Limits of and Alternatives to Multiple Regression in Macro-Comparative Research

Michael Shalev
Hebrew University of Jerusalem
shalev@vms.huji.ac.il

Paper prepared for presentation at the second conference on

This paper offers criticisms of, and alternatives to, the use of multiple regression in cross-national comparisons of welfare states and comparative political economy generally. To set the context for this discussion, it is worth remembering that there was once quite a clear divide between descriptive and prescriptive studies by historians or social policy analysts—the people who knew what they were talking about; and social scientists—the people who saw welfare states as a convenient source of data for testing abstract theoretical generalizations. This division was inextricably linked to the methodological predilections of each camp. The sociologists and political scientists who worked on welfare states were part of the quantitative revolution in comparative studies. Using correlation and regression analysis, they optimistically

1 This is a work in progress, please shower me with comments and criticisms but don’t cite it yet.

I am grateful to the comfortable surroundings and willing ear furnished by Walter Korpi, and a stimulating correspondence with Bruce Western, for rousing me to initially put these ideas on paper. I received extremely helpful comments on earlier drafts from Frank Castles, Peter Hall and Robert Franzese and from participants in the Workshop on Economic Internationalization at Duke University, May 25-27 1997 and a seminar at the Wissenschaszentrum Berlin in February of this year.
hoped to conclusively demonstrate that the size of welfare states is a function of economic development, the political strength of the left, the size of the population in need, or some combination of these master explanations.

This kind of simplistic endeavor has not disappeared altogether, but there has been a compelling trend in quantitative work towards greater sophistication. More complicated forms of regression analysis, and some alternative methods of quantitative analysis, have been proposed. A 1994 volume edited by Thomas Janoski and Alex Hicks is the outstanding source here. Comparativists working on the welfare state and other areas of political economy have also been part of a welcome recent trend towards greater recognition of the limitations of standard applications of multiple regression, and sophisticated attempts to overcome these limitations without sacrificing the power of regression. Young and statistically literate scholars like Robert Franzese (forthcoming), Torben Iversen (1998) and Bruce Western (1994; 1995) are shining examples of this constructive critical awareness.

Meanwhile, the comparative-historical study of welfare states remains an important research tradition. Pioneers like Theda Skocpol and John Stephens and scholars influenced by them continue to utilize historical process-tracing in contrasting cases in order to unravel puzzles posed by the diversity of welfare states. But the proper relationship of these types of endeavors to quantified multivariate comparisons remains contested. For instance while Skocpol eschews quantification altogether, Stephens and his collaborators believe in constructive dialog between comparative history and multi-country regression analysis (Huber, Ragin, Stephens, 1991; Rueschemeyer, Huber and Stephens, 1992).

In this respect, Charles Ragin's 1987 book *The Comparative Method* represents a watershed. Ragin was insistent that comparative scholarship must "not be divided into two parts—those who know something about actual empirical cases and those who know something about multivariate statistical techniques" (Ragin, 1987:xxx). He proposed a method of formalizing the "case-oriented" approach favored by analysts with a qualitative/historical bent, in order that comparative social science could enjoy the rigor of techniques like multiple regression without the burden of their "variable-oriented" epistemology. Qualitative Comparative Analysis (QCA) was presented as a concrete way of fulfilling this potential.

---

2 The *American Journal of Political Science* has emerged as a leading forum for cutting-edge work on how to use MR in more sophisticated ways.
Both the turn to "soft" comparative-historical analysis and the more "rigorous" QCA technique have been subjected to withering criticism in the last few years. The critics stress issues of falsifiability, generalizability and reproducibility that allegedly arise when social scientists utilize the methods and results of historical research (Goldthorpe, 1991 and 1997; Lustick, 1996). Comparative historical sociology has been taken to task for being overly inductive (Kiser and Hechter, 1991) and for failing to emulate the scientific principles embodied in good quantitative analysis (King, Keohane and Verba, 1994). Critics of QCA do not accept the idea that it embodies the best of both case and variable-oriented analysis. Instead they argue that it sacrifices the advantages of probabilistic over deterministic explanation, and deductive over inductive theorizing (Lieberson, 1991 and 1994; Goldthorpe, 1997).

A conspicuous feature of the methodological ferment in comparative social policy and political economy is its confirmation of the hegemony of MR. Thus Janoski and Hicks (1994) presented alternatives like QCA and Event History Analysis, but they also devoted considerable attention to extending MR to cross-national datasets covering multiple points in time ("pooling"). The recent anthology (Mjøset et al., 1997) devoted to the controversy between Charles Ragin and John Goldthorpe makes a fine contribution in clarifying the debate, but the tone of the debate is so general that it leaves intact the usual tendency to conflate the variable-oriented approach with MR. A third, still more telling indication of hegemony comes from Gøsta Esping-Andersen's seminal study of welfare state regimes. It is striking that, after offering a forceful critique of the core assumptions of conventional regression analysis Esping-Andersen himself turned to MR in order to systematically assess the empirical validity of his arguments.

This paper reviews and extends well-known criticisms in principle of MR, but its main focus is on practical examples showing the value of alternative methods. Some of the arguments are of a general nature, but the emphasis is on macro-comparative research. Two principal conclusions will emerge. First, even though technical means are available to deal with many of the limitations of MR, these solutions are either unconvincing or else require such a high technical investment that they offer questionable returns on scholarly investment. Second, dissatisfaction with MR does not necessarily mandate radical qualitative alternatives such as QCA or complete abandonment of numerical data. Low-tech forms of analysis (tables and graphical methods) and multivariate statistical techniques other than MR (such as cluster analysis and multi-dimensional scaling) constitute viable and useful alternatives. The paper's most extensive illustration reanalyzes the data in Esping-Andersen (1990). MR was a poor choice for Esping-Andersen's
purposes, but QCA also suffers from serious limitations in this application. In contrast, factor analysis offers powerful and cogent empirical support for the "three worlds" thesis.

1. The trouble with Multiple Regression

Given that MR is the flagship of variable-oriented research in the social sciences, it is not surprising that most of the difficulties that it poses for comparativists have already been eloquently rehearsed in Charles Ragin’s book *The Comparative Method* (1987). Like most (but not all) other methods of multivariate statistical analysis, MR works by rendering the cases invisible, treating simply as the source of a set of empirical observations on dependent and independent variables. An obvious reason why this is undesirable is that the "cases" which interest us in fields like comparative political economy are limited in number and occupy a bounded universe so they are knowable and manageable. Consequently, leaving "proper names" in the analysis is an efficient way of conveying information and letting readers evaluate it. Moreover, most producers and consumers of comparative political economy are intrinsically interested in specific cases so why not cater to this interest by keeping our cases visible?

Different views of causality are an equally celebrated source of the debate between case-oriented and variable-oriented researchers. Andrew Abbott (1988) has cogently argued that “all too often general linear models have led to general linear reality, to a limited way of imagining the social process”. Abbott notes the constricted theoretical scope of the notion causality underlying linear models, which cannot recognize (or is unlikely to recognize) situations where

3 It is of course debatable just how bounded the research universe is. Boundaries like "rich, capitalist countries with longstanding democratic polities and non-trivial populations" may be theoretically arbitrary and should always be open to challenge. There is a longstanding (although by no means universal) trend towards incorporating Greece, Spain and Portugal after democratization (and more practically, after their inclusion in OECD databases). Other candidates for inclusion in work on comparative political economy of what have until now been known as "the Western nations" might be found in the former Soviet bloc states, Latin America and East Asia. There are good arguments both for and against expanding the universe of comparative studies. For instance, compare Geddes (1990) and Boyer (1997).

4 Even the well-known injunction of Przeworski and Teune (1970) that comparativists should strive to turn the proper names of countries into the abstract names of variables did not entirely reject this view. It is worth pointing out that Przeworski and Teune were railing against the dominance of comparative politics by "area studies" specialists and urging their colleagues to avoid particularizing arguments that could easily strait-jacket both theory and comparison. Many contemporary advocates of case-oriented analysis including Ragin would have no quarrel with this assessment.
the effect of any given causal variable is uneven, contradictory (dialectical), or part of a wider bundle of factors sharing an “elective affinity”. In the social world effects are typically contingent upon their setting, including two types of historical contingency: temporal context (period effects) and time paths (particular historical sequences or cumulations).

