From spurious correlation to misleading association:

The nature and extent of spurious correlation and its implication for the philosophy of science with special emphasis on positivism

Hans O. Melberg

The Philosophy of Social Science,
Dr. Polit course at the University of Oslo
Oslo, 15. September, 2000.

Introduction

In what way is the philosophy of science relevant to a researcher – like me - who is asked to respond to the following question: "What is the best treatment for drug addiction?"

I have no complete definition of the philosophy of science, but among other things it involves questions like what kind of knowledge we should seek, whether it is possible to achieve neutral knowledge and how we best should go about finding reliable knowledge (See, for instance, the list of questions in Hausman (1992: 281-282)). Our researcher - who until now has been blissfully unaware of explicit writings in the philosophy of science – hears about a book trying to answer the questions above and he eagerly runs down to the library hoping to learn how (and if) he should approach the question of finding the best drug-treatment?

Unfortunately, the philosophy of science does not speak with one voice. Different people have grappled with the mentioned philosophical questions and provided conflicting answers. To reduce the confusion slightly one might try to identify a few sets of typical answers which may be honoured with a label like positivism, idealism and so on. Our confused researcher might, for instance, read a book by a positivist/empiricist who claims that the aim of science is to find causal laws and that this should be done by searching for regularities in the empirical facts. Moreover, the positivist argues that the answer thus provided is neutral since there is a distinction between facts and values. Is this good advice?

One way of evaluating the advice would be to examine how it would work out if it were followed. Let us then assume that our newly converted researcher proceeds as follows: He collects data on the success rate of different treatment programs (defined as a reduction in consumption of drugs, the health problems produced by drugs and the level of crime related to drugs). He then conducts the appropriate statistical tests to discover regularities in the data and announces that science has proved that treatment X produces the best result (= the law produced by the research). Now, it would be impossible to evaluate all the different aspects of a positivist approach to the question of drug treatment as exemplified by our imaginary researcher. For instance, I shall mainly ignore the debate about facts and values. I shall also leave out many other interesting comments one might explore (Why would you want to treat people who are using drugs? How do you define the "best" treatment). I will, however, focus

on one particular problem in the story above: The problems that arise when research is based on the assumption that a constant conjunction is a necessary and sufficient condition for making a statement about causal laws. The last phrase is from (Smith,1998: 301) and he labels this position empiricism, as opposed to idealism (constant conjunction is necessary but not sufficient) or realism (constant conjunction is neither necessary nor sufficient).

Based on the discussion above the relevance of the topic should be obvious both to the specific research on drug treatment and more generally for the philosophy of science and the evaluation of positivism/empiricism. If constant conjunction is no good as a guide to causal connections, then the researcher might draw the wrong conclusions – mistakenly arguing that X is best when in fact Y is the best treatment. More generally the positivist philosopher has to face several unpleasant options. He might downplay the degree to which there is a mismatch between constant conjunction and causation in which case he runs the risk of at least sometimes accepting spurious relationships as genuinely causal and – conversely – ignoring true causal relationships that do not manifest themselves in constant conjunctions. Alternatively he might admit that the problem is serious in which case he has to accept the rather disappointing conclusion that the scope for reliable knowledge is not very large. Then, there is the third option of admitting that you were wrong and that you should rather try a different approach to knowledge, like idealism or realism.

The arguments presented in this paper, are that – firstly – the problem of spurious correlation is large in the social sciences. This should lead us to focus on ways to avoid the problem – either by specific tests or by adopting rules of research that reduces the problem of spurious association. Although one may make some progress in this direction, it may be impossible – both for practical and for conceptual reasons – to go very far. Taken together the widespread existence of spurious associations and the limited possibility of revealing the spuriousness, implies that our knowledge is severely restricted. This, in turn, implies - as argued by Jon Elster (1998) - that we would be better off trying to search for mechanisms instead of laws. Lastly, I will not make any large scale pronouncements on whether the problem should lead to the rejection of positivism or empiricism and the adoption of realism. This does not depend on spurious association alone and limits of space and time precludes investigation of more topics than the one I have selected: spurious association and positivism. I will, however, reject the claim that since the positivist definition of knowledge leads to such a limited scope for

knowledge that instead of trying to explain things causally we should rather seek understanding of phenomena based on their symbolic meaning.

The arguments above shape the structure of this paper. First I will try to convince the reader that the problem of spurious association is widespread. Second, I will review some of the ways in which the problem can be reduced and the arguments why these do not go very far. I will then make the connection between this and the conclusion that we should focus on mechanisms instead of laws. Before all this, however, it is necessary to make a few brief comments on terminology.

What is the problem?

One of the first discussions of spurious correlation appears in the writings of Karl Pearson in 1896. John Aldrich (1995) – my main source of information about the history of spurious correlation – writes that Pearson used the term spurious correlation to "distinguish the correlations of scientific importance from those that were not." The problem, according to Pearson, was that some correlations did not indicate an "organic relationship." Although this term is never defined, the examples used suggest that spurious correlation was the same as a correlation between two variables that were not causally connected.

There are problems with both terms in "spurious correlation." The first problem was noted by the next major writer on the topic, Udny Yule. Although he sometimes used the phrase spurious correlation about constant conjunctions that were not causal, Aldrich argues that Yule really preferred to use a different terminology – such as "misleading" or "illusory" correlations. This is in many ways better than spurious correlation because it is not the correlation itself that is spurious, but the inference from the existence of a significant correlation coefficient to the existence of a significant or causal relationship.

The problem with the last term - correlation - is that the correlation coefficient only measures the strength of linear relationships. In the present context however I am concerned with the more general topic of regular conjunctions of all types – linear or non-linear. This should lead us to adopt more general measures of association capable of capturing non-linear associations. For instance, in their article on correlation and causation Ellett and Ericson (1986) present four different measures of association using the concept of probability.[1] Finally, in the litera-

ture it is not only correlations which are labelled spurious but also regressions. This is yet another reason to avoid using a narrow term for what is really a more general phenomenon.

