mathematics

MDPI

Article

Causality Distance Measures for Multivariate Time Series with Applications

Achilleas Anastasiou [1,†], Peter Hatzopoulos [1,†], Alex Karagrigoriou [1,*,†] and George Mavridoglou [2,†]

[1] Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, GR-83200 Samos, Greece; sasd20003@sas.aegean.gr (A.A.); xatzopoulos@aegean.gr (P.H.)
[2] Department of Accounting and Finance, University of Peloponnese, GR-24100 Antikalammos, Greece; g.mavridoglou@teipel.gr
* Correspondence: alex.karagrigoriou@aegean.gr
† These authors contributed equally to this work.

Abstract: In this work, we focus on the development of new distance measure algorithms, namely, the Causality Within Groups (CAWG), the Generalized Causality Within Groups (GCAWG) and the Causality Between Groups (CABG), all of which are based on the well-known Granger causality. The proposed distances together with the associated algorithms are suitable for multivariate statistical data analysis including unsupervised classification (clustering) purposes for the analysis of multivariate time series data with emphasis on financial and economic data where causal relationships are frequently present. For exploring the appropriateness of the proposed methodology, we implement, for illustrative purposes, the proposed algorithms to hierarchical clustering for the classification of 19 EU countries based on seven variables related to health resources in healthcare systems.

Keywords: multivariate time series; Granger causality; clustering; classification; distance; divergence; healthcare systems; pattern recognition

check for updates

Citation: Anastasiou, A.; Hatzopoulos, P.; Karagrigoriou, A.; Mavridoglou, G. Causality Distance Measures for Multivariate Time Series with Applications. *Mathematics* **2021**, *9*, 2708. https://doi.org/10.3390/math9212708

Academic Editor: Vasile Preda

Received: 20 July 2021
Accepted: 20 October 2021
Published: 25 October 2021

1. Introduction

In time series analysis and, generally, in statistics we are interested in the correlation between variables which is often investigated with the aim of discovering the degree and the extent of their association. Such statistical techniques include among others the autocovariance, the autocorrelation, the Pearson correlation coefficient, etc. Generally, the autocovariance depends on the unit of measurement of the variables; thus, it is difficult to measure the dependence of random variables of a stochastic process by using autocovariances. The existence of a high correlation among variables is by no means proof that there is a causal relationship between the variables under investigation. However, in most cases it is hard to detect whether two variables cause one another or are independent of each other or only one is causing the other. The inability of correlations and autocorrelations to capture the underlying mechanisms of stochastic processes and the difficulties of establishing a causal relationship between economic variables led Granger to develop the economic concept of causality known as Granger causality. In this work, we take into consideration the causal relationship measured by the Granger causality and propose new algorithms for measuring the distance (closeness) between two time series. The proposed algorithms are suitable for classification purposes for both univariate and multivariate time series where causal relationships are frequently present. Such techniques are highly useful in cases where financial variables and/or economic indicators are measured across groups (regions, zones, countries, etc.), and the purpose of the analysis is the classification into clusters based on the degree of closeness among groups.

Univariate and multivariate time series techniques have numerous applications in fields ranging from engineering and technical systems to economic, financial, or actuarial data analysis. A field of great significance with considerable economic as well as social impact is the dynamic study of the effectiveness of health related systems (i.e., healthcare systems and surveillance systems). For instance, epidemiological data such as incidence data or rates collected via surveillance systems are often analyzed for the purpose of identifying differences or patterns between geographical areas (regions, communities, etc.) or international comparisons of incidence rates among states. Monitoring, spread prevention and healthcare efficiency are classical issues of importance for health officials. In such studies, multivariate techniques for time series analysis and health indicators play a key role. For instance, efficiency and productivity of healthcare systems have been claimed to have a major impact on healthcare costs and have been a topic of great research activity (see, e.g., in [1,2]).

Comparative studies of multivariate time series involve, among others, the detection and identification of patterns and/or anomalies. Such tasks are explored via time series distance measures. Visual tools and analysis are frequently used to reveal patterns (e.g., visual inspection of electroengephalograms (EEG) [3]). Other examples include surveillance systems for comparing disease behavior in various regions and physiologic databases for comparing regions and identifying temporal patterns [4]. Detection of these complex physiological patterns not only enables demarcation of important clinical events, but can also elucidate hidden dynamical structures that may be suggestive of disease processes. Distance measures in conjunction with visual analysis tools enable analysts to achieve their tasks reliably and accurately including clustering and classification purposes (see, e.g., in [5,6]). For further readings, the interested reader may refer to the works in [7–10].

In this work, we introduce new distance measures based on Granger causality. The fact that necessitates the development of new measures is the understanding that distances should take under consideration any causal relationships existed between the variables involved and, additionally, groups should be combined into clusters according to the degree and extent of their causality. For exploring the applicability and the appropriateness of the proposed methodology, the new distance measures are implemented to healthcare efficiency for combining and classifying via hierarchical clustering, 19 EU countries according to the variables of a survey of the Organization for Economic Co-operation and Development (OECD) covering the period 1999–2016. The classic distance measure of association based on the autocorrelation is also used, for comparative purposes.

The presentation of the paper is as follows. In Section 2, we discuss standard association measures and furnish the concept of Granger causality. In Section 3, we present new distance measures based on Granger causality. Furthermore, in Section 4, we provide the results of the application to healthcare efficiency data from OECD. The last section is devoted to some general concluding remarks.

2. Material and Methods

2.1. Measures of Association

In this section, some classic measures of association are briefly presented and the Granger causality is reviewed [11–13]. New advanced measures will be proposed in the next section.

A useful measure of dependence, as it is dimensionless, is the autocorrelation function (ACF) usually denoted by $\rho(h)$, which measures the serial correlation of a time series with itself, shifted in time:

$$\rho(h) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)Var(X_{t+h})}}. \tag{1}$$

Another classical measure of dependence is the partial autocorrelation function (PACF), which is denoted by $\pi(h)$ and gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all lower lags in a fashion similar to (1).

Other measures frequently encountered in the literature is the classical Pearson's correlation, the cross-correlation which resembles the convolution function or the mutual correlation which is a measure of mutual dependence. For further details, the interested reader may refer to the works in [14–17] which detail a comprehensive reference volume on distances.

