SFORZA, L. L. (1965). A method for cluster analysis. *Biometrics* 362–375.
- ERIKSEN, E., HUSSON, B., SKARET, G., INGVALDSEN, R. B., DALPADADO, P., JOHANNSEN, E., JØRGENSEN, L. L., BOGSTAD, B., DOLGOV, A. V., PROZORKEVICH, D. V., PROKHOROVA, T. A., RUSSKIKH, A. A., STRELKOVA, N. A., TROFIMOV, A. G., PROKOPCHUK, I. P. and FILIN, A. A. (2025). The living Barents Sea response to peak-warming and subsequent cooling. *Scientific reports* **15**.
- FRANSZ, H. G., COLEBROOK, J. M., GAMBLE, J. C. and KRAUSE, M. (1991). The zooplankton of the north sea. *Netherlands Journal of Sea Research* **28** 1–52.
- FRANZE, G., FALKENHAUG, T., ALBRETSEN, J., LERCH, S. J., HÖFFLE, H., HELDAL, H. E., ERSHOVA, E. and GUNDERSEN, K. (2024). North Sea Ecosystem Cruise 2023. Cruise No. 2023002006.
- FUKUCHI, J.-I. (1999). Subsampling and model selection in time series analysis. *Biometrika* **86** 591–604.
- GAO, S., HJØLLO, S. S., FALKENHAUG, T., STRAND, E., EDWARDS, M. and SKOGEN, M. D. (2021). Overwintering distribution, inflow patterns and sustainability of *Calanus finmarchicus* in the North Sea. *Progress in Oceanography* **194**.
- GOTO, Y., KLEY, T., VAN HECKE, R., VOLGUSHEV, S., DETTE, H. and HALLIN, M. (2022). The integrated copula spectrum. *The Annals of Statistics* **50** 3563–3591.
- GOTO, Y., ARAKAKI, K., LIU, Y. and TANIGUCHI, M. (2023a). Homogeneity tests for one-way models with dependent errors. *TEST* **32** 163–183.
- GOTO, Y., ZHANG, X., KEDEM, B. and CHEN, S. (2023b). Residual Spectrum Applied in Brain Functional Connectivity. *arXiv preprint arXiv:2305.19461*.
- GOTO, Y., ARAKAKI, K., LIU, Y. and TANIGUCHI, M. (2023c). Supplement to “Homogeneity tests for one-way models with dependent errors”.
- GUO, W., DAI, M., OMBAO, H. C. and VON SACHS, R. (2003). Smoothing spline ANOVA for time-dependent spectral analysis. *Journal of the American Statistical Association* 643–652.
- HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton University Press.
- HARDISON, S., PERRETTI, C. T., DEPIPER, G. S. and ANDREW, B. (2019). A simulation study of trend detection methods for integrated ecosystem assessment. *ICES Journal of Marine Science* **76** 2060–2069.
- HASSEL, A., ENDRESEN, B., MARTINUSSEN, M., KNUTSEN, T. and JOHANNESSEN, M. E. (2013). *Håndbok for forskningsgruppe Plankton. Prøvetaking og analyse. Prosedyrer for prøvetaking og pre-analyse av dyre- og planteplankton på forskningsfartøy og i laboratorium på land, Version 6.2*. Institute of Marine Research, Bergen.
- HÁTÚN, H., DA SILVA, GUESTA, T. S., SKAGSETH, O. and GRØNKKJÆR, P. (2020). Editorial: Physical drivers of biogeographical shifts in the Northeastern Atlantic - and adjacent shelves. *Frontiers in Marine Science* **9**:1062204.
- HAYTER, A. J. (1986). The maximum familywise error rate of Fisher’s least significant difference test. *J. Amer. Statist. Assoc.* **81** 1000–1004.
- HEAGERTY, P. J. and LUMLEY, T. (2000). Window subsampling of estimating functions with application to regression models. *J. Amer. Statist. Assoc.* **95** 197–211.
- HEATH, M. R., BACKHAUS, J. O., RICHARDSON, K., MCKENZIE, E., SLAGSTAD, D., BEARE, D., DUNN, J., FRASER, J. G., GALLEGU, D. ALEJANDRO AND HAINBUCHER, HAY, S., JONASDOTTIR, S., MADDEN, H., MARDALJEVIC, J. and SCHACHT, A. (1999). Climate fluctuations and the spring invasion of the North Sea by *Calanus finmarchicus*. *Fisheries Oceanography* **8** 163–176.
- ICES (2018). Report of the Workshop on Integrated Trend Analyses in Support to Integrated Ecosystem Assessment (WKINTRA). *ICES Document CM2018/IEASG:15* 23pp.
- ICES (2022a). Third Workshop on Integrated Trend Analysis to support Integrated Ecosystem Assessment (WKINTRA-3). *ICES Scientific Report* **4** 23pp.