Case-oriented analysis easily accommodates such nuances because it assumes from the outset that the effect of any one cause depends on the broader constellation of forces in which it is embedded (“conjunctural causation” in Ragin’s words). If MR models try to emulate this assumption, they will acquire interaction terms which aggravate the familiar “degrees of freedom problem”. MR is also an alien environment for another causal assumption that flourishes in case-oriented analysis, namely that there may be more than one constellation of causes capable of producing the phenomenon of interest. Technically, MR models can handle this by adding more independent variables. But the results will be ambiguous because they offer no way of distinguishing between conditional relationships and multiple causal pathways. Without attending to the cases we cannot know whether all those with Y tend to possess some combination of X1 and X2; or whether the presence of Y is associated in some of them with X1 and in others with X2.

To fully appreciate why MR is such a problematic choice for comparativists, it is worth considering why it is seemingly such a good choice for certain other kinds of social science. Economists are often interested in estimating the marginal effect of one economic variable on another, holding constant the impact of other presumed causes. If prices rise, what will be the likely effect (net of other known influences like the rate of investment and the terms of trade) on economic growth? If people invest in a college degree, what will be the likely effect (net of other known influences like race, gender, and work experience) on their future income stream? MR suits this project perfectly. Estimating marginal effects under conditions of *ceteris paribus* is precisely what it aims to do. Economists, of course, enjoy the added advantage that they typically deal with continuous variables that are conveniently evaluated using a uniform metric, money. In contrast, much of the curiosity of comparative political economists revolves around the presence or absence of certain conditions. Will economic growth be higher in the presence of corporatist trade unions (or a hegemonic social-democratic party, or an independent central bank)? It might be nice to know how much growth results from how much corporatism. But our theories are quite imprecise when it comes to making predictions like that.
The evaluation of marginal effects in macro-comparative research is also dogged by the ambiguity of many of the variables of interest and the difficulty of measuring them precisely. Concepts like corporatism are so contentious that even categorical measures exhibit distressingly little consensus (Shalev, 1990; Janoski, 1997). Naturally continuous variables like the ubiquitous "left party cabinet representation" can be measured precisely but such precision conceals huge reliability and validity problems. Thus, inter-country comparisons of longterm differences are plagued by the difficulty that, for example, a mean fraction of cabinet seats of 50% is consistent with either intermittent left government, stable left participation in cabinet coalitions, or a dominant left (or right) party which is unseated in midstream. Comparison over time is equally problematic, since the numbers alone cannot tell us whether the left’s role in government has shifted between qualitatively different conditions like one-party dominance, wall-to-wall coalitions, junior partnership, pivot party facing a divided right, etc. MR could accommodate such complexity by replacing the continuous measure of left strength with a series of dummy variables, or perhaps by finding an appropriate non-linear functional form to capture discontinuities in the effect of left strength on the phenomenon of interest. But the first solution is "wasteful" of precious degrees of freedom and the second requires either good luck or an unlikely degree of theoretical sophistication.

In the behavioral sub-fields of sociology and political science much of the appeal of MR derives from its comfortable fit with the sample survey method. Because they enjoy a relatively high ratio of cases to variables, survey researchers are able to use MR as a means of introducing "statistical controls". Unlike the economists they may not be motivated by a theoretical view of the world that is inherently marginalist. They use controls in the hope of dealing with causal forces that in the ideal experimental design would have been neutralized by random assignment of subjects to differential "treatments". This approach has been the subject of vigorous debate. In different ways David Freedman (1991) and Stanley Lieberson (1985) have made compelling arguments that proper statistical control would require much more sophisticated and complete causal theories than social researchers can hope to have. In any event, since comparativists

5 The criticism here is not that quantification over-simplifies complex reality. There is always a trade-off between accuracy and parsimony in social research, whether analysis uses quantitative measures or narrative representations. The point is that the use of MR encourages what may well be a mistaken belief that our measures are precise and continuous.

6 Surprisingly few practitioners of MR are aware of the debate over whether statistical controls are actually capable of doing the job. I am indebted to Bruce Western for directing me to this literature.
necessarily deal with a small number of cases it is technically difficult for them to analyze the effect of more than a few independent variables.

Staying with the survey researchers, we can identify a final reason why the appeal of MR outside of comparative research ought not to inspire its use within the field. To economize on resources, analysts of voter opinion or social mobility usually poll only a tiny fraction of their target population. As a result, a fair amount of the immense heterogeneity that characterizes a universe like "American voters" cannot possibly be captured in the typical sample of only one or two thousand. It is quite conceivable that even the most unlikely combinations of the independent variables (let’s say, an unmarried minority mother who earns $100,000 a year as a company CEO and votes Communist) actually do exist in the universe. From this viewpoint, one of the advantages of MR is that using the observations in hand, its coefficients (marginal effects) project relationships across the spectrum. We may worry that our sample artificially truncates the true range, but if that occurs it is because of faulty research design, not because we chose a faulty analytical technique.

In cross-national quantitative research the situation is very different. We often analyze the entire target population, and if not it is usually because of lack of data rather than sampling considerations. For the most part then, if a particular configuration of attributes does not exist in a cross-national dataset, it does not exist at all. To grasp the size of the problem, consider the following hypothetical example using only three independent variables and a crude level of measurement. Social security expenditure as a proportion of GDP is regressed on left party power, exposure to trade and proportion of the population over 65. All variables are measured on a 5-point scale. If we were to construct a multiway table with this dataset, it would have 625 ($5^4$) cells. Since no study of the OECD area can have more than 20 cases, that would mean over 600 empty cells! MR in effect places imaginary countries in some of these empty cells when it seeks out the best linear fit which can be generated for the data at hand. Because it estimates partial parameter effects as if all (linearly-fitting) configurations were possible, MR can easily yield problematic results.

Abbot (1997) has recently offered an elegant formulation of this problem. Variable-oriented approaches “seek to understand the social process by developing linear transformations from a high-dimensional space (of ‘main effects’ and occasionally of interactions between them) into a single dimension (the dependent variable)... Now this strategy ... is useful only if the data space is more or less uniformly filled" (Abbot, 1997).
In this context, it is worth recalling some limitations of the social-democratic model of the welfare state (Shalev, 1983). Comparing the US with Sweden, Andrew Martin (1973) inferred 25 years ago that Social Democratic Party dominance was the crucial difference responsible for Sweden's postwar commitment to the full-employment welfare state and its glaring absence in the US. Numerous correlation and regression studies echoed this argument and went on to confirm its veracity across the whole spectrum of advanced capitalist democracies. Two difficulties arose. One of them was inadequate predictive power for cases that were neither social-democratic nor right-dominated, among other reasons because of the relevance of a different causal complex (the Catholic-conservative dynamic later spotlighted by Esping-Andersen). This is the problem of multiple causal pathways mentioned earlier.

A second difficulty was that the social-democratic model could tell us little or nothing about the "positive" causes of social and labor market policy variation in the US or countries like it. When tested by regression it yielded implausible inferences like the following: a country similar to the US but with one additional decade of socialist rule would likely boast an unemployment rate three points lower and child allowances 40% higher than America's. This is an extreme example of the dangers of generalizing from empty cells when each of our cases is a complex historically-bounded gestalt. On the other hand, one of the tests of a useful causal model is that it be capable of answering counterfactual questions—that is, of filling empty cells with hypothetical data. Thus, it was in effect by asking how US policy would have developed under Swedish conditions that led Martin and others to focus on labor movement strength. However, some "cells" are so unlikely ever to be filled that they should not be part of either our computational space or our predictions. The attributes of societies are not subject to infinite variation in unlimited combination with one another.

From an MR perspective, this problem is not intractable. If a variable capable of explaining differences between Sweden and the US offers no guidance to the contrast between Canada and the US, then our model must be either under-specified or mis-specified. If the problem is under-specification the appropriate response would be to add independent variables capable of accounting for variation among the liberal polities. But with these additional variables in the model, it might become too large to estimate on a small cross-sectional dataset. In response, we might be tempted to enlarge our dataset by combining cross-sectional observations for different years. This would have the added advantage of permitting the investigation of intra-

---

8 I thank Rob Franzese for forcing me to think through this point.
country differences (i.e. within the US as well as between the US and other countries). The limitations of regressions on pooled datasets are taken up in section 4 of the paper.

