



Comparison of Bootstrap Methods for Estimating Causality in Linear Dynamic Systems: A Review

Fumikazu Miwakeichi ^{1,2,*}  and Andreas Galka ³ ¹ Department of Statistical Modeling, The Institute of Statistical Mathematics, Tokyo 190-8562, Japan² Statistical Science Program, Graduate Institute for Advanced Studies, SOKENDAI, Tokyo 190-8562, Japan³ Clinic for Pediatric and Adolescent Medicine II, University Clinic, University of Kiel, 24105 Kiel, Germany; a.galka@pedneuro.uni-kiel.de

* Correspondence: miwake1@ism.ac.jp

Abstract: In this study, we present a thorough comparison of the performance of four different bootstrap methods for assessing the significance of causal analysis in time series data. For this purpose, multivariate simulated data are generated by a linear feedback system. The methods investigated are uncorrelated Phase Randomization Bootstrap (uPRB), which generates surrogate data with no cross-correlation between variables by randomizing the phase in the frequency domain; Time Shift Bootstrap (TSB), which generates surrogate data by randomizing the phase in the time domain; Stationary Bootstrap (SB), which calculates standard errors and constructs confidence regions for weakly dependent stationary observations; and AR-Sieve Bootstrap (ARSB), a resampling method based on AutoRegressive (AR) models that approximates the underlying data-generating process. The uPRB method accurately identifies variable interactions but fails to detect self-feedback in some variables. The TSB method, despite performing worse than uPRB, is unable to detect feedback between certain variables. The SB method gives consistent causality results, although its ability to detect self-feedback decreases, as the mean block width increases. The ARSB method shows superior performance, accurately detecting both self-feedback and causality across all variables. Regarding the analysis of the Impulse Response Function (IRF), only the ARSB method succeeds in detecting both self-feedback and causality in all variables, aligning well with the connectivity diagram. Other methods, however, show considerable variations in detection performance, with some detecting false positives and others only detecting self-feedback.

Keywords: causal analysis; Granger causality; bootstrap methods; multivariate time series; impulse response function



Citation: Miwakeichi, F.; Galka, A. Comparison of Bootstrap Methods for Estimating Causality in Linear Dynamic Systems: A Review. *Entropy* **2023**, *25*, 1070. <https://doi.org/10.3390/e25071070>

Academic Editors: Hector Zenil, Jiang Zhang and Peng Cui

Received: 4 June 2023

Revised: 14 July 2023

Accepted: 16 July 2023

Published: 17 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In many fields of scientific inquiry, the analysis of causal relationships in complex systems represents an important task, which depends on the synergistic interplay between theory and application. An important contribution in this area has been the proposal of Granger Causality (GC) [1], which tests the significance of dynamic influence between pairwise time series. Building upon this foundation, Geweke [2] expanded GC to encompass multivariate time series, introducing conditional Granger Causality (cGC). Bressler et al. [3] and Schiatti et al. [4] further refined cGC, enabling the estimation of causality among multivariate time series using Vector AutoRegressive (VAR) models.

Granger causality and its generalizations represent a statistical method that depends on the comparison of the residual variances resulting from applying different models to the same time series; a wide array of predictive models, including non-linear variants, may be employed. As an alternative approach to quantifying causality from predictive modeling, Impulse Response Function (IRF) analysis has been developed as a tool for examining variable responses to shocks within an identified model.

However, it is worth noting that Granger causality and IRF analysis always depend on the particular chosen model classes; they are not suitable for model discovery since they are lacking invariance theorems, as they exist for the more general concept of algorithmic complexity [5].

Conceptually, the statistical significance of numerically calculated cGC and IRF should be assessed using the asymptotic method. However, real-world data analysis may yield incorrect results due to potential distribution discrepancies. In cases where the distribution function is unknown, such as with partial Granger causality [6], the asymptotic method is inapplicable. The bootstrap strategy serves as a widely used alternative for cGC and IRF significance testing. However, with numerous bootstrap methods available, each with varying sensitivity and specificity, selecting the appropriate method remains a critical challenge when applying time series models to real data analysis.

In a causal analysis of time series data, failing to apply an appropriate bootstrap method can lead to several important issues. First, as causal analysis is based on the estimation of causal relationships, not using a suitable bootstrap method could potentially result in inaccuracies in estimating these relationships, thereby undermining the reliability of statistical inferences. Furthermore, bootstrap methods are often used for estimating confidence intervals, and if an unsuitable method is chosen, the confidence intervals could be under- or over-estimated, leading to inappropriate assessments of uncertainty. Additionally, bootstrap methods are used to estimate the magnitude of causal effects, and without a suitable method, the magnitude of these effects could be inaccurately estimated, which could affect the appropriateness of subsequent actions. Therefore, to avoid these problems, it is important to select and apply suitable bootstrap methods for causal analysis and choose methods based on understanding the characteristics of the time series data and the objectives of the causal analysis.

In this study, we thoroughly evaluate the performance of four bootstrap methods designed for time series analysis. The first method we present is an adaptation of the Phase Randomize Bootstrap (FRB), which generates surrogate data for hypothesis testing through phase randomization in the Fourier transform of original time series data. We propose a modification, the uncorrelated Phase-Randomized Bootstrap (uPRB), which randomizes phases for each variable in the spectral domain, for causality detection among variables. As a counterpart, we consider a technique based on Circular Block Bootstrap (CBB), a method for transfer randomization in the time domain [7]. This approach remedies the under-representation of the beginning and the end of the time series in surrogate data by creating a circular time series. Time-Shifted Surrogates (TSS) [8,9], another technique we discuss, ensures the preservation of all characteristics of the initial signal by maintaining the same state-space trajectory. The Stationary Bootstrap (SB) [10] uses random block lengths for standard error calculations and confidence region construction in weakly dependent stationary observations. Last, we discuss the AR-Sieve Bootstrap (ARSB), a method that approximates the underlying data-generating process by fitting an AR model, useful for handling dependent data [11,12].