- ICES (2022b). Working group on the integrated assessment of the Barents Sea (WGIBAR). *ICES Scientific Reports* **4**.
- ICES (2025). Working group on the integrated assessment of the Norwegian Sea (WGINOR). *ICES Scientific Reports* **7**.
- INGLOT, T. and LEDWINA, T. (2006). Asymptotic optimality of new adaptive test in regression model. *Ann. inst. Henri Poincaré (B) Probab. Stat.* **42** 579–590.
- JENTSCH, C. and PAULY, M. (2015). Testing equality of spectral densities using randomization techniques. *Bernoulli* **21** 697–739.
- KITAGAWA, G. (2021). *Introduction to time series modeling with applications in R*, 2nd ed ed. CRC Press, Taylor & Francis Group.
- KITAGAWA, G. (2023). TSSS: Time Series Analysis with State Space Model.
- KRAUSE, M., FOCK, H., GREVE, W. and WINKLER, G. (2003). Northsea zooplankton: a review. *Senckenbergiana maritima* **33** 71–204.
- LEVIN, P. S., FOGARTY, M. J., MURAWSKI, S. A. and FLUHARTY, D. (2009). Integrated Ecosystem Assessments: Developing the scientific basis for Ecosystem-Based Mangement of the ocean. *PLOS Biology* **7**.
- LUNDBERG, S. A., HARRIS, R. and TROUT, K. (2024). Troubled Waters: How North Sea Countries Are Fueling Climate Disaster Technical Report, Oil Change International.
- MCMURRY, T. L., POLITIS, D. N. and ROMANO, J. P. (2012). Subsampling Inference with K Populations and a Non-standard Behrens–Fisher Problem. *Int. Stat. Rev.* **80** 149–175.
- MEYER, M., PAPARODITIS, E. and KREISS, J.-P. (2020). Extending the validity of frequency domain bootstrap methods to general stationary processes. *Ann. Statist.* **48** 2404–2427.
- MOTODA, S. (1959). Devices of simple plankton apparatus. *Memories of the Faculty of Fisheries Hokkaido University* **7** 73–94.
- NAGAHATA, H. and TANIGUCHI, M. (2018). Analysis of variance for high-dimensional time series. *Stat. Inference Stoch. Process.* **21** 455–468.
- OHLSSON, M. and EKLÖF, A. (2020). Spatial resolution and location impact group structure in a marine food web. *Ecology Letters* **ele.13567**.
- POLITIS, D. N. and ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* 2031–2050.
- POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer.
- SAHAI, H. and OJEDA, M. M. (2004). *Analysis of Variance for Random Models, Volume I: Balanced Data Theory, Methods, Applications and Data Analysis*. Springer.
- SALVANES, A. G. V., AKSNES, D. L., FOSSÅ, J. H. and GISKE, J. (1995). Simulated carrying capacities of fish in Norwegian fjords. *Fisheries Oceanography* **4** 17–32.
- SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 507–512.
- SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. New York: Wiley.
- SHUMWAY, R. H. and STOFFER, D. S. (2006). *Time Series Analysis and Its Application With R Examples*, 2nd ed. Springer, New York.
- SOLVANG, H. K. and ARNEBERG, P. (2024). Flagged observation analyses as a tool for scoping and communication in Integrated Ecosystem Assessments. *PLOS ONE* **19**.
- SOLVANG, H. K. and OHISHI, M. (2023). trec: An R package for trend estimation and classification to support integrated ecosystem assessment of the marine ecosystem and environmental factors. *SoftwareX* **21**.
- SOLVANG, H. K. and SUBBEY, S. (2019). An improved methodology for quantifying causality in complex ecological systems. *PLOS ONE* **14**.
- TEAM, R. C. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- WANG, L. and XU, X. (2014). Global testing method for clustering means in ANOVA. *J Korean. Stat. Soc.* **43** 381–392.
- YU, H., KAISER, M. S. and NORDMAN, D. J. (2023). A subsampling perspective for extending the validity of state-of-the-art bootstraps in the frequency domain. *Biometrika* **110** 1099–1115.
- YU, H., KAISER, M. S. and NORDMAN, D. J. (2024). A blockwise empirical likelihood method for time series in frequency domain inference. *Ann. Statist.* **52** 1152–1177.

