Parsimonious Methodology

Kees Aarts

University of Twente, Netherlands

Introduction

‘Parsimony’ can be interpreted as simplicity, but also refers to being economical or efficient. I chose ‘parsimonious methodology’ as the title of this paper because, as far as I can see, parsimony is increasingly recognised as an important methodological criterion, and its growing importance can also be seen in the literature. In what follows, I start with some examples, and then briefly introduce parsimony as it is usually presented in the methods textbooks. Thereafter I will go into some more recent meanings of parsimony, and in my conclusion I will provide some arguments as to why parsimony in my view deserves more attention in the designing and conducting of research than it has had so far.

Along with many colleagues who teach and do research, I use research methods in a substantive field and I have an interest in research methodology as such. In my case, the substantive field is empirical political science. I believe that talking about research methodology without empirical applications is impossible – therefore, my examples come from political science. I myself have always been convinced that the methodology of research in the empirical social sciences is basically one and the same whether you are in psychology, medicine, sociology, economics, or political science. Research methodology is the language by which we understand each other as researchers. To be sure, there are vast differences in styles of research over, and sometimes within, various academic disciplines. In this respect, the difference between experimental and non-experimental research is very important. But I disagree when colleagues claim a unique methodology for their specific subfield, which cannot be fully understood by ‘the others’. That would just stop scientific debate. First there were questions, then came scientific method, and only after that academic specialisations emerged.

I start with two examples of social research, the implications of which nobody in his or her right mind will believe. The first example is an analysis of the chances of United States senators dying while in office (Wuffle et al 1997). Table 1 depicts the decreasing death rates of incumbent United States senators. The researchers have fitted a logistic regression model to these data, with an ‘impressive’ R², and conclude that: ‘By 2050, estimated Senate mortality will be under one percent. By the year 3000, it will be virtually indistinguishable from zero’. In other words, Senators ‘who manage to retain office will be nearly immortal.’ Selection bias is unlikely because there is ‘compelling evidence that Senators who leave office do die.’

The article from which I quote is a nice and intentional example of the kind of nonsense that can result when a model fitted to specific data is extrapolated to describe unobserved data. But that is an important form of scientific inference, and wasn’t that what science was all about in the first place? Apparently, generalizing statements do not always make sense, even when the model fits the data very well. I will come back to this later.
Table 1 Mortality of sitting senators by decade (Wuffle et al. 1997)

| Decade    | Proportion of Senators who die while in office |
|-----------|-----------------------------------------------|
| 1910-19   | .30                                           |
| 1920-29   | .25                                           |
| 1930-39   | .24                                           |
| 1940-49   | .28                                           |
| 1950-59   | .20                                           |
| 1960-69   | .15                                           |
| 1970-79   | .08                                           |
| 1980-89   | .03                                           |

The second example is an analysis of the causal effect of monarchy as a regime type on societal well-being and democracy (Mayer et al. 1998). Figure 1 shows a summary of regression analyses on country-level variables for 27 Western countries, 11 of which are monarchies. The presence of a monarch seems to benefit almost everything – it even reduces income inequality – and indirectly also democratic performance. I quote a part of the conclusion: ‘As we have demonstrated, monarchy leads to democracy. If every nation crowned just one little monarch, then what a peaceful world this would be!’

The intentional nonsense here lies in the introduction of the variable ‘monarchy’. If this variable is to serve as a proper explanatory variable in a causal model, it should be a matter of pure chance whether a nation has become a monarchy or not. In other words, nations should have been randomly assigned to the conditions of monarchy or republic. When random assignment is not accomplished (and of course it isn’t), introduction of monarchy as an explanatory variable confuses more than it explains. Apparently, adding variables to your model does not always make sense either, even though we have all learnt to control for alternative explanations. I will come back to this later, too.

Nonsensical as these articles are, the research has been carried out with a methodological rigour that most of us, and certainly most introductory textbooks into research methods, would praise. What then is wrong? Different readers will probably use different terms, but my claim is that the fundamental problems with these papers stem from a common source. This is: a lack of parsimony.
What is usually meant by parsimony? The origin of the concept is often (but wrongly) associated with William of Ockham (c.1285 – 1349), the famous philosopher of late scholasticism.

