Time Series, Nonsense Correlations and the Principle of the Common Cause

Julian Reiss

ABSTRACT. This paper examines some recent defences of the principle of the common cause (PCC) against Elliott Sober's famous counterexample. There are two lines of attack: attempts to defuse the counterexample, that is, to show that the scenario described by Sober only apparently conflicts with the PCC; and attempts to demonstrate that the counterexample has no practical consequences. I show in this paper that there are problems with both strategies. In response, I formulate an alternative version of the principle that avoids the known counterexamples and that makes its status as fallible epistemic principle explicit.

1 Introduction

Time series, that is, time-ordered sets of observations on a random variable or random variables, are of fundamental importance for empirical inferences in sciences as diverse as neurophysiology, climatology, epidemiology, astro- and geophysics and many of the social sciences. In this paper I shall argue that a number of particularities of time series pose serious difficulties for one of the most prominent kinds of account of causal inference: probabilistic theories. A core assumption of probabilistic theories is the principle of the common cause, according to which a correlation between two variables is indicative of a causal connection between these variables. "Nonsense correlations"—*i.e.*, correlations that are artefacts of the statistical properties of the variables or that obtain for other non-causal reasons—pose an obvious problem for probabilistic theories.

Despite the recognition of the problem of nonsense correlations, probabilistic methods of causal inference have become increasingly popular in recent years. This has triggered some discussions about the seriousness of the problem. Some authors have tried to show that it is a mere pseudo-problem and that the principle can be saved once the notion of "correlation" is clarified. Others have accepted the genuineness of the problem but denied that it has serious methodological implications.

Here I shall show that the problem of nonsense correlations is both serious as well as ubiquitous in all areas of science where time-series matter. In response, I try to formulate a version of the principle of the common cause that avoids the nonsense correlation issue and, moreover, makes explicit its status as fallible epistemic principle. The lesson to draw from this discussion is that the methodological monism occasionally proposed by defenders of probabilistic theories of causal inference is mistaken: different kinds of situations require methods of inference to be tailored to the specifics of those situations if the researcher aims at inferring correct causal claims.

Let us start with some essential definitions. Time series $X = \{x_1, x_2, ..., x_T\}$ are time-ordered sets of observations on quantitative characteristics of an individual or a structure such as a socio-economic system. At each point in time, the observations are assumed to be drawn from a probability distribution $P_t(X)$.

It is important to distinguish a time series from the stochastic process that generates it. The stochastic process is the world line of the persisting object (a die, a socio-economic structure) itself whereas the time series records measurements or observations on the process made through (usually identical intervals in) time. Quantitative characteristics of an object can assume different values at different points in time and at each point are assumed to be drawn from a probability distribution $f_t(x) = P_t(X = x)$ that satisfies the usual axioms. X is thus a *Random variable*. I will represent a variable by a capital letter X and a value of a variable by a small letter x.

2 The PCC, British Bread Prices and Venetian Sea Levels

The principle of the common cause (PCC) lies at the heart of many accounts of probabilistic causation (*cf.*, **?????**).[1] Simplifying slightly, it can be stated as follows:

PCC. If two random variables X, Y are probabilistically dependent, then either X causes Y, Y causes X or X and Y are the joint effects of a common cause Z.[2]

[1] More recent accounts adopt the related causal Markov condition (CMC) as core principle. Since in the two-variable case the PCC can be shown to follow from the CMC, a counterexample to the PCC is a counterexample to the CMC, too. I will thus not consider it separately here.

[2] In most formulations the PCC contains also the screening-off condition, which states that the (full) common cause Z screens off the dependence between X and Y: $P(Y|X, Z) = P(Y|Z)$. The screening-off condition is controversial itself, and since my discussion focuses on violations of the first part of the PCC, I omit it here.

Two variables X and Y are probabilistically dependent just in case $P(XY) \neq P(X)P(Y)$. A situation in which a probabilistic dependence between two variables is due to non-causal facts about the properties of the variables constitutes an obvious violation of the PCC. The first systematic discussion of the problem is due to G. Udny Yule (**?**). Among philosopher's of science, a counterexample introduced by Elliott Sober has been widely discussed (**?**, p. 332, **?**, pp. 161-2). In this example, X = sea levels in Venice and Y = cost of bread in Britain. Sober assumes the two variables to increase monotonically in time (**?**, p. 334):

Year (t)	British Bread Prices (Y)	Venetian Sea Levels (X)
1	4	22
2	5	23
3	6	24
4	10	25
5	14	28
6	15	29
7	19	30
8	20	31

An intuitive test for whether two variables are probabilistically dependent is asking whether observing one variable is informative about the likely value the other variable will take. This criterion is clearly fulfilled in this case: a higher observed Venetian sea level allows us to infer higher British bread prices and vice versa. And yet, the two variables are *ex hypothesi* not causally connected.

