Suppes' Probabilistic Theory of Causality and Causal Inference in Economics

Julian Reiss, Durham University[1]

1. Introduction

When Patrick Suppes received the Lakatos Award at the London School of Economics (LSE), Nancy Cartwright, who was giving the laudation, listed all the research interests her colleagues in the Department of Philosophy, Logic and Scientific Method pursued: philosophy of physics and foundations of quantum mechanics, general philosophy of science, mathematical and philosophical logic, philosophy of social science and economics, decision theory, foundations of probability and metaphysics. She then joked that the LSE could save a lot of money by firing the entire department and hiring Suppes instead.[2]

Indeed, Suppes made influential contributions to almost all subfields of theoretical philosophy, including all the ones mentioned above and more. Perhaps less well-known, but no less significant, are his scientific writings in psychology and economics. The latter include work on utility theory (e.g., a new axiomatisation of cardinal utility theory), rational choice theory (e.g., an alternative to Savage's theory), experimental economics (e.g., the first experimental measurement of utility), welfare economics, the economics of science and the theory of consumer demand.[3]

In this paper I will address the usefulness of Suppes probabilistic theory of causality (Suppes 1970)[4] as a theory of causal *inference* for economics and draw some lessons from it for empirical economics and contemporary debates in the foundations of econometrics. Specifically, I will argue that the standard method of empirical economics, multiple regression, is inadequate for most but the simplest applications, that the Bayes' nets approach, which can be understood as a generalisation of Suppes' theory, constitutes a considerable improvement but is still subject to important limitations, and that the currently fashionable 'design-based approach' suffers from the same flaws Suppes anticipated a long time ago. I will sketch an alternative in response, one that differs drastically from the formalisms Suppes endorsed but is consistent with his pragmatic general take on science.

---

[1] I wish to thank Pierre-Olivier Bédard, Nancy Cartwright, Samantha Kleinberg and the CHESS Research Group for invaluable comments on a previous version of this paper.

[2] Max Steuer, then a reader in in the LSE Economics Department, quipped in a smaller round later on, "They could indeed fire everyone in the philosophy department and hire Pat instead, but I don't think that that would save them any money…"

[3] For a review of these and a discussion of Suppes' 'methodology of economics', see García de la Sienra 2011.

[4] Subsequent references to this work will cite only the page number when unambiguous.

2. Suppes' Probabilistic Theory of Causality

Patrick Suppes was one of the first postwar philosophers of science to recognise the importance of causality for empirical science, especially social science. While the covering-law model of scientific explanation was already in decline, it took other philosophers another decade or so to put causality back on the agenda and start to work on causal theories of explanation and causal inference (Salmon 1989). In economics, the development was similar but delayed. The use of 'cause' and its cognates was in steady decline from 1930 till the early 1970s in order to rise again, quite steeply, thereafter (Hoover 2004). It is perhaps ironic that Suppes' theory appeared right at the time when interest in causality in economics was at a historical low.

Against the backdrop of (then) widespread belief in the regularity theory of causality, Suppes starts with the observation that the everyday concept of cause is not deterministic in character (7). The statement 'The thunder caused him to get frightened' does not imply that every time he hears a thunder, he does get frightened. That 'changes in the money stock cause prices to increase' doesn't mean that every time the money stock goes up, prices go up in tandem. Suppes argues that one of the main reasons for the probabilistic nature of the ordinary concept of cause is epistemic: we do not normally know all the factors that affect a given outcome; nevertheless, we describe these partial relations using causal language (8).[5]

Suppes is primarily interested in an analysis of causal relations among *events*, using a notion of event from probability theory. Accordingly, an event is a subset of a fixed probability space. Suppes remains deliberately ambiguous about referring to token events or event-types (79). Importantly, he treats them as instantaneous and includes their time of occurrence in their formal characterisation. Thus, '$P(A_t)$' refers to the probability of event A to occur at time t, '$P(A_t|B_{t'})$' to the probability of event A to occur at time t given that event B occurred at an earlier time t' and so on.

With these preliminaries in hand, we can turn to the definition of a *prima facie* cause. Event $B_{t'}$ is a prima facie cause of event A_t if and only if (12):

(i) $t' < t$;
(ii) $P(B_{t'}) > 0$;
(iii) $P(A_t|B_{t'}) > P(A_t)$.

In other words, an event is a prima facie cause of another event if and only if it occurs earlier and raises the probability of the later event. It is important to note at this point that Suppes maintains that causal relationships are always determined relative to a conceptual framework – which can

[5] This characterisation is consistent with a view that maintains that the fundamental laws are deterministic in nature. This isn't Suppes' view, however. Suppes maintains both that the fundamental laws of nature are probabilistic, and that causality is probabilistic in character (Suppes 1984).

be given, for instance, by a particular scientific theory or a particular experiment or series of experiments. We will come back to this point later.

The next two definitions concern the notion of a spurious cause. An event $B_{t'}$ is a spurious cause of event A_t in sense one if and only if $B_{t'}$ is a prima facie cause of A_t, and there is a $t'' < t'$ and an event $C_{t''}$ such that (23):

(i) $P(B_{t'}C_{t''}) > 0$;
(ii) $P(A_t|B_{t'}C_{t''}) = P(A_t|C_{t''})$.
(iii) $P(A_t|B_{t'}C_{t''}) \geq P(A_t|B_{t'})$.

Condition (ii) says that the earlier event $C_{t''}$ screens off A_t from $B_{t'}$; that is, once its occurrence is taken into account, the occurrence of $B_{t'}$ provides no additional information about A_t.[6] Condition (iii) eliminates cases where $B_{t'}$ alone predicts A_t with a higher probability than the joint event $B_{t'}C_{t''}$.