Classical Distance Measures for Time Series

For autocorrelation and partial autocorrelation, distance measures have been proposed in [18] and used in relevant applications [19]. More specifically, autocorrelation distance computes the distance between a pair of numeric time series X_t and X_s based on their estimated autocorrelation (or partial autocorrelation) coefficients. In such settings, a straightforward measure of distance is based on the computation of the autocorrelation coefficients $\rho_t = (\rho_t(1), ..., \rho_t(h))'$ for some h and then is used to define the measure between the two series X_t and X_s as follows:

$$D_\rho(X_t, X_s) = (\rho_t - \rho_s)' W(\rho_t - \rho_s) \tag{2}$$

where W is a weighting function that can be used to assign weights to the coefficients that decrease with the lag. A similar distance $D_\pi(X_t, X_s)$ can be acquired with the use of partial autocorrelation coefficient with $\pi_t = (\pi_t(1), ..., \pi_t(h))'$ in place of ρ_t.

The above distances like any other typical distance measure, e.g., City-Block, Minkowski, Mahalanobis, etc., can be used for clustering purposes in conjuction with hierarchical clustering algorithms.

2.2. Granger Causality

The concept of Granger causality between two time series X_t and Y_t first introduced by [11] and later reformed and formally proposed by Granger and discussed in [12,13], among others, is briefly presented below.

Definition 1. *Assume that X_t and Y_t are two time series and Ω_t is the probability space containing all the information up to time t. Then, X_t is said not to Granger-cause Y_t if for all $h > 0$,*

$$F(Y_{t+h}|\Omega_t) = F(Y_{t+h}|\Omega_t - X_t)$$

where $F(.|.)$ denotes the conditional distribution and $\Omega_t - X_t$ contains all the information except the amount associated with the series X_t. In other words, X_t is said to not Granger-cause Y_t if X cannot help in predicting a future value of Y. For the Granger test for causality, the following autoregressions are considered:

$$Y_t = \mu_0 + \sum_{i=1}^{h} a_i Y_{t-i} + \sum_{j=1}^{k} b_j X_{t-j} + u_t \tag{3}$$

$$X_t = \phi_0 + \sum_{i=1}^{h} d_i X_{t-i} + \sum_{j=1}^{k} c_j Y_{t-j} + e_t \tag{4}$$

where $\mu_0, \phi_0, a_i, d_i, b_j, c_j,\ i = 1, \ldots, h, j = 1, \ldots, k$ with h and k not necessarily equal, are appropriate coefficients and u_t & e_t are the error sequences. Consider also the restricted autoregression associated with (3) where Y_t is regressed only on its past values excluding all X_t terms (to avoid confusion, we use Y^* in place of Y):

$$Y_t^* = \mu_0 + \sum_{i=1}^{h} a_i Y_{t-i}^* + u_t^*. \tag{5}$$

A similar restricted autoregression associated with (4) can be obtained for X_t where all Y_t terms have been removed (as before, we use X^* in place of X):

$$X_t^* = \phi_0 + \sum_{i=1}^{h} d_i X_{t-i}^* + e_t^*. \tag{6}$$

The causality test is presented below:

- If in (3) the coefficients b_j are not statistically significant at a given significance level while in (4) the coefficients c_i are statistically significant then we conclude that Y_t is causing according to Granger, X_t.
- If in (3) the coefficients b_j are statistically significant at a given significance level while in (4) the coefficients c_i are not statistically significant then we conclude that X_t is causing according to Granger, Y_t.
- If all b's and c's are statistically significant at a given significance level, then there is a two-way causality.
- If the coefficients b_j and c_j in (3) and (4) are not statistically significant at a given significance level, X_t and Y_t are independent.

The hypothesis that Y_t is causing X_t according to Granger is tested by using the test statistic F defined by

$$F_{X,Y} = \frac{(SSE^* - SSE)/h}{SSE/(n-k)} = \frac{n-k}{h}\left(\frac{SSE^*}{SSE} - 1\right) \tag{7}$$

where $SSE^* = \sum\left(\hat{X}_i^* - X_i^*\right)^2$ is the restricted sum of squares of residuals associated with the restricted regression (6), $SSE = \sum\left(\hat{X}_i - X_i\right)^2$ is the unrestricted sum of squares of residuals, n is the sample size, h is the number of lags, k the number of parameters of Y_t in (4), and $\hat{X}_i^*$ and $\hat{X}_i$ are the predictions of X_i^* and X_i, respectively. For the hypothesis that X_t is causing Y_t, the test statistic $F_{Y,X}$ is given by (7), where SSE^* is the restricted SSE associated with (5). Under the null hypothesis, the test statistic F follows an F-distribution with h and $(n-k)$ degrees of freedom. For further readings on Granger causality, the interested reader may refer to the work in [13].

Remark 1. *It should be pointed out that the Granger causality is a concept of causality from the statistical point of view, developed to analyze the flow of information between time series. Granger formulated the above-mentioned statistical definition of causality which consists of two aspects, namely, that a cause occurs before its effect and that knowledge of a cause improves the prediction quality of its effect. Thus, Granger causality provides information about the predictive ability of a process and it does not refer to the actual causal relationship between two series. It is under this framework that we proceed below with the definition of the new causality distance measures.*

3. The New Causality Distance Measures

In this section, we introduce new distance measures for time series based on the concept of Granger causality.

The section ends with recalling the hierarchical clustering algorithms that could be used in practice, in conjunction with the distance measures introduced in this section.

3.1. Granger Causality Distance Measures For Time Series

The proposal of the new distance measures by way of the Granger causality test, is based on the idea that variables should be clustered together (classified) as long as the causalities among the variables/elements of a multivariate time series are similar to the causalities of the same variables of another multivariate time series.

3.1.1. The Granger and Generalized Granger Causality within Groups Distances

Consider two $k-$dimensional multivariate time series (MTS) with the following structure:

$$X^{q_1} = (X_{t1}^{q_1}, ..., X_{ti}^{q_1}, ..., X_{tk}^{q_1}) \tag{8}$$

and

$$X^{q_2} = (X_{t1}^{q_2}, ..., X_{ti}^{q_2}, ..., X_{tk}^{q_2}) \tag{9}$$

where

- $X_{ti}^{q_1}$ for each $i = 1, \ldots, k$ is a univariate time series of the first MTS (q_1), $t = 1, 2, \ldots$.
- $X_{ti}^{q_2}$ for each $i = 1, \ldots, k$ is a univariate time series of the second MTS (q_2), $t = 1, 2, \ldots$.
- k : common dimension of each MTS or number of variables (univariate time series).