In principle there are two strategies for saving the PCC in the light of Sober's argument. One could, first, argue that the scenario merely *appears* to be a counterexample to the PCC. Though the data Sober provides makes the underlying variables *look* probabilistically dependent, in fact they are not. We make a fallacious inference from sample statistics to population statistics if we used the PCC in this case rather than a fallacious *causal* inference. A second strategy is to argue that the PCC is indeed violated in the Sober scenario but that it is possible (and indeed, required) to prepare the data prior to analysis in such a way as to avoid the violation of the PCC. I will consider each line of response in turn. (The two strategies can be regarded as complements and used jointly; I'll separate them analytically and examine them one by one.)

3 Fallacious Statistical Inference

In a recent article (**?**), Kevin Hoover argues that although Sober's scenario appears to violate some formulations of the PCC, it does not constitute a counterexample to the spirit of Reichenbach's original idea, which was: "If an improbable coincidence has occurred, there must exist a common cause" (**?**, p. 156). To Hoover, understanding the correct meaning of the word "improbable" is essential here: it means that the observed coincidence must be something out of the ordinary, something unexpected in order to be evidence for an underlying causal connection. That the car that just passes by is a green Volvo, built in 1990, is in some sense improbable—out of all cars that could have driven by, why should it be exactly this green Volvo?—but it is nothing out of the ordinary. If, by contrast, all the members of a theatre troupe develop identical symptoms of food poisoning after a common meal in the theatre refectory, something out of the ordinary has happened.

In order to flesh out the meaning of "improbably" more formally, Hoover distinguishes between two stages of inference from observations to underlying causal relations. In the first stage, statistical inference, the reasoning proceeds from observed sample frequencies to underlying probabilities. In the second, from probabilities to causal relations. The PCC pertains to the second step: it says that once can infer from a fact about probabilities—the probabilistic dependence between two variables—to a fact about causal relations—either one variable causes another or there exists a common case. Sober mistakenly infers from facts about sample frequencies that the two series are probabilistically dependent, which, thus Hoover, they are not.

Statistical inference, according to Hoover, is always conducted against a probability model, that is, a hypothesis about the stochastic process responsible for the generation of the data; that model is accepted, which is most likely to be true, given the data (pp. 530f.). Claiming that Sober's data violate the PCC makes a fallacious inference at this stage. There is no doubt that the two variables are associated at the level of the sample. That can be readily verified by calculating the sample correlation coefficient:

$$(1.1)\quad r_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}},$$

where a bar above a variable denotes the sample mean. For the data Sober provides, $r_{XY} = .99$. But, says Hoover, we cannot readily take this as evidence that the underlying population correlation:

$$(1.2)\quad \rho_{XY} = \frac{E[(x_i - \mu_X)(y_i - \mu_Y)]}{\sqrt{E(x_i - \mu_X)^2 E(y_i - \mu_Y)^2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

where E denotes the expected value of the expression in parentheses, the μ's are the population means of X and Y, σ_{XY} denotes the covariance between X and Y and σ_X and σ_Y are the standard deviations of X and Y, is positive too. ("Correlation" and "probabilistic dependence" are related but not strictly equivalent concepts. If $P(XY) = P(X)P(Y)$, then $\rho_{XY} = 0$ (if X and Y are independent, then they are uncorrelated) but the reverse is not necessarily the case.[3])

In order to understand the details of this argument, a number of concepts from the analysis of time series has to be introduced (*cf.* **?**, pp. 45f. for the following definitions). The jth autocovariance of a variable Y_t of some process (denoted γ_{jt}) is defined as:

$$\gamma_{jt} = E[(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j})], \tag{1.3}$$

In other words, the jth autocovariance of Y_t is the covariance of Y_t and Y_{t-j}. Further, if neither the mean nor the autocovariances of Y_t depend on time t, then the process of Y_t is said to be covariance- or weakly stationary:

$$E(Y_t) = \mu \qquad \text{for all } t \tag{1.4}$$

$$E[(Y_t - \mu)(Y_{t-j} - \mu)] = \gamma_j \qquad \text{for all } t \text{ and any } j. \tag{1.5}$$

A time series is said to be *strictly stationary* if, for any values $j_1, j_2, ..., j_n$, the joint distribution of $(Y_t, Y_{t+j_1}, Y_{t+j_2}, ..., Y_{t+j_n})$, depends only on the intervals separating the dates (*i.e.*, the j's) and not the date t itself. Sober's series are non-stationary if only because the mean of the process increases monotonically with every observation.

