Crisis States Research Centre

Crisis States Research Centre Report

MEASURING POOR STATE PERFORMANCE:
PROBLEMS, PERSPECTIVES AND PATHS AHEAD

Francisco Gutiérrez
with
Diana Buitrago, Andrea González, Camila Lozano
Instituto de Estudios Políticos y Relaciones Internacionales

LSE The London School of Economics and Political Science

UKaid from the Department for International Development

First published 2011 by the
Crisis States Research Centre
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
csp@lse.ac.uk

ISBN: 978-0-85328-460-4

A catalogue record for this report is available from the British Library.

This document is an output from a project funded by UKaid from the UK Department for International Development (DFID) for the benefit of developing countries. The views expressed are not necessarily those of DFID.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BTI:	Bertelsmann Transformation Index
CIFP:	Country Indicators for Foreign Policy
CPIA:	Country Policy and Institutional Assessment
CSP2 :	Crisis State Programme Phase 2
FSI:	Failed States Index
IAG:	Index of African Governance
ISW:	Index of State Weakness
MAT:	The Monopoly-Administration-Territory database
OECD:	Organisation for Economic Cooperation and Development
PII:	Political Instability Index
PSPI:	Poor State Performance Index
STF:	State Fragility Index

ACKNOWLEDGEMENTS

The materials that we circulate, including this report, are the results of an interdisciplinary undertaking that would have not been possible had it not been for the generous and permanent support of the Crisis States Programme and its director, James Putzel. We are also grateful to the CSRC team and to all participants in the Measuring Poor State Performance workshop at LSE in May 2010. We benefited from the great contributions of Mauricio Velasquez, Liliana Narvaez and Sandra Ropero, among others.

Interdisciplinarity is a trade mark of the Instituto de Estudios Politicos y Relaciones Internacionales (IEPRI) at the Universidad Nacional de Colombia, within which this project was developed. Our interaction with Camilo Argoty from the School of Mathematics of the Universidad Sergio Arboleda, was fundamental. Chapter 4 by Francisco Gutiérrez and Camilo Argoty underlies the whole argument here.

FOREWORD

We present in this text results of our investigation of poor state performance indexes (PSPIs), developed within the second phase of the Crisis States Programme. This investigation consists of a 'critical moment' (the analysis of current practices and identification of potential weak and strong points) and a 'constructive moment' (the proposal of potential solutions and of an agenda of research on open questions).

The analysis and evaluation of PSPIs is a small but active field. We have greatly benefited from the overviews provided by Fabra and Ziaja (2009), OECD/DAC (2007), Cammack et al. (2006) and Rice and Patrick (2008), among others. In recent years, innovations have been put forth that basically move in the same direction that we advocate here (see section 3.1). Some of the co-authors have also previously published critical analyses of PSPIs (Gutiérrez and González 2009; Gutiérrez 2009). Here, we hope to have covered all the basic references. Any oversight is completely involuntary.

The core claims of this report are the following:

a) Building social indicators is necessary but difficult. PSPIs are a particular type of social indicator that face especially hard-to-solve predicaments.

b) In their present form, the majority of PSPIs are basically unsound. A substantial number of the rankings and scores that they produce are an artefact of ad hoc decisions that have no substantive justification. In crucial instances they adopt extremely anti-intuitive assumptions. Extant PSPIs have not solved or even acknowledged several of the key problems they face.

c) Some of these problems can be tackled. However, any improvement will necessarily be partial and can only really be considered as an improvement in discursive terms. No such thing as a bullet-proof indicator exists.

Our claims are intimately interrelated. For example, the toolkit of solutions we put forth stems directly from the problems that we identify during the 'critical' phase.

Apart from the operational assumptions, which will be made explicit at the appropriate moment, we lean on two macro-assumptions. The first is that there should be much more systematic interaction between qualitative and quantitative research. This is the rich 'middle path' that Ragin (2008) identified in his seminal reflections. The 'middle path', says Ragin, is *not* 'a compromise' between the quantitative and the qualitative, or the imitation of the quantitative by the qualitative (as King et al. (1994) claim). It is an approach that ideally allows for the redesign of both. Today, this proposal, though as yet unaccomplished, is not as controversial as it was in 1994, except perhaps to hardnosed number crunchers.

Our second assumption is more problematic: that partial corrections and improvements are better than none. It is very easy to find good counter-examples to such a principle. If you live under the rule of a repressive and inept state, you would probably not want it to become efficient. In such a situation, a partial improvement ends in a disaster; inefficiency was a blessing. More generally, in many instances of optimisation, getting close to a specific local optimum may actually take you away from the wider global optimum. Nonetheless, we believe that in the discourse of PSPIs, the partial improvements that we are proposing do not imply a drift away from the global optimum. We flag this assumption because we are focusing on a limited subset of issues, where we believe we can suggest potential improvements. For example, we do not delve into the issue of imperfect counts coming from convenience samples. For excellent treatments of these themes, see work by Ball (1996) and Freedman (2005; 2010).

The report is divided into three chapters. The first is a reflection on the meaning and possible interpretation of a particular group of indexes that we call 'third wave indexes', which include PSPIs. We first show that new indicators that measure complex and multidimensional phenomena by capturing heterogeneous data are constantly being created. In a second section we set out a reminder that some of the criticisms levelled against indexes and quantification are untenable or simply irrelevant. Third, we discuss the problem of order in genuinely multidimensional indexes and databases, and the implications that this holds for the study of aggregation functions. In this we also refer to a quite simple but relevant order result developed elsewhere (see Gutiérrez and Argoty 2010, summarised in chapter 4 here). This shows that a significant number of the rankings produced by an aggregation function that translates a multidimensional database into real numbers basically have no substantive meaning if the choice of the function is unrestricted.[1] Furthermore, the number of reversals in the number of variables of the database grows quickly.[2] After considering normalisation and intrinsic ambiguity, we then concentrate on the evaluation of indicators from the perspectives of validity, reliability and sensitivity (Nardo et al. 2005). We conclude by reflecting on the fact that third wave indexes such as PSPIs face special and relatively new problems.

Chapter 2 is dedicated to the illustration of these problems through the analysis of concrete PSPIs. Since some of the co-authors have done this before, we here provide just a brief illustration. We discuss issues related to validity and reliability (2.1), the capture of information (2.2) ambiguity (2.3), and normalisation and aggregation functions (2.4).

In Chapter 3 we present some of the new approaches that we are proposing.We have no naïve illusion of absolute novelty,

1 For the informal version of the argument, see Gutiérrez (2009).

2 This is somewhat inexact. In reality, it grows in the number of dimensions, that is, the number of linearly independent variables, which is always less than the number of variables.

which in itself should be considered suspect.[3] The first section of the chapter focuses on theoretical background, intellectual precedents and previous improvements in the PSPI domain. Our formal theoretical background comes from fuzzy set theory, and more generally from what is presently called 'soft' (or 'granular') computing and/or 'approximate reasoning' (see, for example, Di Nola and Gerla 2001). As our main intellectual source of inspiration within the social sciences, we recognise Charles Ragin's effort to introduce fuzzy set theory into social analysis in general, and into comparative politics in particular. We pinpoint at least three cases of distinct waves in the PSPI domain. Then we present our new database, MAT (Monopoly of violence, Administration and Territorial reach), and the logic underpinning it. MAT draws heavily on ideas discussed during the Crisis States Research Centre's second phase of research, though naturally we bear the whole responsibility for errors committed.

The next section focuses on our aggregation functions. We present three tools:

a) A fuzzy compensatory method (ie with weights attributed to the variables), but one which:
- **a.** Grounds the imputation of weights to variables onto the data;
- **b.** Does not allow for full compensation between variables (as will be seen, this is a very important point);
- **c.** Takes non-orthogonality into account (for example, the correlation between variables and bundles of variables, which from now on we will call 'boxes');
- **d.** Permits researchers to give a reasonable interpretation of the weights.

b) A non-compensatory (ie without weighting the variables), hierarchical method ('downsets'). According to the reasoning of the first and second chapters, non-compensatory aggregations should be strongly preferred in the context of PSPIs. However, downsets violate an important condition explained in section 1.3.: the independence of irrelevant alternatives. Nevertheless, we show experimentally that the violation may be marginal.

c) A non-compensatory method that produces a hierarchy and a different representation of the countries: not by numbers but by intervals. This interval representation respects all of the desirable characteristics that we would like to impute to an aggregation method, at the cost of sacrificing the assumption of total order. This, we argue, is not a significant cost.

Some researchers have suggested that the multidimensional representation of statehood, or of state fragility, is possible and convenient (Carment et al, 2006; Fabra and Ziaja 2009).

However, the step from this important suggestion to the fully fledged operationalisation of non-numerical objects has not been taken until now. What we do here is to present the tools that enable us to operate with intervals informally (ie without sorting out the mathematical apparatus, and simply narrating what they do). We compare them, add them, subtract them, and indeed regress over them. So, in the following section we present the fuzzy operations that we perform both on numerical and on interval data. We have programmed one regression as used by Hojati et al. (2005), and we have developed another one tailored to operate on intervals. Our regressions can be made more sophisticated, to become crisp regressions, and to take into account several types of variation. For example, they can perform multilevel operations so that they vary across time and countries.

We did not develop interval fuzzy regressions because they seemed particularly groovy (though, frankly, they are). Following Freedman's advice (2010), we have oriented ourselves around the demands of the type of data we are dealing with and not by the esoteric nature of the formal modelling. It was because the best representation of the data seemed to be by intervals – indeed, we claim this representation has a lot of advantages – that we developed the formal tools in this way. Fuzzy regressions are distribution-free (ie no subjacent probabilistic distribution has to be assumed).[4] In general, fuzzy and 'approximate' data management, and interval regressions in particular, have limitations, as does any formal tool. Yet, in the context of dealing with noisy, ambiguous and hard-to-interpret data, they have too many advantages and 'natural' interpretations to be ignored.

In conclusion, we make an inventory of the value added by this work, and of pending questions.