2. Causal Analysis

2.1. Conditional Granger Causality (cGC)

Let $\mathbf{Y}_t = \{y_t^1, \dots, y_t^M\}^T$ denote the vector variables at time t , $1 \leq t \leq N$, where N denotes the length of the time series. The feedback system among the variables \mathbf{Y}_t can be represented by a basic Vector AutoRegressive (VAR) model of order p , defined as

$$\mathbf{Y}_t = \sum_{k=1}^p \mathbf{A}_k \mathbf{Y}_{t-k} + \mathbf{E}_t, \quad (1)$$

where \mathbf{A}_k denotes a set of $M \times M$ -dimensional coefficient matrices, and $\mathbf{E}_t = \{e_t^1, \dots, e_t^M\}^T$ denotes a series of shocks (disturbances), given by white noise vectors with zero means.

Extracting the l th variable of the VAR model gives

$$y_t^l = \sum_{k=1}^p A_k^l Y_{t-k} + e_t^l, \tag{2}$$

where $A_k^l = \{a_k^{l1}, a_k^{l2}, \dots, a_k^{lm}, \dots, a_k^{lM}\}$ denotes the l th row of A_k . Equation (2) represents an autoregressive model with exogenous input (ARX model) that consists of an endogenous part of the l th variable and an exogenous part of all other variables.

The ARX model, excluding the m th variable, is called the restricted ARX (rARX) model:

$$y_t^l = \sum_{k=1}^p \tilde{A}_k^l Y_{t-k}^{(m)} + e_t^{lm}, \tag{3}$$

where $Y_t^{(m)}$ denotes the vector Y_t with the m th element excluded (e.g., $Y_t^{(1)} = \{y_t^2, \dots, y_t^M\}^T$, $Y_t^{(2)} = \{y_t^1, y_t^3, \dots, y_t^M\}^T$ and so on) and \tilde{A}_k^l denotes a set of re-estimated $(M - 1)$ -dimensional coefficient vectors. Suppose that the past values of the m th variable $\{y_{t-1}^m, y_{t-2}^m, \dots, y_{t-p}^m\}$ contribute to the prediction of the l th variable $\{y_t^l\}$; then, the variances of the residuals should satisfy $\text{var}(e_t^l) < \text{var}(e_t^{lm})$. The significance of the difference in variances can be evaluated by a likelihood ratio test:

$$F_{m \rightarrow l} = \log \frac{\text{var}(e_t^{lm})}{\text{var}(e_t^l)}. \tag{4}$$

The null hypothesis of absence of conditional Granger causality from the m th to the l th element given $Y_t^m = \{y_t^1, y_t^2, \dots, y_t^p\}$ states that all a_k^{lm} were zero, i.e., $H_0 : a_1^{lm} = a_2^{lm} = \dots = a_p^{lm} = 0$. This implies that past values of Y_t^m do not improve the prediction of y_t^l . On the contrary, the rejection of this null hypothesis suggests that Y_t^m in Granger causes y_t^l .

2.2. Impulse Response Function (IRF)

There is another representation of the VAR model of Equation (1) using a delay operator L , such that $LY_t = Y_{t-1}$:

$$A(L)Y_t = E_t, \tag{5}$$

$$A(L) = I_M - A_1L - A_2L^2 - \dots - A_pL^p. \tag{6}$$

where I_M denotes the $M \times M$ -dimensional unity matrix. The roots (eigenvalues) λ_j of the polynomial $A(L)$ need to fulfill $|\lambda_j| < 1$ for all j .

Further transformation of Equation (5) yields

$$\begin{aligned} Y_t &= A(L)^{-1}E_t \\ &= \{I_M + \Psi_1L + \Psi_2L^2 + \dots\}E_t. \end{aligned} \tag{7}$$

which demonstrates that a VAR model can be rewritten as

$$Y_t = \sum_{k=0}^{\infty} \Psi_k E_{t-k} = \Psi(L)E_t, \tag{8}$$

where

$$\begin{aligned} \Psi(L) &= A(L)^{-1} \\ &= \{I_M - A_1L - A_2L^2 - \dots - A_pL^p\}^{-1} \\ &= I_M + \Psi_1L + \Psi_2L^2 + \dots \end{aligned}$$

Equation (8) represents the Vector Moving Average (VMA) representation of the VAR model and can be denoted as $VMA(\infty)$.

Partial differentiation of the $VMA(\infty)$ model with respect to one particular shock (disturbance term) E_t yields a set of derivatives

$$\frac{\partial Y_t}{\partial E_t} = I_M, \frac{\partial Y_{t+1}}{\partial E_t} = \Psi_1, \dots, \frac{\partial Y_{t+s}}{\partial E_t} = \Psi_s. \tag{9}$$

The (l, m) th element of Ψ_s is given by $\frac{\partial y_{t+s}^l}{\partial e_t^m}$, which represents the marginal effect (influence) from the m th shock e_t^m to the l th variable y_{t+s}^l . The function defining the time series of such marginal effects is called the Impulse Response Function (IRF) and is denoted by IRF_s^{lm} .

The null hypothesis for the IRF states that a one-time shock to the m th variable has no impact on future values of the l th variable. This can be represented as: $H_0 : IRF_s^{lm} = 0$ for all $s > 0$. The rejection of this null hypothesis suggests that a one-time shock to the m th variable significantly affects the l th variable at some future point in time.

3. Bootstrap Methods

3.1. Uncorrelated Phase Randomization Bootstrap (uPRB)

The Phase Randomization Bootstrap method has been proposed as a technique to generate surrogate time series data for hypothesis testing [13,14]. Using the Fourier transform, the method transforms the original time series data into the frequency domain. Then, the phase of each frequency component is randomized while preserving the original amplitude spectrum. Finally, the surrogate data are obtained by applying the inverse Fourier transform to the randomized frequency components.