A time series can be non-stationary in several ways. For example, they can be stationary around a deterministic trend, as in:

$$Y_t = \delta t + \epsilon_t, \tag{1.6}$$

where $\epsilon_t \sim N[0, \sigma_\epsilon]$. Such a series is called "trend-stationary". Another form of non-stationarity obtains when past errors accumulate, as in:

$$Y_t = Y_{t-1} + \epsilon_t, \tag{1.7}$$

with ϵ as before. A series such as (1.7) is called "integrated". Let the difference operator be $\Delta X_t = X_t - X_{t-1}$. The difference operator transforms variables measured in levels into variables measured changes and can be

[3]Some distributions may have expectations such that $\sigma_{XY} = E(XY) - E(X)E(Y) = 0$ even though $P(XY) \neq P(X)P(Y)$. For the bivariate normal, the concepts are equivalent though.

applied several times: $\Delta^2 X_t = \Delta(\Delta X_t)$. If differencing d times renders an integrated series stationary, it is said to be integrated of order d. More formally, (p. 543, emphasis original):

> Let d be the minimum integer such that $\{\Delta^d X_t\}$ is weakly stationary. Then $\{X_t\}$ is said to be integrated of order d, which is notated I(d). (By convention, a stationary time series is notated as I(0).)

Processes such as (1.7) are I(1) or integrated of order 1 and also called unit-root processes (because the coefficient on Y_t is unity) or random walks.

Now, Sober's makes a mistake in applying the PCC to the data series he provides because inferring from a sample correlation to a probabilistic dependence means that one takes the most likely data-generating process to be stationary.[4] However, given the data, the most likely data-generating process is non-stationary, for example, trend-stationary or integrated. But that means that a sample correlation or association is no evidence for an underlying probabilistic dependence. And if the two variables are associated yet not probabilistically dependent, the antecedent of the PCC is not satisfied, hence the principle does not apply.

How do we know whether two non-stationary time series are probabilistically dependent? If the series are trend-stationary, Hoover says (p. 541): "Principle (P) [the PCC] would naturally be applied to the stationary components of a pair of trend-stationary series". For integrated series, the test is slightly more complicated. If we have two distinct I(1) processes, a linear combination is usually I(1) too. However, in some cases a linear combination of two I(1) series can be stationary. Then the series is said to be "co-integrated" (p. 545):

> Two time series $\{X_t\}$ and $\{Y_t\}$ are cointegrated if, and only if, each is I(1) and a linear combination $\{X_t - \beta_0 - \beta_1 Y_t\}$, where $\beta_1 \neq 0$, is I(0). (Paraphrased from **?**, p. 571.)

In turn, evidence for two series being co-integrated constitutes evidence for their probabilistic dependence (p. 547). His reformulated principle reads as follows (*cf.* p. 548):[5]

PCC* If variables X and Y are probabilistically dependent (for instance, they are *each stationary or trend-stationary and correlated with each other* or *each integrated and cointegrated with each other*), then either

[4] In fact, it needs to be ergodic. But most series that are stationary are also ergodic and vice versa, so the exact details are of no concern here. For a discussion, see **?**, Ch. 19.

[5] I changed Hoover's wording slightly to make it consistent with the remainder of this paper but without, I hope, distorting his intentions.

X causes Y or Y causes X, or X and Y are joint effects of a common cause.

Hoover thus provides an elegant solution to the difficulty raised by Sober's scenario. Nevertheless I would like to suggest an alternative. My worry is the following: defining the PCC in terms of facts about probabilities rather than sample frequencies deprives the principle of much of its inferential power and to some extent betrays the motivation behind it. Let me explain what I mean by this.