The second definition of spurious cause does not demand that there be a single earlier event that screens off A_t from $B_{t'}$ but rather that there be a partition such that every event in that partition screens off the later events. Thus: An event $B_{t'}$ is a spurious cause of event A_t in sense two if and only if $B_{t'}$ is a prima facie cause of A_t, and there is $t'' < t'$ and a partition $\pi_{t''}$ such that for all elements $C_{t''}$ of $\pi_{t''}$ (25):

(i) $P(B_{t'}C_{t''}) > 0$;
(ii) $P(A_t|B_{t'}C_{t''}) = P(A_t|C_{t''})$.

The significance of the difference between the two definitions becomes plain when one considers a situation where an earlier factor affects the strength of a causal relationship between two later factors. Suppose that we run an experiment in which individuals are asked to walk along a street to post a letter in the mailbox. On some occasions, there will be a homeless person next to the mailbox, and after mailing the letter they are asked whether they noticed the homeless person ($H_{t'}$) and whether they felt empathy towards the homeless (E_t), see Figure 1.[7]

In the population, $H_{t'}$ is certainly a prima facie cause of E_t ($H_{t'}$ will predict whether an individual feels empathy for most individuals). As psychopaths tend to be better at focusing on the task at hand, it is plausible to assume that clause (i) of the two definitions of spurious cause is fulfilled – that is, degree of psychopathy makes a difference to the likelihood of an individual noticing the homeless. Moreover, holding fixed information about noticing a homeless, it is informative to learn degree of psychopathy in order to determine empathy, so clause (iii) of the first definition is fulfilled. However, clause (ii) is only satisfied for individuals with a very high degree of psychopathy – to whom seeing a homeless doesn't make a difference. Thus, there *exists* an earlier

[6] The screening-off idea is due to Hans Reichenbach, see Reichenbach 1956.

[7] I describe the example at type-level. An analogous story could be told at the token level.

event that screens off the two later events ('very high degree of psychopathy') but it is not the case that every event in the partition ('for every degree of psychopathy') screens them off.

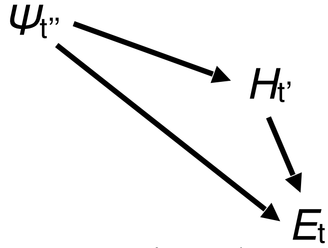

Figure 1

A prima facie cause that is not spurious is genuine (24). Thus, in the example, $H_{t'}$ is a spurious cause of E_t in sense one, but a genuine cause in sense two.

Suppes goes on to define numerous causal concepts such as direct cause, negative cause, and causation between quantitative variables, but they are not relevant for the points that I make below, so I'll ignore them here. Nor will I discuss the counterexamples to the theory, which are widely known and discussed (see for instance Hesslow 1976, Otte 1986, Kleinberg 2012, Reiss 2015a). The reason is that my project here is not to examine whether Suppes' theory is or isn't adequate as a general theory of causality but rather to assess the usefulness of his approach for causal *inference*, especially causal inference in economics. I will therefore turn to a standard method for identifying causes in economics: traditional regression analysis.[8]

3. Simple Multiple Regression and Causality

A widely used approach to addressing causal questions in the social sciences is to collect data on an outcome variable of interest Y, measure a set of determinants $\mathbf{X} = X_1, X_2, \ldots X_n$, and run a regression of Y on $\mathbf{X}$. The functional relation between Y and the X_i's is usually assumed to be linear:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon.$$

A variable X_i is judged to be causally relevant just in case its coefficient β_i is significantly different from zero.

[8] A referee remarked that 'Regressions are not used to identify causes in economics in the sense of determining whether a variable is or is not a cause. They are used, once the direction of causation is assumed to be known (and subject to a lot of well-known caveats with respect to exogeneity), to measure the strength of causes.' This may be true in sophisticated discussions of econometric methodology to which economists of course also contribute. In applications, I often find that cautions are put aside and judgements made without good arguments about exogeneity. I cannot make this point in any detail here but I don't think that it's dramatically misleading to say that there's a difference between practice and econometric theory in this respect, and what I'm describing in the main text is – a good amount of – practice.

Of course, it is well known that 'correlation is not causation'. Social scientists therefore make sure that enough causal variables are included among the X_i's, or they test the set for robustness and judge as causal those variables that have significant coefficients in all or most specifications.

Neither strategy helps, however, if there are causal relations *among* the regressors or the regressand causes a regressor (cf. Kincaid 2012, Glymour et al. 1994). If, for example, the causal structure is as in Figure 2, the coefficient on X_2 in a regression would be positive, even though X_1 doesn't cause Y. If, by contrast, the causal structure is as in Figure 3, the effect of X_1 on Y would be underestimated because only its direct effect is measured, not its indirect effect via X_2.

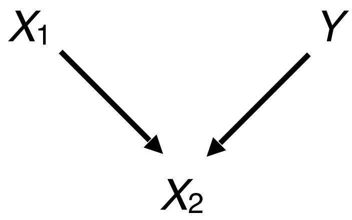

Figure 2

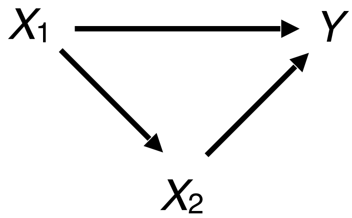

Figure 3

Two solutions to remedy the situation have recently become popular, one by philosophers and computer scientists and one by a new brand of econometricians. Philosophers and computer scientists have developed a powerful methodology for learning causes from data called Bayesian nets, which can be regarded as a generalisation of Suppes' theory.[9] Econometricians have traditionally sought remedy in theory. Theory is meant to do a variety of things, including determining the functional form of the regression and assumptions about the exogeneity or endogeneity of variables. This programme was largely unsuccessful, however, because economic theory tends to be both unspecific and controversial. There is little agreement on the relevant theory, and most theories are not strong enough to make definite prescriptions about the specifics of the regression equation.