William of Ockham was an important and influential political thinker. In those days, political theory was still primarily occupied with the relationship between the worldly and the spiritual powers – the old conflict between pope and king. Ockham’s main philosophical concern was a re-appraisal of the logic of Aristotle (see for example Bertrand Russell’s 1946 *History of Western Philosophy* and its connection with political and social circumstances from earliest times to the present day).

For a paper on parsimonious methodology, Ockham is the natural point of reference because he is presently still primarily associated with the oldest and most forceful statement of the principle of parsimony: ‘Entities should not be multiplied without necessity’. This principle is known as Ockham’s razor. Ockham did not invent parsimony – it had been a defining property of scholastic thinking long before.

Parsimony is not an undisputed criterion. The most obvious objection against it is that there is no reason to believe that the world itself is simple, that ‘nature is pleased with simplicity’. But theories do not necessarily *represent* reality. Theories can be regarded as constructs that help us understand aspects of the world in which we live.

Ockham’s razor and parsimony, or the principle of economy, are concepts which every student in the social sciences is taught about somewhere in the academic curriculum. But in contrast to what might be expected given the familiarity of the term, the meaning of parsimony is not so clear at all.

**Theoretical parsimony**

For example, what can be ‘entities’ and what is ‘without necessity’? Nowadays, in textbooks on research methods, the meaning of parsimony is very limited. Parsimony usually refers to theoretical simplicity. At first sight, there is a clear understanding about the meaning of theoretical simplicity. It is one of the structural components of Karl Popper’s philosophy of science (Popper 1959). Popper rejects the notion that simplicity is merely an aesthetic quality of theories. Instead, simpler theories are to be preferred because of their greater falsifiability. This interpretation of parsimony favours more general theories over more specific theories, or: simpler statements over complex ones. In Popper’s words: ‘because they tell us more; because their empirical content is greater; and because they are better testable’.
This does not solve all conceptual problems. What theoretical simplicity actually means, or how it is measured, is far from clear in concrete cases. Take as an example the debates on Intelligent Design and evolution. Proponents of creationism could – and sometimes do – argue that creationism as a theory of the origins of species is superior to evolutionism because the single ‘as if’-assumption of an Intelligent Creator is more parsimonious than the clumsy set of assumptions that describe the evolution of new species out of old. This is, however, not a common interpretation of Ockham’s razor – the existence of an Intelligent Creator cannot be shown by scientific means, and therefore ‘creation’ cannot be regarded as a scientific alternative to the model of the origins of species.

Multiple criteria for good research

Parsimony in the sense of Popper’s theoretical simplicity requirement is an accepted criterion for judging empirical scientific research. But it is only one of a set of such criteria. Since Aristotle introduced logic and facts as the two pillars of empirical science, we have always had to deal with a variety of criteria that together determine the quality of our research. Other criteria are, for example: validity, reliability, testability, replicability, precision, to name a few of the most important. These criteria sometimes vary over textbooks. For instance, King, Keohane and Verba (King et al 1994) reject parsimony as a general requirement of science because they interpret it as an assumption about the real world which they believe is unwarranted. I already indicated that a different view on what theories try to accomplish is possible. Others would reject replicability as a general criterion because social research very seldom is – and probably also should not be – designed as formal experiments that can be repeated by other researchers.

There is debate on the precise set of methodological criteria, but there is also a widespread consensus about what the core criteria could be. For example, that our propositions should be valid and reliable, can be found in every textbook. In the practice of research, it is often possible to improve different criteria simultaneously. But at some point, sooner or later, different criteria may also conflict with each other. The validity of a research project can be enhanced greatly by introducing more context and more detail, but at some point this will come at the expense of the reliability of the results, simply because these results have become idiosyncratic.