APPENDIX A: SIMULATION STUDY

A.1. Data generating process. To evaluate finite sample performance of the proposed test, we consider two cases: Cases 1 and 2 correspond to independent and correlated groups,

respectively. For each case, we use Processes 1–4, which were considered in [Goto et al. \(2023a\)](#), as disturbance processes of the one-way random effects model defined in (1) with $p = 1$. For both Cases 1 and 2, Processes 1–3 are defined as the moving average model of order 1 processes $e_t = \nu_t + \Psi \nu_{t-1}$, where $e_t = (e_{1t}^\top, \dots, e_{at}^\top)^\top$, and Process 4 is defined as the GARCH process

$$e_{it} = h_{it}^{1/2} \nu_{it}, \quad i = 1, \dots, a, \quad \begin{pmatrix} h_{1t} \\ \vdots \\ h_{at} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + 0.1 \Psi \begin{pmatrix} e_{1t}^2 \\ \vdots \\ e_{at}^2 \end{pmatrix} + \begin{pmatrix} 0.1 h_{1,t-1} \\ \vdots \\ 0.1 h_{a,t-1} \end{pmatrix}.$$

Case 1 Let Ψ be $0.5I_a$. Each component of ν_t , which is independent of the other components, follows the standard normal for Processes 1 and 4, the centered t-distribution with five degrees of freedom for Process 2, and the centered skew normal distribution with location 0, scale 1, and shape 50 for Process 3.

Case 2 Let Ψ be a block diagonal matrix whose block diagonal matrices are given by

$$\begin{pmatrix} 0.7 & 0 & 0 \\ 0 & -0.5 & 0 \\ 0.3 & 0.1 & 0.3 \end{pmatrix}$$

and let $\Sigma^\nu = (\Sigma_{ij}^\nu)_{i,j=1,\dots,a}$ be an a -by- a covariance matrix such that $\Sigma_{ii}^\nu = 1$ and $\Sigma_{j(j+1)}^\nu = \Sigma_{(j+1)j}^\nu = 0.5$ for $i \in \{1, \dots, a\}$ and $j \in \{1, \dots, a-1\}$. The process $\{\nu_t\}$ follows the centered multivariate distribution with covariance matrix Σ^ν for Processes 1 and 4, the centered multivariate t-distribution with five degrees of freedom and scale matrix Σ^ν for Process 2, the centered multivariate skew normal distribution with location $\mathbf{0}_a$, correlation matrix Σ^ν , and shape $50\mathbf{1}_a$, where $\mathbf{0}_a$ and $\mathbf{1}_a$ are a -dimensional vectors whose components are all 0 and 1, respectively, for Process 3.

The MA model is indeed one of the simplest time series models, capable of capturing short-term dependencies and providing easily understandable simulation results. Conversely, the GARCH model is more sophisticated and specifically designed to address conditional heteroskedasticity.

A.2. Empirical size. To investigate the empirical size of our test, we set the number of groups as $a = 3, 9, 15$, sample size as $n = 20, 30, 50, 70, 100$, and subsampling block length $b = \lfloor cn^{1/3} \rfloor$, where $c \in \{1, 1.5, 2, 2.5, 3, 4, 5, 6\}$. Let $p = 1$. Then for given a , n , and b , generate time series z_{it} defined in (1), where $\mu = 0$, $\psi_i = 0$ for all i , and e_t as described in Section and compute the p -value. Iterate 200 times and calculate the empirical size with significance level $\varphi = 0.05$.