In short, were it not for convention I would rather use the term “misleading associations” than spurious correlation. Alternatively one might simply interpret the phrase “spurious correlation” more generally than its literal meaning. In any case, it has to do with a constant conjunction that does not indicate a significant and direct causal relationship. In fact, the problem of course, is even wider than mistakenly accepting misleading correlations. The other half of the problem would be failing to detect causal relationships that are causal but do not manifest themselves in regular associations. This does, however, not have its own name reflecting – perhaps – an undeserved ignorance of a problem that logically must exist in the same way as spurious correlation as long as there is an imperfect relationship between variables that are associated and variables that are causally related.

In order to claim that a constant conjunction is not causal (i.e. spurious or misleading) or that a true causal relation is ignored, we automatically make some kind of judgement as to what to count as a causal relationship. We demand that there be something more than a constant conjunction before we label it causal- at the very least we usually demand that the cause be close to the effect in space and time. The problem, however, is that there is no universal agreement on exactly what more (if anything) we should demand. There are literally dozens of different accounts of causation, ranging from Hume’s definition of constant conjunction, Lewis’ analysis of causation and counterfactuals and Mackie’s famous INUS criteria (See (Sosa and Tooley,1993) for a collection of philosophical views on causation). I will argue, however, that it is possible to sidestep the problem of having to define what causation really is in my debate about misleading association. How is this feat possible?

In defence of my lack of a complete definition of causality, I shall make three arguments brief arguments. First of all, I shall argue that I can come a long way in my arguments without a complete concept of causation. When discussing the meaning of spurious correlation above, I deliberately emphasised the interpretation that focused on the misleading nature of the correlation. A correlation can be misleading from many different perspectives. It can be misleading from a policy perspective; it can be misleading from an explanatory perspective and it can be misleading from the perspective of causality. There is not a one-to-one correspondence here

and it is easier to show by examples that a correlation is misleading from a policy perspective than to show that it is non-causal (see the discussion on the Lucas critique below).

The first argument, I admit, is rather weak. So I will provide a second argument, which is that the concept of causation is a primitive concept i.e. it is a concept which cannot really be fully defined, but we may still use it. G.E.M. Anscombe puts it this way: "…might it not be like this: knowledge of causes is possible without any satisfactory grasp of what is involved in causation." (Anscombe,1993/1971: 91) As an analogy he offers the illustration that we all know about some "long run frequencies" although we are not able to define exactly what a long run frequency is in general. In the same way, causation may be a primitive concept which cannot be broken completely down but it is still sometimes possible to say something about when there is causation and when there is not.

Even with the second argument I still feel rather uncertain. The third argument, however, is more firmly grounded. The argument would be that it is impossible to define causation in general because it relies on normative assumptions. The point is well illustrated if we move outside philosophy and economics, and into the field of causation in history. In a very revealing article W. H. Dray (1978) shows how different historians rely on different paradigms of causal thinking when discussing the causes of World War II. The background is A.J.P. Taylor's arguments to the effect that previous historians had overestimated the role of Adolf Hitler as a cause of the war. Against this another historian, Trevor Roper, argues that Hitler was very important as a cause. As it turns out one important reason for the different arguments is that they define causes as something that is "not normal" and that they have different definitions (as most people do) of normality. A general example is the following: "The storm is said to be the cause of widespread flooding because it upsets normal patterns of drainage …" (Dray,1978: 161-162). Similarly, A.J.P Taylor seems to regard it to be "normal" at that time to desire power and territory, while the Versaille Treaty was "artificial" or "unnatural" – a large nation like Germany could not be denied its "normal" weight in international affairs and most German leaders – not just Hitler – were nationalists. Trevor Roper takes the opposite view and regards Hitler as unnaturally nationalistic and the treaty of Versaille as quite proper. Because of this disagreement on what should count as normal or reasonable they also disagree over the importance of different causes of the War. To me this is a convincing example that causation may involve normative claims that makes it impossible to give a "scientific" definition of causation.

I have so far discussed the terms – spurious, correlation and causation – or more correctly in the case of the last: argued that it is impossible to define it and that it is possible to go some way on the topic without defining it. It is now time to return to the main argument: the problem of misleading association.

How serious is the problem?

How do you measure the seriousness of a problem like spurious correlation? Trying to estimate the share of misleading associations out of all possible associations (genuine and spurious) is both impossible and rather meaningless. We simply do not have the information to do so and even if we had it seems wrong to include all obvious associations. For instance, there is a high degree of constant conjunction between me turning on the electricity and the light that radiates from my lamp. Although my action and the light is causally related this is not really an interesting correlation from a scientific point of view. To include all such cases as examples of successful inference from correlation to causation would not produce a very interesting measure. What we really want is an indication of the amount of spurious correlation in the associations we consider to be interesting. We seek knowledge that is non-obvious and it is the share of misleading correlations in that area that is the valid measure of how serious the problem is.

I do not know of a good way to estimate this. One could, perhaps, pick out the causal claims made based on constant conjunction made in a journal over a year and try to find out how many of those that later turned out to be spurious. As a poor second best, however, I will try to convince the reader about the seriousness of the problem by throwing a large number of examples at him. The reader might, of course, suspect that I am using all possible examples in order to magnify the importance of problem. In order to weaken this suspicion I will try to provide some slightly more general mechanism that create misleading correlations (suggesting that my examples can be multiplied). I will also present a list of how spurious association may arise at all the different stages in the research process. Finally, I will try to convince the reader about the pervasiveness of misleading correlations by challenging him to present a list of (non-obvious) genuine correlations. If this list is not significantly longer than the list I have of spurious correlations, then I claim victory in my quest to convince him of the importance of the problem.