Let’s return to parsimony. Theoretical parsimony is just one of several quality criteria that should be observed in empirical research. And it sometimes conflicts with other criteria. For example, in the study of elections one can come across the proposition that the introduction or re-introduction of compulsory voting would increase electoral turnout. One of the proponents of this view is Lijphart (Lijphart 1997). His defence of the causal relationship between compulsory voting and turnout is built on the strong association between the presence of a form of compulsory voting in a country, and the turnout in elections. The association exists, but the causal proposition is probably invalid because the theory is too simple. Compulsory voting was introduced in many countries long ago, under completely different social and political circumstances from those operating nowadays. Turnout tends to be higher in those countries. But the introduction of compulsory voting in a modern society in the expectation that turnout will rise, is likely to fail. The costs of enforcing people to vote would be enormous, and only considerable fines would prevent mass civic disobedience. In the end, only these fines and not the moral requirement to vote would increase turnout. Lijphart’s theory is an example of what Lieberson has called irreversible causation – a causal effect that only works once, and in one direction (Lieberson 1985). Obviously, a causal theory of compulsory voting and turnout would need to take many other factors into account – including the country’s past experience and the willingness of its people to obey rules of this kind, all at the expense of theoretical parsimony but to the benefit of validity. This example also indicates that not all methodological criteria are always equally important. As a rule, validity in its various meanings is given the highest priority; the other criteria follow.

Now recall the two examples of nonsensical research that I started with. I suggest that it is also parsimony as a methodological criterion, but of a different kind from that which we have seen so far, that should in fact have been applied in these cases.

The first example, about the death rates of incumbent United States senators, suffers from a lack of parsimony in the generalization of the data at hand. The second example, about the relationship between monarchy and
democratic performance, suffers from a lack of parsimony in the specification of the causal model. I will now briefly argue why and how these two new types of parsimony should be applied as criteria for good social research, and start with the latter.

**Parsimony in model specification**

Let’s first take a closer look at the problem of too many variables in a model specification.

Suppose that you are interested in estimating the causal relationship between a Y and an X. When you have collected relevant (nonexperimental, observational) data, you want to rule out that a relationship between X and Y should actually be attributed to other variables. It depends a bit on your methodological upbringing, but I think many social researchers who have learnt the concept of control variables would like to introduce the controls they think are relevant.

But this procedure is only guaranteed to work under a special condition: that is, when your units have been randomly assigned to, or are randomly distributed over, the categories of your control variables. When this is not the case, the selection mechanism that has placed your units in the categories of the control variable by definition has a systematic component, and the control variable could easily result in misleading estimates. The reason is that introducing control variables raises the obvious question whether their effect is or is not dependent on other omitted variables.

As an extreme example of how things may go wrong, think of Simpson’s paradox. Simpson’s paradox shows us that introducing or leaving out a control variable could lead to completely opposite conclusions about a causal effect. A hypothetical example could be the relationship between taking a new medicine (X) and recovering, within a specified period, from an illness (Y). This relationship has been assessed for a sample of 800 patients. The bivariate relationship shows that the new medicine appears to be effective: the recovery rate is higher among the patients who were given the new medicine than among the patients who did not get it.

**Table 2 Effect of a new medicine**

| Did not recover | Recovered | % Recovered |
|----------------|-----------|-------------|
| Did not take medicine | 240       | 160         | 40%         |
| Took medicine        | 200       | 200         | 50%         |

A researcher suspects that this relationship may be affected by the life style of the patient – healthy or not-so-healthy – and consequently controls the results for life style. The results for the subgroups of patients with a healthy and an unhealthy life style may show that in both groups, the relationship between taking the new medicine and recovery is now negative. (The obvious reason is that lifestyle appears to be closely associated with taking the new medicine. When you are the doctor, what do you now suggest to your patients with this disease? Shouldn’t we introduce other variables which explain both the patient’s life style and taking the new medicine? And how would that help us?)
Table 3: Effect of a new medicine, by lifestyle

| Healthy lifestyle | Did not recover | Recovered | % Recovered |
|-------------------|----------------|-----------|-------------|
| Did not take medicine | 30 | 70 | 70% |
| Took medicine | 120 | 180 | 60% |

| Unhealthy lifestyle | Did not recover | Recovered | % Recovered |
|---------------------|----------------|-----------|-------------|
| Did not take medicine | 210 | 90 | 30% |
| Took medicine | 80 | 20 | 20% |

The practice of indiscriminately introducing control variables which should be controlled themselves is still widespread, at least according to what I see in my field of electoral research. In survey research, these control variables typically are attitudes such as political interest, which themselves result from a complex process with many factors involved. Control variables are introduced in order to eliminate selectivity problems in the explanatory variables of interest to the researcher. But, as Lieberson explained at great length in his 1985 book, the control variables themselves bring into the analysis new forms of unmeasured selectivity. As a result, the control variables may not just do a poor job, they may even further obscure what is actually going on.