Because the constant $c = 2.5$ provides the best performance among the considered range, we fix the block length as $b = \lfloor 2.5n^{1/3} \rfloor$. Figure 11 shows the empirical sizes. The horizontal of the subplots corresponds to the time series length. The empirical size tends to get closer to 0.05 as the time series length increases. Even for small time series lengths, the performance of our test is reasonably good. From Figures 1 and 2 of [Goto et al. \(2023c\)](#), we know the classical tests based on S'_n and S''_n exhibit size distortion owing to between-area correlations and the number of areas, respectively, even for $n = 1000$ and $n = 2000$. As expected, our proposed test shows good size control for small-sample time series. Apparently, there is no significant difference in the results for Cases 1 and 2.

A.3. Empirical power. Next, we investigate the empirical power of our test. Setting $a = 6$, $p = 1$, sample sizes $n \in \{20, 30, 50, 70, 100\}$, subsampling block length $b = \lfloor 2.5n^{1/3} \rfloor$, we generate time series z_{it} defined in (1), where $\mu = 0$, $(\psi_1, \dots, \psi_6) := (0, 0, 0, 1, 1, 1)^\top$, and e_t

as described in Section A.1. There are two clusters in this setting. Then we apply our method described in Section 2.3 with significance level $\varphi = 0.05$ and repeat 200 times. Let Area_i denote the area name of $\{z_{it}\}$.

The first and second rows of Figure 12 correspond to Cases 1 and 2, respectively.

The first column of Figure 12 shows the empirical probability of rejecting $H_0 : \psi_1 = \dots = \psi_6$ and dividing the areas ($\text{Area}_1, \dots, \text{Area}_6$) into two groups ($\text{Area}_1, \dots, \text{Area}_3$) and ($\text{Area}_4, \dots, \text{Area}_6$). Our test has power even with a small time series length n , although the power is small when $n = 20$. This is attributable to the variance in the time series and the degree of separation between clusters. We see that the t-distribution provides a smaller power owing to the large fluctuation, which hides the regional effect.

The second column shows, for the data that were correctly divided in the first step, the empirical size for not rejecting the null hypothesis for both groups. The third column shows the number of cases for which the areas were divided correctly in the first step. As n increases, the size tends to approach or slightly exceed 0.05. When n is small, the size is less than 0.05, resulting in fewer trials and smaller observed sizes.

The fourth column shows the empirical probability of reaching the correct clustering results. When n is relatively large, correct clustering results are achieved. We observe that the empirical power for Case 2 is larger than that for Case 1.