We provide below the 4-step algorithm for the evaluation of the proposed *Granger Causality Within Groups (CAWG) Distance* distance (Algorithm 1).

Algorithm 1: CAWG.

```
1. Using (7), calculate for the MTS $q_1$ the $F_{ij}^{q_1}$ value of the Granger causality test
   which tests whether $X_{tj}^{q_1}$ causes $X_{ti}^{q_1}$, $i = 1, ..., k-1$ , $j = 2, ..., k$, $i < j$. The total
   number of $F$ values is equal to $[k(k-1)]/2$.
2. Repeat step 1 for the second MTS $q_2$ and obtain the $[k(k-1)]/2$ values of $F_{ij}^{q_2}$
   for testing whether $X_{tj}^{q_2}$ causes $X_{ti}^{q_2}$, $i = 1, ..., k-1$ , $j = 2, ..., k$, $i < j$.
3. Compute the squared differences between each (corresponding) pair of the $F_{ij}^{q_1}$
   and the $F_{ij}^{q_2}$ values.
4. Compute the summation of the squared differences of step 3 and obtain
   $CAWG(X^{q_1}, X^{q_2})$.
```

The definition of the proposed distance is given below followed by a lemma providing its basic properties. Through the properties which are easily shown, one verifies that the proposed distance is a typical pseudodistance [20].

Definition 2. *The Granger Causality Within Groups (CAWG) Distance between two $k-$dimensional multivariate time series X^{q_1} and X^{q_2} is defined by*

$$CAWG(X^{q_1}, X^{q_2}) = \sum_{1 \leq i < j \leq k} \left[F_{ij}^{q_1} - F_{ij}^{q_2} \right]^2, \tag{10}$$

where F_{ij}^{q} is the value of the test statistic defined in (7) according to which X_{tj}^{q} causes X_{ti}^{q}, with $q = q_1, q_2$.

Lemma 1. *The Granger Causality Within Groups (CAWG) Distance satisfies the following properties:*

1. $CAWG(X^{q_1}, X^{q_2}) \geq 0$.
2. $CAWG(X^{q_1}, X^{q_2}) = 0$ *for* $q_1 = q_2$.
3. $CAWG(X^{q_1}, X^{q_2}) = CAWG(X^{q_2}, X^{q_1})$.

Remark 2. *The proposed measure is based on the intercorrelations of a stochastic nature (like the autoregressions in our setting), of the series involved. The measure proposed in Definition 2 is a classic distance measure as it evaluates the intercorrelations between the components on one series denoted by $F_{ij}^{q_1}$ and comparing it with the associated quantity $F_{ij}^{q_2}$ of the other series. Thus, the Granger causality is used as a tool to measure the overall causality within the components (in pairs) of a multivariate series. The distance between the two series is defined through the classical square difference between the corresponding overall causalities. If this difference is zero, the two series are considered to be close to each other and in terms of clustering, can (and should) be classified into a single cluster. As expected, if the two elements of a pair are interchanged, the causality may not necessarily coincide with the original one. Indeed, the above distance is based on the definition of causality according to which the j^{th} element of a series Granger-causes the i^{th} element of the same series with $i < j$ which does not allow for the reverse Granger causality, i.e., for which the j^{th} element of a series Granger-causes the i^{th} element of the same series with $i > j$. Lemma 2, below, generalizes the distance in Definition 1 in a way that the resulting generalized distance GCAWG, takes under consideration both the above (Granger) causes. Note that we approach in this work, the issue of distance, from the statistical point of view where a satisfactory measure of divergence or* *distance is the one which is not negative with equality to zero occurring when the two arguments of the distance, coincide. Observe though that according to Lemma 1 above and Lemma 2 below, both proposed distance measures also satisfy the symmetry property.*

Lemma 2. *Consider two multivariate time series X^{q_1} and X^{q_2} according to (8) and (9) and define the CAWG distance*

$$CAWG^r(X^{q_1}, X^{q_2}) = \sum_{1 \le j < i \le k} [F_{ij}^{q_1} - F_{ij}^{q_2}]^2, \tag{11}$$

which is the reverse of Definition 1 for which $i > j$. Then, the Generalized Granger Causality Within Groups (GCAWG) defined by

$$GCAWG(X^{q_1}, X^{q_2}) = CAWG(X^{q_1}, X^{q_2}) + CAWG^r(X^{q_1}, X^{q_2}) \tag{12}$$

is a distance measure such that

1. $GCAWG(X^{q_1}, X^{q_2}) \ge 0$.
2. $GCAWG(X^{q_1}, X^{q_2}) = 0$ *for* $q_1 = q_2$.
3. $GCAWG(X^{q_1}, X^{q_2}) = GCAWG(X^{q_2}, X^{q_1})$.

Proof. The proof is immediate by Definition 2, Equation (11), and the application of Lemma 1.

The proposed methodology can be extended to M, $M \ge 2$ multivariate time series of dimension k. Consider the case of $\{q_1, q_2, \dots, q_M\}$ time series each of dimension k, with

$$X^q = (X_{t1}^q, ..., X_{ti}^q, ..., X_{tk}^q),\ q = q_1, ..., q_M \tag{13}$$

representing a $k-$dimensional MTS, $q = q_1, q_2, \dots, q_M$. Then, the generalization of Definition 2 can be achieved by repeating the 4-step CAWG algorithmic procedure for the calculation of all F_{ij} values and their squared differences for each pair of series q_i and q_j such that $i, j = 1, ..., M$, with $i \ne j$. At the end of the algorithm, the resulting *Granger Causality Within Groups Distance $M \times M$ matrix* given by

$$CAWG_M = \begin{pmatrix} 0 & CAWG(X^{q_1}, X^{q_2}) & ... & CAWG(X^{q_1}, X^{q_M}) \\ CAWG(X^{q_2}, X^{q_1}) & 0 & ... & CAWG(X^{q_2}, X^{q_M}) \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ CAWG(X^{q_M}, X^{q_1}) & CAWG(X^{q_M}, X^{q_2}) & ... & 0 \end{pmatrix} \tag{14}$$