Some examples

In his classic article on spurious correlation, G. Yule (1926) reports that the correlation between mortality and proportion of church of England marriages to all marriages is 0.9512. That is, over time there has been a decrease both in the share of marriages in the church of England and the rate of mortality. Despite this Yule does not find it plausible to suggest that the two variables are causally related. It is easy to find similar examples. For instance, David Hendry (1980) reports that the correlation between inflation and cumulative rainfall in the UK is 0.998, yet few would argue that it is the rain that causes inflation; Ole-Jørgen Skog reports that the correlation between the quarterly index of intravenous drug abuse in Stockholm and Wölfer's index of sunspot activity is 0.91 (1965-70), but it is hardly plausible that sun-spots somehow cause addiction (Skog,1988: 570). Now, what is common in all these examples is that we are dealing with time-series and non-experimental data. The easy by which we can find these examples suggest that spurious correlation might be a serious problem for non-experimental time-series.

Spurious correlation over time could become a problem in my research. For instance, I might find that people who leave treatment have a lower consumption of drugs than before they entered treatment i.e. there is a correlation over time between treatment and lower consumption of drugs. It would be tempting, but wrong, to automatically conclude that this indicates a causal relationship (treatment really works!). One reason why the relationship could be spurious is that people enter treatment when they are exhausted after a period of intensive use. Thus, one would expect a reduction in the use even without treatment (assuming the pattern of use goes up and down). Moreover, it is wrong to ascribe all the difference between pre- and post-treatment to the treatment itself since there may be many other variables that affect the outcome. For instance, it might be that the decision to enter treatment itself is the causally important variable, not the treatment the person actually receives. In sum, there is an obvious danger of spurious correlation over time in my research.

The problem, however, is not isolated to time series. Imagine, for instance, that you need information about whether you should send a drug addict to a day-based or residential treatment. You tell a researcher to gather information about the rate of success in these two programs. Imagine the results turn out to be as follows (The example is based on numbers from (Lindley and Novick,1981: 45-46)):

TABLE 1				
	Cured	**Not cured**	**Total**	**Success rate**
Day-care	20	20	40	**50%**
Residential care	16	24	40	**40%**
Total	36	44	80	

From Table 1 one might infer (weakly) that the best treatment is "Day-based." Why is it viewed as the best treatment? The short answer is that the share of people who are cured is higher in day-based programs than in residential programs. More technically, the correlation between "day-treatment" and "cure" is higher than the correlation between "residential" and "cure."

In many cases it may work well to reason as above. This paper, however, is concerned with the problems. Consider, for instance, the possibility that residential programs receive more hard cases than day-care programs. Another researcher takes this into consideration and presents the following information (using the exact same 80 clients):

TABLE 2								
	Easy cases				**Hard cases**			
	Cured	**Not Cured**	**Total**	**Success rate**	**Cured**	**Not cured**	**Total**	**Success rate**
Day-care	18	12	30	**60%**	2	8	10	**20%**
Residential care	7	3	10	**70%**	9	21	30	**30%**
Total	25	15	40		11	29	40	

Based on the information in Table 1, the best choice of treatment appeared to be "day-care" but based on the information in Table 2 the best choice seems to be "residential care." A closer examination also reveals a very non-intuitive possibility. According to Table 2, residential treatment is best *both* for hard cases and easy cases. However, if you put the informa-

tion together in a 2x2 table, it turns out that day-care is best. In short, we have two samples that both point in one direction, but the united information from the two samples point in the opposite direction. This is what is called Simpson's paradox and it is a stark example of how it is possible to get spurious correlation even when we do not have time-series data.

Simpson's paradox is not the only example of spurious correlation in cross-sectional data. In order to impose some structure on the examples, I will list some problems in the order they might appear when we do research.

Some general causes

The problem above was caused by a *selection effect* in the sample in non-experimental data. One type of clients appeared more frequent than others in others in one treatment program. Non-experimental data are often plagued with these selection effects. As an illustration, we might consider a society in which most people do not marry out of love, but allow their parents pick their partners (The example is from (Elster,1993), who in turn has stolen it from Tocqueville). A researcher wants to examine which couples are the happiest; those marrying for love or those who marry whom their parents tell them to marry. He finds that those marrying for love is generally much less happy than the other group and concludes that it would lead to great unhappiness if we introduced a system in which everybody was supposed to marry out of love. One problem with this research, may be that only those who are stubborn enough to go against the tradition actually marry out of love. Hence there is a selection effect: A disproportionate share of stubborn people marry out of love compared to the rest of the population. Now, two stubborn people may not be very happy together and this explains the correlation. Because of this it would be very wrong to conclude that marrying for love for all people would lead to unhappiness. A third example of the selection effect, may be the high correlation between breaking up and living together without being married – as compared to the correlation between divorce and being married. This is sometimes used as an argument to get married – as exemplified by comments when the Prince of Norway chose to move in with somebody without marrying that person. There might be a causal relationship as well here, but there surely is a selection effect in the sense that those who marry might be more religious or more sure that they want to spend the rest of their lives together than those who move in together (Often people who marry have lived together for some time before getting married). If this is the true reason the correlation is no good reason to give the advice that we should all get married.

Marrying for love can also be used to illustrate how the problem of spurious correlation may arise at the next sage in the research. The experiment itself may produce a bias. The classic example of this in experimental studies is the "Hawthorne experiment." In this experiment one "treatment" seemed to be better than the other, but in fact the measured effect probably was more due to the extra attention given to the workers from the researchers, than the actual "treatment." In other words, there is a third variable that causes the correlation. A similar example from non-experimental data could be exemplified by – as mentioned – marriage for love in a society in which this is unusual. Imagine that there is no selection effect (those who marry for love are not unusually stubborn etc). However, one could easily imagine that people who decide to marry for love without parental consent could be ostracised; or at least that they would loose some family support. This, in turn, may make the couple unhappier than the rest of the population (assuming the effect is larger than the pleasure of marrying your loved one). If this is true it is not the fact that the person married for love that caused their unhappiness, but it was the lack of support from and contact with their parents later. These are two – of many – examples of how third variables might make a correlation spurious when we have cross-sectional data, both experimental and observational.