In response to these problems with traditional econometric analysis (also known as structural or Cowles Commission econometrics), a contemporary movement takes experiments as paradigmatic. As certain kinds of experiments are the 'gold standard' for causal inference, they propose to confine econometric analyses to data sets that have been designed in such a way as to

[9] I do not mean to make any claims about originality here. It may well be, and is probably the case, that the Bayes' nets programme arose entirely independently from Suppes. But this does not contradict the systematic connection between the two programmes.

mimic controlled or randomised experiments. The next two section will review these two proposals and discuss their major drawbacks.

4. Bayesian Nets and Economics

A Bayes' net consists of a directed acyclic graph (DAG) whose nodes are variables included in the set **V**, and a probability distribution over the variables in the graph. **V** is normally assumed to be 'causally sufficient' – that is, it is assumed that every variable that influences two or more variables in **V** is itself in **V**. The graphs in Figures 2 and 3 are examples of DAGs, with **V** = {X_1, X_2, Y}. A graph is directed when all edges between nodes are directed (as in Figures 2 and 3), and it is acyclic when there are no loops in which a variable causes itself. For instance, if the arrow between X_1 and Y in Figure 3 pointed the other way, the graph would by cyclic.

The theory of Bayesian nets is closely related to Suppes' theory of probabilistic causality in that it also begins with principles that describe conceptual relations between probability and causality, and these principles are generalisations of some of Suppes' ideas. There are two main principles, the causal Markov condition and Faithfulness. I will now explain what they say and show how they are generalisations of Suppes' theory. To state them, we need the notions of parents and descendants: **Parents**(W) is the set of direct causes of W in **V**; **Descendants**(W) is W together with the set of variables in **V** that can be reached from W by any directed path starting from W. In Figure 2, X_1 and Y are the parents of X_2; in Figure 3, X_1, X_2 and Y are all descendants of X_1.

Causal Markov Condition. Let G be a DAG with variable set **V** and P be a probability distribution over the variables in **V** generated by the causal structure G. G and P then satisfy the Causal Markov Condition (CMC) if and only if for every W in **V**, W is independent of (**V** \ **Descendants**(W) ∪ **Parents**(W)) given **Parents**(W).

In other words, every variable in W in **V** is independent of every other variable in **V** except its own descendants (and, trivially, its parents), given its parents. One implication of the CMC is that common causes screen off their effects from each other – an idea also used by Suppes. Another is that mediating links in a causal chain screen off subsequent from prior causes – an idea Suppes uses in his definition of 'direct cause', which hasn't been discussed above.

Faithfulness Condition. The second fundamental principle is the Faithfulness Condition (FC), which is the converse of the CMC: A DAG G and a probability distribution P satisfy the FC if and only if *only* those conditional independencies are true in P that are entailed by the CMC applied to G. The CMC says that given a causal graph, certain probabilistic independence relations should hold; the FC says that there are no additional probabilistic independencies.

The FC implies, among other things, that causal relations always result in probability changes. In causal structures such as that of Figure 3, it is possible that a (say) negative direct effect of X_1 on Y is counteracted by a positive indirect effect via X_2. If so, the result may be that X_1 and Y are probabilistically independent despite being causally related. Such a probability distribution would be called 'unfaithful' to the graph. In Suppes' theory, this situation is assumed away by the stipulation that every cause must be a prima facie cause. Some people argue that violations of Faithfulness are extremely rare (Spirtes et al. 2000). But they might not be. When systems have backup causes, a cause might appear to be independent from its effect. For example, when multiple genes produce a phenotype, we can inactivate one gene but the phenotype will still be present (Kleinberg 2016: 108). To give an economics example, a policy that effectively stabilises a target quality (such as inflation), will result in an independence of cause and effect that is in fact the result of a successful exploitation of the causal relation (Hoover 2001)!

The Bayesian networks approach is more flexible than standard multiple regression. The latter makes the implicit assumption that one of the measured variables (the 'dependent' variable) is an effect and all other variables (the 'independent variables') are either direct causes or play no causal role. The Bayesian network approach allows many kinds of causal relationships among the variables in **V** provided the set is causally sufficient.

However, the Bayesian networks have their own limitations. According to their proponents, multiple regression should not be used in the absence of strong causal background knowledge about the domain of interest; Bayesian networks, on the other hand, do not require this (Spirtes et al. 2000: 207; see also Glymour et al. 1994):

> In the absence of very strong prior causal knowledge, multiple regression should not be used to select the variables that influence an outcome or criterion variable in data from uncontrolled studies. So far as we can tell, the popular automatic regression search procedures should not be used at all in contexts where causal inferences are at stake. Such contexts require improved versions of algorithms like those described here to select those variables whose influence on an outcome can be reliably estimated by regression.