The style used throughout the text is consciously, openly and unabashedly informal and non-technical. We try by all means to treat in the plainest way possible themes that have many potential technical intricacies. Something is unavoidably lost in the process. This report does not replace more technically oriented presentations, and some of these are circulated simultaneously with this report. At the same time, we believe that it is convenient to have:

a) An overview of our results;

b) An exposition of our main queries;

c) A more general framework to reflect on PSPIs.

3 Many good critical evaluations precede this one. See, for example, Cammack et al. (2006) and Fabra and Ziaja (2009). We will refer to these throughout.

4 Inevitably, this has costs and these are, this time, quite real.

1. INDEXES EVERYWHERE

1.1 OMNIPRESENCE OF INDEXES

Apparently, we live in a world obsessed by indexes and rankings. Musicians, songs, politicians, writers, books (see, for example, www.amazon.com), cars, professional sports people: all are graded and ranked. We run into indexes everywhere. One of the groups of people who grumble most about this obsession is academics, but actually they fully contribute to this. Their lives are structured by ranks and indexes, including:

a) Exams. These are a typical index with certain characteristics. They have a critical cut-off point, below which the student fails and above which the student passes (Ingenkamp 1997). Even more typically, today we also have a wealth of 'meta-ranks': tests that exams should pass to be considered acceptable.

b) Journals, which are marked according to sets of criteria that are not always homogeneous (including, for example, the evaluation process, number of times they are cited, readership etc.). Authors are granted incentives by their institution to publish in journals that are more highly ranked. Journals are ranked according to their readership, the number of times their articles are cited, difficulty in getting published in them and so on.

c) Evaluation by students. In many universities, teachers are marked by their students.

d) Seniority, depending on performance. In almost all universities, staff are paid according to a rank-based system of incentives, which can depend on the amount of money brought to the institution, publications and teaching quality. The formula to establish seniority can become extremely complicated (see, for example, the cases of 'Rule 12 seniority, layoff, and resignation' and 'Starting salary and progression through seniority points' (Podgursky and Springer 2007)).

e) Universities themselves. Of late, universities are globally ranked as institutions. These ranks depend on aggregated or accumulated patrimony. The comparison of all the universities of the world according to a certain set of criteria has proliferated.[5] For example, the Academic Ranking of World Universities marks universities in general but also according to areas of knowledge (mathematics, computer science, social sciences etc.) (ARWU 2010).

It may not be an exaggeration to claim that academics are tinkering with indexes and rankings on a daily basis. Typically, almost all of these – at least those in which the teacher is rated, not necessarily those in which the speaker does the rating – are generally considered obnoxious. They are perceived as spuriously exact and flawed in some deep sense, and with good reason, as will be claimed throughout this text.

The craze for ranks is not limited to the academic world. Competitive sports are actually driven by indexes. They are inconceivable without measuring and without building hierarchies of performance. This is the reason why sport professionals have very sharp insights into measuring (see, for example, the South African Airways ATP Rankings, the Elo rating system. and the FIFA/Coca-Cola World Ranking). Something similar can be said about popular music. For musicians and entrepreneurs it is essential to identify potential hits. Consumers are also better off if they can convey their preferences to producers. Yet the world's passionate interest in hierarchical classification goes well beyond even this. There is a webpage entirely dedicated to rankings, which hosts evaluations, based on the votes of visitors to the webpage, such as 'the most overrated band ever', 'most hated bugs', 'worst politician ever' and the 'country most likely to surrender quickly' (Rankopedia 2010). All of this sounds rather flippant, but in fact Rankopedia is a fabulous resource that in some ways illustrates quite well why ranks are so attractive and, at the same, so problematic.

Of more interest for this report is that the state as a regulator of social life is a major index producer. One of its social functions is in fact to generate indexes, which allow decision makers – whether or not they belong to the state – to position themselves in the world. Examples of indexes and rankings that have played a decisive role in state-building processes include those developed:

a) When progressive tax was instituted. It was necessary to categorise people into levels of property and income (Blank and Blinder 1985).

b) On ethnicity, where 'superior' and 'inferior' races have been treated differentially by the state (MacLaughlin 2001; Bell and Freeman 1974). For some states, ethnicity is still a very important classificatory criterion in efforts to measure performance in overcoming ethnic discrimination.

c) When states that enforced conscription had to decide about the target population, based on criteria of gender, age and citizenship (Aleinikoff and Klusmeyer 2002). There were many boundary cases. The way in which these were solved was frequently highly consequential for the social and political incorporation of broad sectors of the population (Haggard and Kaufman 1995).

d) For bids and tenders managed by the state. For many countries this is legally mandatory.

e) For policy making. Some kind of index construction is a de facto precondition of policy making, which is based upon the capacity of the state to identify target populations in social and physical space, and in time. For example, a housing programme has to establish forms of marking and/or ranking the eligibility of the aspirants based on income and a series of socio-demographic characteristics. Another example: special subsidies for internally displaced persons (IDPs) depend upon

5 Not all appear in the rankings, but these are designed in such a way that all could.

the capacity of the state to establish diverse forms of identification to avoid false positives (people who are not IDPs but want access to the subsidies) and false negatives (people who are IDPs but are denied access).

It is impossible to describe adequately what the modern state is, and pretends to be, without understanding in detail this relentless day-to-day activity of capturing and retrieving information. States consume information insatiably and their ever expanding regulatory functions demand that they constantly refine and broaden their informational mechanisms.

The reader may have noted that we have offered examples of indexes both in the academic and in the policy domains that have a long and venerable trajectory, and others that are rather new. Exams were institutionalised many years ago and have been a matter of interest for pedagogues and, more recently and typically, for specialists who want to understand the workings of expert opinion and ranks. We will come back to this theme several times. On the other hand, formal global rankings of universities according to their research and pedagogic excellence are quite new, although informal rankings have existed on this subject for a long time. The Shanghai rank, for example, was created in 2003 and is already a reference point for hundreds of decision makers around the world. The same can be predicated about regime and state indicators. GDP measures started to be produced in the developed world in the 1940s and gradually spread to the periphery. Homicide counts preceded this, and already provided a good lead about how the state should be defined. This categorisation of information was further refined when the state had the incentive and the means to do so. Homicides became classified by author, by victim and by place.

In the last two or three decades, a whole new family of rankings has been developed. These were new in two ways. First, they were not produced by the state or by one of the traditional rank producers (such as sports-regulating agencies) but by transnational associations that had a voice and wanted to be heard by a global community. Second, they depended heavily upon computers and the internet in their production and use. These:

a) Permitted states, NGOs, universities and researchers to store and retrieve enormous masses of information very efficiently.

b) Made statistics available to an army of analysts, bureaucrats and intellectuals. To produce what we consider today a relatively simple regression in the 1950s involved the intense use of high cognitive capabilities and was flagged as an operation of particular merit (Hald 2003). Democratising number crunching changed both the social sciences and policy making.

c) Allowed for massive queries and information retrieval on the web, so that in situ information gathering is no longer a prerequisite for data manipulation or, in particular, for index building.

This technological advance went hand in hand with two major political changes:

1. States no longer held the monopoly in the international arena. New global agencies – multilateral, human rights, pro-democracy – flourished. States ceded part of their sovereignty.

2. In parallel with this, new global audiences were created. The obvious example is human rights supporters, who believe that there is a global political community to which all human beings belong. This may sound as if it is simple common sense, but it should not be taken for granted (Finnemore 2003). The proliferation of special audiences goes way beyond the obvious examples, though we acknowledge the importance these hold. Some are fairly specialised. Take risk-investment rating groups, for example: that of Standard and Poor's has enormous power because the marks it attributes to each country may have a significant impact on that country's economy (Standard and Poor's 2010).

In short, technological change plus the globalisation of capital and politics have led to a new wave of indexes. They are not created or administrated by the state, nor by the old index-oriented interest groups such as sports associations. They are interwoven with global audiences and frequently produce marks that have an impact on states and state building. Actually, they explicitly want to orient states in a particular direction: for example, to make them more open to investment, or more democratic or more respectful of human rights. States here act as students, and the indexes as exams. Sometimes this is quite explicit, as in the certification processes institutionalised by the United States (USAID 2010). More often than not states feel obliged to respond to those global audiences and grading agencies in one way or another. Many states conform – willingly or not – to global policies and dynamics, which constitute yet another crucial variable. The newness of this should certainly not be exaggerated, as the following quote suggests: 'from this time forth, History becomes a connected whole: the affairs of Italy and Libya are involved with those of Asia and Greece, and the tendency of all is to unite' (Polybuis 1923: 2). However, it is a fact that today states have to count upon transnational agencies and networks (and a couple of dominant states) that have the economic, political and technological capacity to push forward agendas that act as severe objective constraints for all actors in the international arena. As in previous historical processes, policy formulation necessarily entails the creation of criteria and instruments for follow-up, evaluation, adjustment and correction. One or other will be used by decision makers within the agencies that develop the policies, by opponents and by third parties. Sometimes – where the state is strong enough to create its own count and evaluation structure – these will be internalised by the state.

If this coarse sketch is correct, then it is legitimate to speak about 'third wave indexes'. First wave indexes were basically counts. Second wave indexes were produced by extensive bureaucracies, involved the aggregation of several counts and were related to ever complex policies, developed still exclusively or at least mainly within the framework of the national state. Third wave indexes are related to a technological revolution, the globalisation of capitalism and politics, the creation of global audiences and the development of global policies that, as any other policy, must be followed, evaluated and adjusted. Table 1.1 illustrates all of this. The counting of taxes, bullion, homicides and men in arms began with the building of the modern state; the Gini index was developed in 1912; and national counts came almost thirty years later. Typically, and rather importantly for this narrative, the concept of GDP was created by Simon Kuznets in the context of the US war effort in 1942.[6] It was supposed to help measure wealth and productivity at a time when it was crucial to be able to compare the potential output of a state with the output of its adversaries.