Things can, of course also go wrong when analysing data. One particularly strong reason to suspect that many relationships reported are spurious is data-mining. That is, the researcher will simply keep going until he finds a relationship that is significant. He may either try out new variables or playing around with the form of the relationship assumed until he "discovers" something that it statistically significant. It is of course, only the last and significant result that is reported. This method is likely to generate many spurious relationship because one is bound – sooner or later – to find a variable that is associated with another maybe for no other reason than accident.

When can it go wrong?

The problems described above can be categorised based on when in the research process the mistake occurs that later creates the misleading association. Assume an empirically oriented positivist scientist is given a question (Is day-treatment or residential treatment best for drug addicts?). He then lists the variables believed to be important (whether they receive day-treatment or residential-treatment, sex, age). To examine whether the variables really are important he collects data (empiricism; experimental or non-experimental data). Based on the

strength of the correlations in the data he concludes with a causal probabilistic law (the causal relationship between day-care and success is strongest because it has the highest correlation). This knowledge is then used to give policy advice. Things can go wrong at all these stages which in turn may produce spurious associations: He may forget some important variables or mechanisms (or he may include some irrelevant variables or mechanisms); he may get a biased sample (e.g. sorting mechanisms as described above); the experiment itself may induce biased results (e.g. the Hawthorne effect); the recording out the experiment may be inaccurate (measurement errors); he may use inappropriate statistical theory when analysing the data. And if the possible mistakes so far has not produced a misleading association, the researcher will make sure that he finds one by re-specifying the relationship or looking for different variables until something "significant" turns up. Indeed, he may even be wrong in the sense that there are no causes and no laws and no "neutral" empirical observations to be collected.

Success stories

The list above is not meant to be completely nihilistic about the possibilities of using correlations to help develop reliable and useful knowledge. There are, indeed, some examples where the discovery of non-obvious correlations has suggested (and verified) important causal relationships. David Freedman (1999) mentions at least two such examples in an article entitled "From Association to Causation: Some remarks on the History of Statistics." The first example concerns the spread of cholera. Before 1855 some believed cholera was caused by imbalances in the humors of the body, some said it was bad air (miasma) others claimed it was poison in the ground. Then in 1855 John Snow produced convincing evidence to the effect that cholera was in fact a waterborn disease. Her did this by showing the high correlation (visually) between people who used the same water source and the occurrence of the disease.

Another example of an important correlation is the following. In 1867 a doctor named J. Lister published a paper which showed that surgery was much safer when the environment was sterilised (a correlation). Previously the Hungarian doctor I. Semmelweis had been ridiculed for suggesting that there was a connection between the dirty hands of hospital staff and infections caught by women after childbirth. Although there was a correlation there was no "scientific reason" to support it. Pasteur then provided the causal mechanism by demonstrating how bacteria in the air or on hands could cause the disease. After Lister's paper hygienic standards were raised and the occurrence of the disease decreased drastically

There are also, of course, countless examples of important small discoveries made in experimental situations based on correlation in the physical and chemical sciences (discovery of new materials). Although one might question whether we are here talking about "constant conjunction" being the source of the causal suggestion. In fact, often it seems possible to be mainly convinced after only one try: We might try different materials to produce the correct amount of resistance to produce the thread in a light bulb. When we find one that is just right (strong enough not to break; weak enough to glow) natural scientist do not go on to try the same material 30 times before they are convinced. They are often quite sure (and rightly so) after the first try. In this sense it would be wrong to argue that the natural sciences exemplify the importance of constant conjunction for revealing causation.

Even if it could be established that in the natural sciences correlation is a good indicator of causation, it does not imply that the same is the case in the social sciences. In fact, there is one important distinction which makes correlation more likely to be misleading in the social sciences then the natural sciences. In the natural sciences the object being study do not learn about the conclusions of the study, while in the social sciences the agents will learn and exploit a causal law as soon as it is discovered. The general problem is sometimes referred to as self-reflexivity and in economics the problem is often labelled "the Lucas critique." Basically this implies that a correlation can be misleading in the sense that if the policy-makers try to exploit it (using the "law" for policy purposes) they will fail because as soon as the policy makers act, the "law" is changed.

One example of the problem discussed above could be the correlation behind the Phillips curve. At one stage it appeared as if there was a trade-off between inflation and unemployment because of a historically relatively stable correlation between the two. But as soon as the policy-makers tried to exploit this trade-off (and people understood that this was what the policy makers were doing), then people changed their behaviour (raising their expectations of inflation and demanding correspondingly higher wages) and the policy failed – the old law of constant conjunction was no longer valid. Two things should be noted about this example. First one might question whether this really is spurious correlation. In fact, the correlation between unemployment and inflation in the past may reflect a causal relationship, but – and this is the important point – the size of the correlation is misleading if used as input in the policy-making context. This is why I in the introduction emphasised "misleading" instead on "non-causal" correlations and argued that it was possible to some extent to discuss misleading

correlation even if I did not define causation. If one accepts this terminology, the second point is simply that the existence of self-reflexivity makes misleading correlation a more serious problem in the social sciences compared to the natural sciences. For instance, assume there is a strong and positive association between crime and the level of punishment. Does this imply that lower punishment will translate into lower levels of crime? An alternative interpretation would be that people believe that strong punishment deters crime so when a problem becomes serious, they increase the punishment. In the same way it does not seem too convincing to suggest that strict laws against drugs makes the number of drug users increase (historically there may be a correlation between more users and stricter laws). Instead, the higher the level the use, the more problematic it is considered by some and the higher punishment they implement.