We can understand the PCC (as stated) either in a metaphysical reading or in an epistemic reading. As a metaphysical principle, it would help to *define* the notion of causation.[6] The problem with the metaphysical reading is that, Sober's scenario aside, there are a number of clear-cut counterexamples, such as:[7]

- *Colliders.* When two variables cause an effect, the two can be probabilistically dependent conditional on the effect even though they are unconditionally independent.
- *Mixing.* When populations from different probability distributions are mixed, dependencies can arise even though the homogenous populations are probabilistically independent (see for instance **?**).
- *Laws of coexistence.* Frank Arntzenius has pointed out that many laws of physics can be read as laws of coexistence without posing the need for a causal explanation. He mentions Maxwell's equations, Newtonian gravity, the Pauli exclusion principle and relativistic laws of coexistence (**?**).
- ...

The PCC thus cannot serve as a metaphysical principle in a definition of causation. Indeed, few philosophers would attempt to define causation in such a way today. It is rather used as an epistemic principle for causal inference. As such, however, the PCC as formulated is both too strong and too weak.

Using probabilistic dependencies rather than empirical correlations deprives the principle of much of its inferential power. One reason is that

[6] The PCC by itself obviously gives only a necessary, not a sufficient condition for causation. The reverse condition, in some versions called "faithfulness", is less plausible than the PCC, and I won't discuss it here.

[7] Hoover is fully aware that there are situations in which the PCC does not hold. See for example **?**.

probability distributions do not always exist (as Hoover is aware, see pp. 531f.). They arise rather in fairly special circumstances, in what Ian Hacking called a "chance set-up" or Nancy Cartwright's "nomological machines" (**??**). A chance set-up or nomological machine is essentially a persistent structure that can operate unimpededly and thus allow the generation of probabilities. But there is no reason to restrict the PCC to such special situations. Indeed, at least in some of the examples that are usually given to motivate the principle no such arrangement seems to be in place: the theatre troupe (we suspect a common cause to be responsible for the simultaneous appearance of symptoms of food poisoning in all the members of a theatre troupe after taking a joint meal completely independently of whether or not they regularly eat in the same place or such poisonings occur regularly); two students handing in the exact same term paper; correlations between phenotypic traits in evolutionary biology.

Second, using the PCC as formulated above presupposes that statistical inference is always prior to causal inference (perhaps in a temporal sense, but definitely in an epistemic sense: we need to know probabilities in order to use the principle for inferring causal relations). But such neat division into two stages of inference, and such that one is prior to the other, is not always possible and surely not always the most efficient way to do things. Although I don't think many would disagree (and I know Hoover would not disagree), I would like to point out that background knowledge, including *causal* background knowledge often plays a role in inferring probabilities. In this sense knowledge about probabilities cannot be prior to knowledge about causal relations. Judging whether or not probabilities exists is a case in point: we can determine whether a chance-set up or nomological machine is in place on the basis of the causal knowledge about the situation. We can use that kind of knowledge for determining the potentially relevant variables. And we can use that kind of knowledge in order to determine whether or not time series are stationary.

Suppose we would like to determine the causal efficacy of a training programme and therefore need to determine average test results X. Our data are $X = (x_1 = 153, x_2 = 157, x_3 = 161, x_4 = 168, x_5 = 175, x_6 = 183)$. Suppose further that a greater subscript means "measured later in time" and thus that the data constitute a time series. It is obvious that whether or not we can use this time series for certain kinds of inferences depends on how the measurements were made. If, for example, the data are the results of a test the students sat on a specific date, and the different times are the times at which we finished marking the test (say, at $t_1 =$ 10:25, $t_2 =$ 10:27, $t_3 =$ 10:29 *etc.*) we are entitled to draw an average over time and use it for causal inference. We simply happened to mark the test of

the worst-performing student first, then that of the second-worst performing student and so on. Whether data are arranged in this order or in another order does not make a difference. If, however, the relevant times are $t_1 = 2000, t_2 = 2001, t_3 = 2002$ *etc.* and data record a student's performance on subsequent (though identical) tests, we have to respect the order and drawing an average would not be meaningful. This is because, as we know, students' performance may change over time and the effectiveness of training programmes may accumulate. But there is nothing in the data themselves that tells us this. Moreover, causal background assumptions will often solve the statistical and causal inferential problem at once. If we judge on their basis that sea levels and bread prices cannot be causally connected, it simply does not matter whether they are correlated a sample or population level.