This process generates surrogate time series data with the same power spectrum as the original data but with disrupted temporal correlations. By comparing the original data's non-linearity measures or other statistical properties to those of the surrogate data, researchers can test hypotheses and detect the presence of non-linear dynamics in the original time series data. This method has been applied in various fields, including the study of physiological signals [15], economics and finance [16], climate data [9], and so on.

In order to preserve all linear auto-correlations and cross-correlations, the Phase Randomization Bootstrap adds a common random sequence $\varphi(f)$ to the phases of all variables. Thus, since this method cannot detect causality among variables, we prepare random sequences independently for each variable. In this paper, we call this method uncorrelated Phase Randomization Bootstrap (uPRB).

The procedure for generating surrogate datasets by uPRB is as follows.

Step 1 Transform the original time series by applying Fourier transform to each variable as

$$\mathcal{Y}^l(f) = \mathcal{F}\{y_t^l\} = A^l(f)e^{i\phi^l(f)},$$

where l is the index of the variable, and $A^l(f)$ and $\phi^l(f)$ denote the amplitude and the phase, respectively.

Step 2 For each frequency f , add an independent random value $\varphi^l(f)$ following a uniform distribution throughout the interval $[0, 2\pi)$ to $\phi^l(f)$, while satisfying the symmetry property $\varphi^l(f) = -\varphi^l(-f)$. That is,

$$\tilde{\mathcal{Y}}^l(f) = A^l(f)e^{i[\phi^l(f)+\varphi^l(f)]}.$$

Step 3 Transform the spectral domain representation back to the time domain by applying the inverse Fourier transform to each variable as

$$y_t^{*l} = \mathcal{F}^{-1}\{\tilde{\mathcal{Y}}^l(f)\} = \mathcal{F}^{-1}\{\mathcal{Y}^l(f)e^{i\varphi^l(f)}\}.$$

Step 4 Repeat Steps 2–3 for all variables.

Step 5 Repeat Steps 2–4 a large number of times, thereby generating a set of surrogate datasets.

3.2. Stationary Bootstrap (SB)

The conventional Non-overlapping Block Bootstrap (NBB) has been proposed by Carlstein [17]. By improving this conventional method, Künsch [18] has proposed a Moving-Blocks Bootstrap (MBB). This method is useful, especially for small sample data having a wider range of blocks than the conventional method. However, in the process of random sampling, there is an edge effect of the uneven weighting of the selection at the beginning and end of the data. To compensate for this shortcoming, Politis and Romano [7] proposed a Circular Block Bootstrap (CBB) that concatenates the start and end points of the original data. In the block bootstrap method, the stationarity of the sample is an important assumption. However, surrogate data obtained by resampling using the above-mentioned method are not necessarily stationary. Therefore, the Stationary Bootstrap (SB) method has been proposed, in which the surrogate data are also stationary. This method is similar to the TSS method, except that the block width is not fixed but is resampled as a random variable following a geometric distribution [10]. The procedure for generating surrogate data by SB is as follows.

- Step 1** Set the mean block width to w . Then, in the geometric distribution used in Step 3, we have $p = 1/w$.
- Step 2** Duplicate the original data and merge it to the end of the original data such that $Y_{N+1} = Y_1, Y_{N+2} = Y_2, \dots, Y_{2N} = Y_N$.
- Step 3** Generate a sequence of natural random numbers $L_1^{*b}, \dots, L_K^{*b}$ corresponding to each block width following a geometric distribution so that the probability of the event $L_l^{*b} = r$ is $(1-p)^{r-1}p$ for $r = 1, 2, \dots$. Here, the value of K is determined to satisfy the condition $K = \min\{k : \sum_{l=1}^k L_l^{*b} \geq N\}$.
- Step 4** Generate a sequence of natural random numbers $I_1^{*b}, \dots, I_K^{*b}$, corresponding to the index of the starting point of the block, following a uniform distribution over the interval $[1, N]$.
- Step 5** The blocks, $\{\xi^{*b}(I_1^{*b}, L_1^{*b}), \dots, \xi^{*b}(I_K^{*b}, L_K^{*b})\}$, constructed according to Steps 1 and 2, are arranged in the order in which they were extracted, and a pair of resamples is obtained with the first N elements as $Y_1^{*b}, \dots, Y_N^{*b}$.
- Step 6** Repeat Steps 3–5 a large number of times, thereby generating a set of surrogate datasets.

In the case of multivariate time series, there are two ways to perform Steps 3–5. One way is to use the same blocks $\{\xi^{*b}(I_1^{*b}, L_1^{*b}), \dots, \xi^{*b}(I_K^{*b}, L_K^{*b})\}$ for all variables for data shuffling, as described above, and the other is to perform Steps 3–5 for each variable independently. In this paper, we refer to the former as correlated SB (cSB) and to the latter as uncorrelated SB (ucSB).

3.3. Time Shift Surrogates (TSS)

While PRB is a method for generating surrogate datasets by randomizing the phases in the frequency domain, we evaluated the performance of TSS as a method of randomizing the phase in the time domain. This method corresponds to the special case of the CBB, where the number of blocks is limited to 1. The procedure for generating surrogate datasets by TSS is as follows.

- Step 1** Duplicate the original data and merge it to the end of the original data such that $Y_{N+1} = Y_1, Y_{N+2} = Y_2, \dots, Y_{2N} = Y_N$.
- Step 2** Generate a natural random number s following a uniform distribution over the interval $[I_a, I_b]$, extract a sequence of the data $\{y_s^l, \dots, y_{N+s}^l\}$, and use it as surrogate dataset for the l th variable.
- Step 3** Repeat Step 2 for all variables.

Step 4 Repeat Steps 2–3 a large number of times, thereby generating a set of surrogate datasets.