So far I have only tried to convince the reader that the problem of spurious association. Of course, these problems would not be too worrying if it was possible to develop tools to distinguish between genuine and misleading associations. This is the topic for the next section.

Can the problem be solved?

I shall discuss two ways of reducing the problem of misleading associations. First of all there is the question of whether it is possible to create test or manipulate the data in a way so as to reveal whether the association is misleading. Second there is the possibility that more general rules about design and methodology may reduce the risk of accepting misleading associations. In conclusion, however, I shall argue that there are both practical and conceptual reasons why it is no possible to go very far in this direction.

Tools for revealing spuriousness

The ideal would be if we could simply discover some kind of statistics – lets call it the Hans's test of misleading association – which produced a number that indicated the probability of the association being spurious. This is clearly not possible (what, for instance, is the probability of forgetting a variable that is important?). The general impossibility of such a test does however not imply that it is impossible to reveal all kinds of spurious correlation. The best example of this is spurious correlation between serially correlated time-series. In 1926 Yule argued that one cause of this problem was serial correlation i.e. that the observations in each series were not independent, but strongly related. In 1974 Granger and Newbold re-issued the warn-

ing and the classic status of this article may indicate that Yule's warning was ignored in the period between (Granger and Newbold,1974). In any case, the important point to note is that standard rules for statistical inference assumes that we are dealing with independent observations. In a regression it is possible to test for this by finding the correlation between the errors in the regression. One way of doing so is to take what is called the Durbin Watson statistics – which is essentially a measure of the degree to which the observations are independent. Depending on the value of this statistics we may conclude that the observations do not appear to be independent in which case the standard way of testing whether there really is an association is mistaken. In short, there is a "number" which sometimes will reveal whether the correlation really is statistically significant and this can be used to eliminate some cases of spurious correlation (especially in time series).

A second method by which we might try to reveal whether the correlation is spurious, would be to filter the data. For instance, in a discussion of spurious correlation, Ole-Jørgen Skog (Skog,1988) suggests that a possible solution would be to eliminate the autocorrelation using what is called the Cochrane-Orcutt model (Essentially to "eliminate" the problem - autocorrelated errors - by estimating and explicitly including the autocorrelation in the model.) . In appendix 1 I show why I do not think this is a good suggestion. Skog also suggest, as does many others, that we could try to difference the data and see whether the correlation still holds. This is indeed a possibility, but I am slightly less optimistic about the procedure than Skog. Economic theory usually has little to say about how the difference of variables are related. Typically, economic theory tells us something about the long-run relationship - the equilibrium to which the economy is supposed to gravitate toward. For example, monetarist think that inflation - at least in the long run - is always caused by the money supply. In the short run, however, the relationship is not so clear-cut. Hence, a regression of differences need not reveal a significant relationship, while a regression in levels would. The same problem appears in the consumption function. In the long run we may consume a constant proportion of our income, but in the short run consumption often deviates from this desired ratio. A regression in differences may solve some technical estimation problems, but these models are not inspired by theory, they do not always have obvious theoretical interpretations, and we loose information about levels. Skog, of course, is also aware of the increased danger of ignoring true causal relationships when we use correlation in differences (Type II error).

A third solution would be to test the data for parameter constancy and exogeneity. Recall that I labelled a correlation misleading for policy purposes when it did not express a law that the authorities could act upon without destroying that law (the Lucas critique). To get around this problem one could simply try to test whether the relationship was invariant to policy changes in the past. If it was, then one might be more willing to believe that it is stable even for future changes. There are several ways of doing this, but I have relegated the technical discussion to a footnote. The point, in any case, is that it is to some extent possible to test for whether the Lucas critique is relevant.[2]

General ways of reducing the risk of being mislead

In addition to manipulating the data you already have, it is possible to reduce the risk of accepting spurious correlation by bringing in more information and adopting more general rules for research. I will explore some of these here.

One of the most obvious ways of reducing the risk of spurious correlation would be to actually introduce the supposed confounding variable to see whether the relationship under question is still significant. This is the essence of the so-called Simon-Blalock's approach to spurious correlation according to Asher (1983). Asher himself prefers a slightly more advanced version on the same theme called path analysis (Still introducing the supposed confounding variables to see whether there is still a relationship, but this time using regression and not partial correlation analysis).

It is obviously good advice to try to control whether your claimed relationship holds under a variety of circumstances using a variety of variables before you conclude that you have discovered a meaningful association. For instance, recall our labouring researcher trying to find out whether day-treatment or residential treatment is best. As mentioned there are some obvious reasons why we should not compare the success rates directly, but to adjust for the toughness of the cases and other possible confounders. In Table 3 I have calculated some of the relevant values based on my information from 400 drug addicts in my dataset and it turns out, as expected, that there is indeed a significant difference between the clients sent to residential and day-based treatment. The table also indicates that I might be less concerned about sex and as a confounder in this particular comparison (because it is very equal in both institutions).

Table 3: Possible confounders when trying to identify the best treatment

	Day-based	*Residential*
Average number of years injecting drugs	6,9	10,2
Months in previous treatment	11,0	23,6
Age	27,4	29,7
Percentage of male	65,0%	67,4%
Number of clients	100	307

Although obviously correct, the method of introducing control variables has several limitations. First there is the problem of multicollinearity between the controls and the variable you want to explore. This is a technical problem which makes standard statistical procedures invalid. Second, there is the problem of limited data sets. Sometimes we want to control for a factor which is non-quantifiable or for which we have no data. Finally there is, of course, no way of knowing whether you have included all (and only) the relevant variables and confounders. In a sense the method requires us to know exactly what we do not know (but want to test for) – that is the existence of important variables that we have forgotten. The advice that we should just include these variables is then not very helpful since the problem has arisen precisely because it is a variable which we do not know about.