Using causal background knowledge for statistical inference would only be a problem if the contention was that knowledge about probabilities is *always* prior to knowledge about causes (say, because we wanted to use probabilities for a *definition* of causation). But we could subscribe to a more modest claim: there are situations in which after having successfully used our (causal and non-causal) background knowledge for statistical inference, we can use the PCC for further causal inference. Of course, I agree. But even in such situations using the PCC may be unnecessary and cumbersome. Often, our causal background knowledge will allow us to make the causal inference without worrying about probabilities. To use an example of Hoover's (p. 547), it is a very unlikely coincidence that his daughter should have been born on the day the Challenger space shuttle blew up. But are we looking for a common cause here? Of course not. Our causal background knowledge tells us immediately that this is a mere coincidence (rather than a genuine co-occurrence of events that warrants the search for causal relations). We do not detour via judgements about probabilities in such cases.

The principle as stated by PCC* is also too weak because the problems for the metaphysical reading of course double up as problems for the epistemic reading—unless one qualifies the latter. Colliders for instance are a serious problem for practical causal inference because we often collect data with a specific purpose in mind. But this may mean that everybody in that population has a specific characteristic, say Z. Now, if X and Y are both causes of Z, they are probabilistically dependent *conditional upon* Z even though (let us suppose) X and Y are unconditionally independent. The problem is that often we may not notice that we sampled only members of the populations in which Z is present. Importantly, this is a problem at the level of populations, not samples. And: it may obtain for stationary as well

as non-stationary variables. Similarly, problems regarding heterogeneous populations are a serious problem for practical inferences.

Before presenting my own proposal for a reformulated PCC, let me discuss the second strategy to deal with Sober's scenario, data preparation.

4 Data Preparation

In a discussion note on Hoover's paper Daniel Steel disagrees with Hoover's analysis that the Sober scenario is a problem for statistical rather than causal inference. With Spirtes, Glymour and Scheines he thinks that the problem in Sober's case is just a special case of mixing. He provides the following argument (**?**). Central to his demonstration is the so-called mixing theorem, which can be applied to time series (p. 310). For the simple case of $T = 2$ it reads:

Mixing Theorem. Let $P(XY) = nP_1(XY) + mP_2(XY)$, where n and m are real numbers greater than zero such that $n+m = 1$. Let $P_1(XY) = P_1(X)P_1(Y)$ and $P_2(XY) = P_2(X)P_2(Y)$. Then $P(XY) = P(X)P(Y)$ if and only if

$$P_2(X)P_2(Y) + P_1(X)P_1(Y) = P_1(X)P_2(Y) + P_2(X)P_1(Y)$$

An important corollary is the following:

Corollary. Let $P(XY) = nP_1(XY) + mP_2(XY)$, where n and m are real numbers greater than zero such that $n + m = 1$. Let $P_1(XY) = P_1(X)P_1(Y)$ and $P_2(XY) = P_2(X)P_2(Y)$. Then $P(XY) \neq P(X)P(Y)$ if and only if $P_1(X) \neq P_2(X)$ and $P_1(Y) \neq P_2(Y)$.

As briefly mentioned above this shows that dependencies can arise whenever populations from different probability distributions are mixed, even though the homogenous subpopulations are independent. It is important to note that what matters here is that the mixed population is *probabilistically* heterogeneous independently of whether or not it is *causally* heterogeneous. Some arguments to the effect that this case isn't one of mixing because two processes may be causally identical (because time is not a cause) and yet does the problem associated with the Sober scenario arises are somewhat beside the point (this argument has been made for instance by **??**).

What is wrong with the Spirtes, Glymour and Scheines/Steel proposal in my view is that it suggests the wrong methodological lessons. The natural response to a problem posed by the heterogeneity of populations is, to require that populations be made homogenous prior to analysis. We can, for example, regard each x_t (for all $t = 1, 2, ..., T$) in the time series as drawn

from a different variable X_t (for all $t = 1, 2, ..., T$) and analyse only contemporaneous probabilistic relationships.[8] Alternatively we can condition each variable on its past and analyse the probabilistic relations between the conditional variables $X_t|X_{t-1}$ *etc.*[9] Another alternative would be to difference the data prior to analysis or to use another preparation method in order to homogenise the variables. This seems to be exactly what the proponents of this solution have in mind. Clark Glymour, for one, says (**?**, pp. 73f., emphasis added):

> Applying the program [that incorporates a version of the PCC as core assumption] to real data requires a lot of adaptation to particular circumstances: variables must often be transformed to better approximate normal distributions, decisions made about modeling with discrete or continuous variables, *data must be differenced to remove auto-correlation*, and on and on.