3.4. AR-Sieve Bootstrap (ARSB)

The AR-Sieve Bootstrap (ARSB) method generates surrogate datasets by feeding a VAR model, which employs re-estimated parameters, with residuals from modeling [11,12]. The procedure for generating surrogate datasets using ARSB is as follows.

Step 1 Fit the VAR model of Equation (1) to the original time series, obtaining estimates for the parameters and the residuals \hat{E}_t .

Step 2 Compute centered residuals $\hat{E}_1 - \bar{E}, \dots, \hat{E}_N - \bar{E}$, where

$\bar{E} = N^{-1} \sum_{t=1}^N \hat{E}_t$, and generate bootstrap residuals E_1^*, \dots, E_N^* by shuffling the indices according to a sequence of natural random numbers $\{J_1^{*b}, \dots, J_N^{*b}\}$, which was drawn randomly with replacement from a uniform distribution over the interval $[1, N]$.

Step 3 Compute the surrogate time series recursively by

$$Y_t^* = \sum_{k=1}^p \hat{A}_k Y_{t-k}^* + E_t^*$$

where $(Y_1^*, \dots, Y_p^*) = (Y_1, \dots, Y_p)$.

Step 4 Re-estimate the VAR parameter matrices A_1, \dots, A_p based on the bootstrap time series.

Step 5 Repeat Steps 2–4 a large number of times, thereby generating a set of surrogate datasets.

Similar to the SB method, there are two ways to process Step 2 in the ARSB method. One is to use the same sequence of random values $\{J_1^{*b}, \dots, J_N^{*b}\}$ for all variables for data shuffling, as mentioned above, and the other is to perform Step 2 independently for each variable. In this paper, we refer to the former as correlated ARSB (cARSB) and to the latter as uncorrelated ARSB (ucARSB).

4. Simulation

In order to verify the performance of the methods for causal analysis and of the bootstrap methods, as discussed above, we prepared a simulation model by modifying a model proposed by [19]. We set up seven oscillatory variables as depicted in Figure 1, five of which generated their stochastic oscillations locally from driving white noise via self-feedback; furthermore, the seven variables were coupled directly or indirectly within a global Vector AutoRegressive (VAR) process, such that a dynamic feedback system results. We fixed the frequency of each variable to 0.1 Hz and randomly selected a damping coefficient for each variable from a normal distribution. The sixth and seventh variables were isolated from the other five. For instance, the first and third variables were directly connected, while the first and fourth variables were indirectly connected through the third variable. The fourth and fifth variables were directly and bidirectionally connected.

The sixth variable generated a local oscillation at 0.1 Hz via self-feedback, like the other five, while the seventh variable did not have any self-feedback, but was driven only by the sixth variable. If causality would be detected between either the sixth or seventh variables and any of the first to fifth variables, this would represent a false positive.

We set the connectivity among variables as shown in Figure 1 and generate a simulated time series of length $N = 2000$ through the VAR process of Equation (1), with $p = 2$. The nominal sampling frequency was 10 Hz, and the series of shocks (disturbances) E_t was sampled from a 7-dimensional multivariate Gaussian noise distribution with zero mean and diagonal covariance matrix given by $\Sigma_E = 0.1I_7$.

Table 1 displays the parameters of the VAR model for the simulations. Figure 2 displays the simulated time series.

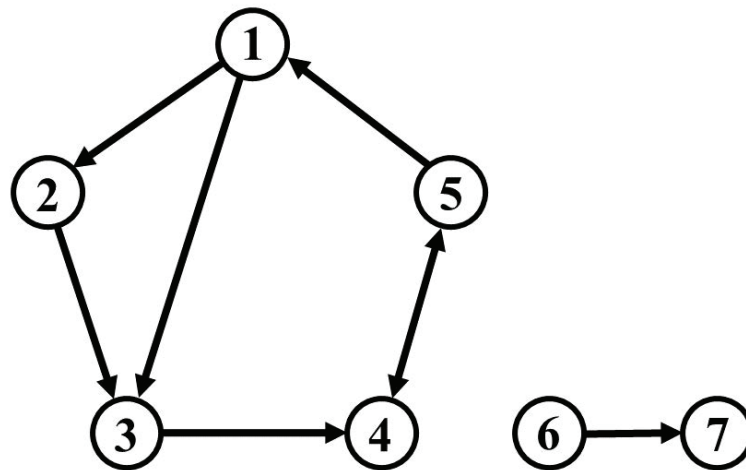


Figure 1. Connectivity diagram of the simulation model.

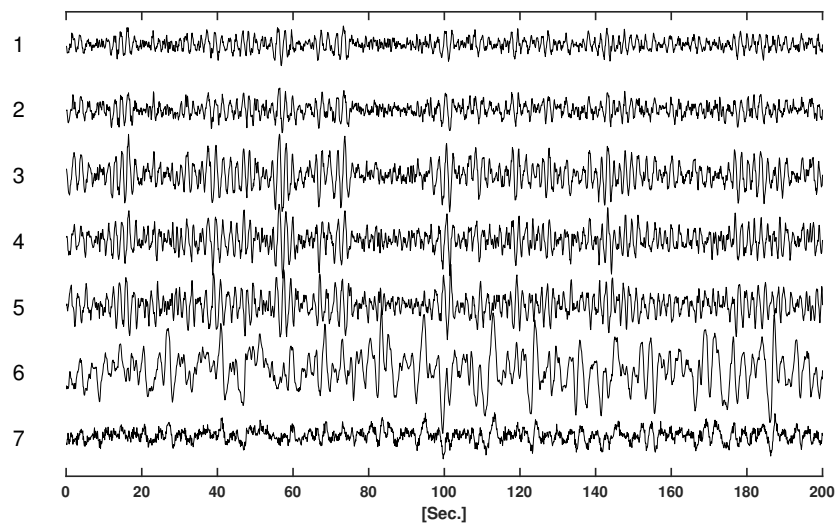


Figure 2. Simulated time series.