Instead of adding variables (claimed confounders) to a model we believe is true, we might start with the most general model possible (both in terms of which variables to include and the specification of the relationships) and allow the data to simplify the model. David F. Hendry has proposed this as a general research rule and he used the label "General-to-Simple" about this approach as opposed to the traditional "Simple-to-General" approach in which you start with a simple model and then revise it to "solve" problems (like autocorrelation) and increase its "fit" with the real world (The best non-technical source on Hendry's methodology is (Gilbert,1986)). This is good advice for a technical reason (the concept of statistical significance is more meaningful after a simplification search), a methodological reason (whereas S-t-G can lead two researchers using the same dataset to two very different models, the G-t-S is more likely to lead to the same model), and – finally - it seems sensible since we reduce the risk of spurious correlation because we reduce the probability of forgetting variables that might be important. Of course, the downside is that we also increase the risk of including irrelevant variables. One might also accuse the model of being applied "data-mining" since we use the data to simplify the most general model (i.e. eliminating variables that explain little

and so on). Hence adopting the G-t-S method alone need not imply a reduced risk of spurious correlation.

Hendry is, of course, aware that he might be accused of data-mining. To reduce the danger of including irrelevant variables and data-mining, he argues for another methodological rule: Allow only variables and specification that can be justified by a theory! This could be interpreted as demanding that we are not allowed to include variables unless we can also give a plausible story as to why that variable should be included. As Jon Elster (1998: 49) argues: "Understanding the details of the causal story reduces the risk of spurious explanations (i.e., of mistaking correlation for causation). Also, knowing the fine grain is intrinsically more satisfactory for the mind." Or, in my words, knowing *why* there is a correlation instead of just know that there is one makes me more sure that the correlation is not spurious.

Given the existence of some tests that reveals spurious correlation and some methodological rules that reduces the risk of accepting spurious relationships, one might ask whether the problem disappears. Hendry, for instance, has written that "The ease with which spurious results could be created suggested [that econometrics is] alchemy, but the scientific status of econometrics was illustrated by showing that such deceptions are testable." (Hendry,1980: 403). Although I certainly agree that it is sometimes possible to reduce the problem, I will, however, argue that there are also great practical and conceptual problems that limits the scope for revealing spurious correlation by statistical test.

Impossible?

The first problem is very mundane: We simply do not have enough data to be anything near certainty about causal laws. This is well illustrated by a paper by David Little (1998:, chapter 11) who has conducted what he calls "an experiment in causal reasoning." He starts with an assumed true causal structure (formulated as a truth functional law) in which revolution is caused by six different variables - for instance, hunger, war, weak institutions and so on (some necessary, some sufficient – by assumption). The question is then how many cases you need to have in order to find the original truth functional law from the data alone. This, of course, depends on the assumptions you make about the nature of causation. For instance, if you employ Mill's comparative method and we assume both exceptionless causation and causal closure (i.e. that we have listed all the relevant variables – all we need is to "recover" the law from the data), we have 64 logically possible combinations. (Each variable can be

"on" or "off" and there are 6 variables, so the number of possible combinations are $2^6 = 64$). Now to recover the law we would need a real world case that coincided with these 64 different cases and whether the outcome was a revolution or not. Already at this stage we understand that even with the strong assumptions of exceptionless causation and causal closure, the required amount of data is formidable. Of course, a hypothesis can be falsified with less data, but to find the true causal structure we need information on all 64 cases.

It becomes even worse when we relax the assumption of exceptionless causation by introducing probabilities. As a though experiment Little assigns each cause a "true" probability. The task of the researcher would then be to find these probabilities and the causal structure based on just observations from the world. How many observations would they need? To explore this he produces two datasets. One with 51 observations and one with 700 cases. He conclusion is that "the correlation matrix does provide suggestive evidence for causal relations" (the necessary causes appears to "stand out" and it is possible to recover some of the probabilities), but the "data does not permit us to sort out causal dependence among the variables" (Little,1998: 226-227). Once again the conclusion is that it is possible to use the data to falsify some hypothesis, but that the data alone cannot reveal the causal story even when there is large amounts of data.

By way of criticism I would note that Little does not relax the other assumption – causal closure. Hence, when he concludes that the correlation matrix sheds valuable light on the causal structure, this is partly the result of the fact that he only includes the correct variables. There is no way of generating a spurious correlation in his frame since all the variables are causally relevant (by assumption). Without the assumption of causal closure we would be les sure that all correlations were meaningful and revealed causally relevant variables. In any case, Little does demonstrate that the data required to recover a causal structure from the data is formidable.

Needless to say we often do not have the data required and we may doubt whether we will ever get the data needed. This is so for a variety of reasons: It is difficult to conduct experiments with the whole economy; ethical reasons make it difficult to withhold treatment e.g. for a disease, some variables are non-quantifiable and some are non-observable. There is, however, a more fundamental problem that makes me doubt whether it is possible to derive unique causal laws from data alone. I shall label this "the problem of judging what is similar

to something else". It can be given more technical interpretations (e.g. in terms of exchangeability, see (Lindley and Novick,1981)), but this is probably not the place to do so (but see appendix 2).

So far I have treated a correlation as something neutral that just jumped out of the data. In fact, before we can say that two things are correlated, we must assume that we agree on the classification of phenomena. Little's example, for instance, assumes that we must agree on what should count as a revolution. This is not obvious as is illustrated by the eternal debate on whether what happened in Russia in October 1917 (not February) was a revolution or a coup d'etat. Here is another example loosely based on information from Conlisk (1996): Assume you want to buy a plane ticket, but that you are extremely risk averse so before you make up your mind you want to determine which airline is the safest, here defined as fewest death per flown mile. You start collecting data, but you soon encounter the following problems: How far back do you go? (Is the accident rate in the 1950s relevant?); Do you distinguish between the rate of accidents in the season you are flying (say, winter only) or do you use overall accidents as an indicator? Do you distinguish between nigh and day flights and weather conditions? In short, before you can determine the strength of an association you have to exercise a great deal of judgement as to what kind of data should go into the calculation of the association. This questions the possibility of deriving causal statements from the data alone because before we can derive the correlations we must make some judgements as to which variables are causally relevant. The data do not "speak for themselves." Hendry may be right in one context – revealing one form of spurious correlation in time-series – but in general I would conclude that we cannot test the problem away.