Similarly, Steel writes (**?**, p. 314):

> [T]he above discussion illustrates how researchers interested in drawing conclusions from statistical data can design their investigation so that counter-examples like Sobers are not a concern. For instance, if the series is non-stationary but transformable into a stationary one via differentiating with respect to time, then differentiate. Then PCC can be invoked without concern for the difficulty illustrated by the Venice-Britain example.

The idea seems to be that data can always be suitably prepared (by, say, conditioning on time or on series' past or by prior differencing or detrending) before using the PCC for analysis. Thus, we can reformulate the PCC as follows:

PCC** If two *suitably prepared* random variables X, Y are probabilistically dependent, then either X causes Y, Y causes X or X and Y are the joint effects of a common cause Z.

Unfortunately, data preparation does too much and too little at the same time. Regarding only contemporaneous statistical relations, conditioning on the past of variables and differencing, detrending *et al.* all result in the loss of important long-run information that a prudent statistician should make use of (see for instance **?**, Sect. 7.4). The statistical concepts of cointegration, which Hoover discusses at length, and co-breaking (see for instance **?**) were developed specifically in order to deal with Sober-like situations while retaining the long-run information contained in the time series. Consider the following. Unit-roots processes are sometimes said to

[8] An anonymous referee, for example, wrote: "What two variables [this relates to my X and Y]? What we have is two times series: $X_t, X_{t-1}, X_{t-2}, ..., X_{t-n}$ and $Y_t, Y_{t-1}, Y_{t-2}, ..., Y_{t-n}$. There are $2n + 2$ variables!"

[9] This seems to be suggested by Frank Arntzenius, see **?**, section 2.3.

be subject to *stochastic* shifts because the error term accumulates over time. By contrast, when *deterministic* shifts occur, parameters of a process such as its mean, variance or trend change. In recent econometrics, methods have been developed that exploit information about common shifts in two or more series for causal inference. Through differencing, however, this information can be lost. Consider the following series:

$$X_t = \alpha_X + \beta_X \delta t + \epsilon_{X_t} \tag{1.8}$$
$$Y_t = \alpha_Y + \beta_Y \delta t + \epsilon_{Y_t}, \tag{1.9}$$

where the δ denotes a common trend. When there are shifts in the value of that coefficient, these common "breaks" can be detected by statistical methods. The breaks will disappear, however, when the series are differenced. In other words, if the non-stationarity of a series is due to shifts in deterministic coefficients, series should not be differenced prior to causal analysis (even though such series may be, as Steel demands, non-stationary and transformable into a stationary one via differencing). The point is that prior data preparation of the kind discussed here ignores that we are dealing with continuous and persisting processes here, and not merely with contemporaneous events or changes, and the analysis methods we use should reflect just that (**?**, 165).

While data preparation will often result in information loss, it may not solve the problem either. Differencing can remove some sources of nonsense correlation but is often inapplicable. Of course, it is an analytical truth that differencing removes unit roots. If we have two independent random walks:

$$X_t = X_{t-1} + \epsilon_{X_t} \tag{1.10}$$
$$Y_t = Y_{t-1} + \epsilon_{Y_t}, \tag{1.11}$$

then a regression of the differenced series such as

$$\Delta Y_t = \beta \Delta X_t + \nu_t \tag{1.12}$$

will, correctly, find a zero regression coefficient. But unit roots constitute only one source of non-stationarity, and non-stationarity is only one source of nonsense correlation.[10] Differencing is ineffective when nonsense correlation arise in stationary series. That is, even stationary time series can appear correlated even though they are *ex hypothesi* causally independent. Indeed, in his original article on the problem Yule did not discuss the problem in the context of non-stationarity. For example, for two series of the

[10] On the pitfalls of differencing see the papers **??**; in defence of *a priori* differencing, see **?**.