could be useful for classification purposes and especially in connection with hierarchical clustering for classifying the M groups into clusters. Observe that due to Lemmas 1 and 2 the matrix associated with the Generalized CAWG Distance is simplified as follows:

$$GCAWG_M = \begin{pmatrix} 0 & GCAWG(X^{q_1}, X^{q_2}) & ... & GCAWG(X^{q_1}, X^{q_M}) \\ GCAWG(X^{q_2}, X^{q_1}) & 0 & ... & GCAWG(X^{q_2}, X^{q_M}) \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ GCAWG(X^{q_M}, X^{q_1}) & GCAWG(X^{q_M}, X^{q_2}) & ... & 0 \end{pmatrix}. \tag{15}$$

with $(GCAWG_M)_{i,j} = (GCAWG_M)_{j,i}$. □

3.1.2. The Granger Causality between Groups Distance

In this section, we present the Granger Causality Between Groups (CABG) Distance where the causality is related to the association between the corresponding elements (components) of two multivariate time series. Consider the two $k-$dimensional MTS (8) and (9) of the previous subsection. For the calculation of CABG Distance, the quantities F^{q_1,q_2}

and F^{q_2,q_1} will be used which represent the Granger causalities between the corresponding elements (components) of the MTS q_1 and q_2.

A 5-step CABG Algorithm is provided below for the calculation of the proposed CABG Distance (Algorithm 2).

Algorithm 2: CABG.

```
1. For the series $q_1$ and $q_2$, calculate the $F_i^{q_1,q_2}$ value of the Granger causality test which tests whether $X_{ti}^{q_2}$ (the $i$ component of $q_2$) causes $X_{ti}^{q_1}$ (the $i$ component of $q_1$).
2. Repeat step 1 for each component of the series and obtain a total of $k$ values of $F_i^{q_1,q_2}, i = 1, 2, ..., k$.
3. For $q_1$ and $q_2$ calculate the $F_i^{q_2,q_1}$ value of the Granger causality test which tests whether $X_{ti}^{q_1}$ causes $X_{ti}^{q_2}$.
4. Repeat step 3 for each component of the series and obtain a total of $k$ values of $F_i^{q_2,q_1}, i = 1, 2, ..., k$.
5. Compute the inverse of the summation $F_i^{q_1,q_2}$ and $F_i^{q_2,q_1}$, over $i = 1, \ldots, k$
```

The definition of the CABG distance together with its properties are presented below in Definition 3 and Lemma 3.

Definition 3. *The Granger Causality Between Groups (CABG) Distance between two multivariate time series X^{q_1} and X^{q_2} is defined by*

$$CABG(X^{q_1}, X^{q_2}) = \frac{1}{\sum_{i=1}^{k}(F_i^{q_1,q_2} + F_i^{q_2,q_1})} \tag{16}$$

where $F_i^{q_1,q_2}$ is the value of the test statistic defined in (7) according to which $X_{ti}^{q_2}$ causes $X_{ti}^{q_1}$ and $F_i^{q_2,q_1}$ is the corresponding value according to which $X_{ti}^{q_1}$ causes $X_{ti}^{q_2}$.

Lemma 3. *Assuming the convention $CABG(X^{q_1}, X^{q_2}) = 0$ for $q_1 = q_2$, the Granger Causality Between Groups (CABG) Distance satisfies the following properties:*

1. *$CABG(X^{q_1}, X^{q_2}) > 0$, for $q_1 \neq q_2$.*
2. *$CABG(X^{q_1}, X^{q_2}) = CABG(X^{q_2}, X^{q_1})$.*

The above algorithm can be extended to any number of $k-$dimensional multivariate time series. Indeed using the notation (13) of the previous subsection, the *Granger Causality Between Groups Distance $M \times M$ matrix* is easily obtained by repeating steps 1–5 of the above CABG Algorithm for any pair of series from $\{q_1, q_2, \ldots, q_M\}$:

$$CABG_M = \begin{pmatrix} 0 & CABG(X^{q_1}, X^{q_2}) & ... & CABG(X^{q_1}, X^{q_M}) \\ CABG(X^{q_2}, X^{q_1}) & 0 & ... & CABG(X^{q_2}, X^{q_M}) \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ CABG(X^{q_M}, X^{q_1}) & CABG(X^{q_M}, X^{q_2}) & ... & 0 \end{pmatrix}. \tag{17}$$

with $(CABG_M)_{i,j} = (CABG_M)_{j,i}$.

As in the case of CAWG and GCAWG algorithms, the CABG algorithm results could be useful for classification purposes and especially in hierarchical clustering for classifying groups into clusters. Recall that hierarchical clustering techniques for cluster analysis, with a widespread use in practice, are based on a series of successive groupings (Agglomerative algorithms) or successive divisions (Divisive algorithms). As for measuring the distance between groups, standard methods like the nearest neighbor, furthest neighbor, etc. are used. For details see in [21]. Observe that CAWG and GCAWG distances join multivariate time series with similar causalities (either weak or strong) among their components while in the CABG distance the series are joined together as long as elementwise the causalities are strong.

Remark 3. *It should be noted that for the above distance measure we adopted an idea different from the one used for the first two distance measures. More specifically, in this case, the interrelations are evaluated (in pairs) among the corresponding components of the two multivariate series involved. Thus, the overall amount of interrelation is large if each of the components of each series is causing the corresponding component of the other series. Then, in terms of clustering, the two series should be clustered together under the same class, since they are considered to be closely related. Thus, under Definition 3, large values imply closeness while small ones imply separation (which is the opposite of Definition 2). To accommodate this idea into the statistical analysis, we have chosen to use the inverse in the definition of CABG, so that if the value is close to zero (for large interrelations) the series will be grouped together and if the value increases (tends to infinity, for small interrelations) the series will stay in separate clusters/groups. As the new distance is undefined if the two series coincide, we set, by convention, CABG = 0 for* $i = j$.

4. Application

In order to better understand the properties of the new measures, an application to state health data was designed and implemented. The application and the results are presented in this Section.