By the 1960s, there was already a well-developed toolkit of economic indicators that measured and allowed the regulation of productivity, inequality and demographic trends, among other issues. With very rare exceptions, political indicators appear later. There is a more or less pristine relationship between the transformation of the concept of sovereignty and the increasing need of states to justify themselves vis-à-vis transnational audiences, and the development of political indexes. There is nothing capricious or haphazard here. Note that the 1963 National Capability Index (NCI; see Table 1.1) was not designed to make states respond to any audience but rather to measure state strength in a 'realistic' (in international relations theory jargon) fashion. For example, a state's ability to launch or to stave off a military attack is evaluated. The NCI was supposed to be the political equivalent of GDP.

Table 1.1 also reveals quite clearly that third wave indexes are new not only from the point of view of the conditions that enabled or made necessary their development but also in the ways in which they were built. Note that first and second wave indexes were mainly counts, the second wave incorporating increasingly more data and more complex ways of manipulating it.[7] Third wave indexes are different in many regards:

- They tend to incorporate many more variables. Let us consider some examples of the 'first wave'. The Gini index (1912) includes one variable, as does the Stock Market Index (1923). In the 1940s we can already observe a certain inflation of information within existing indexes. The University of Michigan's 'Consumer's Sentiment Index' had three variables, and the Purchasing Managers Index five: this is quite baroque. The GDP is of course the winner in these terms, although *au fond* it is really a one-variable tool (for wealth, counted in units of a given currency, such as dollars).[8] In the 1960s, the Composite Index of National Capability reached the heights of six variables. Now compare this with current state performance indexes. The Country Indicators for Foreign Policy has 83 variables. The CPIA, 16; the Failed States Index, 12; the Index of State Weakness in the Developing World, 20; the State Fragility Index, 8.[9]

- They have a more complex structure. Some of them have an intermediate level between the variables and the final aggregation (the boxes). With only one exception, every PSPI is composed of boxes, which in turn contain variables.

- They use sources of information other than traditional sources: global polls, global samples of expert opinion, in-house coding (which is a variant of expert opinion), web sampling, statistics produced by transnational agencies. Thus, they deal with several types of data.

- At the same time, they aspire to cover much more terrain and to process much more information.

- They are not (necessarily) counts. They try to quantify phenomena that range from the moderately subjective to the wildly subjective (for example, happiness or life satisfaction).

- They try to quantify multi-layered, multidimensional and highly contested concepts, such as democracy.[10]

The obvious question is whether these third wave indexes correspond to some 'objective need', like the first or second wave ones, or whether they constitute a rather short-lived trend that will eventually seek refuge in Rankopedia (if, of course, Rankopedia accepts them: bugs and bands are much more interesting). In short, how important are they? Can we live without them? Are they really indispensable? The answer depends on what we consider to be indispensable. The world would probably not end if all the third wave indicators were wiped away.[11] On the other hand, there are three very good reasons for wanting to preserve them:

1. For any student of the state, or at least for those driven by the interest of strengthening some states, it seems fairly inconsistent to reject *en masse* the project of index construction. As has been suggested above, indexing and classifying reality is a natural and routine function of the state, and indexing is a necessary condition for the existence

6 Leontieff's 'input-output matrices' were also created at this time. Of course, both GDP and input-output counting have precedents in the first part of the 20th century.

7 It must be noted that the Gini index is a quite neat example of capturing a social concept mathematically.

8 Accounting operations take place in the equivalent of a spreadsheet, but are never multidimensional.

9 Our own index has 10. The notable exception is the Bertelsmann Transformation Index (BTI), which only has two variables.

10 Note that corruption belongs in a rather different category. It is a hard concept, exactly in the inverse sense to happiness. The latter is difficult to define, but easy (at least on a personal basis) to observe. Corruption is easy to define, but until now it has not been demonstrated that it can be observed reasonably and systematically.

11 It is not clear how well states would manage without first or second wave indicators. It is probable that all of their key functions would be severely affected.

of a modern state. If states are to interact on a daily basis with powerful global agencies, we can expect that third wave indexes will become increasingly important for them.

2. Indexes not only describe reality but also constitute it (Bouyssou et al. 2000). They do so by isolating specific windows of interest, by creating authoritative forms of counting and by developing the criteria that choose winners. There are many classical examples of this taken both from politics and from sport. When relevant decision makers decided that it was a priority to avoid dull draws in football, for example, they chose to award winners three points, against the two they got in the past.[12] This, of course, damaged the interests of the teams that privileged defence. Another way of expressing the idea is that indexes are also systems of incentives. This being the case, there must be stronger and weaker incentives, and thus better and worse indexes, which once again means that they should not be entirely rebuked.

3. Third wave indexes are a product of large-scale technological, economic and political change. This change has enabled several actors – powerful states, transnational agencies and networks – to produce, follow and administer global public policies. These policies are an important reality in the contemporary world (and certainly cannot be described in black and white terms). Policy formulation and administration are historically linked to bureaucratisation and quantification. This deep link will not go away.

There is yet another criterion – in fact the most important one – that strongly supports the active study and development of indexes. To be able to discuss it in detail, however, we have to sketch out the main criticisms levelled against third wave indexes, the main lines of defence of the index administrators, and what can be gauged from such debate.

1.2 WHAT SHOULD, AND SHOULD NOT, BE DEMANDED FROM AN INDEX

Until this point we have spoken of indexes, ranks and measures in quite general terms. It is high time that we introduced some definitional clarity. The following terms define the PSPI (and, more generally, index) environment:

Core concept: The process, phenomenon or state of the world that the index administrators intend to quantify. The definition of the core concept may refer to auxiliary concepts also.

Database: 'A comprehensive collection of related data organized for convenient access, generally in a computer' (Dictionary 2010). For our purposes, a database will always be represented as a rectangular array of data, where the rows are constituted by cases and the columns by variables. A *variable* is 'The characteristic measured or observed when an experiment is carried out or an observation is made' (Upton and Cook 2008). The latter we will frequently call simply 'countries'.

Data operations: To be used for classificatory purposes, the data in a database must be transformed. Political scientists and economists subject their data to numerous transformations (such as the deletion of outliers or smoothing, for example). Some of these are optional. However, there is one operation that is indispensable: *aggregation*. Informally, an aggregation function is an operation that combines all of the available country data (the variables for each country) to produce a single attribute (almost always numerical, but not necessarily),[13] which can be treated as a distinct entity. Aggregation is a crucial step not only because it impacts heavily on the results (see below) but also because it makes sense of the data. An incoherent set of variables is difficult both to read and to interpret because the variables themselves are not the same categories with which policy makers, public opinion, state builders and analysts operate. All of these people want to know what happened to criminality, or wealth or inequality, not necessarily to one or another of their constitutive components.[14]

The aggregation function can produce a rank (for example, an ordering of countries) or a mark (a value that may entail a substantive interpretation, for example the level of democracy or of state fragility). If marks are numerical or at least ordinal they will imply a rank. Thus the former are more informative than the latter. Marks and ranks are particular types of indexes, which will be treated here as the more general term.

Second in order of importance is the normalisation function, which, once again informally, expresses all of the variables in a single scale.

A third data transformation of consequence is missing data imputation (Little and Rubin 2002).

We will henceforth treat an index as the following set: a core concept, a universe of cases (the countries), a set of variables, a set of possible values that each variable can have, a set of data transformation functions and a final output.[15] An index provides a quantitative specification of reality, which is encapsulated by the set of tools that constitute it. As such, indexes are used to orient comparisons, evaluations and incentive provision. Another crucial use is as input to probabilistic models, where indexes can play the role of independent or of dependent variable. Throughout this text, we will concentrate on poor state performance indexes. The common characteristic of PSPIs

12 Participants in matches that ended in a no-score draw got 1 point under the past system, as today.

13 For example, it can be linguistic, such as 'failed'. Not all countries will have the attribute. If a country has a missing datum in some variable, the aggregation function may fail to produce a result.

14 These may make sense in one or another conjuncture. Precisely because of this, the frequently offered solution of dropping aggregation and operating on single variables is a non-starter.

15 Sometimes we will use the word to denote the set, sometimes simply the output.

is that they are intended to produce a measure of state weakness, vulnerability or breakdown.[16] Before considering the problems of PSPIs, it seems appropriate to set the limits of the discussion by specifying what should not be demanded from indexes in general, and from PSPIs in particular.

We start with the obvious. Indexes are supposed to be (extreme) simplifications of reality. The demand that they describe the context or the complex historical trajectory of concrete countries is incorrect. This is the forte of qualitative research and indexes cannot, and should not, try to imitate this. Indexes are powerful precisely because they are (aggressively) simple and shed a substantial amount of context in their calculation. This is what makes them tractable, and enables us to use them to perform relatively abstract, but systematic, comparisons. Abstraction and isolation are a source of both strength and weakness. Even apparently bullet-proof indexes can fall victim to counter-examples if the context is adequately chosen. Take the following example: it is well known that in medical research a corporal mass above 40 is classified in the category of morbid obesity. Morbid obesity is a severe condition, which leads to reduced physical mobility and high risk of organ (heart and lung) failure and of strokes, among many other dangers. Yet despite the fact that sumo fighters are well above that weight limit they are supple, fit and are not exposed to the risks of people with similar corporal mass (WHO 2010). Their lifestyle and body structure are different (Japan Sumo Association 2010). If we diagnose a person with a corporal mass of 40 based only on this datum and ignore the fact that he/she is a sumo wrestler, we will completely miss the point.