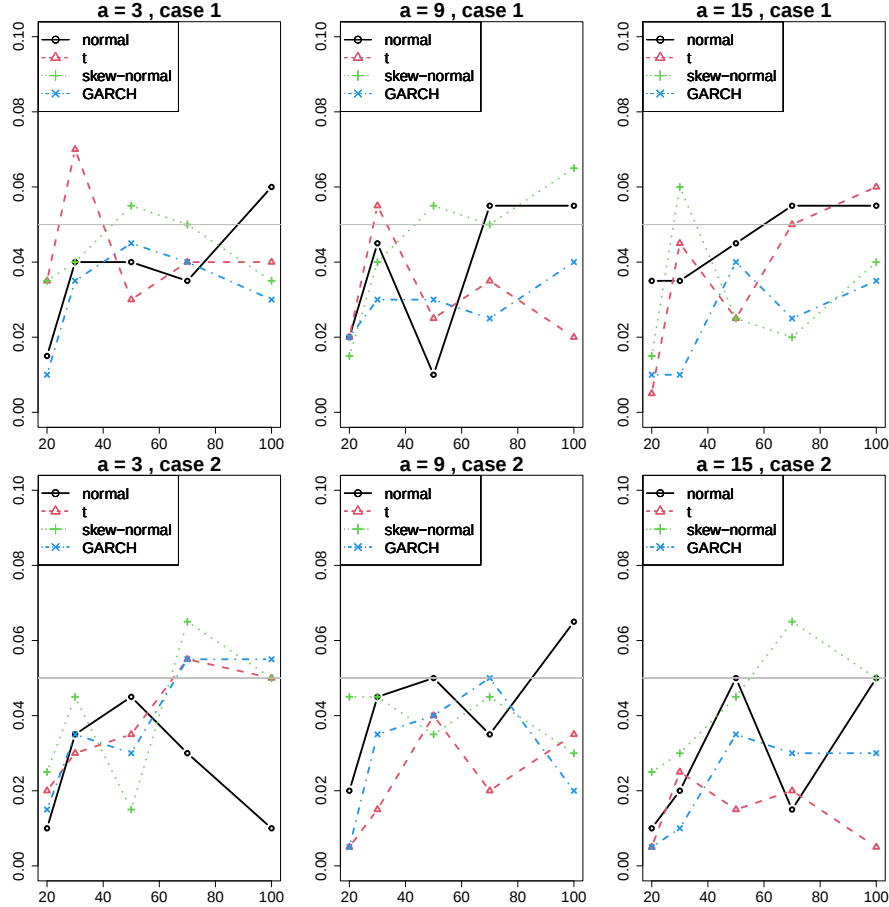


FIG 11. Empirical size of the proposed test. The horizontal and vertical axes correspond to time series length n and empirical size, respectively. The left, middle, and right columns correspond to $a = 3, 9, 15$, respectively. The top and bottom rows correspond to Cases 1 and 2, respectively.

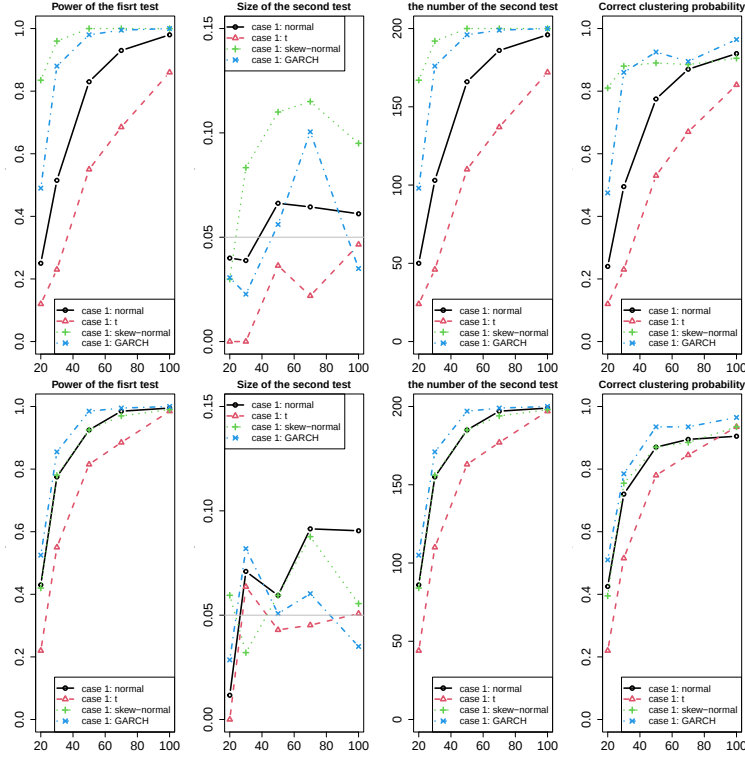


FIG 12. The empirical probability of rejecting the null and dividing the areas correctly in the first step (first column), the empirical size for not rejecting the null hypothesis for both groups for data that was correctly divided in the first step (second column), the number of cases for which the areas were divided correctly in the first step (third column), and the empirical probability of reaching the correct clustering results (fourth column). The horizontal and vertical axes correspond to time series length n and probability, respectively. The top and bottom rows correspond to Cases 1 and 2, respectively.