Table 1. VAR parameters for the simulation.

$A_1 =$	0.828	0	0	0	0	0	0
	0.541	0.651	0	0	0	0	0
	0.74	0	0.744	0	0	0	0
	0	0	0.456	0.73	0.3	0	0
	0	0	0	−0.4	0.859	0	0
	0	0	0	0	0	1.752	0
	0	0	0	0	0	−0.120	0
	−0.172	0	0	0	0.17	0	0
$A_2 =$	0	−0.107	0	0	0	0	0
	0	0.238	−0.139	0	0	0	0
	0	0	0	−0.134	0	0	0
	0	0	0	0	−0.185	0	0
	0	0	0	0	0	−0.810	0
	0	0	0	0	0	0.430	0

5. Results

In our pursuit of accurately estimating the 99% confidence intervals (CI) for both conditional Granger Causality (cGC) and Impulse Response Function (IRF) analyses, we

generated 2000 surrogate datasets using each of the previously discussed bootstrap methods. For the SB method, we determined the optimal mean block width w , selecting values of 5, 10, 20, and 40. As illustrated in Figure 3a, the cSB method consistently yielded correct causality results, irrespective of the chosen mean block width. In contrast, the ucSB method produced correct causality results for $w = 5$, but the detection rate of self-feedback diminished as w increased (see Figure 3b,c).

While the uPRB method provided accurate causality results concerning variable interactions, it failed to detect the self-feedbacks of the first through fifth variables, yielding results akin to ucSB(20) and ucSB(40) (see Figure 3c). If the time shift s in the TSS method was excessively small or large, the data closely resembled the original data, consequently minimizing the phase randomization effect. The selection of interval $[I_a, I_b]$ in the Time Shift Surrogates (TSS) method has potential implications on the statistical characteristics of the generated surrogate data, particularly its correlation structure. For causality detection between variables, it is necessary to set the interval large enough beyond the lag where cross-correlation occurs. In this simulation, we set $[I_a, I_b] = [3/N, 0.9N]$.

Notably, the TSS method's detection performance is inferior to that of the ucSB(20) and ucSB(40) methods, failing to detect causality from the fifth to the first variable. Importantly, our study identified no false positives within any bootstrap method. Table 2 concisely summarizes the sensitivity and specificity for each combination of causal analysis and bootstrap methods.

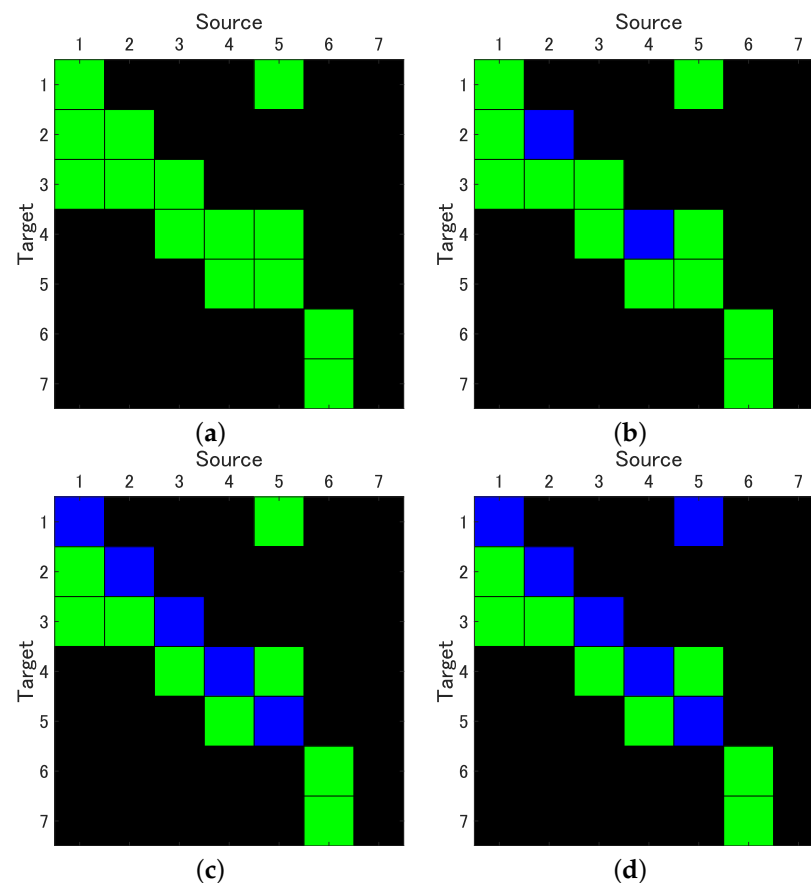


Figure 3. Results of evaluation of Granger Causality by correlated Stationary Bootstrap (cSB) with mean block width $w = 5$ (a), and uncorrelated Stationary Bootstrap (ucSB) with mean block width $w = 10$ (b) and $w = 40$ (c), and by Time Shift Surrogates (TSS) (d); variables in columns represent sources, variables in rows represent targets ($p < 0.01$); green: true positives, blue: false negatives.

Table 2. Sensitivity and specificity of detected causality by Granger Causality and Impulse response function, according to different bootstrap methods. Numbers in parentheses indicate mean block width w for correlated and uncorrelated Stationary Bootstrap (cSB and ucSB, respectively).