form:

$$X_t = \theta_X X_{t-1} + \epsilon_{X_t} \tag{1.13}$$
$$Y_t = \theta_Y Y_{t-1} + \epsilon_{Y_t}, \tag{1.14}$$

where $|\theta_i| < 1$ and the ϵ_{it}'s are i.i.d. and zero mean one can show that nonsense correlations obtain regularly (**?**, p. 899). As one can see easily, the mathematical form of the differenced series is exactly the same as that of the original series. Although in this context, the problem arises to a somewhat lesser extent—when θ is 0.75, significant correlations obtain in about 30% of the cases—the problem does obtain despite the fact that the series are stationary. Something similar happens when time series are moving averages, as in the following:

$$X_t = \sum_{j=0}^{k} e_{X,t-j} \tag{1.15}$$
$$Y_t = \sum_{j=0}^{k} e_{Y,t-j} \tag{1.16}$$

Even if k is only 5, nonsense correlation results obtain in about a third of cases (**?**, p. 902).

Lastly, serial correlation may persist even after differencing, and in fact will in general persist. In his 2001 paper, Sober discusses an example from evolutionary biology in which similar developmental sequences evolve independently in two lineages (pp. 335ff.). This is also common in time-series analysis. Series that are integrated of orders higher than 1 will have to be differenced several times before achieving stationarity. There are series that are fractionally integrated (*i.e.*, they have a non-integer order of integration), which can lead to spurious results whenever the orders of integration sum up to more than 0.5 (**?**). Such fractional orders of integration can obviously not be removed by differencing.

Other *a priori* data preparation methods do not fare better. An alternative method to remove non-stationarity is detrending, *i.e.*, subtracting a linear deterministic trend before the analysing data. However, detrending too can yield spurious results, namely when the processes are unit roots (see **?**, Sect. 4.3). Lesson: there are no data preparation methods that can be used prior to systematic statistical (and therefore causal, see above) analysis. Hence, PCC**, too, fails.

5 Non-Statistical Sources of nonsense Correlations

There are also various non-statistical sources of nonsense correlations. For example, when variables are connected because of logical, mathematical and conceptual links, they may be correlated but the correlation is not due to a

causal connection. Non-causal physical laws may provide yet another source of correlation. These problems are not peculiar to time series and they are well recognised in the literature (for a valuable overview, see **?**, pp. 52ff.). For the sake of completeness, let us briefly consider each issue in turn.

When variables have logical links, correlations are not indicative of causal connections. "Day" may be perfectly negatively correlated with "Night" but this is because of a logical, not a causal relation. A variable and a function thereof are highly correlated. Time-series econometrics often uses logarithms of variables such as money, income and prices. The logarithms are correlated with the original variables but not for causal reasons. More serious are conceptual links between variables. Many economic variables are linked because they have interdependent measurement procedures. For example, the measurement of some variables is based on the same national accounts.

There may be other non-causal constraints between variables, for example, when two variables are related by budget constraints. Consumption (C) will be correlated with savings (S) not because they are causally related but because a third variable, income (I) constrains them by the mathematical relation $C + S = I$. Problems of this kind and conceptual relations can be very serious in time-series analysis, especially in social science applications.

6 Evidence, Eliminative Induction and the PCC

Clearly, there is a core of truth in the PCC. But what is it? In some cases it does indeed work, just think of Salmon's famous examples: the theatre company all of whose members out of a sudden get violently ill; the students who hand in the exact same term paper; the twin quasars (**?**, p. 158f.). Can we formulate the principle in a way as to avoid the counterexamples discussed here?

The core of it, in my view, is that an empirical or sample correlation between variables sometimes provides evidence for the hypothesis that these variables are causally connected. In many cases, of course, the correlation arises for reasons other than causal connectedness: the sample is small; there is selection bias; there are logical, mathematical or conceptual relations between them; they are generated by unit-root processes. If we can rule out these non-causal accounts for the correlation, then the causal account is probably true.

In order to formalise this idea to some degree, recall Patrick Suppes' probabilistic theory of causation (**?**). He defined an event A as a *prima facie* cause of another event B if and only if A precedes B and $P(AB) > P(B)$. Not every *prima facie* cause is, however, also a genuine cause. Thus he defined as *spurious* cause an event A that is a *prima facie* cause of

event B but such that there is a partition π prior to A such that every element C_i in that partition renders A and B probabilistically independent: $P(B|A, C_i) = P(B|C_i)$(for all i). A genuine cause is a *prima facie* cause that is not spurious.