APPENDIX B: PROOFS

B.1. Proof of Theorem 2.1. This theorem can be proved in the same way as [Goto et al. \(2022, Proof of Theorem 4.2\)](#). The essential tool is [Politis, Romano and Wolf \(1999, Proposition 7.3.1\)](#), which implies that $\sup_{x \in \mathbb{R}} |H_{n,b}(x) - H(x)| = o_p(1)$, where

$$H_{n,b}(x) := \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} \mathbb{I} \left\{ \frac{b}{1-\frac{b}{n}} \sum_{i=1}^a (e_{i.,b,t} - e_{..,b,t})^\top (e_{i.,b,t} - e_{..,b,t}) \leq x \right\}$$

and H is the c.d.f. of the asymptotic distribution of T_n under the null with $e_{i.,b,t} = \sum_{j=t}^{t+b-1} e_{ij}/b$ and $e_{..,b,t} = \sum_{i=1}^a \sum_{j=t}^{t+b-1} e_{ij}/(ab)$. This yields that the proposed test has the asymptotic size φ .

Next, we show the consistency of the test. Note that

$$\begin{aligned} T_n &= n \sum_{i=1}^a (\psi_i - \psi_{\cdot} + e_{i.} - e_{..})^\top (\psi_i - \psi_{\cdot} + e_{i.} - e_{..}) \\ &= n \sum_{i=1}^a (\psi_i - \psi_{\cdot})^\top (\psi_i - \psi_{\cdot}) + 2n \sum_{i=1}^a (\psi_i - \psi_{\cdot})^\top (e_{i.} - e_{..}) \\ &\quad + n \sum_{i=1}^a (e_{i.} - e_{..})^\top (e_{i.} - e_{..}) \\ &= n (\zeta_{\psi,\psi} + 2\zeta_{\psi,e_n} + \zeta_{e_n,e_n}) \end{aligned}$$

and

$$T_{n,b,t} = \frac{b}{1-\frac{b}{n}} (\zeta_{\psi,\psi} + 2\zeta_{\psi,e_{b,t}} + \zeta_{e_{b,t},e_{b,t}}),$$

where

$$\begin{aligned} \zeta_{\psi,\psi} &:= \sum_{i=1}^a (\psi_i - \psi_{\cdot})^\top (\psi_i - \psi_{\cdot}), \quad \zeta_{\psi,e_n} := \sum_{i=1}^a (\psi_i - \psi_{\cdot})^\top (e_{i.} - e_{..}) \\ \zeta_{e_n,e_n} &:= \sum_{i=1}^a (e_{i.} - e_{..})^\top (e_{i.} - e_{..}), \quad \zeta_{\psi,e_{b,t}} := \sum_{i=1}^a (\psi_i - \psi_{\cdot})^\top (e_{i.,b,t} - e_{..,b,t}), \end{aligned}$$

$$\text{and } \zeta_{e_{b,t},e_{b,t}} := \sum_{i=1}^a (e_{i.,b,t} - e_{..,b,t})^\top (e_{i.,b,t} - e_{..,b,t})$$

with $e_{i.,b,t} = \sum_{j=t}^{t+b-1} e_{ij}/b$, $e_{..,b,t} = \sum_{i=1}^a \sum_{j=t}^{t+b-1} e_{ij}/(ab)$. It holds that

$$\begin{aligned} p_n &= \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} \mathbb{I}\{|T_{n,b,t}| > |T_n|\} \\ &\leq \frac{1}{n-b+1} \\ &\quad \times \sum_{t=1}^{n-b+1} \mathbb{I} \left\{ \frac{b}{1-\frac{b}{n}} |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_{b,t}}| + \frac{b}{1-\frac{b}{n}} |\zeta_{e_{b,t},e_{b,t}}| > n |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_n}| - n |\zeta_{e_n,e_n}| \right\} \\ &= 1 - H_{n,b} \left(n |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_n}| - \frac{b}{1-\frac{b}{n}} |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_{b,t}}| - n |\zeta_{e_n,e_n}| \right) \end{aligned}$$

$$= 1 - H \left(n |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_n}| - \frac{b}{1 - \frac{b}{n}} |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_{b,t}}| - n |\zeta_{e_n,e_n}| \right) + o_p(1).$$