If mis-specification is the problem then the solution would be to find an explanation sufficiently general that it could accommodate a wider range of variation—between the US and Canada as well as vis-a-vis Sweden. Here the argument against MR turns on the readiness of case-oriented researchers to assume that distinctive causal trajectories apply to different "families of nations" (Castles, 1993). If MR is obviously not the best way of testing plural explanations, what is? This issue will be discussed at some length in section 5, in the context of Esping-Andersen’s thesis of three distinctive policy regimes. There previous attempts to reassess the empirical support for this thesis will be reviewed, with special emphasis on the promise and limitations of QCA, and an alternative to both MR and qualitative approaches will be illustrated.

Before proceeding to the relatively sophisticated techniques for dealing with the problems of "too many variables and not enough cases" (section 4) and plural causation (section 5), I will discuss two concrete illustrations of the deficiencies of quotidian uses of MR. These illustrative examples were chosen with an eye to countering two possible strategic responses to the general critique of MR that has been offered so far. One of these would be to lower our expectations of MR, using it more as a means of partitioning empirically-driven variance than of strictly testing hypothesized causal relationships, assuming that it offers no more than a rough guide to model plausibility. Section 2 examines an example where MR was used as in this "loose" sense, and finds it wanting. Alternatively, one might argue that the causal status of regression coefficients should indeed be treated tentatively, but that our confidence would be

---

9 An example is Theda Skocpol's argument that "state capacities" and "policy legacies", not working class power resources, explain policy variations between the US and other countries (e.g. Weir and Skocpol, 1985). State-centered theory has however proven untestable (except very superficially) in quantitative comparative research. Partly because of theoretical slipperiness and partly due to an absence of good indicators, most of its independent variables cannot be either quantified or simplified. Straightforward aspects of the structure of states may be measurable, e.g. the constitutional provisions incorporated into the work of Huber, Ragain and Stephens (1993) on welfare state variations. But the framing of political action and agendas by state capacities, policy legacies and the autonomous initiatives of state managers has not been given serious consideration except in non-formal historical research. An exception is Amenta and Poulson's (1996) regression and QCA study of the American states. This exception proves the rule, however, since the measurement of such concepts as "administrative strength" was possible in this study only because it compared sub-national units of a uniform national entity.
strengthened if alternative methodologies yielded convergent findings. An illustration of this triangulation strategy will be found in section 3.

2. "Causal arguments" or mere "summaries"?

With multidimensional data sets, regression may provide helpful summaries of the data. However, I do not think that regression can carry much of the burden in a causal argument. (Freedman, 1991:xxx)

David Freedman is a statistician who believes in the power of numbers but has made it his mission to disabuse social scientists of their exaggerated belief in statistical inference as a tool of causal analysis (Freedman, 1985 and 1991). The essence of the argument made by Freedman (see also Leamer, 1983) is that statistical hypothesis-testing requires that the researcher have a well-developed theory in hand, and a hands-off relationship with the data, prior to the point at which testing is carried out. Since sociologists are unable or unwilling to follow these canons, the most they can legitimately do with MR is to use it to summarize multivariate datasets.

I know of no comparativist who has followed this advice literally, but some have treated MR as something between a formal hypothesis-testing device and an economical method of sustaining broad empirical claims. My example of this "low expectations approach" is drawn from Bo Rothstein's (1990) study of cross-national variation in union membership from a "new institutionalist" perspective. Under the so-called "Ghent system" unions bear responsibility for administering unemployment insurance, with the consequence that in periods of economic crisis or transformation their membership is unlikely to be eroded and may even increase. For theoretical reasons Rothstein wishes to demonstrate that the highest levels of unionization have been reached only in countries where this system is in place. Union density figures for 18 OECD countries in the mid-1980s reveal that Ghent is indeed present in all of the countries with the highest rates of union penetration, and only these countries. Unless Ghent is but a spurious understudy for the real star of the causal show, it has been a necessary condition for rates of more than 70 percent unionization. Of course this does not mean that Ghent is a sufficient condition for union hyper-success. It might be that Ghent merely amplifies the effects of other favorable conditions.
There are thus two possibilities that a simple table showing union membership alongside Ghent presence/absence can not address: spurious association (alternative explanations) and additional causes (complementary explanations). Following convention, Rothstein seeks to lay both issues to rest by executing a multiple regression that takes into account other probable influences on cross-country differences in unionization. These are left party participation in government and potential union membership (the absolute number of employed and unemployed wage-earners).

Rothstein’s model was re-estimated for this article using a modified version of his dataset which excludes picayune Iceland and replaces his left strength indicator with a more sensible alternative. Following the original, the coefficients are standardized *betas*.

Union Density = \(0.47\) (Ghent) + \(0.28\) (Left Government) - \(0.34\) (Log of Potential Membership)

All coefficients are significant at conventional levels (although Left Government only marginally so) and the adjusted R-squared is 0.73. The metric coefficient for the Ghent variable reveals that the net difference between Ghent and non-Ghent systems is a striking 27 percentage points.

Notwithstanding these indications of success, it can be argued that Rothstein’s use of MR is inappropriate and in part misleading. Rothstein is content, in his words, to show “that all three variables have an independent explanatory effect of about the same standardized size” (Rothstein, 1990). A prerequisite for these "explanatory effects" to be taken seriously is that the model be theoretically plausible. Yet earlier in the article Rothstein himself casts doubt on this, when he describes the argument for the significance of potential membership size as logically

\[ ^{10} \text{Rothstein sagely recognizes a more fundamental limitation of comparative statics, conceding that in order to prove his point he must go beyond cross-sectional analysis to provide concrete historical evidence. My concern here is only with Rothstein's use of MR to make the cross-sectional case for his argument.} \]

\[ ^{11} \text{It seemed inconsistent that Rothstein would include Iceland (with only 80,000 potential union members) but not Luxembourg, Monaco or Andorra in his analysis. The problem with Rothstein's left party representation indicator, which he borrowed from Wilensky (1981), is that it covers the entire 1919-79 period. Disruptions and discontinuities due to non-democratic interludes and wars largely vitiate the pre-1945 data. At the other end of the spectrum, the unionization data reveal that cross-national differentials stabilized after about 1965. It therefore seems reasonable to treat the first two postwar decades as the politically formative period. Figures for average left cabinet strength in this period were taken from the dataset used by Korpi and Shalev (1980). Neither of these modifications detracted from the results Rothstein himself obtained using MR. On the contrary, using the modified data the beta coefficient for the Ghent variable rose at the expense of that for left strength.} \]
indefensible and the left-government argument as guilty of what statisticians would call simultaneity bias. In addition, while the standardized coefficients indeed suggest that Ghent has as much empirical weight as rival explanations, because countries are invisible the results do not speak to Rothstein’s core claim that it is Ghent, not left strength or small size, which differentiates between the most unionized countries and all the rest. True, this claim would have been negatively ruled out had the Ghent effect disappeared once the other variables were added to the equation. But the regression could not make a positive case for Rothstein’s argument.

Beyond these specific limitations of MR in Rothstein’s case, it should be noted that the specification of his model rests on typical but questionable assumptions. One is the usual supposition that the explanatory variables exert independent rather than interconnected effects (which is obviously doubtful in this and the majority of cross-national models). The other is that none of the effects is assumed to be conditional on the value of other variables—i.e. no interaction effects are anticipated.