Implications for the philosophy of science: An extended conclusion

What is the upshot of the discussion so far? I have argued that the problem of spurious association is widespread, that it is sometimes possible to prevent or reveal it, but that both practical (data limitations) and conceptual limitations (no neutral criteria of what constitutes "similar cases") imply that measures of association are often unreliable as a basis for causal statements. What are the implications for this with respect to the questions raised in the introduction of the paper: Is it good advice to tell the researcher searching for the best treatment to

try to discover neutral causal laws that can be used for policy formulation and that the right way to do so is to find constant conjunctions in the data?

First of all, if one accepts that spurious association is a widespread problem, then one must also accept that our knowledge is severely limited. One could, of course, argue that this should lead us to redefine knowledge in a way that makes it possible to achieve some knowledge. Say, for instance, that we do not see explanations (using causal laws), but that we seek understanding (based on the concept of meaning). This is a major theme in Martin Hollis' (1994) book *The philosophy of social science*, but I agree with Elster (1983) that understanding is no alternative to explanation. Two simple examples will suffice. First, consider a person who colours his hair green. How would one explain this? One possible explanation would be to argue that in that particular culture green hair had a particular meaning (being a rebel), and the person wanted to express this meaning. On my understanding of the terms the exploration of the "meaning" of green hair fits naturally as a part of the causal explanation. The second example is less artificial. When Norway was occupied by Germany during World War II, many Norwegians started to wear bow-ties instead of regular ties. To explain this (causally) we certainly need to know the symbolic meaning of bow-tie (a protest against the Germans). Once again the meaning of something fits nicely as an element in a causal explanation. In short, the problem of spurious correlation and the resulting conclusion that knowledge is limited, does not lead me to claim that the hermeneutic approach promises more knowledge and hence should be adopted. I think this wrong in the sense that knowing the symbolic meaning of something is a part of the causal explanations so the hermeneutic approach is not an alternative to causal explanation.

It does, however, lead me to the conclusion that it is very difficult to establish a law ("treatment of type X is the best treatment") that can be used for policy-purposes. What is more likely, is that it is possible to find some mechanisms that increase our capacity to explain treatment outcomes (without being able to predict it). As an example of a mechanism consider the following: Some children of alcoholics sometimes become alcoholics. However, some children of alcoholics also become fervent teetotallers. In both cases we suspect that there is a causal connection. Those who become teetotallers may do so in reaction to their parents behaviour and – similarly – those who become alcoholics are sometimes said to end up like that because of their unpleasant childhood experiences. There is no law to the effect that "all children of alcoholics become alcoholics/teetotallers", but there is a common causal pattern. Of

course, we would like to isolate the factor which would enable us to distinguish the variable that was important in making some alcoholics and other teetotallers. However, the reason for the difference may be very difficult to find – a very small difference in some variables might be decisive and we might not be able to distinguish. Hence, all we have is a mechanism that allows us to explain ex-post (which is valuable in itself given that we are curious), but which does not allow us to predict in advance.

Finally, should our scientist reject empiricism and become a realist? If one by a realist means somebody who believes that constant conjunction is neither necessary, nor sufficient for statements about causality, then the argument about spurious association is a good reason to become a realist. There is, however, much more to realism than this and to argue that our scientist should become a realist would require me to discuss all of this. The topic for this essay, however, was only spurious association and since this is only one of many issues involved in the debate I can make no general conclusion about realism.

Appendix 1

Why the Cochrane Orcutt model is not a good way to solve spurious correlation

Autocorrelation in the error term could be estimated by the following regression:

(6) $\hat{e}_t = p\,\hat{e}_{t-1} + v_t$

What should be do when faced with autocorrelation (a significant p in the model above). Assume we have the following structure:

(7) $y_t = b\,x_t + e_t$

(8) $e_t = p\,e_{t-1} + v_t$

By substituting (8) into (7) we get:

(9) $y_t = b\,x_t + p\,e_{t-1} + v_t$

From (7) we know that:

(10) $y_{t-1} = b\,x_{t-1} + e_{t-1}$

Manipulating this we get:

(11) $e_{t-1} = y_{t-1} - b\,x_{t-1}$

If we substitute (11) into (9) we have:

(12) $y_t = b\,x_t + p\,y_{t-1} - p\,b\,x_{t-1} + v_t$

In this model there is no problem with autocorrelation - the e's are gone and v is uncorrelated with the other independent variables. We could then estimate p by - for example - equation (6). This is what Wonnacott and Wonnacott (Wonnacott and Wonnacott,1970) presents as one solution to autocorrelation.

The problem with this approach is that it imposes restrictions on the model without justification based on economic theory. For example, there is no reason to assume that

the coefficient of x_t should have the same value as the coefficient of x_{t-1} (namely b). Without restrictions the general model would be:

(13) $y_t = b_1 y_{t-1} + b_2 x_t + b_3 x_{t-1} + v_t$

And the restrictions necessary to get (12) from (13) are that $b_1 = 1$ and $b_2 = - b_3$ (or, in short: $b_1 \; b_2 = b_3$). Before we impose them we should at the very least test whether they hold. Also, for reasons explored in the paper it is best to start with a model without these restrictions (general to simple methodology) and such a methodology is also less likely to generate relationships (with some further assumptions described in the paper i.e. that we demand theoretical justification).

Appendix 2

A short note on exchangeability, based on (Lindley and Novick,1981).