Since $\zeta_{\psi,e_n} = O_p(1/\sqrt{n})$, $\zeta_{\psi,e_{b,t}} = O_p(1/\sqrt{b})$, and $\zeta_{e_n,e_n} = O_p(1/n)$, there exists $M > 0$ such that for all $n \in \mathbb{N}$,

$$P(\sqrt{n} |2\zeta_{\psi,e_n}| > M) < \epsilon, \quad P(\sqrt{b} |2\zeta_{\psi,e_{b,t}}| > M) < \epsilon, \quad \text{and } P(n |\zeta_{e_n,e_n}| > M) < \epsilon.$$

Therefore, we obtain, for any $\epsilon' > 0$,

$$\begin{aligned} & P(p_n > \epsilon') \\ & \leq P \left(1 - H \left(n |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_n}| - \frac{b}{1 - \frac{b}{n}} |\zeta_{\psi,\psi} + 2\zeta_{\psi,e_{b,t}}| - n |\zeta_{e_n,e_n}| \right) > \epsilon' \right) + o(1) \\ & \leq P \left(1 - H \left(n \zeta_{\psi,\psi} - n |2\zeta_{\psi,e_n}| - \frac{b}{1 - \frac{b}{n}} |\zeta_{\psi,\psi}| - \frac{b}{1 - \frac{b}{n}} |2\zeta_{\psi,e_{b,t}}| - n |\zeta_{e_n,e_n}| \right) > \epsilon' \right) + o(1) \\ & \leq P \left(1 - H \left(\left(n - \frac{b}{1 - \frac{b}{n}} \right) \zeta_{\psi,\psi} - \left(\sqrt{n} + \frac{\sqrt{b}}{1 - \frac{b}{n}} + 1 \right) M \right) > \epsilon' \right) + 3\epsilon + o(1). \end{aligned}$$

Since there exists $N \in \mathbb{N}$ such that for $n \geq N$,

$$P \left(1 - H \left(\left(n - \frac{b}{1 - \frac{b}{n}} \right) \zeta_{\psi,\psi} - \left(\sqrt{n} + \frac{\sqrt{b}}{1 - \frac{b}{n}} + 1 \right) M \right) > \epsilon' \right) < \epsilon,$$

we obtain $p_n = o_p(1)$. $\square$

B.2. Proofs of the convergences in (5). From the proof of Theorem 2.1, p_n converges in distribution to a uniform distribution on $(0,1)$ under H_0 . Also, $p_n - \varphi_n$ converges in distribution to the same, provided $\varphi_n \rightarrow 0$, and thus $P(p_n < \varphi_n) \rightarrow 0$ as $n \rightarrow \infty$.

Next, we consider the case under H_1 . First, we note that Lemma 1 of [Ingolot and Ledwina \(2006\)](#) yields, for $a \geq 2$ and $u > a$, that

$$P(\chi_a^2 \geq u) \leq \frac{1}{\sqrt{2\pi}} \frac{\sqrt{a}}{u - a + 2} \left(\frac{\exp(1)}{a} u \right)^{\frac{a}{2}} \exp\left(-\frac{u}{2}\right).$$

Let c_n denote the quantity

$$c_n := \left(n - \frac{b}{1 - \frac{b}{n}} \right) \zeta_{\psi,\psi} - \left(\sqrt{n} + \frac{\sqrt{b}}{1 - \frac{b}{n}} + 1 \right) M,$$

which is $O(n)$ as $n \rightarrow \infty$. Then, from the proof of Theorem 2.1, we have, under H_1 ,

$$\begin{aligned} P(p_n > \varphi_n) & \leq P(1 - H(c_n) > \varphi_n) + 3\epsilon + o(1) \\ & \leq P \left(\frac{1}{\sqrt{2\pi}} \frac{\sqrt{a}}{c_n - a + 2} \left(\frac{\exp(1)}{a} c_n \right)^{\frac{a}{2}} \exp\left(-\frac{c_n}{2}\right) > \varphi_n \right) + 3\epsilon + o(1), \end{aligned}$$

which, provided the condition $\varphi_n / \left(c_n^{\frac{a}{2}-1} \exp\left(-\frac{c_n}{2}\right) \right) \rightarrow \infty$ as $n \rightarrow \infty$, is $o(1)$. $\square$