Analogously, we can define e as *prima facie* evidence for hypothesis h if and only if e stands in an appropriate relationship with h. In the present case, the appropriate relationship is explanatory: e is *prima facie* evidence for h if and only if h, if true, explains e: a causal relation between two variables (whether direct or due to a common cause) explains the correlation between the variables.[11] e is spurious evidence for h if and only if e is *prima facie evidence* and h is explained by an alternative hypothesis h_i^a. If e is *prima facie* evidence and not spurious, it is genuine evidence.

Thus the PCC reads:

PCC*** The proposition e = "Random variables X and Y are (sample or empirically) correlated" is *prima facie* evidence for the hypothesis h = "X and Y are causally connected". If all alternative hypotheses h_i^a (*e.g.*, "the correlation is due to sampling error", "the correlation is due to the data-generating processes for X and Y being non-stationary", "X and Y are logically, conceptually or mathematically related") can be ruled out, then e is genuine evidence for h.

There are various advantages of this formulation. First, and foremost, it is very explicitly formulated as *epistemic* principle. There is no way to misread the principle as saying that all correlations must have causal explanations for instance. Unlike previous versions, this formulation makes evident that the principle is fallible. It nowhere says that the evidence is conclusive or that the evidence entails its hypothesis. Evidence provides a reason to believe, not metaphysical certainty. Second, it makes evident that causal inference is context-dependent. I haven't said a lot about where the alternative hypotheses h_i^a come from. Usually our background knowledge about the situation under investigation will determine what plausible candidates there are, and how much effort needs to be taken to rule them out. Third, unlike Hoover's PCC*, PCC*** is very widely applicable. In particular, it is applicable to cases of empirical correlation where no probability distributions exist or where statistical inference may be difficult or cumbersome. Fourth, unlike Steel's PCC**, it does not rely on data-preparation techniques that frequently do more harm than good.

[11] There are also cases where evidence explains the hypothesis for which it is evidence, and where a third statement c explains both evidence and hypothesis (see **?**). Moreover there are cases where the evidential relationship is not explanatory (for a discussion, see **?**, Ch. 1). My definitions are thus intended to apply only to the present case.

There may be a residual worry that I throw out the baby with the bath water. What happens if one of the alternative hypotheses h_i^a is true but X and Y are causally connected nevertheless? But this isn't a problem as I make no suggestion to the effect that empirical correlations can be the *only* evidence for causal connections. Of course, this is absurd. When samples are small and statistical tests lack power, one can try to physically investigate the units. When time series are non-stationary, one can try to exploit "structural breaks" in the series for causal inference and use tests of the kind Hoover (**?**) and David Hendry and his collaborators (*e.g.* **?**) have developed. In other cases we may be able to improve the quality of the data and thus ameliorate problems. In yet other cases we may be able to use one of a host of qualitative methods such as ethnographic methods. Violations of the PCC are only a problem if one thinks that all correlations need a causal explanation and if one thinks that the principle is the only or only appropriate or most important or "core" principle of causal inference. But it isn't. It is one of many such principles and has it's own advantages and drawbacks. What is important, though, is to keep its limitations in mind.

Thus let me end with an irony. When investigating what kinds of systems do we have good reason to believe that most or all of the alternative hypotheses are false? As mentioned above, shifts that render a time series non-stationary can be of two kinds: "stochastic" and "deterministic". Stochastic shifts obtain when error terms accumulate; deterministic shifts, when deterministic parameters (such as coefficients on trends) change. Systems where neither kind of change is likely are systems that lack internal dynamics and that are shielded from outside influences or "closed". Moreover we want to rule out chance associations and thus require that the static and closed system persists for a while so that sample sizes are sufficient. We also want to make sure that the populations in the system are homogeneous, that variables are well-measured and so on. Now, aren't these characteristics the characteristics of experimental systems? But if they are, why do we need the PCC to draw causal conclusions?

Acknowledgements

I'd like to thank Nancy Cartwright, Damien Fennell, David Hendry, Elliott Sober, three anonymous referees as well as audiences in Kent and Vancouver for helpful suggestions. Special thanks go to Kevin Hoover who provided extensive and extremely valuable comments on several drafts of this paper. Some disagreements remain but the paper profited enormously from the discussion.

Julian Reiss
Department of Logic and Philosophy of Science, Complutense University,

28040 Madrid, Spain.
and
Centre for Philosophy of Natural and Social Science, London School of Economics, Houghton St, London WC2A 2AE, United Kingdom.
jreiss@filos.ucm.es