4.1. Preliminaries

In the years following World War II, in all OECD countries, a significant share of the economy was used to improve or preserve the health of the population. On average, 8.8% of a country's GDP was dedicated to health in 2018 [22], from 3.5% in 1970. The annual increase in per capita health care costs outpaced the average annual economic growth of the last twenty years in all OECD countries [23]. According to Mueller and Morgan, on average 71% of health expenditures is funded government revenues generated from taxes and social insurance contributions. According to OECD projections, health expenditures will reach 9% of GDP by 2030 and 14% of GDP by 2060 [23]. The major driving forces behind the continuing rise in health care costs are new medical technology, health care services price inflation, rising income, and population aging.

Policy-makers have expressed the view that continued increases in health care spending may be "unsustainable", particularly in light of current and projected government budget deficits [24]. Today, health care is one of the most complex expenditures area and health and budget officials face the challenge to develop policies (a) for risk-sharing between the state and citizens, (b) an effort to increase the efficiency and effectiveness of funding ([5].

Several scientific methods have been proposed for measuring and providing ways to improve the efficiency of health systems [25–27]. Operational research proposed a large variety of such models. Their main feature was the need to compare homogeneous systems and policies. Many of them are based on comparing efficiency between countries. However, comparing countries with different systems and health policies can lead to underestimating or overestimating efficiency. For best results, the comparison of similar states is required. Cluster analysis helped to classify countries according to their health system characteristics. As the health expenditures are 'dynamic' and evolve, the classification should not be based on 'static' measures. Measures based on the correlation and causality of time series may be more useful [5].

As mentioned before, the logic of the proposed distance measures is different; in the GAWG measure, countries with the same internal causality between the variables, namely, countries following the same health policy will be in the same cluster. On the other hand, the GABG measure classifies together the countries that follow the same "development path" for their health systems.

4.2. Data

To explain cost increasing, many studies try to diagnose the underlying factors such as an ageing population, increased social expectations, broader insurance coverage, supplier-induced demand, and relative prices that may affect the utilization and costs of healthcare services. Many researchers examine the effect of these risk factors to the health system, such as the study of the effect on the mean length of stay (mls) in the hospitals [28] or techniques to estimate future spendings [29].

The data have been retrieved from the Organization for Economic Co-operation and Development (OECD) and EUROSTAT [30,31] and concern seven main health indices (Table 1) for 19 EU countries (Table 2), for the period 1999–2016. The survey focused on EU countries without taking into account the healthcare system in these countries.

Table 1. Variables used in the survey.

Variable	Variable Name
V1	Total health expeditures as a share of GDP
V2	Government spending as a share of total spendings
V3	Out of pocket as a share of total spendings
V4	Pharmaceutical spending as a share of total spendings
V5	Doctors per 1000 inhabitants
V6	Nurses per 1000 inhabitants
V7	Beds per 1000 inhabitants

For a brief description of the variables involved in the analysis, the reader is referred to supplementary material at the Laboratory of Statistics and Data Analysis of the Univ. of the Aegean at https://labstada.weebly.com/publications.html (accessed on 25 July 2021).

Table 2. Countries in the survey.

Country	Code	Country	Code	Country	Code
Austria	AUT	Greece	GRC	Poland	POL
Belgium	BEL	Hungary	HUN	Portugal	PRT
Czech Republic	CZE	Ireland	IRL	Slovakia	SVK
Denmark	DNK	Italy	ITA	Spain	ESP
Finland	FIN	Luxembourg	LUX	Sweden	SWE
France	FRA	Netherlands	NLD	United Kingdom	GBR
Germany	DEU				

For handling missing values, we proceeded with imputations using the linear interpolation technique. The percentage of missing values was approximately 10%. The data analysis was conducted with the R free software, and figures were built via *Tableau for Teaching* and *Rawgraphics* [32].

4.3. Data Analysis

For the implementation of the proposed methodology to the dataset, both CAWG and CABG distances have been used in this section. The distance among all possible pairs of countries in each dataset has been calculated according to the algorithms in Sections 3.1.1 and 3.1.2. For comparative purposes, the ACF distance has been applied to the dataset. Among the hierarchical algorithms, the agglomerative one has been chosen for the analysis so that at the beginning of the classification process, each country forms a singleton cluster/class. For the distance between clusters, the complete linkage method has been used. Note that the agglomerative algorithm and the complete linkage method have been used for illustrative purposes. Equally effective would have been the divisive algorithm in conjunction with any other well-known distance between clusters (single linkage, average linkage etc.).

Hierarchical algorithms provide clustering results in a form of a dendrogram, for any possible number of clusters and the researcher is free to choose which choice is more appealing to him/her. In what follows, and as this application is for illustrative purposes, we provide the results of the agglomerative algorithm only for 2 and 3 clusters according to each of the above-mentioned methods. Note that these choices happened to coincide with those recommended by popular techniques available in the literature for various purposes, such as silhouette and elbow methods ([33–35]).

The geographic representation of the classification results is depicted in Figure 1. The differences in classification observed are due to the different ways of measuring association. Indeed, ACF is associated with the correlation of lags, while Granger causality is an intercorrelated mechanism. As a result, the two measures which have been defined in this work and are based on the Granger causality, are expected to arrive at different classifications. One though could notice that a relative large number of countries are classified in a single cluster irrespective of the method used (see and compare Figure 1 and Table 3).

Table 3. Countries classification.

Country	ACF	CAWG	CABG
Austria	2	2	2
Belgium	2	2	1
Czech Republic	3	3	3
Denmark	2	2	3
Finland	2	2	3
France	2	2	3
Germany	2	2	1
Greece	3	2	3
Hungary	1	2	1
Ireland	3	2	1
Italy	2	2	3
Luxembourg	1	1	3
Netherlands	2	2	1
Poland	3	2	2
Portugal	3	2	3
Slovakia	3	2	2
Spain	2	2	3
Sweden	2	2	3
UK	2	2	3

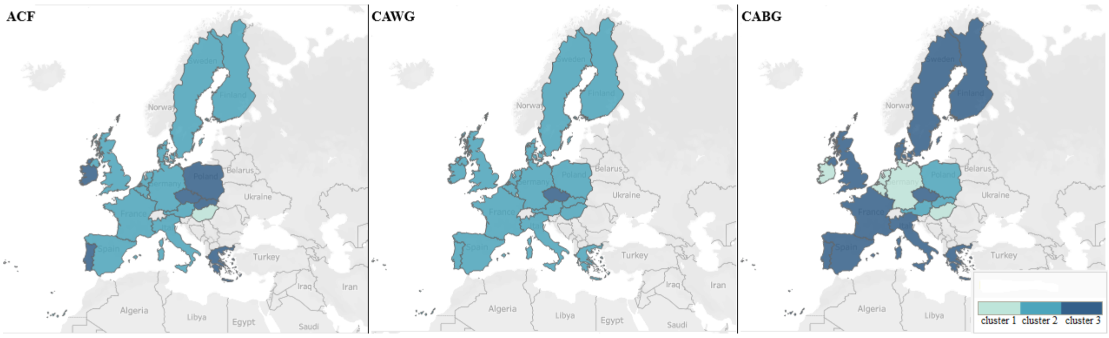

Figure 1. Geographic representation of classification comparison of ACF, CAWG, and CABG (in 3 clusters)—Luxembourg cannot be seen due to its size.