In its worst form, introducing control variables resembles the so-called Casablanca approach. In this approach, every variable which has been spotted in suspect places in the past, or which has a reputation of being a control variable, is included in the model. As in the movie, the order is to ‘round up all the usual suspects’. But as in the movie, the one variable that should have been included, the key explanation of the phenomenon that we are interested in, is left out.

Introducing control variables in models should therefore be done with care; they do not always behave as expected. This was stated in an alternative, but extremely forceful way a few years ago by Chris Achen. Achen condemns the current practice in political science research in which basically simple research questions are obscured in empirical analyses with multivariate models fitted to a given data set. In his words: ‘The key point is that no one can know whether regressions and MLEs actually fit the data when there are more than one or two independent variables. These high-dimensional explanatory spaces will wrap themselves around any dataset, typically by distorting what is going on.’ (Achen 2002)

There are serious problems with empirically estimating models with many variables. But what is the solution? Achen himself proposes a strict, but in my view very attractive rule – a rule of three, or ART: ‘A statistical specification with more than three explanatory variables is meaningless’. If you think you need more controls in your model, then there is too much going on in your data set to allow reliable inferences. Instead of using more controls, you should focus on parts of your data set – those parts for which you can find meaningful causal relationships. Other observations should be discarded. Achen argues for careful, cross-tabulation-like analyses of subsets of the available units, in order to make more sense out of the data.

Thus, the discussion of model specification and control variables has so far resulted in two recommendations of parsimony:

1. Limit the number of explanatory variables to a maximum of three;
2. Limit your analyses to the subset of data units for which it makes sense.

Parsimony in an empirical domain

Secondly, we take a closer look at the problems that may arise when an empirical finding is, so to speak, overstretched to cover an empirical domain for which it is clearly not valid. This was the case with the
example of the immortality of incumbent United States senators by the year 3000. It is in fact a very common procedure to transplant research results obtained in one setting to completely different settings.

Looking from a slightly more formal perspective to the example discussed earlier, we note that the data cover the years 1910 until 1989. The variable ‘year’ in the data set is thus bounded by this minimum and maximum. What we then do when making predictions, is to apply the model fitted to the 1910-1989 range of ‘year’ to values of that variable that are far outside of this range (actually: 1990-3000). This is the common procedure of extrapolation when making predictions. It is based on strong assumptions. The model fitted to the data for one range of a variable should also be valid for another range: this can only be expected when the model is extremely reliable, valid and robust. When does that happen with our models?

In my own field, you can see how different ideas about the functioning and development of democracy, developed in Western Europe and the United States, have since 1989 been increasingly applied to new democracies. We have already referred to the relationship between compulsory voting and turnout. Lijphart’s argument that I referred to are practically all obtained for countries with a turnout somewhere between 60 and 95 percent. The United States is an outlier in models of turnout. So it is far from likely that introducing compulsory voting in the United States will increase turnout within this 7 to 16 percent range. What will happen cannot be predicted – on the one hand, there is a lot of room for increasing turnout in the United States; on the other hand, it may be doubted that citizens will let themselves be forced to behave according to these formal democratic standards.

Should we find a non-compulsory-voting country with a 5 percent turnout, or with a 95 percent turnout, how confident would we be applying the findings obtained in a set of ‘normal’ democracies? The more extreme the value of units on the key variable, the less confident we would probably be. This means that we should be very careful in setting up counterfactuals (‘If the United States introduced compulsory voting, turnout would go up with 10 per cent’) for cases which are clearly different on the dependent variable. And since the effect of compulsory voting is not an isolated phenomenon, but also depends on other political and social variables, careful consideration of the values of these other variables is required as well.

Being careful is perhaps not the most helpful advice one could get. Is it possible to be more concrete? In a recent article, Gary King and Langche Zeng provide an overview of the problems of what they call ‘extreme counterfactuals’, plus a number of solutions (King and Zeng 2006). Extreme counterfactuals include, for example, the estimated death rate of United States senators in the year 3000, or the estimated electoral turnout in the United States when compulsory voting is introduced. These counterfactuals are extremely dependent on a model fitted to data that are far removed from them, and are therefore probably wrong. Drawing on the idea that extrapolation is a more demanding activity than interpolation, they develop a decomposition of the bias in estimated causal effects: the bias is partly attributed to interpolation and extrapolation. Extrapolation bias – the type of bias we have referred to so far – appears to be important and hard to correct. In the end, the best remedy is to collect more data, which can make the counterfactual more plausible. When this is not feasible, King and Zeng offer well-meant advice to change the population of inference (and thus the research questions). ‘It may be disappointing, of course, to know that the desired questions have no good answers in available data, but it is better to know this than to ignore it’ (King and Zeng 2006: 152).