Bootstrap Method	Granger Causality		Impulse Response	
	Sensitivity	Specificity	Sensitivity	Specificity
cSB(5)	1.0	1.0	0.94	0.93
cSB(10)	1.0	1.0	1.0	0.89
cSB(20)	1.0	1.0	1.0	0.89
cSB(40)	1.0	1.0	1.0	0.93
ucSB(5)	1.0	1.0	0.41	1.0
ucSB(10)	0.86	1.0	0.41	1.0
ucSB(20)	0.64	1.0	0.41	1.0
ucSB(40)	0.64	1.0	0.41	1.0
uPRB	0.64	1.0	0.76	0.78
TSS	0.57	1.0	0.41	1.0
cARSB	1.0	1.0	1.0	1.0
cuARSB	1.0	1.0	1.0	1.0

As shown in Equation (9), the IRF is denoted as IRF_s^{lm} , which represents the impact of the m th shock e_t^m on the l th variable y_{t+s}^l at time s . We evaluated the significance of IRF_s^{lm} using each bootstrap method. Figure 4 provides a two-dimensional representation of significant IRF_s^{lm} .

In the IRF analysis, only the ARSB method, whether correlated or uncorrelated, successfully detects both self-feedback and causality among variables in their entirety. Figure 4a displays the IRFs for the impact assigned to each variable ($m = 1, \dots, 7$), with ucARSB evaluating the significance. This outcome aligns with the connectivity diagram depicted in Figure 1. The cSB method demonstrates results were almost identical to the ARSB method for the first through fifth variables. However, it detects a response from the sixth to seventh variable, constituting a false positive, which was observed for all tested values of the mean block width. The uPRB method primarily detects causality between directly coupled variables (see Figure 4c). Additionally, false positives emerge where responses occur between uncoupled variables, such as the reciprocal response between the first and seventh variables and between the fifth and sixth variables. For all tested values of the mean block width, the ucSB method exhibits the weakest detection performance among the evaluated bootstrap methods, detecting only self-feedback (see Figure 4d).

To summarize the results of these IRFs, such as GC, we compared the true IRFs generated through the VAR parameters in Table 1 with the IRFs detected by the respective bootstrap methods and summarized the sensitivity and specificity in Table 2. The time interval of the IRFs used for comparison was the maximum time step before feedback occurs, i.e., the IRFs up to three time steps before the impact on the third variable propagates to the first variable to the fourth and fifth variables (see Figure 1).

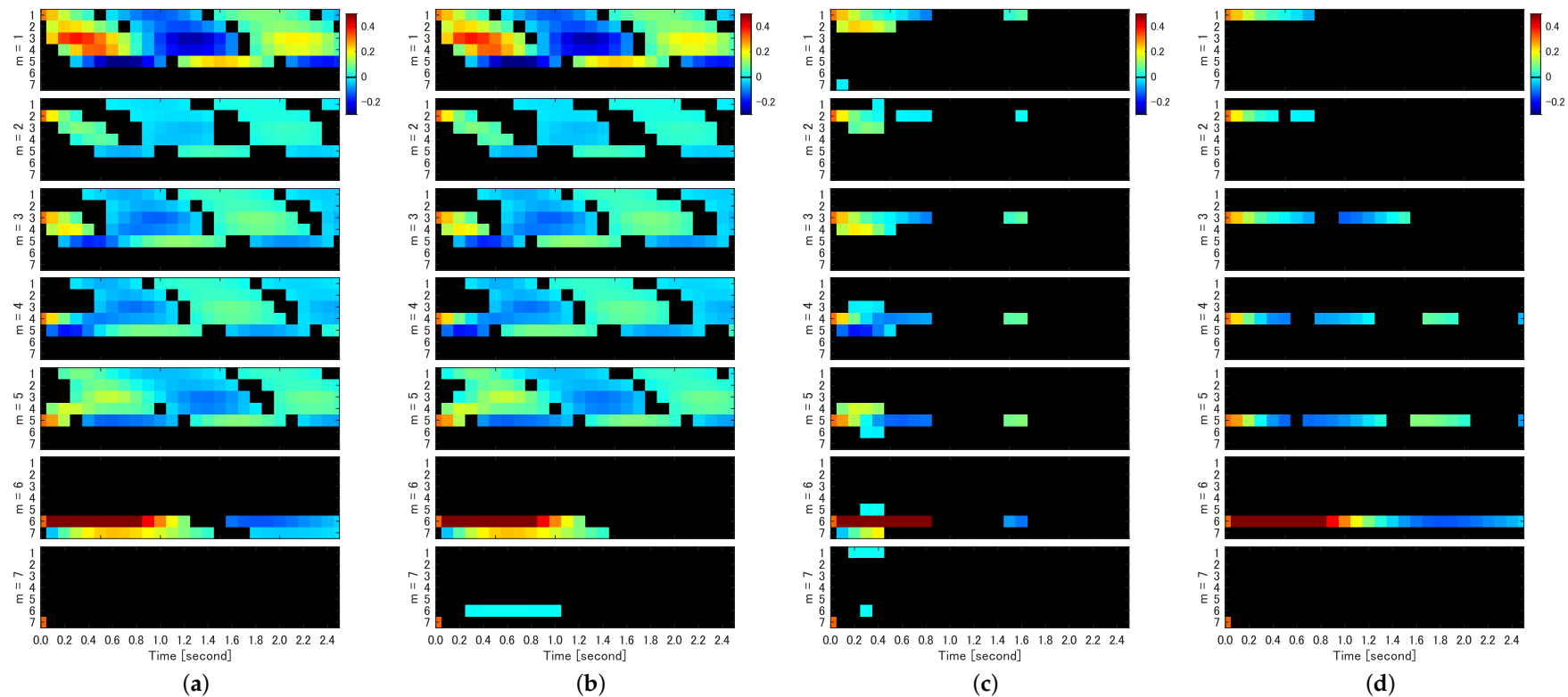


Figure 4. Significant impulse response function (IRF_s^{lm}) evaluated by uncorrelated AR-Sieve Bootstrap (ucARSB) (a), correlated Stationary Bootstrap (cSB) with mean block width $w = 40$ (b), uncorrelated Phase Randomization Bootstrap (uPRB) (c), and uncorrelated Stationary Bootstrap (ucSB) with mean block width $w = 40$ (d). Responses exceeding the significance level ($p < 0.01$) are displayed using color coding (see colorbars). Non-significant parts of impulse response functions remain black. Each column of subfigures displays IRF_s^{lm} for $m = 1, \dots, 7$ (subfigures from top to bottom) and for each value of m for $l = 1, \dots, 7$ (rows within each subfigure).