A straightforward way to address these problems is to summarize causes and effects in a descriptive tabular format that identifies different combinations of conditions (causes) with the countries that “carry” them. This requires some forethought because Rothstein’s model refers to three different causal variables and the dependent variable, unionization, is not easily collapsed (it is distributed fairly evenly across a broad spread). The proposed solution is a simple multiway flow chart or “tree” showing exact values of unionization for different “clusters” of countries. These clusters were created very simply, by cross-tabulating the presence or absence of Ghent with categorical versions of Rothstein’s two other causal variables.\footnote{Potential membership was simply dichotomized after exploratory charts revealed that it had an evident threshold effect on unionization. With the exceptions of only Switzerland and the Netherlands, all small countries (no more than 5 million potential members) had more than 50 percent density, while all the large countries (10 million and up) scored less than 50 percent. Within these two categories no relationship was discernible between the two variables. Left strength was grouped into four categories that reflect breaks in its distribution. “None” were cases with zero or trivial (up to Japan’s 4%) left party representation in cabinet; “weak” 7-15%; “medium” 22-29% plus an intermediate case (the UK) with 36%; “strong” 45% or more.}
Chart 1: Reanalysis of Rothstein

Large Countries (max. 5 million) (Mean=30)
- No Left
  - USA 18
  - Japan 28
  - Germany 31
  - Canada 38
- Weak or Med. Left
  - France 15
  - Italy 36
  - UK 43

Small Countries (min. 10 million) (Mean=62)
- Weak Left
  - Switzerland 34
  - Ireland 68
- Medium Left (Mean=58)
  - Netherlands 29
  - Australia 51
  - Belgium 74
  - Finland 80
- Strong Left (Mean=71)
  - Austria 57
  - Norway 58
  - Denmark 83
  - Sweden 86

No Ghent System (Mean=39)
- Ghent System (Mean=81)
The results (Chart 1) offer interesting evidence of nested causal effects. This is apparent, firstly, from the systematic difference between extant and non-existent configurations. Substantial left party representation was only attained in small countries, and only countries with a substantial left had the Ghent system. In the case of the affinity between Ghent and left strength, we cannot know which way the causal arrow points without branching into historical research. (Indeed, the same is true of the relationship between unionization and both Ghent and left strength.) But we can say that it is the syndrome of smallness, “leftness” and Ghent that is associated with the highest rates of unionization. The results also hint at an interaction. The “Ghent effect” seems to be stronger in countries with medium left strength than in the fully-fledged social democracies.

This “unsophisticated” method of presenting the data vindicates Rothstein’s thesis much more effectively than his regression. It makes clear precisely the things he wanted to demonstrate: that the Ghent effect is large and not spurious, and that it comes into play in countries where other conditions are broadly favorable to unions. But these results do something else important, which is to point the interested researcher to the most fertile questions for selective case comparisons that might help nail down how important Ghent really is. In particular, it must be questioned whether the Ghent system alone can explain the very large differences in density between the members of two otherwise well-matched pairs of countries: Belgium and the Netherlands, and Sweden and Norway.

---

13 On the other hand, left strength discriminated only weakly between the unionization rates of small countries, and not at all between the large ones (except perhaps for the British case).

14 The importance of these kinds of anomalies for scientific progress has recently been illustrated by Rogowski (1995). Note that the data also point to anomalies that may be irrelevant to Rothstein’s concern with the Ghent system but ought to interest students of unionization. One is the wide gap between the US and Canada. Another is Ireland’s unexpectedly high union density.

15 Visser (1992) has pointed out that most of the vast difference between Belgian and Dutch unionization can be attributed to the fact that Dutch unions have no presence in the workplace. The origins of the Norway-Sweden gap are less clear, but they may be traceable to the Norwegian union movement’s lesser effectiveness in some of the sectors that grew from the 1960s, when Norway’s density plateaued while Sweden’s entered a long period of growth. Data collected by D’Agostino reveal substantial gaps in union density favoring Sweden in the following (partly interrelated) categories: women, private sector trade and services, and white-collar workers.
The visibility of the relationship between variables and cases in the simple diagrammatic presentation favored here may thus draw attention to anomalous cases which reveal limitations in the theoretical model. Attending to outliers from a regression analysis is sometimes also a way of identifying anomalies, but not of the kind discussed here—namely, countries that don’t “make sense” when they are viewed in relation to other similar cases. Tabular or graphical presentation of the dataset with the observations named permits this; inspection and diagnostic testing of regression residuals does not.

3. Building confidence through triangulation

Peter Hall and Robert Franzese (1998) are the authors of an exceptionally thorough study that questions one of the most influential tenets of modern political economy, that independent central banks have a powerful favorable effect on economic performance. They argue that while independent banks are always anti-inflationary, there is a risk that their impact on unemployment will be far less salutary. Unless wage-setting is centralized and coordinated the bargainers will fail to internalize bank “signals”, and the result will be higher rather than lower unemployment.

In testing their argument Hall and Franzese proceed in three stages. First, they demonstrate its plausibility by referring to the paradigm case of West Germany. Second, they use data for 18 OECD countries over the postwar (1955-90) period presented in a simplified tabular format. Finally, they use MR to test a more elaborate model at several levels of aggregation ranging from full-period means (pure cross-section) to pooled annual data. The results of each one of these analyses is consistent with their argument that the impact of central bank status is conditional upon the structure of wage bargaining.

In their tabular analysis Hall and Franzese collapse measures of both bank independence and wage coordination and cross-tabulate them, first at the level of three-point scales and then as dichotomies. The results clearly confirm the hypothesized interaction effect. However this effect could in principle be an artifact, the result of some confounding influence like countries’ wealth, economic openness, or government composition. But the initial result survives the application of MR using controls for key economic, political and institutional variables. Conditional parameter estimates show that the interaction between independence and coordination is substantively, as well as statistically, significant. And diagnostic testing shows that these results do not depend on the presence of any particular case.
While the authors' conduct their regression analysis with professionalism, it is heroic to imagine that 18 observations can sort out the effects of a total of 9 “independent” variables (including interaction terms). Hall and Franzese appear to stand on solid ground here, because (a) some of their regressions are conducted using pooled data which obviate the degrees of freedom problem, and (b) the regression builds on the clear-cut (if “uncontrolled”) result of their tabular analysis.

The one element missing from Hall and Franzese's results is the same one that proved to be crucial in our reanalysis of Rothstein's study: identification of countries. In addition, as in any analysis of collapsed data, the result is dependent on decisions about how to aggregate the raw figures. A variety of criteria may be used: substantive knowledge of cases, aggregation into categories of similar size, or tailoring categories to “breaks” in the distribution of observations. Inspection of the underlying data on coordination did not suggest any “natural” basis for grouping any of the 5 original categories. The data for bank independence are continuous but skewed. After visually inspecting the raw scores and finding two apparent breaks, I converted them to low (<.4), medium and high (> .6) scores. The first panel of Table 1 shows the countries that occupy each cell, and the second indicates the level of unemployment.

The first panel of the table reveals that 6 of the 15 possible cells are empty. Fortunately the gaps are not bunched too much, but empty cells necessarily make it harder to infer continuous relationships. Such inference is also complicated by another feature of the data: in three of the cells occupied by a pair of countries, they have widely divergent unemployment rates (shown separately).

The marginal totals in the bottom panel are only partially consistent with Hall and Franzese's expectations. The bottom row shows that unemployment does tend to fall as coordination rises, although there are a few exceptions. But the rightmost column indicates that the highest average rate of unemployment is found in countries with medium rather than high levels of bank independence. Inspection of the individual columns suggests that bank independence makes little or no difference at any level of coordination. A mild interaction is however suggested when we look across individual rows: the tendency for coordination to lower unemployment is clearest among the five countries with the most autonomous central banks.

---

16 All coordinated economies (.75 or 1) except Denmark had unemployment rates under 3.5%; all uncoordinated economies (0 or .25) except New Zealand had unemployment rates above 3.5%.
Table 1: Reanalysis of Hall and Franzese

<table>
<thead>
<tr>
<th>Coordination of Wage Bargaining</th>
<th>Central Bank Independence</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>1</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low (&lt;.4)</td>
<td>ITA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NZL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>SWE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>IRE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UKG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>AUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DEN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>JAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High (&gt; .6)</td>
<td>CAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SWI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Mean</td>
<td>6.3</td>
<td>3.8</td>
<td>4.9</td>
<td>2.8</td>
<td>2.0</td>
<td>3.9</td>
<td></td>
</tr>
</tbody>
</table>

Note: Mean values are rounded to one decimal place.
This weakness of the evidence for the anticipated conditionality of the bank effect on wage bargaining arrangements might of course be the result of ignoring the theoretically extraneous influences that Hall and Franzese control for in their regressions. Given the criticisms of MR earlier in this paper, I would of course be reluctant to accept the evidence from the MR analysis at face value. An alternative approach is to return to the top panel of Table 1, where the 18 countries are named. The authors’ expectation that in the absence of coordinated wage bargaining, central bank independence would likely exacerbate the unemployment-inflation tradeoff may well explain the wide gap between unemployment rates in “coordinated” Austria, Switzerland and Germany versus the two “uncoordinated” North American states. Hall and Franzese need to persuade us that it is indeed coordination, rather than factors like economic openness or the political representation of labor, which explains this specific gap.