4.4. Comparison of the Results

The three competing techniques (ACF, CAWG, and CABG) are represented by the three layers of the dendrogram (Figure 2), starting from the left for the ACF classification and moving to the CAWG classification in the middle and to CABG classification at the far right. Note that the extra layer to the far right is the palette where each color represents a country or a group of countries. Each cluster for each technique is represented by a number (1 through 3), with the countries included in each cluster reported at the far right layer of the figure. For a better reading of the dendrogram, consider for instance cluster 2 according to the ACF (far left layer), with 11 members that correspond to red, blue and orange countries (with Luxembourg and Hungary creating cluster 1—upper left corner—and all other countries—including the Czech Republic (in green)—forming cluster 3 at the lower far left corner of the dendrogram). Moving to CAWG classification in the middle of the dendrogram, we observe that Hungary (brown) splits from cluster 1 (leaving Luxembourg-yellow as the only member of cluster 1) and joins cluster 2. At the same time, the Czech Republic-green stays by itself in cluster 3 with all other members of the cluster, joining (together with Hungary) cluster 2. The classification according to CABG appears in the right layer of Figure 2.

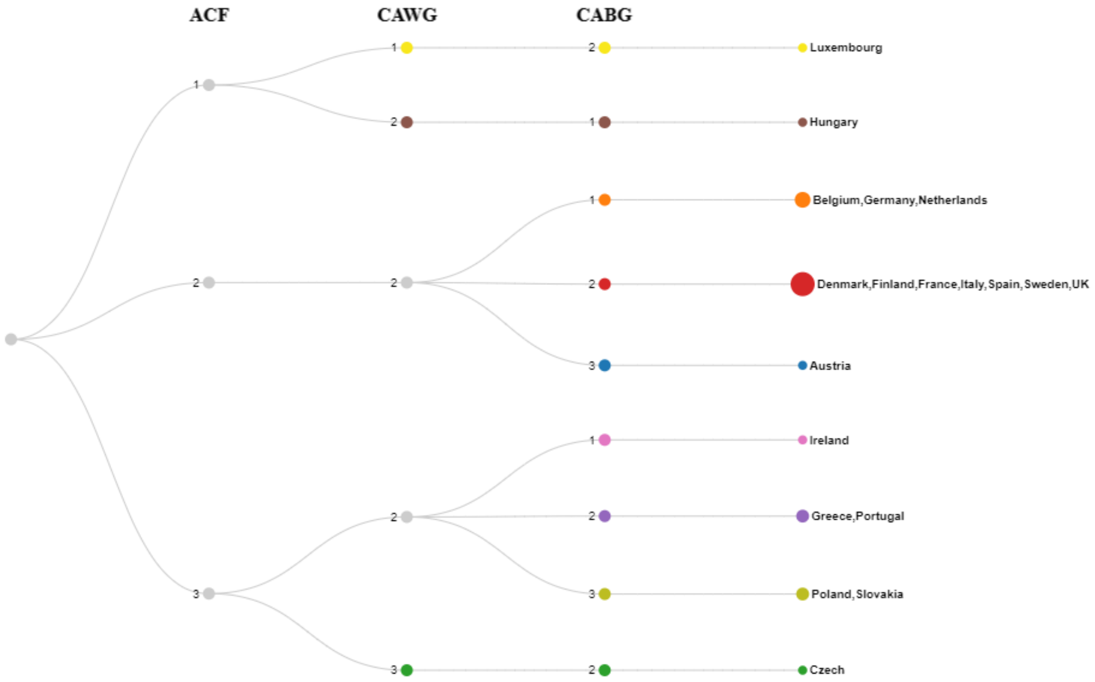

Figure 2. Classification comparison of ACF, CAWG, and CABG (in 3 clusters).

ACF and CAWG appear quite similar as 13 out of 19 countries are classified in the same clusters. On the other hand, the dissimilarity between CAWG and CABG is approximately 60% (12 out of 19 countries are not classified into the same clusters).

For the evaluation of the accuracy (agreement) of the partitions (classifications) produced by ACF, CAWG, and CABG distance measures, the well-known Rand Index ([36]) has been calculated. Recall that if a set S consists of n elements, $o_1, .., o_n$ and two clustering algorithms A and B produce r and s clusters, respectively, then the Rand Index between the A and B partitions is given by

$$RI(A,B) = \left(\frac{\alpha + \beta}{\binom{n}{2}} \right)$$

where α is the number of pairs of elements in S that are in the same cluster in A and in the same cluster in B and β is the number of pairs of elements in S that are in different clusters in A and in different clusters in B. It is reminded that the index ranges from 0 to 1 where higher values indicate a higher degree of similarity (agreement) between the two clustering algorithms. The index for pairwise classification comparisons takes values in the range 0.51–0.55 except in the case of the input dataset for the 2 new distance measures for which the index is equal to $RI(CABG, CAWG) = 0.38$ (Tabel 4).

Table 4. Rand Index table for the clustering prerformance of ACF, CAWG and CABG.

	ACF	**CAWG**	**CABG**
ACF	1	-	-
CAWG	0.51	1	-
CABG	0.55	0.38	1

These results indicate that the partitions carry a similar degree of agreement.