6. Discussion

In this study, we present a thorough comparison of the performance of four different bootstrap methods (encompassing twelve variations, including different values of mean block width, as well as correlated and uncorrelated derivations) to evaluate the results from causal analyses.

Our findings reveal that, for the cSB method, the Confidence Interval (CI) width decreases as the mean block width increases, which is particularly evident in Figure 5a,b. A possible explanation is that a shorter block width introduces more discontinuities in the surrogate time series, causing the dynamic properties to deviate further from those of the original time series due to the random shuffling of blocks disrupting temporal forward/backward relationships. This disruption may lead to falsely detecting causality from the seventh to sixth variables in the IRF analysis, as seen in Figure 4b. In contrast, the CI width for the ucSB method remains wider than that of the cSB method, irrespective of the mean block width (see Figure 5a,c), or even increases with the mean block width (Figure 5b). This stems from each block's starting point being determined independently for each variable, causing unnecessary destruction of causality among variables in the surrogate time series. Consequently, no causality among variables is detected in the corresponding IRF analysis (see Figure 4d).

The TSS method accurately detects significance for the Granger causality $F_{1 \rightarrow 3}$, but the CI widths for $F_{5 \rightarrow 1}$ and $F_{2 \rightarrow 2}$ are so large that they include zero, resulting in false negatives. This could be attributed to the (1, 5)th element of A_2 having the smallest value among the VAR parameters corresponding to causality among variables, and the (2, 2)th element of A_2 having the smallest value among the VAR parameters corresponding to the attenuation rate of self-feedback. A prime example is $F_{2 \rightarrow 5}$, which should be numerically ignorable since our simulation model does not contain causality from the second to fifth variable (see Table 1). Although the TSS method correctly assesses non-significance, the CI width is substantially larger than that estimated by the other bootstrap methods. These results suggest that the TSS method may not provide stable estimates of bootstrap statistics, compared to other methods, when the statistic value of the original time series is small.

In contrast to other methods, the uPRB method's algorithm does not require a priori tuning of parameters, such as a mean block width for the SB method. Upon providing the original time series, surrogate time series can be generated almost instantaneously. The uPRB method is characterized by its simplicity and the ability to generate surrogate time series devoid of discontinuous time points, while preserving stationarity. However, a limitation of this method is its inability to detect self-feedback (see Figure 5b). Although it accurately identifies causality among variables, some of the Confidence Intervals (CIs) have considerable width, which questions its reliability (see Figure 5a). By constraining the interval of phase randomization in Step 2, the CI width may be reduced, thereby enhancing performance; however, adjusting the degree of restriction remains arbitrary. A key distinction between the TSS and uPRB methods is that the former shifts the phases in all frequency bands simultaneously, while the latter does so for each frequency band. As the uPRB method can selectively disturb the phases corresponding to a specific frequency band, it may prove valuable for evaluating causality in the spectral domain.

For both cGC and IRF analyses, the ARSB method outperforms the other bootstrap methods, regarding detection performance. For this method, determining the VAR model order p is essential, which can be accomplished using the Akaike Information Criterion (AIC). As outlined in Section 3.4, there are two derivations: correlated ARSB (cARSB), which randomly shuffles the residuals across all variables synchronously, and uncorrelated ARSB (ucARSB), which shuffles the residuals independently for each variable. Our simulation study demonstrates that both derivatives yield nearly identical results. However, observational errors (artifacts) may be concurrently superimposed onto the data when analyzing actual time series, rendering ucARSB potentially more effective for canceling noise correlation among variables.

Finally, we emphasize that beyond Granger causality, as discussed in this paper, other approaches to causality estimation have been proposed; as an example, we mention Convergent Cross Mapping, which has been shown to be effective in deterministic non-linear dynamical systems [20]. Furthermore, Transfer Entropy has been defined as an extension of mutual information, representing the causality from one random variable to another [21]. Moreover, Zenil et al. [22] have developed a very general framework for model discovery, which may overcome the inherent limitations of statistical methods based on predictive models and entropy-like measures; their method is based on algorithmic probability, i.e., on decomposing observations into the most likely algorithmic generative models. Future research should investigate whether this framework can compensate for or replace the incompleteness of Granger causality.