Even in disciplines where measurement is relatively clear cut, context plays a decisive role. Given that the very nature of index building is abstracting and producing context-free tags (numerical, verbal or any other type), their inability to capture context is an intrinsic limit. Indexes are not designed to take the context (fully) into account. Their function is to simplify and isolate. This, by the way, suggests that the standard line of defence of index builders – repeated again and again in the interminable debates around the quantitative/qualitative boundary – does not hold. When quantitative defenders argue that at least they make their assumptions explicit, they are clearly expressing excessive optimism.[17] Only rarely, if ever, does a researcher succeed in making all relevant assumptions explicit; simplifying reality is a very complex business (no pun intended). It is difficult to be aware of all the connective tissue that you are cutting off. This is easily shown in the domain of PSPIs, where several hidden assumptions (including the existence of total order in the range of the aggregation function and the existence of substitution rates between variables and boxes) creep in (Gutiérrez 2009). Without such assumptions, the whole enterprise falls apart. Unfortunately, not only are they hardly credible, but they have not been discussed in the codebooks of the indexes and associated literature.

16 There is a whole dictionary of analogous words used in the literature. See Cammack et al. 2006.

17 On the other hand, the fact remains that informal and implicit procedures can hide very big biases.

The other side of the coin should also be highlighted. The systematic loss of information that index building entails allows, at least in principle, for operations such as large-scale (in time, number of cases and number of variables) aggregation, comparison and generalisation. It is not reasonable to pretend that these big comparisons can be performed only verbally or informally, and without the aid of computers. Simply put, human beings are not programmed to do so. Even if they were, they should rather use these cognitive capacities in better undertakings than number crunching. Both common sense and elementary principles of cognitive economy suggest that there are operations that are done better with algorithmic procedures and automated aids. In sum, the losses and benefits of isolation and simplification are both significant. This is one of the strong reasons why qualitative and quantitative research should not be seen as substitutes but as complements.

It is frequently asserted that indexes incur the error of summing up apples and oranges. The statement contains more than a grain of truth (as we show below) but is based upon a misunderstanding. One of the functions of indexes is precisely that they enable us to compare apples and oranges. They cannot be criticised for doing so; this is a fundamental part of their strength. Are oranges and apples comparable? Indeed, in many respects. For example: their weight, their volume, their colour. This is not meant to imply that every comparison based on indexes is acceptable in equal degree. Take for example taste, or the intensity of the sound that apples and pears make when they fall.[18] Here we are not sure that we are actually measuring because the result depends in part on the observer (and, indeed, this dependence may be systematic so that the result of the measurement is stable but biased. For more, see sections 2.4.2 and 3.3).[19] The question that the comparison is based upon may be even murkier. So, in this case, are oranges better or worse than apples; in what sense; for whom, and for what purpose? This question is already near nonsense, because the marks that oranges and apples will get in this sort of exercise depend not only on the differential perception of the coders (for example, respondents in a survey) but on their differential understanding of the question. In sum, there are many ways to interrogate reality and, depending on the one we choose, summing apples and oranges can either make sense or not. Every measure is imprecise, even in the macro-world.[20] However in the first type of measure we can restrict, even if arbitrarily, the degree of imprecision (at least if the setting in which the measurement takes place is good enough). In the second type there is a significant residue of

18 We consciously refer here to an example that has good pedigree in measurement debates. In 1891, the British Commission for the Advancement of Science summoned a committee to debate the possibility of measuring the (subjective) intensity of sound. Typically, the debate produced no clear conclusion, and divided the members of the committee into two factions those who supported the notion that such measurement was possible, and those who staunchly denied it. For a recount of this fascinating precedent see British Science Association (2010). This debate remains far from settled.

19 Note that measurement of physical properties is also not observation-free. The measurement of weight, for example, will vary depending on the hour, the scale, etc. But the variations are marginal, and can be further minimised with finer instruments. In some contexts, though, even this minimisation programme has a bound.

20 This caveat is necessary to avoid the intricacies of the nano-world, which are a fantastic theme beyond the scope of this report.

intrinsic subjectivity. It is not clear if the third type makes sense at all. It should not be requested of indexes that they do not sum apples and oranges, because that is their business. It should be demanded of them that they make sense (ideally measurable sense) of questions that are precise or that allow for a type of vagueness that can be coped with. Finally, they should produce the instruments to manage and tame such vagueness.

Indexes do not tell the whole story – but nor are they supposed to. Actually, it is good that they refrain from doing so. The great danger of indexes is 'conceptual stretching' (Sartori 1970), or trying to measure more than one thing at a time. As psychologists know too well, this inevitably produces noise (Collier and Levitsky 1997). It is wrong to demand from indexes that they encapsulate reality because indexes are not reality. They are a radical simplification of it for the purposes of abstraction and manipulation. So how radical is good? The golden standard was established by Einstein: things should be as simple as possible, but not more than that (100000 Quotes 2010). This, of course, is easier said than done. In the social sciences, the understanding of how much simplification is good enough might allow for a rich interaction between qualitative and quantitative, and for the identification of areas where simplification has gone too far, as long as it is understood that simplification and isolation are integral to index building. For example, GDP has been taken to task over its measurement of 'standard of living' because it does not speak about inequality. Here the criticism should be rebuked: inequality is a different concept from wealth or economic growth, and a complex one at that (Atkinson and Bourguignon 2000). There are two very strong arguments in favour of the separation of the notions of growth and inequality: operationally, trying to measure both at the same time might produce a sloppy result that measures neither properly; and conceptually, researchers and policy makers are better off if they can evaluate empirically how different concepts are correlated. This is a very strong argument put forth by Przeworski (2004) in debates related to the operationalisation of democracy. If the concepts of growth and inequality are collapsed into a single measure then we will not be able to see how they covariate. In definition we will have precluded the possibility of observing their interaction in the world, and thus lose one of the great potential contributions of quantitative research. Combining knowledge of GDP with an adequate measure of inequality makes possible a pretty clear understanding of the landscape of wealth and social justice for any given country in any given year. From this it is possible to study the conditions under which growth and increased equality come together.

If indexes are to be useful, then, they must try to isolate reality and be as accurate as possible (but not spuriously accurate: see section 1.6). So what is the minimum level of accuracy that we ought to demand? Socio-economic data are always messy, and more so in poor, unequal or conflict-ridden settings. Social researchers are harassed by a kind of inverted neo-Lamarckian law: the more you need the data the more difficult it will be to gather it. Everybody knows that even our so-called 'hard data' on unemployment or GDP is imperfect. Admitting these imperfections, though, is not equivalent to believing that anything goes. The calculation of the GDP of Afghanistan may be problematic but the uncertainty that stems from the lack of quality of the data can be reasonably tamed. For example, the World Bank complements its point GDP estimation with a categorical estimation from 1 (very low) to 4 (very high). This contains very few obvious errors, if any, and has few boundary issues (for example: does country x fit within 2 or 3?) Hence, problematic data is not always an insuperable hurdle. Correct representation and the avoidance of spurious exactitude of problematic data should be demanded. The bottom line is that a good indicator ought to produce defensible hierarchies. Defensible does not mean exact. We know, for example, that the official homicide rate in Colombia was 39+ in 2010 (Ministerio de Defensa 2010), but we have a fair amount of uncertainty about the real figure. It could be less than 39, or more, and it is not clear how big the interval of possible values is. It will be difficult to make a tenable assessment of the concrete figure. However, we do know that it is higher than Norway's homicide rate. As another example, the Democratic Republic of Congo (DRC)'s GDP is lower than Bolivia's, though in both countries the production of the figure may have technical problems. So what can be said about the GDPs of DRC and Afghanistan? They are both very low and more or less the same (World Bank 2006): this verbal evaluation can be systematically transformed into a numerical representation on an ordinal scale.

Defensible does not mean acceptable for experts on its specific details either. Only very rarely are good classification machines able to put all cases in the correct pigeonholes (see section 2.1). What we can demand of a 'well behaved' indicator is that: (a) it separates the extremes adequately; and (b) it provides reasonable criteria for comparison between the intermediate cases. We will go into the details of this in section 2.4 (and see 3.3.3).

Thus, indexes cannot be attacked because they isolate reality. Rather, this is part of their strength and contribution. The issue of data quality is fundamental but not always insuperable. What should be demanded of indexes is that they isolate reality with the maximum of clarity, and squarely address the relevant data problems. Once again, this is easily said and not so easily done.

The contestable nature of the data used is one of the focal points of the wholesale rejection of indexes. Since these are based on hopelessly poor information it is argued that they are little more than a source of noise. What this position inconsistently misses is that any type of research is open to exactly the same observations. For example, empirical material coming from in-depth interviews may vary depending on, for example, the characteristics of the interrogator (including experience, empathy and preferences), the physical setting in which the answers were elicited and the system of incentives that generated the interview. Of course, it is perfectly legitimate to wonder if a 'formal decision system' (Bouyssou et al. 2000), for example in the guise of a 'multi-attribute' database (Stelios et al. 1998; Kahraman 2008; Ehrgott and Gandibleux

2002) is necessary or convenient in a given context. Often an informal decision-making procedure will suffice. The 'external' answer to such a question would be to point out that historically there is a strong link between policy making and evaluation on the one hand, and index building and administration on the other. This link is likely to deepen, not to disappear. With the emergence of global audiences, policies and demands, indexes related to these are more or less inevitable. The 'internal' answer would ask how large scale comparisons would be established in the absence of a formal decision-making tool. If we are to make dense, context-oriented comparisons between two, five, ten or fifty countries, then how should we proceed? How should we consider a country's performance over time: t, $t+1$ *to* $t+k$? Please note that this 'internal' issue is completely independent of the 'external' one. Suppose that no global audience or demands existed. States would have to aggregate tens, hundreds and sometimes thousands of sub-national units anyway as this is a fundamental pre-condition for policy design and resource allocation. So how will this be done? In many countries developing formal criteria to do it is not only necessary but also legally mandatory.