Spirtes et al.'s point of view is controversial. For instance, many of the algorithms they use implement the Faithfulness Condition in one way or another: the idea that probabilistically independent variables are not causally connected. As noted, this assumption may not always be met. There is also the problem of using the data to infer that variables are independent since no algorithm, when fed with real data, can determine whether two variables are probabilistically independent. It can measure a correlation and test, against a probability model, whether the correlation is or isn't significantly different from zero. If the test says that the null hypothesis of no difference cannot be rejected, then any of three interpretations can be made: a) the two variables are indeed probabilistically independent; b) something 'rare' has happened (how rare is determined by the level of significance); or c) the probability model is false.[10]

How frequently a), b), and c), respectively, is the correct interpretation of the test result obviously depends on the quality of the probability model we are using. Often the probability model assumes that data are multinomial or normal but economics data are rarely multinomial or

[10] This is, in my view, true even when non-parametric methods are used. Non-parametric measures of association such as Spearman's rho are still measures of correlation and are only valid under appropriate assumptions about the data generating process.

normal. Most variables are in fact time series, and correlations are rarely indicative of causal connections in time series (Reiss 2007). Cross-sectional data often mix populations, and correlations in mixed populations suffer from the same problem (Cartwright 2001).

It is therefore not surprising that real-world applications of this methodology may lead to nonsense results. Humphreys and Freedman 1996, for instance, discuss a case analysed in Spirtes et al. 1993, but not fully reported there, where the graph produced by TETRAD, a programme that implements specific algorithms of the Bayes' nets approach, represents race and religion as causes of region of residence, even though the latter is a dummy variable that is one for 'South' and zero otherwise – which makes neither sociological nor mathematical sense.[11]

One person who would not have been surprised by this is Suppes. While Bayesian networks arguably constitute a development and generalisation of his probabilistic theory of causality, they ignore that the theory made causal judgements always against a framework given by a scientific theory. Purely empirical applications, such as those intended by the developers of Bayesian networks, fall outside its scope. And this is for good reason: unless a scientific theory justifies the making of assumptions such as the CMC and the FC and assumptions about the probability model, there is little reason to believe that they should hold. Suppes comments on the requirement of causal sufficiency made by this approach (Suppes 1994: 363):

> I am myself reluctant ever to commit to having a causally sufficient set of variables. I am too skeptical a sometime Bayesian to think that it is possible to make such an identification. I would certainly agree that when a specific theory is formulated and the random variables and their distributions are given on the basis of fundamental theoretical assumptions, then it is possible, relative to such a theory or, in even more restricted cases, to particular models of such a theory, to identify in a completely explicit and exact way causally sufficient sets of random variables. But this is not the kind of thing to which the analysis of Glymour, Spirtes and Scheines is directed at all. They are really concerned with highly empirical situations for which there is no overriding theory guiding and, in fact, fixing the causal structure. In such highly empirical situations I have skepticism about this notion based upon lots of examples to be found in every area of science.

5. Design-Based Econometrics

The other alternative to simple multiple regression (and structural econometrics) that has recently gained in popularity as well as influence especially in policy circles is similarly as empirical as the Bayesian network approach. It starts from the presumption that randomised experiments are the gold standard or benchmark for causal inference. Thus, in many fields of applied micro economics there is a move towards testing scientific hypotheses experimentally, or, should that not be feasible, to employ quasi-experimental designs. Especially development

---

[11] I should mention that Spirtes et al. have written a forceful response (Spirtes et al. 1997), as have Kevin Korb and Chris Wallace (Korb and Wallace 1997). Spirtes et al. claim relations among the regressors in the example discussed by Humphreys and Freedman are irrelevant to the point they were trying to make, which doesn't address the charge that the algorithms can produce nonsense results. They also suggest that one cannot trust algorithm results blindly and should, for instance, test results for robustness using different levels of significance. This is good advice, certainly, but does not make my claim that the algorithms work only under strong – causal *and* statistical – and controversial assumptions untrue. See also Humphreys and Freedman's rejoinder (Freedman and Humphreys 1999).

economics has seen a flurry of work that uses randomised field experiments in order to evaluate the 'impact' of a policy on an outcome variable of interest. The experimental approach has been said to 'have the potential to revolutionize social policy in the twenty-first [century]' just as it 'revolutionized medicine in the twentieth century' (Duflo and Kremer 2005) and to have brought about a 'credibility revolution in empirical economics' (Angrist and Pischke 2010).

There are numerous quasi-experimental designs such as instrumental-variable studies, regression-discontinuity analyses and differences-in-differences methods. Of these, instrumental variables are probably the most important and most widely used. An instrument is a variable that: (a) causes the putative cause (or independent) variable; (b) causes the putative effect (or dependent) variable, if at all, only through the putative cause variable (and not through a mechanism that bypasses the latter); and (c) isn't caused by the putative effect variable or any of its other causes. Under suitable sets of assumptions, instrumental variable studies can be proved to yield causally correct conclusions (see for instance Pearl 2000, Reiss 2008).[12]

It is easy to see that an instrument constitutes the observational equivalent of randomisation. In a randomised experiment, a randomisation device is responsible for the allocation of subject to treatment groups; the randomisation therefore causes treatment status. As the result of the randomisation (and thus treatment status) is masked from experimental subjects, experimenters, analysts etc. randomisation shouldn't affect the outcome except through the treatment. Finally, randomisation is genuinely exogenous and cannot be caused by either outcome or any of its other causes.