The moment that this contrast is specified, two reservations immediately spring to mind. First, each one of these sub-groups of countries constitutes a cultural, political and economic region with numerous internal similarities but which is worlds apart from the other. Regression's assumption of the independence of characteristics that cluster so strongly appears particularly misplaced in such a context. Secondly, we know that by the 1990s these two sub-regions had experienced opposite shifts in their respective past performance regarding unemployment. Replication using more recent data would provide an obvious test of the model's robustness. Unless changes in the control variables have created disparities of similar magnitude, it is questionable whether this test would be passed.

In sum, the weight of Hall and Franzese's comparative analysis is carried by their multiple regressions. But insofar as the validity of their argument turns on five critical cases, this should be explicitly recognized by carrying out reasoned and disciplined comparisons of (named) countries and/or periods. Even a simple tabular analysis that fails to name names delegates the task of generalization to a “black box” (in this instance, mean values based on anonymous cases). Accordingly, triangulation cannot be regarded as complete without moving to the case level, where researchers' knowledge of individual cases and creative use of selective case comparisons may play a decisive role in reinforcing or questioning the plausibility of higher-level generalizations.

3. Is pooling a panacea?

Some readers might view my comments on the two articles discussed so far as just another illustration of a well-known problem: that because comparativists have “too many variables
chasing too few cases”, MR can only be applied either crudely (Rothstein) or implausibly (Hall and Franzese) in standard cross-sectional designs. My alternative approach might be criticized as a dishonorable retreat to rendering descriptive summaries of the data that are all too dependent on arbitrary decisions about how to group and present them. These critics would doubtless reject my argument that regression is fundamentally unsuited to macro-comparative analysis, and would prefer to focus their creative energies directly on solving the problem of insufficient cases (cf. King, Keohane and Verba, 1994).

In this spirit, John Goldthorpe has recently argued that “au fond the small-N problem is not one of method at all but rather of data”. Goldthorpe specifically recommends emulating the large number of researchers who “have ‘pooled’ data for the same set of nations for several different time-points. Observations—and degrees of freedom—are in this way increased…” (Goldthorpe, forthcoming). However, there are well-established reasons to believe that the most likely consequence of a turn to pooling is to muddy the causal waters still further.

What does pooling entail? Traditionally, quantitative macro-level research analyzed either snapshots of different countries at a single moment in time (cross-sectional data), or else period-to-period data for a single country (annual figures or sub-period “panels”). In contrast, pooled datasets “stack” timeseries or panels for multiple countries one on top of the other. Hence, they embody both comparative variation between countries and dynamic variation over time. As a result, analysts must contend with the technical complications characteristic of both cross-sectional and timeseries estimation. Not surprisingly, practitioners face a bewildering

17 Goldthorpe recommends even more strongly that researchers widen the “geographical and sociocultural range” of their research. In this matter however it cannot be said (as it can of pooling) that the recommended solution is a popular one. It is not only (as Goldthorpe concedes) that data quality and availability are limited outside of the bloc—the OECD countries—that interests his intended audience (and mine). Rather, it is widely understood that what might be called the “specification costs” of going beyond the OECD (additional casual factors and alternative causal paths) usually seem to outweigh the potential benefits. In any case, it is far from certain that in practice larger samples would overcome the difficulties of treating cross-country comparison as an exact science. It is sobering to consider the example of the economics of growth. Even in a theoretically developed field where it was possible to gather comparable data for a stunning 119 countries, Levine and Renelt (1992) found themselves hopelessly unable to use cross-national regressions to adjudicate between rival theories.

18 For an early and non-technical exposition, see Stimson (1985). For more complex elaborations, as well as references to numerous applications of pooling in comparative social policy and political economy, see Hicks (1994) and Beck and Katz (1995).
range of technical choices and pitfalls, rapidly changing fashions and fads, and contradictory advice. I shall not touch on these problems here. Rather, my concern echoes the conclusion of the most comprehensive “how to” guide, that pooling may be counter-productive “if thoughtful consideration is not given beforehand to the meaning of the aggregations in the pool” (Sayrs, 1989).

Most comparative researchers who do pooling have been motivated by the same agenda that inspires cross-section regressions, i.e. an interest in explaining enduring differences between countries. For these researchers, pooling is essentially a way of multiplying the number of cross-sectional “snapshots” at their disposal. They are willing to treat each “snapshot” as just one more view of the same between-country variability. As a result of the greatly increased number of observations in a pooled dataset, they hope to overcome the severe limitations on multivariate causal analysis that are imposed by the small number of countries in the universe of OECD states.

However, pooled datasets could also be useful to researchers interested in explaining cross-national differences in dynamic processes. For instance, two decades ago researchers (including this writer) became interested in comparing the causes of fluctuations in strike activity across countries (Hibbs, 1976; Shalev, 1979). Divergent results from running the same timeseries regressions for different countries were treated by some scholars simply as an antidote to exaggerated generalizations (Paldam, 1982). But others saw these divergences as exemplifying the predictable effect of contextual variations (Snyder, 1975). As I explain below, this has been the tack followed by the most thoughtful analysts of pooled datasets, Larry Griffin, Larry Isaac and their associates (Griffin, 1986; Griffin, 1989).

A second a priori difficulty for the pooled design is that the effect of a given independent variable may be quite different in timeseries and cross-section “because the underlying causal structures differ” (Firebaugh, 1980). Walter Korpi and I observed just such a divergence in our study of strikes. While temporal fluctuations in strikes followed an economic logic, with falling unemployment stimulating labor militancy, the cross-sectional variance followed a political-economic logic, with lower unemployment operating as a disincentive to strong labor

---

19 Recently, Beck and Katz (1995) discredited many of the leading studies in this genre, some by famous researchers with prodigious polimetric skills. Beck and Katz’s conclusion that the best approach in many cases is standard OLS with the lagged dependent variable included on the right-hand side flies in the face of the conventional wisdom. See, for example, Sayrs (1989).
movements to employ the strike weapon (Korpi, 1980). In this spirit, Alex Hicks (1994) has urged those who pool to undertake “systematic comparisons of cross-sectionally and longitudinally varying causal forces”.

Despite these reasons not to treat pooled data simply as more data, the multivariate analysis of pooled datasets all too often proceeds as if the data are driven by a single invariant set of determinants. But the practical consequences of this assumption are well-established. In the earliest and still the best exposition of pooling, Griffin et al. (1986) used annual data for 12 nations and 16 years to explore the effects of six economic and political variables on countries’ expenditure on income maintenance. Their first finding was that the bulk of the variation in most of their independent variables was concentrated in either the time or cross-country dimension. This alone suggests that it wouldn't have made sense to use a single model to explain both dimensions. And indeed, Griffin et al. found that “the average cross-national slopes and the average timeseries slopes… have very little in common” (p.116). Even within either the time or the space dimension, the contextuality of causal relations could not be ignored. The results of annual cross-sections proved to be “extraordinary unstable across years”, even contiguous years (p.111). While country-specific timeseries estimates were more stable, they nevertheless seemed to “evolve markedly different [national] processes underlying transfer outlays” (p.115).

Perhaps it is not necessarily the case that all pooled datasets are riddled with discontinuities as severe as those discovered by Griffin et al. But researchers who use multiple cross-sections simply in the hope of buying more degrees of freedom ought at least to preface their analysis by empirically confirming the uniformities across time and space which they need to assume.

———

20 In his own attempt to do this, Hicks (1994) utilized different periodicity (annual data versus 5-year averages) and methods (Parks/GLS versus “variance components”) for what he called “shortrun” and “medium-run” models. However if I understand his article correctly (and it may well be that I did not), Hicks did not actually compare cross-sectional and timeseries results. It seems that he also did not test for either cross-national or temporal parameter instability.