1.3 ORDER AND IMPOSSIBILITY THEMES

In this section we discuss the problem of order and some of the implications of the fact that PSPIs are an act of multidimensional, multi-attribute decision making. The first part of the section is dedicated to the preliminaries: setting the stage, presenting some basic definitions and putting forward some measurement issues. The second part is dedicated to the problem of order. Databases exist in a 'partially ordered set' (Gutiérrez and González 2009), which means that some (or possibly many) cases are incomparable. One of the results of aggregating the variables into a single number is to impose on the data the assumption of total order. Sometimes this assumption is tenable, but sometimes it is not. So what are the consequences of imposing total order? The third part of this section explores the analogy of index building and voting systems. Though the analogy is not perfect, it offers several important insights into the objective limits that index building holds. The concrete implication of this is that there may be two desirable properties that indexes cannot have at the same time. For example, it may be the case that they cannot simultaneously fulfil the very important condition of 'independence of irrelevant alternatives' (IIA) and capture the implicit hierarchy that pre-exists in the domain. We have to choose between the two. The fact that index building finds objective limits (not that we have failed to find a solution to the problem but rather that we know for sure that the solution does not exist) has several simple but important implications:

- We will never have a bullet-proof index. This does not condemn index building as a social and intellectual enterprise. Arrow's discovery of objective limits to voting (and, more generally, welfare function building) did not condemn elections or non-consensual decision making, but rather placed them in a specific conceptual setting that allowed us to read them properly (Arrow 1950). The fact that the quest for a perfect index is a wild goose chase is *not* synonymous with the notion that everything goes.

- Some indexes are better than others. This takes us directly to the following two sections. We believe that, all in all, this is still a rather open area of application, where there are many powerful precedents and tools and where new developments would produce non-negligible marginal benefits.

1.3.1 PRELIMINARIES

As seen in section 1.2., a database can be viewed as a rectangular array where columns correspond to variables and rows to countries. An *aggregation function* is a rule that attributes an object to the collection of numbers that represent each country (its values in the variables). This may be a number or a linguistic tag (for example: 5, 0.5, pi, or 'failed', 'fairly good'). The collection of numbers is called a *vector*. The function imputes one, and only one, of these objects to each country vector. The vectors (collections of numbers) constitute the domain of the aggregation function, and the objects constitute the range (or co-domain).

The variables can be of two types. On the one hand, they may be characteristics that can be reduced to a single, quantifiable unit of measure (a *numeraire*). For example, if one is going to buy a car, then it is reasonable that the attributes of the car are expressed in their monetary value (Lootsma 1997). Indeed, the act of buying a car can be seen as the revelation of the consumer's preference about the relative weights of car characteristics, given his budget constraints. On the other hand, variables may be irreducible to a single counting unit. This is a well-known situation in the 'multi-attribute decision making' literature (Kahraman 2008; Lootsma 1997), where aggregation functions that assume that there are tradeoffs between the variables are called '*compensatory*'. The others, in a flight of imagination, are called '*non-compensatory*'. In economics and in other social disciplines, the assumption that one can always rely on a counting unit (money, or von Neumann utilities (Neumann and Morgenstern 1944)) for all purposes has produced a strong focus on compensatory methods, which entail the assumption of the existence of substitution rates. Substitution rates between goods 'A' and 'B' are the price in terms of units of good A that a consumer is willing to pay for one unit of good B. A proper utilisation of substitution rates requires at least that:

a) There is a common counting unit.

b) There is a concrete decision maker, a real individual or an ideal type: *the* consumer, *the* politician. In the absence of this figure, all rationalistic musings lose their power (Przeworski 2004).

c) There is a clearly delimited institutional context where the decision maker makes choices, from where his/her preferences can be deduced using the principle of revealed preferences. This is not necessary but highly desirable.

We will be at pains throughout this report to show that it is necessary to open the black box: the assumption of the existence of substitution rates between variables and boxes. The substitution rates that can be deduced from PSPIs seem contrived and unsound, and sometimes plainly meaningless (Gutiérrez 2009; Fabra and Ziaja 2009; section 1.5 here). It is easy to understand why. First, there is no common counting unit. More or less obviously, the value of democracy as such has not, and possibly cannot, be put in monetary terms (Dahl 1990). The same can be said about, say, the rights of women weighted against those of ethnic minorities.

There is an even more serious underlying issue here. What can a unit of democracy possibly mean? Even if we had joules of democracy, how could we compare them to coulombs of fragility? When we are measuring in monetary value we have a good yardstick that allows us to establish universal comparisons that are independent of the concrete characteristics of the object: a Mercedes is worth seven-ninths of a Porsche. We also have a good theoretical link between money and utility: modulo details. For individuals in our societies, marginal utilities are decreasing in money: for a poor person US$100 has much more value than for a rich person. Counter-examples can be found, but in general this framework is sensible enough. When we are measuring the relative weight of a bag of apples and one of oranges we are in an even better position. Also, when making regressions, the interpretation of the associations with these units is more or less clear cut. One dollar more of income means x days more of life expectancy; one kilo more of weight will increase in y the probability of facing a serious illness after fifty. Indeed, even purely ordinal scales can be added, the typical example perhaps being the rating scales of psychologists (Pfeiffer and Jarosewich 2007).

The problem of measurement is still open, it is not clear how the scales utilised by PSPIs are proper measurements. For decades, social scientists, especially psychologists, have been trying to create constructs that are equivalent to physical measurement. However, the results are still open to debate.[21] According to the prevalent, though far from consensual, metric theory – the so-called 'representational theory of measurement' (Boumans 2007) – measurement consists of the establishment of a correspondence between a qualitative domain and a numerical system. This correspondence must have certain properties.[22] In particular, this representational theory has established that ordinal scales do not translate directly into quantitative, numerical properties (see, for example, Michell 1997). In terms relevant for us, it is not enough to have ordinal scales; additional structure is needed for quantitative properties and proper measurement (Smelser and Baltes 2002). It is not evident that PSPIs and other third wave indexes have this structure,[23] and indeed they may even lack ordinal structure. Take the full Polity Scale (which ranges from -10 to 10).[24] Zero here[25] is in reality neither bigger than -1 nor less than 1, as it should be if Polity were fully ordinal. 0 enjoys a special status, nearer to 'we do not know' than to 'a degree of democracy between -1 and 1'. It is used to define not the situation of the political regime but rather a notion close to that of 'state failure' that other indexes try to capture (see Table 1.3). Finally, it is not clear what it means to arithmetically manipulate numerical tags that are not even fully ordinal.

All in all, those using PSPIs should be extremely careful about the specific meaning of the weights they impute to variables and boxes. In no case should they use compensatory methods where the weights have been decided on an ad hoc basis, or where they suggest meaningless relations between variables and boxes.

PSPIs are not endowed with an 'ideal type' – a universal decision maker. In economic theory we have *the* consumer; in political theory we have *the* politician. These constructs have been contested (Dahl 1990; Green and Shapiro 1994) but they are still backed by substantial theoretical reflection. They are grounded on assumptions that stem from the concrete realm of human activity that they are supposed to represent. So, *the* consumer wants to maximise his/her budget; *the* politician wants to know his/her probability of getting elected, or of remaining in office. What reasonable theoretical interpretation can the weights of a PSPI have? Take the two variables that appear in two different boxes in the Fund for Peace Fragile States Index: on the one hand, 'Brain drain of professionals, intellectuals and political dissidents fearing persecution or repression'; and, on the other, 'Outbreak of politically inspired [as opposed to criminal] violence against innocent civilians'. These variables have the same worth.[26] What can we make of this? Does it mean that the repatriation of one (or several) scientist(s) compensates for, say, one massacre? If this is not the concrete implication of that weighting, then how can it be read?[27] What does it mean if you take away one joule of brain-drain and throw in one of politically inspired violence: will you have the same number of coulombs of fragility? The aggregation function of the Fragile States Index says as much. This is the implication of attributing the same weight to both variables. Needless to say, no decision maker or theoretician would be willing, or able, to support this based on plausible argument. Not only is there no literature to support such a claim but it is highly probable that different people (starting with those involved in a massacre or in repatriation, for example) would give different weights to these events. Note that the line of defence of

21 This fits well with our everyday intuition. For example, there is still genuine debate about the nature of IQ. Does it really measure intelligence? Or does it capture something else? Does it even establish a proper measurement?

22 We persist with using a very informal style of exposition. In particular, the correspondence is a homomorphism (though more exacting authors demand that it be an isomorphism, such as Boumans (2007) and Narens (1985).

23 A manuscript further exploring this is in preparation.

24 Polity's democracy is a definitional part of statehood in several PSPIs. See sections 2.1. and 2.2.

25 Before this it was -77.

26 Note, by the way, that each of these measures different things and that these might decrease or increase in different senses. For example, some types of violence against civilians may go up while simultaneously others go down. This is also an obvious violation of ordinality.

27 In this case it is fortunate that the unit of analysis is hazy. 'Outbreak of politically inspired violence' against civilians puts in the same basket many different types of violence (massacres, threats, imprisonment). How do you compare different patterns of state violence and repression? For example, in Cuba you will hardly find a massacre, but there is no liberty of expression. Colombia goes the other way. How are you going to substantiate the mark for each country in this variable? This is a typical case of erroneous summation of apples and oranges, and a good example of ordinal numbers that cannot be easily transformed into quantitative attributes proper.

revealed preferences does not work here either. States do not take decisions about their modalities of fragility so that the tradeoffs between, say, administrative prowess and monopoly of violence can be calculated after observing their behaviour on some kind of market. Fragility, contrary to preferences, is not a subjective state or a disposition of a set of individuals, but supposedly an objective characteristic 'out there'. From this perspective, by establishing ad hoc weights, PSPIs are only revealing some of the preferences of their builders, and even in this they are doing so quite inexactly.

This discussion strongly suggests that there is a lower boundary below which the dimensionality of PSPIs cannot be reduced. Since there is not a common counting unit, the aggregation function does not go from a totally ordered set to a totally ordered set, but from a partially ordered set (from now on poset) to a totally ordered set.[28] This entails further challenges and problems.

1.3.2 ORDER

As noted in previous publications by some of the co-authors of this text, one routine but crucial assumption of PSPIs is that they can impose a total order (Gutiérrez 2009). So, for example, for any two countries, *A* and *B*, either *A* is better off than *B*, or *B* is better off than *A*, or they are equivalent. The assumption is trivially true in three cases:

- *Case 1* (some first wave indicators). There is only one variable, so the aggregation function goes from a totally ordered set to a totally ordered set (like the Gini index).[29] Here, no order is being imposed by the aggregation function; it exists by construction.