That randomised experimentation isn't a panacea has been argued a long time ago (e.g., by Heckman 1992). The list of criticisms is long: randomisation means that there is uncertainty about treatment condition, and since individuals differ in their preferences towards taking on risk, more risk averse subjects may not enter the experiment to begin with, exit prematurely or get the preferred treatment elsewhere; sample sizes tend to be small so that treatment and control group are likely to be accidentally unbalanced; blinding is practically impossible in social science applications; even if initial blinding is successful, subjects may learn about treatment status if the treatment is effective; randomised experimentation can address only a narrow range of research questions, and to regard it as the 'gold standard' often means that hypotheses are changed so as to allow testing by the method; the method is more costly than alternatives; as a consequence it tends to focus on the short run and ignores important long-run effects.

Quasi-experimental methods such as the analysis of natural experiments and instrumental-variable studies certainly do not create artefacts that are due to the deliberate interventions of experimenters or randomisation. But it has been argued by structural econometricians that they are highly unlikely to yield interpretable and useful results unless the empirical specification is backed up by economic theory (e.g., Deaton 2010). Let me repeat the main points here. A theory

[12] Pearl derives this result from assumptions that include the Causal Markov Condition and the Faithfulness Condition (called 'Stability' by Pearl); Reiss uses similar assumptions but does not rely on the 'screening off' property.

is a fully specified system, and so if a variable appears in the reduced form, we know the mechanism through which it affects the left-hand side variable. There are no other mechanisms than those given by the equations. By contrast, in a purely empirical specification, it is likely that an instrumental variable affects the outcome variable through a variety of different mechanisms, which, in turn, are unlikely to be constant over the different units or structures on which the data are measured. The results obtained are average results over different units/structures, and therefore unlikely to be predictively informative except for new populations that are constituted by the exact same mix of units/structures. In the kind of observational studies that typically use these techniques, this variation can't be thought of as random variation that is uncorrelated with anything else of interest. Moreover, if there is heterogeneity, the result depends on the exact choice of instrument. So what we measure is not the average effect across all units/structures in the population of the study but rather the average among those units/structures that were induced by the specific instrument to assume the value of the putative cause variable that they have. In the absence of detailed information about the population over which the average is drawn here, this is not a very useful and potentially misleading quantity to learn (Heckman and Urzua 2009). Finally, the instrumental variable estimator will be biased if effect size is correlated with the value of the putative cause variable, which, again, we do not have good reason to suppose is not the case outside of a specific theory.

Now, while I am not aware that Suppes ever commented on this debate between 'design-based' and 'structuralist' econometricians, it is probably safe to assume that he would side with the structuralists. If anything, my guess would be that Suppes would urge economists not just to use economic theory but develop theories that are strong enough to have implications about all aspects of an empirical study that need to be addressed, including independence relations, functional form, error terms and so on, or at least implications that are strong enough so that we have a good reason to believe that tests of the statistical assumptions of lower-level empirical models yield informative results.[13]

At any rate, causal judgements are to be made relative to a theoretical framework because outside such a framework there is little reason to believe that the assumptions necessary to validate the judgements hold. Design-based econometrics isn't as neatly laid out as the series of definitions and principles one finds in Suppes theory and in the theory of Bayesian networks, but it equally depends on a variety of assumptions connecting causality and probability, the causal structure of the system studied as well as the functional form of the relationships, the properties of the error terms and so on.

At this point we might think that the arguments of Suppes and the structuralists are at an impasse. The problem is that design-based econometrics itself grew out of a reaction against structuralist econometrics which, in the opponents' view, relies too heavily on economic theory because economic theory is not 'credible'. The problem is not that there isn't any theory, but that there is far too much of it, and economists have a hard time agreeing on which bits are good ones

[13] I take this to be one of the significant conclusions of the seminal paper 'Models of Data' (Suppes 1962).

to use in empirical applications. Neoclassical microeconomics can be said to command wide assent in the profession, but this is at least partly due to the fact that it is so flexible as to accommodate an enormous breadth of empirical phenomena. About the case of development economics Dani Rodrik, for instance, writes (Rodrik 2007: 29):

> The main point I take from these illustrations is robust to these fallacies, and has to do with the "plasticity" of the institutional structure that neoclassical economics is capable of supporting. All of the above institutional anomalies are compatible with, and can be understood in terms of, neoclassical economic reasoning ("good economics"). Neoclassical economic analysis does not determine the form that institutional arrangements should or do take.

Nor do I think that there are too many macroeconomic principles the are both strong enough to constrain empirical implications and widely agreed among economists. In other words, there are good reasons for empirical economists to be sceptical about econometric methodologies that essentially rely on economic theory.

It now seems that we are stuck between a rock and a hard place. 'Purely empirical' methods such as Bayesian networks or design-based econometrics don't work – for all the reasons Suppes anticipated a long time ago. However, the theory-based methodology that he advocated and many still defend today doesn't work in a discipline such as economics either, because there is no good theory. Does this mean that causal inference in economics is impossible? There is no space here to develop the idea in acceptable detail, but in the next section I will sketch a procedure for learning causes from data that works, essentially, by reducing the need to make strong background assumptions to a minimum and instead making causal arguments on the basis of a large variety of independent pieces of evidence, none of which would on its own clinch a conclusion.

6. An Alternative: Statistical Minimalism

In this section I want to defend the idea that very simple patterns in the data – simple measures of dependence, ratios, temporal coincidences and sequences etc. – can often give us great inferential leverage when they are based on a deep substantive understanding of the domain at hand. The technically more sophisticated methods of Suppesians, structuralists, Bayesian networkers, design-based econometricians can all be said to produce conditional causal inferences: *if* the scientific theory or statistical and causal or other background assumptions used in the derivation of the empirical specification are correct, *then* we have good reason to believe that the putative cause-variable actually causes the putative effect-variable. In the absence of well-established theories and reasons to trust the background assumptions made, these kinds of inferences are often of limited usefulness.