Thus, our discussion of parsimony in an empirical domain leads to two recommendations:

(1) For any inference on the basis of your data, take the model-dependency of your results into account;
(2) When you are tempted to say something about cases that are far removed from your original data, make your population of inference smaller in order to avoid extreme counterfactuals.

Summary and discussion

My argument so far can be summarised as follows. In the centuries since William of Ockham formulated the principle of parsimony, in the philosophy of science this principle has had a primary meaning of theoretical
simplicity. In Popper’s terms, parsimony thus enhances the falsifiability of theories. But parsimony as a criterion of good science can be applied to other aspects of the research process than theory construction. I have pointed to two other applications of the principle of parsimony. First, parsimony in model specification warns us against the dangers of introducing control variables in our model which perhaps should be controlled in their turn. Because too many variables in a model can easily lead to confusion rather than insight, it has been suggested that the number of explanatory variables in any model be limited to a maximum of three. Also, specific analyses may only be conducted on a subset of units for which these analyses make sense. Secondly, parsimony in choosing the empirical domain is important because, when making inferences based on collected data, we often stretch our estimated model as if it would fit data from a completely different context. This can be avoided by adjusting the population to which you want to make your inference – in time, or in space.

Simple theories, fewer variables, analyses on subsets of your units, and a better delimited population – these are in short the recommendations of parsimonious methodology.

This summary leads to the final question, relevant for the context in which this paper was written: is parsimonious methodology also methodologically innovative?

Of course, the idea of parsimony as a good quality in scientific activities is old – at least as old as scholasticism. But that does not mean that researchers always live up to this principle. Much of the research published in political science journals, for example, is hardly more than the mechanical testing of hypotheses on (survey) data, preferably with sophisticated statistical models and a number of controls. Following our earlier argument, it is extremely unlikely that this type of research will reveal anything interesting about the political world. Finding interesting results requires focus, and focusing means that you have to decide what, and which aspects of it, to look at.

Parsimony in model specification and in an empirical domain does not make your research more easily generalizable. Thus, these meanings of parsimony may conflict with the theoretical parsimony that Popper had in mind. That is in itself not a problem: it also applies to various interpretations of validity, and we know that good research has to fulfil different criteria simultaneously.

Parsimony as a methodological criterion is a conservative criterion: it recommends not constructing overly complex theories, not specifying more variables than you can really grasp the microtheory of, and not overstretching findings beyond a reasonable empirical domain. But conservatism can be innovative too. Innovation always involves a comparison with existing practices. In this sense, parsimony is a methodological innovation.

References

Achen, C. (2002) ‘Toward a new political methodology: Microfoundations and ART’. Annual Review of Political Science, 5: 423-450.

King, G. and Zeng, L. (2006) ‘The dangers of extreme counterfactuals’. Political Analysis, 14: 131-159

King, G., Keohane, R. and Verba, S. (1994) Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton: Princeton University Press

Lieberson, S. (1985) Making It Count: The Improvement Of Social Research And Theory. Berkeley: University of California Press.

Lijphart, A. (1997) ‘Unequal participation: Democracy’s unresolved dilemma’. American Political Science Review, 91: 1-14.

Mayer, J. and Sigelman, L. (1998) Zog for Albania, Edward for Estonia, and monarchs for all the rest? The royal road to prosperity, democracy, and world peace. Political Science and Politics, 31: 771-774
Popper, K. (1959) The Logic Of Scientific Discovery. New York: Harper & Row.

Russell, B. (1946) History of Western Philosophy. London: Allen & Unwin.

Wuffle, A., Thomas, B. and Koetzle, W. (1997) ‘Death, where is thy sting? The Senate as a Ponce (de Leon) scheme.’ Political Science and Politics, 30: 58-59