- *Case 2* (taken from economics). There are more than two variables, but there is a common currency that allows somebody (consumers, or politicians, or an external observer) to establish a trade-off between the variables. The obvious example is GDP.

- *Case 3*. There are many variables, but a technique of dimensionality reduction, such as principal component or factorial analysis, can chop down the number of dimensions to 1.

If, however, it is not legitimate to suppose that we find ourselves in any of these three situations, then aggregation functions become opaque in at least two senses. The first is that in a poset, and for any two cases, *A* and *B* can be 'comparable' or 'non-comparable'. If they are comparable and, say, *A* is preferred to *B*, then *A* is equal to or bigger than B in all of the variables. Then the function should place *A* not lower than *B*, independently of the concrete values it bestows on each. This key property is called the Pareto condition, and note that it is definitional. An aggregation function proper should behave like this. It should also respect the so called 'boundary conditions'. If a country gets the highest mark in all variables, it should get the highest aggregated mark. If it gets the lowest mark in all variables, then it should get the lowest aggregated mark (Beliakov et al. 2007).

If they are not comparable – if *A* is better than *B* in some senses, and *B* is better than *A* in some others – then what should the function do? The answer is: anything, as long as it orders the comparable cases well. In other words, there is no reasonable restriction on what aggregation functions can do apart from ascribing the correct values (those that preserve the order in the domain) of the comparable pairs.

It can be demonstrated that, for any well-formed aggregation function *F* from a lattice, or more generally from a poset, to real numbers, and for any two countries *A* and *B* that are non-comparable, there is another well-formed aggregation function, *G*, that reverses the way in which *F* orders *A* and *B* (Gutiérrez 2009). For details and full proof of this, please refer to Gutiérrez and Argoty (2010) and Chapter 4: A curse of excess here. The consequences of this simple proposition for the PSPI domain are not trivial. As long as:

a) The existence of substitution rates (or of dimensionality less than or equal to two), or

b) The superiority, substantial or formal, of one aggregation function over the rest

have not been demonstrated, then the rankings of all the non-comparable cases are a simple artefact of the choice of the function, and from the point of view of the real information they convey they are vacuous.[30] We illustrate this with a simple example. We have a four-country database: Norway, Colombia, Venezuela and Haiti. We know that Norway is less fragile than the other three, and that Haiti is at the bottom (ie worse off than the rest). How will we aggregate the values of Colombia and Venezuela so as to be able to rank them? The proposition discussed in the previous paragraph demonstrates that for any conceivable *F* that puts, say, Colombia over Venezuela, we can build an impeccable *G* that puts Venezuela over Colombia.[31]

It is worthwhile pausing to ponder this result for a moment. Let us start with the simplest setting: two binary variables. Here, the only incomparable cases have the form *(0,1)* and *(1,0)*. The only way in which a country *(1,0)* will receive a higher or equal mark than *(0,1)* is when the first variable gets at least the same weight as the second. This is the '*F*'. We can now build our '*G*' by simply giving a slightly bigger weight to the second variable. Then *(0,1)* will fall above *(1,0)*, the boundary conditions will be respected, and the function will indeed behave monotonically. We have succeeded in constructing a good *G*, that reverses the hierarchy between *(1,0)* and *(0,1)*. Now let us complicate matters a bit further. Suppose we only

28 In section 3.4., we relax this condition, reducing dimensionality but creating functions that go from posets to posets.

29 Actually, if there are also two variables no major problems should be expected either, despite the fact that R2 is not a totally ordered set.

30 The point is argued in detail in Gutiérrez 2009.

31 Quite obviously, this would not happen without multidimensionality.

have two variables but each of them is a rank going from 0 to *k* (*k* being bigger than 1). Take the most unbalanced example of unordered cases: *(k,k-1)* and *(0,k)*. The first country gets a higher mark than the second. Can we still reverse this ordering with a correct *G*? Yes, if the weight of the second variable is sufficiently high. In particular, the weights w1 and w2 for variables 1 and 2 must have the values

$$\{\{w1 \to \frac{1}{1+k}, w2 \to \frac{k}{1+k}\}\}.$$

It can easily be seen that, if the order of even these two very unbalanced countries can be reversed, all the other ones also can.

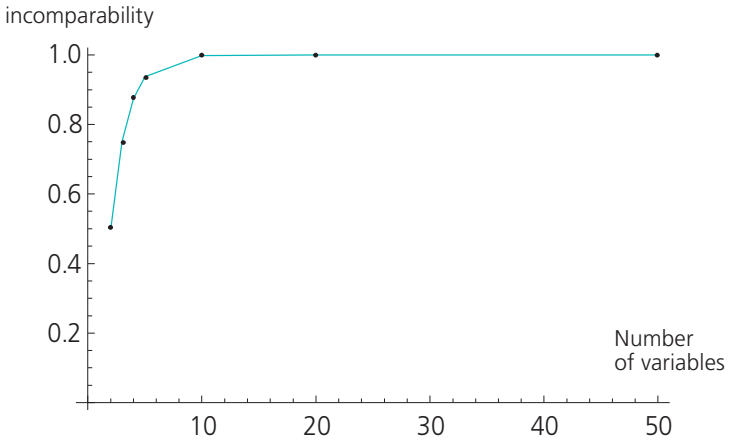
Is there an answer for the general case? Gutiérrez and Argoty (2010) provide this. Every single aggregation function from a lattice (and more generally, a poset) to the real numbers (the integers) can be reversed in incomparable pairs. Furthermore, the number and type of reversals grows awfully fast. If the biggest number of mutually incomparable cases is *n*, then the lower bound of different functions (from the point of view of the orderings they yield), up to linear transformations, will be *n*! This is quite big. For example, if there are ten mutually incomparable cases in the dataset, then there are at least three and a half million alternative and perfectly reasonable rankings. The case is worse with concrete values, which can be produced by the same number of correctly built functions and then made subject to arbitrary linear transformations (so that the 'distance' between one country and another is much more arbitrary than a hierarchy). Let us add that it is unlikely that in each year a PSPI will have only ten mutually incomparable cases. We conducted a small experiment with random values between 0 and 1 for 200 cases and 1,000 replications for each number of variables to see how much incomparability would grow in the number of variables. The answer was that, after 20 variables, the proportion of incomparable pairs grows asymptotically to 1 (see Table 1.4 and Figure 1.1). Thus, if the variables are not correlated and there is no argument for any aggregation function in particular, after 20 variables the ranking is simply an ad hoc choice of the investigators. It is true that, since there is a high correlation between variables and dimensions in PSPIs, the proportion of incomparability grows much more slowly than in the experiment (because there will be linear dependencies between the variables). Gutiérrez and Argoty (2010) show that the minimum number of incomparable cases will correspond to the minimum dimensionality of real vectors in which the poset can be embedded. However, it will be seen in Chapter 2 that far more than half of the pairs are incomparable (with the exception of the BTI).[32] As illustrated by the experiment, this is not an accident.

If the number of incomparable cases was substantially less than 10 and the variables were highly correlated, then the damage caused by the arbitrariness of the choice of the aggregation function would be manageable (at least with respect to orderings).

32 The very interesting BTI has only two dimensions, monopoly of violence and administrative prowess. Thus, it is invulnerable to the impossibility themes that we will discuss below. Typically, the incomparability proportion in the BTI data base is very, very low.

Figure 1.1: 'Incomparability' grows asymptotically to 1 very fast

Source: Authors' own calculations



What possible solutions are there to this problem, apart from drastically reducing the dimensionality of the database (at the cost of a radical loss of information)? We put forth four possibilities:

1. Create a specific theory, with additional substantive reasons according to which, given what we know in *this* domain of knowledge, *F* is better than any other alternative aggregation *G*;

2. Ground the aggregation on the data, so that, given clear reasons for extracting certain parameters from it, only one (or only one family of) aggregation function(s) can be implemented, and thus at least the hierarchy is unique;

3. Weaken the assumption of total order in the range, obtaining in turn the flexibility to express at least partially the incomparability of the domain;

4. Combine any of these solutions.

Of course, if we were able to find an aggregation function that were superior 'in general' (for all situations and purposes) to all others, then the problem would disappear. Now we proceed to a review of the rich thread of literature that suggests that we will never be able to find that miraculous function.

1.3.3. IMPOSSIBILITIES?

When the database domain cannot be reduced to a totally ordered set, indexes can be looked at as if they were voting systems, where the 'voters' are the variables and the 'candidates' the objects (in our context, the countries). This perspective has produced a very rich literature, which is brilliantly synthesised by Bouyssou et al (2000). Another strong analogy that formally boils down to the same structures as those of voting systems is to look at aggregation functions as social welfare functions, devised to

evaluate how well a country is doing (according to some kind of 'absolute' measurement or in relative terms) in the database – for example, the Gini coefficient says how unequal a country is; the GDP how wealthy; PSPIs how fragile the state is. As in any voting system or social welfare function, the 'preferences' of the variables (the mark or rank that they confer to each country) must be aggregated in some reasonable form so as to produce a 'social decision' (an aggregated ranking that adequately reflects the 'decision' of the variables), which is the final product of the index.

The analogy is not perfect. For example, we would demand of a voting system that it is anonymous because it should treat each citizen equally. This basic criterion of fairness has no appeal in the world of index building. In reality, we do not have the least interest in trying to 'represent' adequately the 'preferences' of variables. If we are interested in anonymity (as we are – see section 3.3.3), it is for reasons very different from 'representativeness'. From a very fundamental perspective, however, the analogy makes a lot of sense (Bouyssou and Perny 1992; Bouyssou and Vansnick 1986; Carbone and Hey 1995; Dubois et al. 1997). As seen above, a PSPI embodies a typical 'welfarist' view of society. It is crafted to answer a question of this type: from a specific perspective (for example the quality of the state), who is better off: the citizens of Rwanda or those of Uganda? Note that, in a country where crony capitalism is prevalent and the state is captured by private agents, the state works very well for people who are well connected. However, we want to evaluate if it works well in general for all of society (or for an enlarged, global society). We want to adopt the perspective of somebody who aggregates social goods across the population. This is precisely the 'welfarist' perspective.