Moreover, I think that there are good reasons to believe that the problem is here to stay. Economic phenomena are too complex and uncontrollable, research agendas are too varied and shifting, and economic concepts, hypotheses and theories are too deeply infused by values for there being a high chance that a consensus on a broad theoretical framework (one that has any substance) will emerge soon. Situations that are more favourable to statistical analysis can arise in experimental set-ups, but control is often lacking even in typical economic experiments and if it can be achieved, the relevance of the experimental result to economic behaviour outside the laboratory is questionable. Suppes, himself an early practitioner of experimental economics, clearly saw the limitations of the approach (Suppes and Carlsmith 1962: 60 [emphasis original], 77, and 78):

> It is very likely the case that economists will not take the theory too seriously as an *economic* theory. This we are prepared to accept. The evidence for the theory comes from a highly structured, highly simplified experimental situation.
>
> It may be argued that the data were collected in such a highly structured and oversimplified situation that they have little relevance to economic behavior. This is possibly true, and it would of course be desirable to collect similar data in the field. But then one runs into the problem of lack of control of many variables which are obviously relevant, but which are not as yet incorporated into the theory. It was primarily for this reason that we chose to do the experiment in the laboratory, where it was possible to control many more of these variables.
>
> Finally, it is certainly true that the theory is at present far too simple to describe much actual economic behavior. The direction of generalization seems clear, although problems, both experimental and mathematical, are present.

There's a tendency for econometric models to get more and more sophisticated from a formal or statistical point of view – no doubt in part thanks to the growing computational power of desktop computers – at the expense of substantive engagement with the subject matter at hand. 'Form beats substance', in a manner of speaking. Many of the younger generation of 'stars' in the profession – Josh Angrist, Steffen-Jörn Pischke, Steven Levitt, to name but a few – seem to be proponents of methodologies to address any economic question (in fact, any socio-economic issue whatsoever) rather than labour or international or public or monetary or ecological economists.[14]

A convincing causal argument cannot be made, however, without very detailed domain-specific background knowledge.[15] Dani Rodrik provides a beautiful thought experiment illustrating this idea (Rodrik 2007: Ch. 1). He has us imagine an intelligent Martian and ask him to match the growth record of the developing countries against the list of policy proposals that jointly make up the Washington Consensus.[16] Rodrik writes that the Martian would find almost a negative correlation between adherence to the list and economic success: many countries that did exceptionally well (such as the Asian tigers) score low on the list and many of those who closely

[14] Self-professedly so in Levitt's case, of course. Chang 2014 laments that economics has taken this turn.

[15] For a detailed argument along these lines, see Reiss 2015b.

[16] The original Washington Consensus consisted of the ten items: Fiscal discipline, Reorientation of public expenditures, Tax reform, Interest rate liberalisation, Unified and competitive exchange rates, Trade liberalisation, Openness to direct foreign investment, Privatisation, Deregulation, Secure property rights. It was later augmented by a number of additional items such as anti-corruption measures, flexible labour markets and social safety nets, but there was less of a consensus on these items. See Rodrik 2007: 17.

followed the Washington policies did quite poorly (such as some of the Latin American countries past 1980).

Perhaps this is not so surprising, even from the point of view of someone who has no domain-specific knowledge but knows some of the literature on causality published in the past 50 or so years. Causes are often INUS conditions.[17] That is, causal factors require the right constellation of background conditions in order to be causally efficacious. The Martian's expectation that closer adherence to the list should be followed by higher rates of growth, would be analogous to assuming that the more of the following conditions: presence of a match; presence of oxygen; the match being struck with sufficient force; absence of water; the match being heated up to a certain point; absence of strong wind etc. would be followed by a higher rate of lighting. To reason this way would be silly, even for a Martian.

At best, we should expect that some subsets of the Washington list are perhaps sufficient, but only when they are all present. Maybe there are several different such subsets, such that whenever these factors are co-present, growth will be ignited. Maybe some factors affect not the presence of growth as such but rather its rate (or other characteristics) once it's there. Maybe some factors are causes in some contexts but inhibitors in others. It is at this point that the Martian would indeed need domain-specific knowledge in order to know what he should expect about growth under the hypothesis that the Washington list is a list of causes of growth.

But Rodrik's main point is a different one. He argues that a number of items on the Washington list – protection of property rights, contract enforcement, market-based competition, appropriate incentives, sound money, debt sustainability – should not be understood as concrete causal factors but instead as abstract capacities or abilities of an economy that can be implemented in a variety of different concrete institutions, and in order to work, these institutions have to respond to local constraints and opportunities. If that is so, it is neither immediately obvious whether a causal factor such as 'secure property rights' is even present unless one has the requisite knowledge about what that would mean in the given case, nor would one know what to expect about its growth effects without that knowledge.

The reliability of a causal argument, then, derives from substantive background information and not from the formal properties of inference techniques that are being used (cf. Norton 2003 on induction). An alternative to the currently fashionable (strongly[18]) conditional inferences is to reduce the need for making formal assumptions to a minimum and proceed on the basis of a substantive understanding of the subject matter at hand. Often it is *very simple patterns* in the data – simple measures of dependence, ratios, temporal coincidences and sequences etc. – that can

---

[17] Insufficient but non-redundant parts of sufficient and non-necessary sets of conditions. See Mackie 1974.

[18] All inference is conditional on a body of background information; to accept this truism is one thing, to make inferences dependent on the validity of strong and controversial assumptions such as the Causal Markov Condition is quite another. Hence the qualifyer.

give us great inferential leverage when they are based on a deep substantive understanding of the domain at hand.