Remark 4. *Note that as clustering is an unsupervised machine learning algorithm the data do not contain ground truth labels making it hard to test the extent of the classification error.*

4.5. Concluding Remarks

According to ACF distance measure results, the second cluster consists of economically strong EU countries while in the third, not so economically strong countries are clustered together. The CAWG distance indicates that the causalities between the variables obtained are more or less similar among almost all the countries included in the analysis (cluster 2), except Luxembourg (singleton cluster 1) and the Czech Republic (singleton cluster 3). Thus, the effect of the overall cause is more or less the same among the European countries implying that the causal relationship between the variables (the elements of the multivariate time series) affecting the health system is quite similar across Europe. The CABG distance combines together countries of North and South Europe leaving on their own; Central European countries are to be combined in a separate entity. This observation implies that the extend of causality between Northern and Southern European countries is of the same magnitude, with causalities lower than those between central European countries.

5. General Conclusions

In this work, we proposed three new distance measures for measuring the distance (closeness) between multivariate time series by way of causal relationship. The measures defined together with the associated algorithmic procedures proposed are suitable for classification purposes for both univariate and multivariate time series where causal relationships are frequently present. A measure that takes into consideration some (any) kind of causal relationship like the one introduced by Granger and used in this work for proposing CABG, GCAWG and CAWG distance measures, is therefore recommended for clustering (unsupervised classification) purposes to ensure as accurate and as precise as possible decision making across groups with similar causalities. The contribution of this work lies on the proposal of the new distances CABG, GCAWG, and CAWG which are based on the idea that groups, regions, or countries should be combined into clusters as long as the causalities among the elements of two multivariate time series are similar.

The methodology proposed in this work is highly useful among other fields, in international or cross-national economics where financial variables and/or economic indicators are measured across groups (regions, zones, countries, etc.) and the purpose of the analysis is the classification into clusters based on the degree of closeness among groups. Furthermore, due to rapid integration of international economic markets, causal relationships are considered to be vital in the international economy as the identification of a possible impact could be used to alter or void economic policies, prevent socio-economic crises or enforce the same economic or financial decisions to groups with similar causal relationships. In this work, the implementation of the proposed methodology to healthcare systems shows the applicability and the usefulness of the new distance measures based on Granger causality.

Based on the results of the present analysis, we observe that the classification is strongly related to the distance/similarity measure considered. Different measures give rise to different classifications. It is clear from the analysis that the distance measure plays an important role so that the investigator should choose the one that is preferable according to the issue under investigation. In datasets such as the ones considered in this work, the relation between the variables involved, in terms of autocorrelation, partial autocorrelation or Granger causality, is quite common. It is therefore expected that the distances consider in this work are directly connected to the above association function in order to incorporate into the classification methodology the special characteristics of time series data. Moreover, it is natural to explore and implement more than one distance measures like the ones considered in this work. If though, more than one measure fits satisfactorily the needs of the investigator, then a comparison between the methods should be made by quantifying and evaluating the appropriateness of each method via a proper loss/error function. Thus, the proposed methodology offers an extra, user-friendly toolkit in the researcher's toolbox. The fact that various tools exist (or could be introduced) each based on a special distributional feature, provides the researcher with a great flexibility in choosing from the toolbox, that tool that fits better his/hers needs.

The applicability of the proposed methodology goes beyond healthcare systems, economics, finance, or actuarial science ranging from political sciences with regional conflicts and their causal relationship with political, institutional, and economic factors [37–39] to medicine, epidemiology and biology as well as to psychology and behavioral sciences where causal relationships play a fundamental role in understanding social behavior or identifying disease causation for the purpose of administering proper and effective behavioral or therapeutic treatments [40–42]. Furthermore, causal relationships and causal research in general, are particularly useful in business and management as for example in increasing customer retention or effective advertising ([43–45]. Further work with the focus on some of the above fields is expected to unfold at least some more of the numerous advantages of the proposed methodology.

Author Contributions: Conceptualization, A.A., P.H., A.K. and G.M.; Data curation, P.H. and G.M.; Formal analysis, A.A. and A.K.; Methodology, A.A. and A.K.; Software, A.A. and A.K.; Supervision, P.H., A.K. and G.M.; Writing—original draft, A.A.; Writing—review and editing, P.H., A.K. and G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data recievied from the OECD-iLibrary. https://www.oecdilibrary.org/social-issues-migration-health/health-spendingindicator/english_8643de7e-en (accessed on 21 July 2021).

Acknowledgments: The authors wish to express their appreciation to the anonymous referees, the Guest Editor, and the Associate Editor for their valuable comments and suggestions. This work was completed as part of the research activities of the Laboratory of Statistics and Data Analysis of the University of the Aegean. The authors wish to thank K. Tsilika of the University of the Thessaly and K. Ntotsis of the University of the Aegean for bringing to their attention the visual capabilities of Tableau and RawGraphics.