Voting systems and welfare functions have intrinsic limits. The justly celebrated Arrow impossibility theorem (Arrow 1963; Tullock and Campbell 1970; Bossert and Peters 2000) has produced a robust and steady stream of important results that explain why none of them can have a set of very elementary desirable properties simultaneously. The method of demonstration is axiomatic, as fits a formal procedure.[33] In particular, no voting system can be simultaneously anonymous, symmetrical, universal, transitive, unanimous, independent of irrelevant alternatives and non-dictatorial. One of the key implications of the impossibility theorem for our discussion of indexes is that in many situations and, importantly, when candidates are not comparable, the social choice does not depend on the intrinsic characteristics of the candidates but on the method of choice.[34]

We illustrate this with a concrete example and three perfectly reasonable methods of election: plurality (the most voted-for candidate arrives first), the Borda count (you deposit *n* votes for your preferred candidate, *n-1* for the second, and so on), and the run-off. In the 1998 elections in Colombia there were three main candidates: Andrés Pastrana, Horacio Serpa and Noemí Sanín. According to all the opinion polls, before the first round the preferences were distributed as shown in Table 1.5.[35]

The plurality rule, which was used before the 1991 Constitution, would have given the presidency to Horacio Serpa. The run-off rule, approved in 1991, gave it to Pastrana (because Noemí's voters turned to him in the second round). An elementary calculation shows that the Borda count would have led to the appointment of Noemí (Saari 2000). So who really represented the preferences of the Colombians in 1998? The question cannot be answered easily. The election result was an artefact of the method of choice. Note the strong similarity of this with the Gutiérrez and Argoty chapter, and demonstration, at the end of this report. In the first case, we see the proliferation of well-formed aggregation functions that reverse rankings of non-comparable cases.[36] In the second case, we see that different procedures of decision making, all apparently innocent and based on common sense, produce different rankings whenever the number of candidates is bigger than two and the candidates are not Pareto-ordered (ie when it does not happen that every voter prefers one candidate over another).[37]

The Arrow impossibility results and those that followed have prompted two types of response. Some researchers have tried to relax some of the properties demanded by the voting mechanism, especially IIA or transitivity (see, for example, Mas-Colell and Sonnenschein 1972; Campbell and Kelly 2000). Others have claimed that the impossibility would disappear if more information were allowed to flow into the decision system (see, for example. Sen 1970; Sen 1979; Bowles, 2004).[38]
Let us come back to the indexes, and to the PSPIs. Take the aggregation function *F*. What should we demand of it? Which characteristics of *F* do we consider indispensable, or at least highly desirable? *Prima facie*, the following seem beyond reasonable controversy:[39]

- Unanimity (or the Pareto criterion). If for any two countries *A* and *B*, *A* is over *B* in all the variables, then *F* should allocate *A* above *B*.

- Monotonicity. If country *A* is improved in one of its variables (call the improved version *A'*) then the final aggregation will be at least as good as that of *A*. In symbols, given an *A'* such that *a1=a1'*, *a2=a2'*… *ai<ai'*, … *an = an'*, then *F(A) ≤ F(A')*.

These are the basics. Anonymity, which 'implies that every individual [variable] is treated in the same way' (Craven 1992: 98), might be desirable, especially in a context in

33 There are many different versions of Arrow's theorem, and powerful impossibility follow ups. We skip here the whole issue of translating voting into index language, though it is clear that a careful consideration of impossibility results offers rich insight into the problems of index building. A brilliant treatment of some of the key themes can be found in Bouyssou et al. (2000). However, their concerns are very strongly related to eliciting consistent preferences from decision makers. Below, we introduce some considerations that are inspired detailed tinkering with some of these issues, without flagging them or going into the technical details.

34 We obviously maintain here a very informal style of exposition.

35 We simplify and stylise here somewhat, to avoid irrelevant detail.

36 Note that even here it is not the case that anything goes. A rule that gives the presidency to the candidate who gets fewer votes is worse than the run-off in many respects.

37 So, the two magic numbers of this chapter are 1 and 2. If the number of variables is bigger than 1, you lose total order. If it is bigger than 2, you can fall into choice issues.

38 There is yet another very important route: the 'fuzzyfication' of the decision system, which may solve some of the Arrovian dilemmas (see, for example, Richardson 1998).

39 We translate the properties into 'index language'. It is obvious that the present informal discussion cannot replace an adequate axiomatic treatment, and it does not pretend to do so.

which multidimensionality is indisputable. Symmetry, which 'implies that all the alternatives are treated in the same way' (Craven 1992: 98) is difficult to sacrifice, as we would prefer in principle that all of the countries are treated on the same standing. Let *F* be Pareto, monotonous, symmetric and possibly anonymous. Should it be transitive? Transitivity means that if *F(A) > F(B)*, and *F(B) > F(C)*, then *F(A)PF(C)*.[40] Transitivity is considered a benchmark of rationality. On the other hand, we may want to relax transitivity in the presence of:

- Ambiguity: the classic and early example provided by Raiffa (1979) is the following. Suppose you prefer to take your coffee with three spoons of sugar than to take it with four. At the same time, you are indifferent to the difference between having it with three spoons and having it with three spoons and one additional grain. Then, by adding one grain at a time, you can build a chain of relations where you are indifferent at each step, but you end up preferring the initial situation (three) to the final one (four).[41]

- Corrupted data.

In sum, transitivity cannot simply be discarded but it can be relaxed. What about universality of domain (where *F* should be able to rank any two countries)? This can be relaxed as well. Actually, it would not be too bothersome if our *F* withdrew when data were too muddled to produce a sensible comparison, and came back with a 'who knows?'[42] Of course, we would not like to get this answer too frequently. *F* should produce a clearly interpretable result in the majority of cases.

Independence of irrelevant alternatives means the following: if *F* allocates *A* above *B* when aggregating data of *A* and *B*, then it will allocate *A* above *B* when aggregating data of *A*, *B* and a third country *C*.[43] IIA has a special status. In the literature on Arrow's theorem, a very common assumption is that it is the less respectable property – something like Euclid's fifth axiom (Arrow 1950; Taylor and Pacelli 2008). But the main product of an index and especially of a PSPI is an ordering of countries: a rank. If IIA is violated, this means that the rank is unstable. Even worse, since the set of countries that PSPIs rank is always a proper subset of all countries (because data is not available, etc.), if *F* does not respect IIA then one has the right to wonder if the outcome would not be different if some cases had not been excluded. So IIA seems pretty crucial in our context. On the other hand, precisely because *F* is a function from a poset to a totally ordered set, it would be highly desirable that it captured the underlying hierarchy that exists in its domain. However, it can be said, informally, that it is impossible to have both properties at the same time, that is, you cannot preserve IIA with a function that captures, beyond monotonicity, the underlying ordering of the set. (Bouyssou and Perny 1992). So here we have a tough choice. The moral of this seems to be the following: IIA should not be relaxed without a very powerful reason (and a very strong trade-off). If it is relaxed, this should go hand in hand with a demonstration that the violations are only marginal.[44]

Until now, we have roughly guided ourselves by the axioms of the Arrow impossibility theorem. As seen above, in the context of indexes, some of these can be relaxed. On the other hand, there is a property of *F* that seems undesirable in the PSPI context – it being compensatory. A typically compensatory method, though not the only one, is weighted averages. For PSPIs, the problem with compensatory methods is that it is not clear if they yield an interpretation that is substantive and reasonable, at least to a minimum standard. The assumption of 'total compensation' is particularly implausible, where a loss in one dimension, however radical, can be outweighed by gains in the others. As Fabra and Ziaja (2009) note, this underlies all PSPIs.[45] Weighted averages, which are the default choice of aggregation functions in numerous contexts, are particularly vulnerable. These are not anonymous if the weights are unequal, compensatory and based on substitution rates.

So there will probably be a trade-off between having good hierarchical indicators that do not respect IIA and compensatory methods that are IIA but that are based on moot assumptions. There is no easy way out of this. In the excellent *Handbook on Constructing Composite Indicators* (Nardo et al. 2005), it is suggested that the so-called 'deprivation index' (geometric aggregation) is a solution that at least does not equate countries with a very different profile but gives them an equal average. However, the geometric aggregation has several highly undesirable characteristics and it is as compensatory as a common average.[46] The authoritative discussion of this by Bouyssou et al. concludes that:

> The pervasive use of simple tools such as weighted averages can lead to disappointing and/or unwanted results. The use of weighted averages should in fact be restricted to rather specific situations that are seldom met in practice (Bouyssou et al. 2000: 247).

In particular, we believe that the use of compensatory methods should be tamed by a detailed and reasonable substantive interpretation of the weights.[47]

40 We treat '>' as the everyday notion of 'preferred to', skipping also the technical details of the type of relation it is. There is a wealth of very good technical treatments of the subject. See for example Bustince et al. (2007).

41 This goes way beyond simple bad taste (good coffee should be taken without sugar!). Similarity relations are not transitive (Peters 2004).

42 This is a luxury that voting systems cannot afford!

43 This is the simplest of definitions, and there are many alternatives. See Arrow (1950) and Taylor and Pacelli (2008) for example.

44 Not necessarily in general; only for the database to which *F* is being applied.

45 In the PSPI context we have the additional task of having to provide substantive interpretations of the compensations between variables and between boxes.

46 It does not behave well with values between 0 and 1, and the weights are assigned ad hoc.

47 This is also true of the scale in which the variables are formulated. But 'if we want to characterise the weighted sum itself (not the derived ranking), then we need to impose additional conditions that make a distinction between the weighted sum and all the increasing transformations of the weighted sum (square root of the weighted sum, etc.)' (Bouyssou et al. 2000).