I want to illustrate this idea with a piece of a causal argument given by Ed Leamer in support of the claim that the Fed's manipulation of interest rates can cause recessions (Leamer 2009). The first thing to note about Leamer's book is that he spends about two thirds of the book (the first thirteen of 20 chapters) not making any causal argument at all but rather telling us about the U.S. economy. This exercise would be to a large extent pointless if credibility derived from design or metaphysical principles rather than substance. But it doesn't, so it isn't.

The causal argument itself is then based on the understanding that the housing market and the market for consumer durables play an important role in the explanation of the business cycle. Housing and durables are both very good predictors of recessions and important enough in size that they could be implicated in causing recessions (Leamer 2009: §15.3). He also notes that the Federal Funds rate is a good predictor of housing and durables. Of course, this doesn't mean (yet) that it's also a cause of the latter or of recessions, as the Federal Funds rate might itself respond to one or more factors (e.g., unemployment) that also cause housing and durables, and through them, recessions. Leamer then considers the orthodox solution of statistically controlling for such common causes. But that wouldn't be a good idea (248):

> That's like including the cloud coverage in a model that predicts rain on the basis of the rain forecast. If after controlling for the clouds, we find that the weather forecasts help to predict rain, then that must be a causal effect, right? Wrong. That merely suggests that the forecasts and the rain both depend on some other common variables, like the barometric pressure. It only gets worse as we layer in more and more assumptions about how expectations are formed and what kinds of delays are present in the responses, and what is observable by whom. At this point, the usual solution is to roll out some very heavy and very loud econometric cannon to shoot at the causal target. The noise and the brilliance when these cannon are fired utterly disguise the fact they widely miss their mark. But this econometric performance is only a slideshow. The Priests of the Orthodox Economics Church meditate so deeply in pursuit of Enlightenment about the Causal Structure of the Economy that they do not even hear the noise or notice the cannon flashes. Meanwhile, the Fed's belief in their own importance is not at all affected by the cannon blasts or the meditative chants of the Priests.

There's little point in trying to control for confounders in the absence of a good theory that tells us what all the factors might be that influence an outcome, much less using heavy econometric artillery. What does Leamer propose instead? He first tells us to look at the long-term interest rates instead of the Federal Funds rate, because housing and durables will be influenced by them, not, or not directly, by the Federal Funds rate. Changes in 10-year Treasury bonds also predict recessions, but the Fed has little influence over them.

The core of the argument is given by an account of the significance of the relation between long- and short-term interest rates. Banks make long-term loans (such as mortgages and personal credits) and refinance short-term (through, for instance, deposits and inter-bank loans). They make money when short-term rates are relatively low and long-term rates are relatively high. The 'relatively' is important here, though: what matters is that and how much long-term rates are above the short-term rates, not their absolute levels. As long as the spread is large, banks will happily give loans, even to borrowers with a high risk of default – as everybody could observe on a grand scale in the run-up to the 2007- crisis. But when the yield curve flattens or inverts, banks must be much more selective in their lending decisions, if indeed they are prepared to make any new loans at all.

It turns out that every single recession in postwar-U.S. history was preceded by a flat or inverted yield curve. The decision to increase the Federal Funds rate causes a recession when that increase leads to a flattening or inversion of the yield curve, this brings about a credit crunch, the credit crunch, in turn, a decline in housing, and the latter spreads to the remainder of the economy.

Let me emphasise that in a sense, Leamer's causal inference is conditional too: conditional on historical and institutional background information about the U.S. banking system and other aspects of the U.S. economy. What it does not depend on is general theoretical principles about the connections between probability and causality such as the Bayesian-network methodology, strong identification assumptions such as design-based econometrics or economic theory such as the structuralist approach. It is a theory-free, but not a background information-free approach. Though I don't have sharp criteria to make the three-fold distinction theory – strong and controversial causal/statistical assumption – substantive background knowledge, there are important differences. Substantive background information is more readily testable and should therefore more readily command assent. Whether or not the Causal Markov Condition is true makes no difference to any possible observation. If it seems to fail, it is always possible to blame an omitted common cause. Claims about the structure of the U.S. economy don't suffer from this deficiency. There may well be differences in opinion especially when the claim is a quantitative one, but there is no reason to believe that evidence cannot in principle settle such differences in opinion. Moreover, controversies about substantive background information will be fruitful. They will involve appeals to evidence and further substantive background information from which we will learn more about the economy. I don't see the same collateral benefit in the philosophical debates about the adequacy of the Causal Markov Condition and its companions.

7. Conclusions

Would Suppes endorse 'statistical minimalism'? In one reading, the approach is radically un-Suppesian. There are no axioms, no formalism, no theory. It is not surprising that Suppes own empirical and conceptual work in this area was confined to experimental set-ups in which human behaviour was constrained in such a way that certain theoretical and probabilistic assumptions could be expected to be met. Outside of such constraints, in the real world outside of psychological and economic experiments, and in particular in the world described by macro economics, such assumptions are invalid.

In a different reading, though, statistical minimalism captures the spirit of Suppes' pragmatic outlook of many of his works in the philosophy of science such as, for example, *Probabilistic*

Metaphysics (see also Suppes 1988; 2010 among many others). While Suppes did like to axiomatise, applications it had to be based on a detailed knowledge of what is going on in the situation the axioms are supposed to describe. Pragmatism tells us to reject ideas that are not practicable or not useful – in the prevailing circumstances. And in the circumstances of contemporary economics, statistical minimalism is the best we can do.