21 Researchers often contemplate (and the LSDV model apparently requires) the use of “dummy variables” to capture country or period effects. However, what the dummies actually capture are differences in the intercept or “baseline value” of the dependent variable. (As a consequence, they are liable to “steal” the effects of institutional variables like corporatism.) Interaction terms, far more costly in degrees of freedom, would be required to test country or period differences in slopes. A compromise that is more sensitive to context but less exhaustive of degrees of freedom, is to permit both intercepts and parameters to vary across generalized contexts. An innovative study by O’Connell (1994) has followed
think it makes more sense, however, to recognize that longrun differences between countries, and shortrun variation over time, pose different questions and invite potentially different explanations. That need not rule out focused and “responsible” uses of pooling. Even if our primary concern is to explain national variability, we may turn to pooled data to see whether patterns of causation alter over time, or alternatively, to isolate historical turning-points when cross-country differences first became “fixed” or were decidedly altered. For those mainly interested in explaining dynamic processes, on the other hand, pooling makes it possible to contemplate multiple explanations tailored to different contexts. The dynamics characteristic of a country (or group of countries) might be seen as both indicative of, and caused by, longrun (structural) differences.

Griffin and his colleagues proposed a systematic methodology for researchers interested in this kind of problem. They suggested that timeseries parameters be estimated in regressions for individual countries. Then, in a second round, these parameters would be treated as dependent variables to be explained cross-sectionally, by broad-brush differences between countries (Griffin, 1986). While the technique produced suggestive results (Griffin, 1989), the credibility of these results depends on our belief in the first-round timeseries estimates. Given the paucity of observations involved (over anything but a rather short period, we would have to suspect causal instability), it is hard not to be skeptical.

Recently Bruce Western (1996) has suggested how this limitation might be overcome. Western is substantively interested in explaining the dynamics of unemployment. He wishes to demonstrate that institutional factors like the presence or absence of corporatism can explain cross-country differences in the effects of fluctuations in variables like government composition and social expenditure. Assembling a pooled dataset for 18 OECD countries between 1964 and 1990, Western deliberately selects a mix of “dynamic” and “institutional” explanatory variables. He advocates a Bayesian approach to estimation that allows for possible contextual differences in causal dynamics, but differs in an important respect from Griffin’s two-stage method. Western’s technique permits estimates for individual countries to “borrow strength” from the whole sample. The implications of this are quite profound. It seemingly allows the analyst to take advantage of the more numerous observations and greater diversity afforded by pooled

this strategy by including interaction terms for “corporatist” versus “non-corporatist” countries and “crisis” versus “non-crisis” periods.
datasets, without having to assume identical causality for time and space. Pooling would then be freed of most of the objections I have raised and, as Western explains, the issue of whether comparativists ought to generalize within or beyond specific contexts would become a tractable empirical question rather than an epistemological conundrum.

Western’s success in this regard is best assessed by considering the results of his illustrative analysis of unemployment. Impressively, he is able to demonstrate corporatism’s implications in both the long and short-run. Over the long run, corporatist countries experience significantly lower rates of unemployment. At the same time, in corporatist countries the dynamic impact of wage growth on unemployment is also lower, confirming the claim that corporatism safeguards employment by improving the short-run tradeoff between wages and jobs. Once again, though, the credibility of statistical conclusions needs to be checked against the cases and in this instance, the outcome is not encouraging. Consider the timeseries coefficients for individual countries (Western’s Table 4). The Scandinavian countries (half of the corporatist group) exhibit positive effects of wage growth on unemployment; equally oddly, half of the countries with negative coefficients are not corporatist.

Western’s findings for the dynamic effects of shifts in government composition are even more puzzling. They appear to show that in corporatist countries and other settings where collective bargaining is widespread, increases in left party power cause unemployment to rise. To get to the bottom of this counter-intuitive result, Chart 2 reproduces Western’s estimates of the dynamic effect of changes in left cabinet representation for each country. To highlight possible institutional effects of the type Western is interested in, the countries have been grouped into several broadly similar contexts. (“Unregulated” labor markets are those in which no more than half of the workforce is covered by collective agreements. Note that by this criterion, all of the corporatist countries except Switzerland are “regulated”.)

At first sight, the chart seems to confirm the finding that in corporatist/regulated labor markets, “social democratic governments tend to raise unemployment” (Western, 1996). However, it is also apparent that without two outliers—Japan and Finland—this tendency would not amount to much. The problem is that in neither of these two countries are Western’s dynamic coefficients

22 Western also proposes using “fat-tailed” distributions to improve accuracy by including outliers but downplaying their effects. I am concerned here only with his use of hierarchical modeling to estimate contextual effects on timeseries parameters.

23 More precisely, 3 of the 4 Scandinavian countries have positive coefficients, the other (Finland) is zero.
Chart 2: Western's Hierarchical Model of Unemployment

Dynamic Left Effect
plausible. Finland experienced few significant shifts in the left’s overall role in government (it had one of the lowest coefficients of variation, only .23). What did vary in Finland was the relative role of the communist and socialist parties, but of course Western’s measure was oblivious to that. Japan is an even worse candidate to exemplify trends, since in the relevant period its left party representation was an unvarying zero.24

I am inclined to conclude that the dream of simultaneously harnessing the wealth of information in pooled datasets while respecting and even exploiting the difference between synchronic and diachronic causation remains unattainable. It will be recalled that the mechanism which was supposed to make reconcile these two objectives was described as “borrowing strength”. In Western’s (1996) words, “Information from other countries will help provide an estimate for a coefficient in a particular country where, say, a given independent variable shows no variation.” This procedure would be just another variation on regression’s standard theme of inferring causal relationships into non-existent data, but for one thing. In Western’s hierarchical model, non-existent dynamic linkages become the raw material for real tests of institutional effects. It is hardly surprising that an analysis which relies on imaginary data for its input risks generating implausible output.

This is a pity, because Western’s strategic premise was inviting. It would make our lives easier if the issue of whether causation is contextual (proper names are indispensable) or general (proper names surrender to variable names) could be settled on empirical grounds. But of course, such an undertaking predisposes that our concepts can be operationalized in ways that permit generalization. Sufficient variability is one requirement. To again belabor the obvious, if there never has been and may never be any socialists in the Senate, then socialist representation in government is of little use an explanandum of fluctuations in unemployment or social expenditure in America. Even given enough variability, however, as Przeworski and Teune emphasized we need good (i.e. comparable) measures. Left party share of cabinet seats is typical of the problems this poses for comparativists. On the one hand, it is uniformly and fairly easily measurable across countries. On the other, a given percentage-point shift in left share from one year (or sub-period) to another may mean something quite different, depending

24 Western (1996) concedes that the left government variable for Japan was constant and counsels against “substantive interpretation” of the Japanese result. The problem is that whatever he or I may make of this result, the Bayesian hierarchical model necessarily relies on it to establish contextual influences on social democratic dynamics.
on the country and period. These differences may be perfectly amenable to conversion to multiple generalized concepts, but in practice that will lead us back to the degrees of freedom problem. Once again we see that done properly, pooling cannot escape this problem.

4. Testing the “regime” approach

Gøsta Esping-Andersen’s *Three Worlds of Welfare Capitalism* raises a poignant question for comparativists who, like him, are committed to (1) recognizing that there may be striking causal discontinuities across different contexts; (2) informing their hypotheses about relationships between variables by drawing on their knowledge of cases; and (3) using quantitative indicators to systematically test propositions across the entire universe of cases. Like the first two studies reviewed in this paper, Esping-Andersen adopted a two-stage approach. He developed indices of “universalism”, “decommodification”, and “stratification” and used tables to show that his 18 OECD countries tended to fall into three distinct subgroups (Esping-Andersen, 1990: Tables 2.1, 3.3, 4.3). He then utilized MR to perform a causal analysis of cross-country variation in more than a dozen indicators, some raw and some aggregated, which were regressed on political variables and in some cases also demographic and other control variables (Esping-Andersen, 1990). I believe that Esping-Andersen’s first technique was unnecessarily “soft”, while the second explicitly conflicts with his epistemological stance. After explaining these claims, I propose an alternative analytical strategy.

Esping-Andersen’s tabular analysis relied very heavily on the author’s judgement—both in the construction of aggregate indices and the identification of country clusters. No systematic test was carried out of whether the ensemble of indicators of welfare state regimes actually do “hang together”; and if they do, whether countries actually do cluster in three distinct subgroups on specific policy dimension(s). It would have been a logical step to subject these claims to techniques like factor analysis, cluster analysis, or multidimensional scaling.