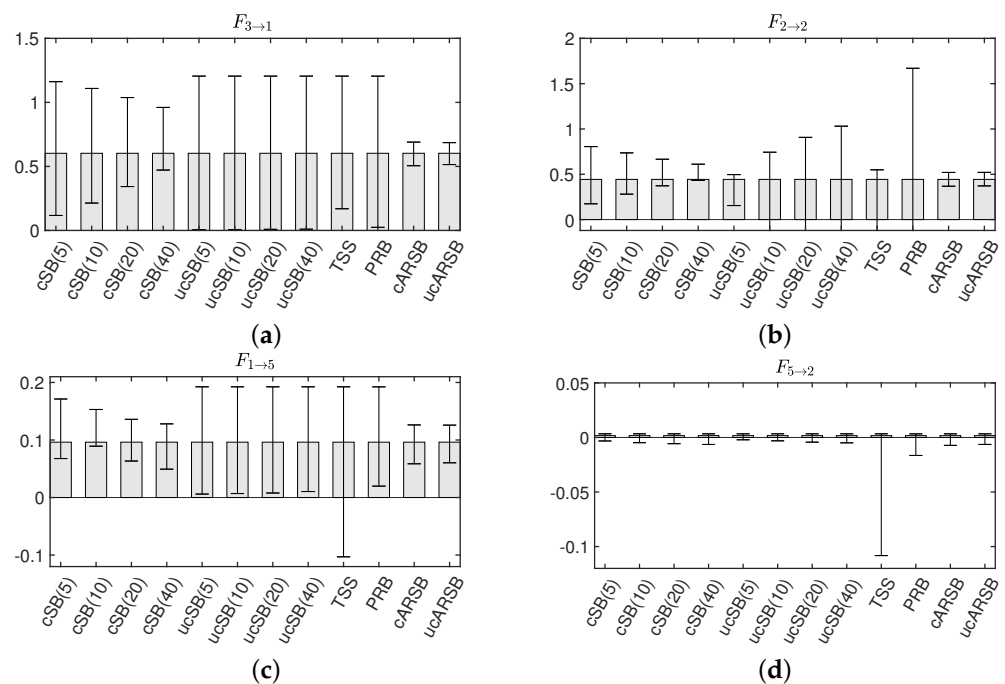


Figure 5. Representative examples of 99% confidence intervals of Granger Causality $F_{l \rightarrow m}$, estimated for each bootstrap method.

Author Contributions: Methodology, F.M. and A.G.; Investigation, F.M.; Writing—original draft, F.M.; Writing—review & editing, A.G.; Visualization, F.M.; Project administration, F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI Grant Numbers 19K12212 and 21H04874, and “Strategic Research Projects” grant from ROIS (Research Organization of Information and Systems).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438. [[CrossRef](#)]
2. Geweke, J. Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.* **1984**, *79*, 907–915. [[CrossRef](#)]
3. Bressler, S.L.; Seth, A.K. Wiener-Granger causality: A well established methodology. *Neuroimage* **2011**, *58*, 323–329. [[CrossRef](#)] [[PubMed](#)]
4. Schiatti, L.; Nollo, G.; Rossato, G.; Faes, L. Extended Granger causality: A new tool to identify the structure of physiological networks. *Physiol. Meas.* **2015**, *36*, 827–843. [[CrossRef](#)] [[PubMed](#)]

5. Zenil, H.; Kiani, N.A.; Tegnér, J. Low-algorithmic-complexity entropy-deceiving graphs. *Phys. Rev. E* **2017**, *96*, 012308. [[CrossRef](#)] [[PubMed](#)]
6. Guo, S.; Seth, A.K.; Kendrick, K.M.; Zhou, C.; Feng, J. Partial Granger causality—Eliminating exogenous inputs and latent variables. *J. Neurosci. Methods* **2008**, *172*, 79–93. [[CrossRef](#)] [[PubMed](#)]
7. Politis, D.N.; Romano, J.P. A circular block resampling procedure for stationary data. In *Exploring the Limits of Bootstrap*; LePage, R., Billard, L., Eds.; John Wiley & Sons: New York, NY, USA, 1992; pp. 263–270.
8. Quiroga, R.Q.; Kraskov, A.; Kreuz, T.; Grassberger, P. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys. Rev. E* **2002**, *65*, 041903. [[CrossRef](#)] [[PubMed](#)]
9. Ashkenazy, Y.; Baker, D.R.; Gildor, H.; Havlin, S. Nonlinearity and multifractality of climate change in the past 420,000 years. *Geophys. Res. Lett.* **2003**, *30*, 2146. [[CrossRef](#)]
10. Politis, D.N.; Romano, J.P. The stationary bootstrap. *J. Am. Stat. Assoc.* **1994**, *89*, 1303–1313. [[CrossRef](#)]
11. Bühlmann, P. Sieve bootstrap for time series. *Bernoulli* **1997**, *3*, 123–148. [[CrossRef](#)]
12. Berg, A.; Paparoditis, E.; Politis, D. A bootstrap test for times series linearity. *J. Stat. Plan. Inference* **2010**, *140*, 3841–3857. [[CrossRef](#)]
13. Theiler, J.; Eubank, S.; Longtin, A.; Galdrikian, B.; Farmer, J.D. Testing for nonlinearity in time series: The method of surrogate data. *Phys. D* **1992**, *58*, 77–94. [[CrossRef](#)]
14. Prichard, D.; Theiler, J. Generating surrogate data for time series with several simultaneously measured variables. *Phys. Rev. Lett.* **1994**, *73*, 951–954. [[CrossRef](#)] [[PubMed](#)]
15. Stam, C.J.; Breakspear, M.; van Walsum, A.M.v.C.; van Dijk, B.W. Nonlinear synchronization in EEG and whole-head MEG recordings of healthy subjects. *Hum. Brain Mapp.* **2003**, *19*, 63–78. [[CrossRef](#)] [[PubMed](#)]
16. Soofi, A.S.; Galka, A.; Li, Z.; Zhang, Y.; Hui, X. Applications of methods and algorithms of nonlinear dynamics in economics and finance. In *Complexity in Economics: Cutting Edge Research*; Faggini, M., Parziale, A., Eds.; Springer: Cham, Switzerland, 2014; pp. 1–30.
17. Carlstein, E. The use of subsample values for estimating the variance of a general statistic from a stationary sequence. *Ann. Stat.* **1986**, *14*, 1171–1179. [[CrossRef](#)]
18. Künsch, H.R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **1989**, *17*, 1217–1241. [[CrossRef](#)]
19. Baccalá, L.A.; Sameshima, K. Overcoming the limitations of correlation analysis for many simultaneously processed neural structures. *Prog. Brain Res.* **2001**, *130*, 33–47. [[PubMed](#)]
20. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.h.; Deyle, E.; Fogarty, M.; Munch, S. Detecting causality in complex ecosystems. *Science* **2012**, *338*, 496–500. [[CrossRef](#)] [[PubMed](#)]
21. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461. [[CrossRef](#)] [[PubMed](#)]
22. Zenil, H.; Kiani, N.A.; Zea, A.A.; Tegnér, J. Causal deconvolution by algorithmic generative models. *Nat. Mach. Intell.* **2019**, *1*, 58–66. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.