Bibliography

Angrist, Joshua and Jörn-Steffen Pischke (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2):3-30.

Cartwright, Nancy (2001). "What's Wrong With Bayes' Nets?" *Monist* 84(2):242-264.

Chang, Ha-Joon (2014). Economics: The User's Guide. London: Penguin.

Deaton, Angus (2010). "Instruments, randomization, and learning about development." *Journal of Economic Literature* 48(2):424-455.

Duflo, Ester and Michael Kremer (2005). "Use of randomization in the evaluation of development effectiveness." In *Evaluating Development Effectiveness, Vol. 7*, ed. G. Pitman, O. Feinstein and G. Ingram. 205-231. New Brunswick (NJ): Transaction Publishers.

Freedman, David and Paul Humphreys (1999). "Are There Algorithms That Discover Causal Structure?" *Synthese* 121(1-2):29-54.

García de la Sienra, Adolfo (2011). "Suppes' Methodology of Economics." *Theoria* 72:347-366.

Glymour, Clark, Peter Spirtes and Richard Scheines (1994). "In Place of Regression." In *Patrick Suppes: Scientific Philosopher, Volume 1. Probability and Probabilistic Causality*, ed. P. Humphreys. 339-366. Dordrecht: Springer.

Heckman, James (1992). "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, ed. C. F. Manski and I. Garfinkel. 201-230. Boston (MA): Harvard University Press.

Heckman, James and Sergio Urzua (2009). Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. Secondary Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. Secondary ---. Boston (MA). 14706.

Hesslow, Germund (1976). "Discussion: Two Notes on the Probabilistic Approach to Causality." *Philosophy of Science* 43:290-292.

Hoover, Kevin (2001). *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Hoover, Kevin D. (2004). "Lost Causes." Journal of the History of Economic Thought 26(2): 149-164.

Humphreys, Paul and David Freedman (1996). "The Grand Leap." *British Journal for Philosophy of Science* 47:113-123.

Kincaid, Harold (2012). "Mechanisms, Causal Modeling, and the Limits of Traditional Multiple Regression." In *Handbook of Philosophy of Social Science*, ed. H. Kincaid. 46-64. Oxford: Oxford University Press.

Kleinberg, Samantha (2012). *Causality, Probability, and Time*. Cambridge: Cambridge University Press.

--- (2016). Why: A Guide to Finding and Using Causes. Sebastopol (CA): O'Reilly.

Korb, Kevin and Chris Wallace (1997). "In Search of the Philosopher's Stone: Remarks on Humphreys and Freedman's Critique of Causal Discovery." *British Journal for Philosophy of Science* 48:543-553.

Leamer, Edward (2009). Macroeconomic Patterns and Stories: A Guide for MBAs. Cambridge (MA): MIT Press.

Mackie, John (1974). The Cement of the Universe: A Study of Causation. Oxford: Oxford University Press.

Norton, John (2003). "A Material Theory of Induction." *Philosophy of Science* 70(4):647-670.

Otte, Richard (1986). "A Critique of Suppes' Probabilistic Theory of Causality." *Synthese* 48:167-189.

Pearl, Judea (2000). Causation: Models, Reasoning and Inference. Cambridge: Cambridge University Press.

Reichenbach, Hans (1956). *The Direction of Time*. Berkeley (CA): Unversity of California Press.

Reiss, Julian (2007). Time Series, Nonsense Correlations and the Principle of the Common Cause. Causality and Probability in the Sciences, 179-196,

--- (2008). Error in Economics: Towards a More Evidence-Based Methodology. London: Routledge.
--- (2015a). Causation, Evidence, and Inference. New York (NY): Routledge.
--- (2015b). "A Pragmatist Theory of Evidence." *Philosophy of Science* 82(3):341-362.
Rodrik, Dani (2007). One Economics, Many Recipes: Globalization, Institutions, and Economic Growth. Princeton and Oxford: Princeton University Press.
Salmon, Wesley C. (1989). *Four Decades of Scientific Explanation*. Pittsburgh: University of Pittsburgh Press.
Spirtes, Peter, Clark Glymour and Richard Scheines (1993). *Causation, Prediction, and Search*. 1st. Cambridge (MA): MIT Press.
--- (1997). "Reply to Humphreys and Freedman's Review of*Causation, Prediction, and Search*." *British Journal for Philosophy of Science* 48:555-568.
--- (2000). *Causation, Prediction, and Search*. 2nd. Cambridge (MA): MIT Press.
Suppes, Patrick (1962). "Models of Data." In Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress, ed. E. Nagel, P. Suppes and A. Tarski. 252-261. Stanford: Stanford University Press.
--- (1970). A Probabilistic Theory of Causality. Amsterdam: North-Holland.
--- (1984). Probabilistic Metaphysics. Oxford: Blackwell.
--- (1988). "Pragmatism in Physics." In *The Role of Pragmatics in Contemporary Philosophy*, ed. P. Weingartner, G. Schurz and G. Dorn. Vienna: Holder-Pichler-Tempsky.
--- (1994). "Comments by Patrick Suppes." In *Patrick Suppes: Scientific Philosopher, Volume 1. Probability and Probabilistic Causality*, ed. P. Humphreys. 362-365. Dordrecht: Springer.
--- (2010). "The Nature of Probability." *Philosophical Studies* 147:89-102.
Suppes, Patrick and J. Merrill Carlsmith (1962). "Experimental Analysis of a Duopoly Situation from the Standpoint of Mathematical Learning Theory." *International Economic Review* 3(1):60-78.