As noted, Esping-Andersen’s turn to MR was intended to provide support for his causal claims. One way he tried to do this was by demonstrating that his preferred (political) variables garner stronger empirical support than rival (e.g. demographic) explanatory variables. This exercise had limited credibility though, given that it entailed regressions with 5 or 6 “independent” variables across only 18 cases. The larger purpose of the regression analysis was to offer support for hypotheses regarding the political antecedents of the policy regimes. But this approach imposed an obvious mismatch between Esping-Andersen’s claims and his methods. The book’s key causal argument is that *countries cluster on policy because they cluster on*
The regression approach, however, treats both policy and politics as continuous variables—not as a limited number of qualitatively different configurations with distinctive historical roots.

It is hard to exaggerate the fundamental incompatibility between MR and Esping-Andersen’s regime approach. It is difficult to imagine him claiming that, say, any discrete increment of Catholicism or absolutism ought to yield a discrete and uniform increment in pension “corporativism”. Only countries that are predominantly Catholic and/or have an absolutist past were expected to exhibit the corporativist policy profile. By the same token, he also would not claim that the social policy of any given country may be understood precisely as the combined effect of Catholicism, absolutism, and working class mobilization. (As in, “to make a loaf of bread combine one part yeast, 2 parts water, and 10 parts flour…”) On the contrary, his purpose was to show us how socialist, Catholic-Conservative and liberal political worlds have generated different worlds of welfare. In his hands, MR was simply a blunt instrument for tapping gross differences between groups of countries that could have been conveyed by the use of tables and charts without the implication of constant linear effects across countries.

How might Esping-Andersen have gained the benefit of his quantitative data without falling back on the conventional statistical paradigm which the spirit of his analysis (and his explicit criticisms of earlier work) flagrantly contradicted? Three earlier studies have offered interesting suggestions. Ragin (1994) carried out an elaborate QCA analysis using 7 different independent variables. Ragin limited himself to pension policy as the dependent variable and performed separate analyses for indicators of each regime type. Kangas (1994) compared the performance of QCA and cluster analysis with traditional regression techniques for testing a simplified political model of the “quality” of sickness insurance. A third study, by Castles and Mitchell (1992), used simple tabular data to build an alternative typology of (four) overall worlds of welfare capitalism.

As we can see, researchers have been divided on two points, how to define the dependent variable and what method of analysis to use. Regarding the first of these issues, both the Ragin and Kangas studies chose (for reasons of their own) to focus on limited aspects of the social policy regime. However, one of Esping-Andersen’s most important and original claims was that social policy regimes are *multidimensional*: they are defined by both the size of the social wage and the quality of social rights; they concern the state’s commitment to employment, not just income maintenance; and they depend not only on the activities of the state but the balance between public and private provision. Castles and Mitchell shared Esping-Andersen’s quest for
a comprehensive classification scheme, but challenged him on what it should encompass. In their study Esping-Andersen’s focus on the decommodifying aspect of social rights was replaced by consideration of the distributional consequences of the social wage. Castles and Mitchell took the important step of adding taxation to the regime mix, but their analysis left out the division between public and private provision and the state’s commitment to employment.

The three studies also diverged from a methodological viewpoint, beginning with how they measured the dependent variable. Castles and Mitchell followed Esping-Andersen’s lead and presented simplified tabular data in support of their typology of policy regimes. In contrast, both Ragin and Kangas utilized cluster analysis to assign countries to regimes. These efforts were innovative but problematic. Each country was forced to fit into one regime, thereby predetermining an issue in need of empirical exploration. Serious practical problems arose. Kangas had difficulty finding the “liberal” countries and Ragin was placed in the awkward position of having to assign one third of his countries to a “spare” category which automatically excluded them from his causal analysis (QCA). Neither researcher capitalized on the potential strength of clustering and other multivariate techniques as mechanisms of data reduction. By sidestepping Esping-Andersen’s contention that welfare state regimes are multidimensional, Kangas and Ragin forfeited the opportunity to test that contention by searching for the different dimensions.

Recall that Esping-Andersen identified four different dimensions of welfare states: social spending, social rights, the public/private division, and commitment to employment. He assembled a multiplicity of indicators on all four dimensions and claimed that these indicators associate in characteristic patterns, forming three distinctive policy constellations. I will call these the institutional (universalistic, decommodifying), residual (minimalist) and corporativist (status-preserving) regimes. The second leg of Esping-Andersen’s theory posits a causal link between each regime and the political milieu that is said to have spawned it: respectively, socialist, liberal and Catholic-conservative. I propose to test these two arguments directly. Do

\[\text{\footnotesize 25 Castles and Mitchell pointed out that Esping-Andersen’s measures of decommodification and universalism had as it were “penalized” Australia and New Zealand for their reliance on means-testing. Their counter-argument was that in practice means-testing excludes only the advantaged, and that in combination with reliance on progressive taxes for the financing of benefits, it represents an egalitarian thrust that Esping-Andersen’s measures for the Antipodean cases could not capture. In deference to this criticism of Esping-Andersen’s treatment of means-testing, in my own analysis I refrain from utilizing his indices of decommodification or his scoring of universalism in unemployment and sickness programs.}\]
policy indicators cluster along the lines of the three ideal-typical regimes? If so, is there evidence of the expected association between policy regime and political milieu?

The clustering of policy indicators might be investigated by a variety of methods. I experimented with both cluster analysis and multi-dimensional scaling to attain groupings of countries. The results were broadly consistent with Esping-Andersen’s expectations, but in some isolated cases countries received odd and even capricious assignments. This is the price of trying to fit the cases into mutually exclusive subgroups, when the conceptualization of policy regimes is actually (and properly) ideal-typical. I therefore switched strategies, seeking to link indicators instead of cases, and to identify underlying policy dimensions rather than groups of countries. I reasoned that analytically, Esping-Andersen’s triplet of regimes rested on two dimensions of policy—one representing Titmuss’ (1974) dichotomy of residual versus institutional principles, and the other capturing the fragmented, hierarchical and status-preserving measures that are anathema to both socialist and bourgeois forces.

If Esping-Andersen is right about there being three ideal-typical worlds, we should be able to parsimoniously characterize real existing welfare states in terms of these two dimensions of institutionalism and corporativism. This characterization should be particularly informative for “mixed” cases, rather than making us wish they would go away.

My reanalysis of 13 of Esping-Andersen’s policy indicators provided striking confirmation of his view of welfare state regimes. Multidimensional scaling (not reproduced here) located all of the

---

26 It was precisely the innovation of *Three Worlds* that it recognized that not all welfare state variation can be described by the polarity between worker-friendly “institutional” policies and the “residual” welfare state favored by economic liberals on the right. Three-way comparison between Austria, Germany and Sweden revealed that the impact of labor movement strength was offset in Austria by a strong right, Catholic influence, and a history of attempts by conservative regimes to use social policy to divide the working class (Esping-Andersen, 1984) That is why Austria’s welfare state looked more like the German than the Swedish variant. This insight was foreshadowed by references in earlier work to the influence of Catholicism and the unity of the political right (Castles, 1978; Korpi, 1978; Stephens, 1979). What Esping-Andersen did was to crystallize the “corporatist” policy dimension by concretely locating it in a specific European milieu.

27 The data were obtained as follows, references are to Esping-Andersen (1990). Social wage (Table 5.1, source data from the author); universalism (Table 3.1, data disaggregated by program from the author); decommodification (Table 2.1); uniformity (Table 3.1); “poor relief” (Table 3.1); the public-private division in health (Table 3.1) and pensions (Table 4.3); “full-employment performance” (Table 5.9, data from the author). Active manpower program expenditures relative to GDP (c. 1975) and public employment as a
indicators along the perimeter of a sphere (a “simplex”), indicating that they belong to a single underlying spectrum. This appears to strengthen Esping-Andersen’s case for a broad conceptualization of social policy. The distribution of specific indicators followed the hypothesized two dimensions as expected. Since a “clean” solution could be found using a method which imposes few restrictions on the spatial distribution of the objects and does not require a “high” (interval) level of measurement, there were grounds for going one step further and performing factor analysis. Factors are economical linear combinations of variables. They are generated in such a way that there is strong correlation between the variables with the highest “loadings” on a given factor, but minimal correlation between factors (ideally they are “orthogonal”).