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

References

1. Pelat, C.; Boëlle, P.Y.; Cowling, B.J.; Carrat, F.; Flahault, A.; Ansart, S.; Valleron, A.J. Online detection and quantification of epidemics. *BMC Med. Inform. Decis. Mak.* **2007**, *5*, 29. [CrossRef] [PubMed]
2. Kalligeris, E.N.; Karagrigoriou, A.; Parpoula, C. On mixed PARMA modeling of epidemiological time series data. *Commun. Stat. Case Stud. Data Anal. Appl.* **2020**, *6*, 36–49. [CrossRef]
3. Jing, J.; Dauwels, J.; Rakthanmanon, T.; Keogh, E.; Cash, S.S.; Westover, M.B. Rapid annotation of interictal epileptiform discharges via template matching under dynamic time warping. *J. Neurosci. Methods* **2016**, *274*, 179–190. [CrossRef] [PubMed]
4. Saeed, M.; Lieu, C.; Raber, G.; Mark, R.G. MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput. Cardiol.* **2002**, *29*, 641–644.
5. Cinaroglu, S. Clustering of OECD countries out of pocket health expenditure time series data. *Res. Appl. Econ.* **2016**, *8*, 23–38. [CrossRef]
6. Lefevre, T.; Rondet, C.; Parizot, I.; Chauvin, P. Applying multivariate clustering techniques to health data: The 4 types of healthcare utilization in the Paris metropolitan area. *PLoS ONE* **2014**, *9*, e115064. [CrossRef]
7. Basalto, N.; Bellotti, R.; De Carlo, F.; Facchi, P.; Pantaleo, E.; Pascazio, S. Hausdorff clustering. *Phys. Rev. E* **2008**, *78*, 046112. [CrossRef] [PubMed]
8. Basalto, N.; Bellotti, R.; De Carlo, F.; Facchi, P.; Pantaleo, E.; Pascazio, S. Hausdorff clustering of financial time series. *Phys. A* **2007**, *379*, 635–644. [CrossRef]
9. Ferreira, L.N.; Zhao, L. Time series clustering via community detection in networks. *Inf. Sci.* **2016**, *326*, 227–242. [CrossRef]
10. Dau, H.A.; Keogh, E.; Kamgar, K.; Yeh, C.M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C.A.; Chen, Y.; Hu, B.; Begum, N.; et al. The UCR Time Series Classification Archive. 2019. Available online: www.cs.ucr.edu/~eamonn/time_series_data_2018/ (accessed on 25 July 2021).
11. Wiener, N. The theory of prediction. In *Modern Mathematics for Engineers*; Beckenback, E.F., Ed.; McGraw-Hill: New York, NY, USA, 1956; pp. 165–190.
12. Granger, C.W.J. Investigating causal relation by econometric and cross-sectional method. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
13. Siggiridou, E.; Kugiumtzis, D. Granger causality in multi-variate time series using a time ordered restricted vector autoregressive model. *IEEE Trans. Signal Process.* **2016**, *64*, 1759–1773.
14. Esling, P.; Agon, C. Time-series data mining. *Acm Comput. Surv. (Csur)* **2012**, *45*, 1–34. [CrossRef]
15. Fu, T.-C. A review on time series data mining. *Eng. Appl. Artif. Intell.* **2011**, *24*, 164–181. [CrossRef]
16. James, N.; Menzies, M.; Azizi, L.; Chan, J. Novel semi-metrics for multivariate change point analysis and anomaly detection. *Phys. Nonlinear Phenom.* **2020**, *412*, 132636. [CrossRef]
17. Deza, M.M.; Deza, E. *Encyclopedia of Distances*; Springer: Berlin, Germany, 2009.
18. Galeano, P.; Pena, D. Multivariate analysis in vector time series. *Resenhas* **2000**, *4*, 383–403.
19. Montero, P.; Vilar, J.A. TSclust: An R Package for Time Series Clustering. *J. Stat. Softw.* **2014**, *62*, 1–43. [CrossRef]
20. Toma, A.; Karagrigoriou, A.; Trentou, P. Robust model selection criteria based on pseudodistances. *Entropy* **2020**, *22*, 304. [CrossRef]
21. Everitt, B.S.; Landau, S.; Leese, M.; Stahl, D. *Cluster Analysis*, 5th ed.; John Wiley and Sons Ltd.: Hoboken, NJ, USA, 2011.
22. Mueller, M.; Morgan, D. *Focus on Public Funding of Health Care*; OECD: Paris, France, 2020.
23. Mueller, M.; Morgan, D. New insights into health financing: First results of the international data collection under the system of health accounts 2011 framework. *Health Policy* **2017**, *121*, 764–769. [CrossRef] [PubMed]
24. Paolucci, F. *Health Care Financing and Insurance: Options for Design*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 10;
25. Wang, J.; Jamison, D.; Bos, E.; Preker, A.; Peabody, J. *Measuring Country Performance on Health—Selected Indicators for 115 Countries*; Health, Nutrition, and Population Series; World Bank: Washington, DC, USA, 1999.
26. Bem, A.; Ucieklak-Jeż, P.; Predkiewicz, P. Measurement of Health care system efficiency. *Manag. Theory Stud. Rural Bus. Infrastruct. Dev.* **2014**, *36*, 25–33. [CrossRef]
27. Bekaroglu, C. A Multi-Stage Efficiency Analysis of OECD Healthcare and the Impact of Technical Change. Ph.D. Thesis, University of Connecticut, Storrs, CT, USA, 2015. Available online: http://digitalcommons.uconn.edu/dissertations/977 (accessed on 25 July 2021).
28. Livieris, I.E.; Kotsillieris, T.; Dimopoulos, I.; Pintelas, P. Decision support software for forecasting patient's length of stay. *Algorithms* **2018**, *11*, 199. [CrossRef]
29. Livieris, I.E. Forecasting economy-related data utilizing weight-constrained recurrent neural networks. *Algorithms* **2019**, *12*, 85. [CrossRef]
30. OECD-iLibrary. Available online: https://www.oecdilibrary.org/social-issues-migration-health/health-spendingindicator/english_8643de7e-en (accessed on 21 July 2021).
31. Eurostat Health Database. Available online: https://ec.europa.eu/eurostat/web/health/health-statusdeterminants (accessed on 25 July 2021).
32. Halkos, G.; Tsilika, K. Programming Correlation Criteria with free CAS Software. *Comput. Econ.* **2018**, **52**, 299–311. [CrossRef]
33. Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]

34. Amorim, R.C.D.; Hennig, C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Inf. Sci.* **2015**, *324*, 126–145. [CrossRef]
35. Thorndike, R. Who Belongs in the Family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
36. Rand, R.W. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850. [CrossRef]
37. Gassebner, M.; Luechinger, S. Lock, stock, and barrel: A comprehensive assessmentof determinants of terror. *Public Choice* **2011**, *149*, 235–261. [CrossRef]
38. Krieger, T.; Meierriecks, D. What causes terrorism? *Public Choice* **2011**, *147*, 3–27. [CrossRef]
39. Couttenier, M.; Soubeyran, R. A Survey of the Causes of Civil Conflicts: Natural Factors and Economic Conditions. *Revue D'économie Politique* **2015**, *125*, 787–810. [CrossRef]
40. Elwood, J.M. *Causal Relationships in Medicine: A Practical System for Critical Appraisal*; Oxford University Press: Oxford, UK, 1989.
41. Yeung, S.; Griffiths, T.L. Identifying expectations about the strength of causal relationships. *Cogn. Psychol.* **2015**, 76, 1–29. [CrossRef]
42. Woodward, J. Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biol. Philos.* **2010**, *25*, 287–318. [CrossRef]
43. Ang, L.; Buttle, F. Customer retention management processes: A quantitative study. *Eur. J. Mark.* **2006**, *40*, 83–99. [CrossRef]
44. Varian, H.R. Causal inference in economics and marketing. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 7310–7315. [CrossRef] [PubMed]
45. World Health Organization. *World Health Report*; World Health Organization: Geneva, Switzerland, 2000.