Assume you are given the following information about a test for a disease

Table 4

	Has the disease	*Does not have the disease*	*Total*
Positive test result	16	24	40
Negative test result	4	56	60
Total	20	80	100

Define P(.) as the frequency probability and p(.) as your subjective probability

Now, you have 100 "units." The question is whether you can apply the information on the 100 to another unit. Essentially: Do you judge the new unit "similar" enough to those you already have.

More formally: To go from the frequency probability to the subjective probability you need to assume exchangeability i.e. to say that p(.) = P(.)

Exchangeability defined "A number n of units is termed exchangeable in X if the joint probability distribution $p(X_1, X_2, \ldots, X_n)$ is invariant under permutation of the units."
Origins of concept: de Finetti
In our example we can regard a new patient as:

1. full exchangeability (both test result and disease)
2. exchangeable both in test result and disease
3. exchangeable in test result given disease
4. exchangeable in test result given disease, but not "not disease"

Why don't we just work in terms of sub-populations? Well, that concept does not allow the same degree of flexibility when making the distinctions above, nor does it have a formal criteria. When we say that something is "exchangeable" we basically say that it is similar enough for us to generalize our results to a new person. We are in fact making a judgement about relevant causes – that existing information has not ignored relevant causes.

References

Aldrich, John. 1995. Correlations genuine and spurious in Pearson and Yule. *STATISTICAL SCIENCE* 10 (4):364-376.

Anscombe, G.E.M. 1993/1971. Causality and Determination. In *Causation*, edited by E. Sosa and M. Tooley. Oxford: Oxford University Press.

Asher, Herbert B. 1983. *Causal Modelling, Quantitative applications in the social sciences*. Beverly Hills, CA: Sage University Papers.

Conlisk, J. 1996. Why bounded rationality. *Journal of Economic Literature* 34:669-700.

Dray, W. H. 1978. Concepts of causation in A.J.P. Taylor's account of the origins of the Second World War. *History and Theory* 17 (May):149-174.

Ellett, Frederick S., and David P. Ericson. 1986. Correlation, Partial Correlation, and Causation. *Synthese* 67:157-173.

Elster, J. 1998. A plea for mechanisms. In *Social mechanisms*, edited by P. Hedström and R. Swedberg. Cambridge: Cambridge University Press.

Elster, Jon. 1983. *Explaining technical change :a case study in the philosophy of science, Studies in rationality and social change*. Cambridge Cambridgeshire: Cambridge University Press.

———. 1993. *Political psychology*. Cambridge England: Cambridge University Press.

Freedman, David. 1999. From association to causation: Some remarks on the history of statistics. *Statistical Science* 14 (3):243-258.

Gilbert, Christopher L. 1986. Professor Hendry's Econometric Methodology. *Oxford Bulleting of Economics and Statistics* 48 (3):283-307.

Granger, C.W.J, and P. Newbold. 1974. Spurious regression in econometrics. *Journal of Econometrics* 2:111-120.

Hausman, Daniel M. 1992. *The inexact and separate science of economics*. Cambridge ; New York: Cambridge University Press.

Hendry, David F. 1980. Econometrics - Alchemy or Science. *Economica* 47:387-406.

Hollis, Martin. 1994. *The philosophy of science*. Cambridge: Cambridge University Press.

Lindley, D.V., and Melvin R. Novick. 1981. The role of exchangeability in inference. *The Annals of Statistics* 9 (1):45-58.

Little, Daniel. 1998. *Microfoundations, method, and causation : on the philosophy of the social sciences, Science and technology studies*. New Brunswick, N.J.: Transaction Publishers.

Maddala, G. S. 1992. *Introduction to econometrics*. 2nd ed. Englewood Cliffs, NJ: Prentice Hall.

Skog, O.J. 1988. Testing Causal Hypotheses about correlated trends: pitfalls and remedies. *Contemporary Drug Problems* Winter:565-606.

Smith, Mark J. 1998. *Social Science in question*. London: SAGE Publications in association with The Open University.

Sosa, Ernest, and Michael Tooley, eds. 1993. *Causation*. Oxford: Oxford University Press.

Wonnacott, Ronald J., and Thomas H. Wonnacott. 1970. *Econometrics*. New York,: J. Wiley.

Yule, G. Udny. 1926. Why do we sometimes get nonsense-correlations between time series? *Journal of the Royal Statistical Society* 89:1-69.

Endnotes

[1] The four concepts of correlation in Ellett and Ericson (1986) (and examples of authors using that concept) are as follows:
a. $P(AB) - P(A)\,P(B) > 0 \rightarrow$ positive correlation (P. Kendall & P.F. Lazarsfeld)
b. $P(B|A) - P(B) > 0 \rightarrow$ positive correlation (Hans Reichenback & Patrick Suppes)
c. $P(B|A) - P(B|\tilde{A}) > 0 \rightarrow$ positive correlation (Wesley Salmon)
d. $[P(AB)\,P(A^cB^c) - P(AB^c)\,P(A^cB)] / [P(A)\,P(A^c)\,P(B^c)\,P(B)]^{1/2}$ (H.M. Blalock, H.B. Asher, H.A. Simon)

[2] The basic answer to the Lucas critique is to distinguish between different types of exogeneity and use tests to see which concept applies in a concrete situation. Leamer, for example, has suggested that we should distinguish between exogeneity in the sense of predeterminedness (i.e. when the variable is independent of the contemporaneous and future errors in the equation) and strict exogeneity (predeterminedness plus independence of past errors too). Engle, Hendry and Richard want to divide the concept of exogeneity into three: weak, strong and superexogeneity. The distinction follows from the view that exogeneity is only relevant if we first ask "Exogenous for what?" (which variables). The concepts are somewhat technical (but relatively easy. For more see Maddala (1992: 392-393)). The important point to note, however, is that only weak exogeneity is required for efficient estimation, while superexogeneity is required if we want to use our model to conduct policy predictions. We may then use various test to determine what kind of exogeneity we have, such as the Hausmann test.