The results of an unrotated principal component analysis are reported in Chart 3. The fact that the first two factors together accounted for the majority (nearly 60%) of the variance is itself good news for Esping-Andersen’s hypothesis. The first factor, running from East to West in the chart, evidently captures his residual/institutional dimension. It exhibits high positive loadings on public employment, active labor market expenditure, benefit equality and social security spending; and strong negative loadings on poor relief and indicators of the scope of private health and pension provision. The second (North-South) factor signifies the corporativist dimension of policy: high positive loadings on the number of pension schemes and the prominence of civil service pensions, and a high negative loading on the role of “citizen pensions” (social security). The factors are not completely orthogonal, but the areas of overlap are intelligible. For instance, the results confirm that both corporatist and institutional regimes are repelled by occupational pensions. They also imply that employment performance (low unemployment and high job creation) was strong for the institutional regime but weak for the corporatist one.

We are now ready for the “causal” phase of the analysis, reported in Chart 4. When the 18 nations in Esping-Andersen’s study are arrayed in accordance with their scores on our two factors, the evident linkage between contexts and policies generates an illuminating cross-

---

27 Thus the researcher hopes that each item will load high on only one of the factors. The procedure known as factor “rotation” is designed to force this to happen, but I opted for the more difficult test of an unrotated analysis.
national mapping. While Esping-Andersen sometimes blurred the distinction between political contexts and policy regimes, his text makes it clear that he conceived the three contexts as follows:

- **Socialist**: The Scandinavian social democracies, characterized by levels of working class mobilization almost without peer in other Western nations.
- **Catholic-Conservative**: European nations—Italy, France, Belgium, Austria and Ireland—which share an absolutist past, relatively late-blooming democracy, and a largely Catholic population.
- **Liberal**: The USA, Canada, Switzerland and Japan—in which working class mobilization is very weak and the conservative heritage is largely or wholly absent.

The remaining five countries in Esping-Andersen’s study are more difficult to classify. Their state traditions are either close to the conservative group (Germany and the Netherlands) or liberal in character (the UK, Australia and New Zealand), but they have also experienced moderate levels of working class mobilization.

The fit between the political clusters and the factor scores is striking, albeit imperfect. The liberal states and Australia have the most negative “institutionalism” scores, while the Scandinavian states and New Zealand have the highest positive scores. The remaining countries, which have intermediate scores for institutionalism, divide neatly between conservative states with high scores on the corporativism factor, and three mixed cases that score close to zero on both factors and are effectively outside the scope of the theory. The chart also reveals a difference between the non-corporatist countries: some are indifferent to corporativism, while others (the bastions of social democracy and some liberal states) are “repelled” by it. Here too lies an interesting challenge for further development of Esping-Andersen’s model.

5. Conclusion

Despite considerable methodological debate and innovation among comparativists in recent years, MR remains by far the predominant mode of numerical data analysis and most its critics see qualitative analysis (whether formal or conventional) as the only real alternative. This paper

---

29 Note that with one exception (Finland’s assignment to the social-democratic group), the classification I inferred from Esping-Andersen’s discursive analysis is perfectly reproduced by clustering countries on his three political indicators: working class mobilization, absolutism/late-democratization, and Catholicism.
rejects both of these tendencies. The emergence and gradual diffusion of QCA is certainly a major development for comparativists. At its most basic, QCA is a means of rendering explicit the kind of "causal pathways" arguments typical of classical comparative-historical research in the genre of Weber, Moore, Rokkan and Skocpol. This desire to systematize the evidence for such arguments is not new (Somers, 1971). But Ragin (1987) is the first to have offered a formal procedure for doing so, by using Boolean algebra to make a parsimonious identification of the regularities underlying a series of case configurations.

QCA is certainly not a “qualitative” method in the sense of relying on the interpretive skills of analysts wading knee-deep in thick description. If anything, as Griffin and Ragin (1994) have insisted, QCA is more like MR: both apply rules that are independent of the researcher, and both treat cases as "discrete, multiple instances of more general phenomena". QCA has great advantages in principle because of its fidelity to principles of case-oriented analysis. But in contrast to MR, QCA is restricted to categorical (usually binary) data and it uses logical rather than statistical methods to generalize from the cases. Even advocates of QCA concede that the empirical claims that it yields must be treated with caution. In particular, because the method seeks out single or combined causes that are logically necessary and/or sufficient, its results are vulnerable to both measurement error (including erroneous determinations of presence/absence) and specification error (there is no QCA equivalent to the error term). Some of these limitations were evident in my discussion of implementations of QCA in the literature on welfare state regimes.

What QCA embodies which is especially valuable in the context of small-n macro-comparisons is visibility of, intimacy with, and dialog with the cases—precisely the advantages that MR lacks. But it is questionable whether, for those unwilling to give up quantification altogether, these advantages may only be had using QCA. My reanalysis of four MR-based studies suggests a number of alternatives to both QCA and MR. In closing, I incorporate these suggestions into a summary statement of the three major options open to quantos who are aware of the limitations of MR.

1. Refinement. This is the "optimistic" approach best represented in the present survey by Bruce Western's approach to pooling. Western is a good representative of the refinement strategy because pooling holds such potential promise for comparativists, and because he is a member of a new generation of political economists whose statistical expertise is incomparably superior to nearly all members of this author's generation. The problem of intelligently disaggregating cross-sectional and dynamic causal effects is only one of the
numerous issues in MR analysis for which sociologists and political scientists have been seeking inspiration from their technically more advanced counterparts in economics and statistics. But the limitations and tradeoffs discovered from a close reading of Western's solution to the problem of causal heterogeneity offer some grounds for pessimism. This impression is powerfully reinforced by a recent review of controversies among econometricians and their counterparts in political science by Gregory S. Maddala, one of the acknowledged doyens of the profession. Maddala (1997) makes it politely but unmistakably clear that he regards many of the current fads in the practice of high-end MR by political scientists as the result of either mistaken or misguided emulation of econometric fads.

2. **Triangulation.** John Stephens and his colleagues have gone furthest in rationalizing and implementing the use of multiple analytical techniques. They argue that both non-formal comparative histories of limited numbers of cases, and broad-spectrum quantitative comparisons using MR, have costs and benefits that are essentially complementary. In practice, they adopt a clear epistemological hierarchy in which the results of MR are confronted by theory and knowledge of cases and the resulting causal anomalies are put to the test of historical process-tracing. Few researchers, especially the vast majority who work alone or in limited collaboration with their colleagues, are likely to emulate this ambitious *modus operandum*. Triangulation more typically boils down to utilizing a variety of different techniques to analyze the same data or problem and seeing if they generate similar results. The study by Kangas cited in section 4, and a more recent work by Ebbinghaus and Visser (forthcoming) on comparative union membership trends, have combined MR with QCA. As with Hall and Franzese's combination of single-case and tabular analysis with MR, this kind of hybrid analysis is a welcome innovation, especially if it results in a genuine dialog between cases and generalizations. While such a dialog is facilitated by both QCA and tabular methods, the restrictive policies of journal editors and other factors can easily limit this potential in practice.

3. **Substitutes.** The third and most radical strategy for dealing with the problems of MR without abandoning quantitative analysis is to opt for alternative techniques altogether. Sections 2 and 3 of this paper presented some simple tables and flowcharts in which countries were identified and numeric data collapsed or averaged. It was shown that these techniques can effectively overcome some of the most unattractive limitations of MR while incorporating key elements of the case-oriented approach. On the one hand, they were able to convey
complex analytical ideas (i.e. syndromes and hierarchies of causes). On the other, they drew attention to cases deserving of additional more focussed comparative scrutiny—something that no other method could claim. I have also suggested that, provided that they offer a better fit with our ontological assumptions, there is no reason why multivariate statistical methods other than regression should not be exploited by comparativists. The utility of factor analysis in clarifying the evidence for Esping-Andersen's approach to welfare state diversity is the illustration offered here, but cluster analysis, multidimensional scaling, and the methods of data reduction developed by Dirk Berg-Schlosser and his collaborators (e.g. Berg-Schlosser and De Meur, 1994) are other mechanisms by which the traditional handicap of small-\(n\) and many-\textit{indepvars}\ might be turned into an asset.
Bibliography