In this section we have discussed some Arrovian themes in an informal manner. Since indexes are a social welfare function of sorts, we would expect them also to be affected by impossibility. At some point, they will fail to have some of the highly desirable characteristics that we would want them to have. Consideration of the axioms that give origin to impossibility suggests that to escape aggregation problems we must relax some of them. There is, however, another strong constraint: other than in rather specific situations (such as when it is possible to offer a strong and reasonable substantive interpretation of the weights), we should not resort to compensatory methods.

1.3.4. CONCLUSIONS

We have discussed here in a very informal way the challenges that a lack of total order poses for PSPIs. We claim that the question of measurement remains open, and that it should be squarely addressed.[48] Since the existence of a *numeraire* cannot be assumed, multidimensionality in the domain has to be addressed. After discussing the consequences of this, we used themes of impossibility to reflect on limits to the formulation of an index-aggregation function and the desirable characteristics it should have. A consequence of this discussion that should be stressed with full force is that no PSPI is complete without a detailed discussion of its aggregation function(s) and their strengths and weaknesses (or at least limitations). Such discussion seems indispensable. We here give Boyssou et al. the floor once more:

> Devising an aggregation technique is not an easy task. Apparently reasonable principles can lead to a model with poor properties. A formal analysis of such models may therefore prove of utmost importance. (Bouyssou et al. 2000: 6).

1.4. NORMALISATION

As seen in the previous sections, index building frequently operates under the critical assumption that it is possible to perform arithmetic operations over the variables. The assumption is critical because in its absence it would not be possible to add apples and oranges, which is precisely what indexes should do.

Databases can contain the following types of numbers:

- Integers (when one is counting, for example, deaths in a civil war);
- Rational numbers (when one is performing different operations (for example, division) over integers);
- Ordinal numbers (tags in a scale): for example, when respondents to a questionnaire are asked to evaluate the economic performance of the government, their answers are coded as 3 (good), 2 (average), 1 (under average) or 0 (bad). Ordinal numbers represent the place of the case on a scale. Ordinal numbers *cannot* be added, divided or multiplied unless some basic conditions are met: they are like temperature scales.[49] Proportions with ordinal numbers are also tricky (though not necessarily wrong). For example, we cannot say that if the average temperature in Cali is 30 degrees Celsius, and in Bogotá it is 15, then Cali is two times warmer than Bogotá. The same operation carried out in Farenheit would produce another proportion.

The most general case is when the database has variables of the three types, with ordinal variables that represent very different scales. As we will see in 2.2, the majority of PSPIs have ordinal data. Then the following simple problem arises: how can all of the variables be expressed in the same scale?

The apparently straightforward answer is, through normalisation. Let us focus on this operation. To unpack its meaning, we propose a simple example. Suppose our PSPI is composed of two variables: democracy (as measured by Polity) and development, using the World Bank categorical GDP. Our aggregation function is straight average.[50] Our problem is to how to add numbers that describe the position of each country on a 0 to 10 scale (Polity's democracy) with others that come from a 1 to 4 scale (the World Bank's categorical GDP). In this case, we simply choose a transformation that puts both values on a common scale: for example the real numbers between 0 and 1 (a totally ordered set). The most common normalisation function is

$$\frac{x - \text{var}\,[\text{min}]}{\text{var}\,[\text{max}] - \text{var}\,[\text{min}]} \qquad \text{f(1.4.1)}^{51}$$

Where *x* is the value of the country, and VAR [max] and VAR [min] are the maximum and minimum values of the given variable. So, in our example, if the most democratic country gets a 10, and the least democratic one a 0, and the country we are marking has a 5, then its normalised value will be *(5-0)/(10-0)=0.5*.

This seems to be innocent enough. However, it illustrates well why normalisation is yet another dangerous operation. As seen in section 1.3, it is not clear what the unit of analysis is after normalisation, or how to interpret it, for example, in the context of a regression. Any normalisation is valid for linear transformations. This means that I can use the 0–1 benchmark, or 0–10, or 0–100, or 0–51,344. The last seems less natural but is equally correct. In this context, the routine interpretation of a regression result – an increase of a unit in independent variable x will produce a change of two units in the dependent variable y–seems rather capricious.

This is not the only serious issue related to normalisation. For example, the standard normalisation function described above can change the sign of a regression (for example it is not stable in time) and violates IIA. Let us consider the first aspect. Table 1.6

48 There is a very rich literature on measurement. See, for example, Suppes and Zinnes (1962) and Dan and Machover (2000).

49 This is why psychologists go to such lengths to build scales ('summated scales' (de Vaus 2002)) that can be operated on.

50 As used in many PSPIs.

51 Hann and Kamber (2000).

contains a hypothetical set of four countries and two variables (a 1-5 and a 1-10 scale) in two successive years. In the raw data, every single country improved in both variables in Year 2. In figure 1.2, the reader can observe the 'ascending trend' where red points correspond to Year 1, blue points to Year 2. This produces a positive and highly significant regression coefficient. In the normalised data, every single country descended in both variables in Year 2. In figure 1.3, the reader can observe the 'descending trend', which will produce a negative and highly significant regression coefficient. There is no mystery here. What happened is that everybody became a better performer, but the worst countries grew a little more quickly than the rest. This produces the apparently strange behaviour that we have described above.

Figure 1.2: A hypothetical set of four countries – Regression

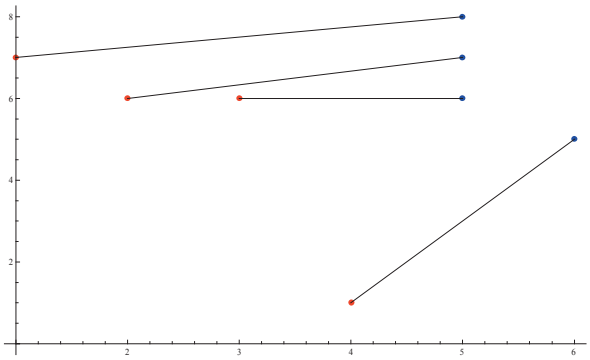
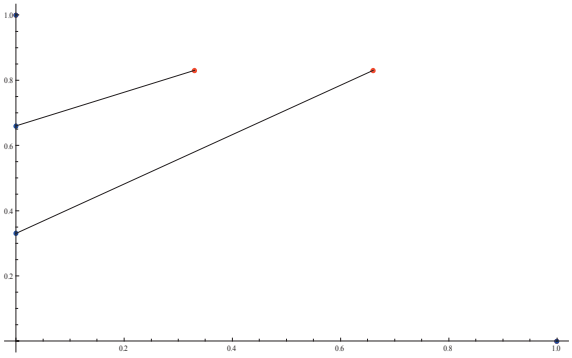


Figure 1.3: A hypothetical set of four countries – Normalised data



Standard normalisation does not behave any better with respect to the IIA criterion. Table 1.7 shows the type of problems that can inadvertently pop up. First, note that normalisation, combined with the popular straight average aggregation function, can change the rankings relative to the originals in the raw data even in the absence of new cases. Now include 'country', which behaves very well in Variable 2 but terribly in Variable 1. The relative positions of the countries change substantially.[52]

In sum, we can say that standard normalisation does not necessarily preserve hierarchies. Normalised ranks can diverge, sometimes strongly, from raw data ranks. Several normalisations – including those used most – do not necessarily preserve trends. The sign, and not only the significance and size, of the parameter can change after normalisation.

The lesson emerging from all of this is *not*, of course, that normalisation is wrong per se, or that standard normalisation has some intrinsic flaw. It is rather that normalisation is not self-explanatory. Just like aggregation (to which it is younger brother) it must be tested, evaluated and substantiated.

1.5. EVALUATING AGGREGATION FUNCTIONS: 'EXTERNAL' CRITERIA

1.5.1. HOW GOOD IS AN AGGREGATION FUNCTION?

Thus far we have discussed aggregation and normalisation functions, finding some problems and illustrating them. We have proposed some criteria of evaluation, taking into consideration the problem of order and looking at indexes as social welfare functions that are designed to measure when a state is better off, using a particular issue as determinant. We have arrived at two main conclusions. First, there is no bullet-proof function, or thus index. Second, depending on the intent of the researcher or policy maker and on the nature of the dataset at hand, some indexes are better than others. In particular, we prefer those indexes that:

a) Fulfil Pareto and monotonicity, are symmetric, are non-compensatory, and do not violate (or violate only in the margins) IIA;

b) Fulfil Pareto and monotonicity, are symmetric, are compensatory, but at least impute weights that are 'reasonable' in some clearly established sense.[53]

Note, however, that observing indexes as social welfare functions only takes their 'internal' workings into account: the way in which they behave from an axiomatic point of view.[54] There are also 'external' criteria of evaluation. The three most important are validity, reliability and sensitivity. *Validity* is 'the extent to which any measuring instrument measures what it is intended to measure'' (Carmines and Zeller 1979: 17). *Reliability* 'concerns the extent to which an experiment, test, or any measuring procedure yields the same result on repeated trials' (Carmines and Zeller 1979: 11). *Sensitivity* is generally shown in terms of the sensitivity measure of each input source to uncertainty. These sensitivity measures 'represent how much the uncertainty in the composite indicator for a country would be reduced if that particular input source of uncertainty were reduced' (OECD 2007: 35).

Note that these concepts are 'external' but are related to the index proper and not to their use, for example, in probabilistic models. Once the database is set up, there are families of methods

52 Note that it is the violation of IIA that explains why the normalised final scores of countries 1 and 2 differ, despite their having equal values (only in different variables) and the aggregation function being straight average.

53 One of the main issues being that it is at least not *fully* compensatory.

54 Here also we shed the axiomatic apparatus, and indulge in an informal exposition.

of dimensionality reduction. For example, factor or principal components analysis, or structural equation modelling, may allow the researcher to find formal constructs that underlie the associations between the variables. These have to be chosen first.[55]

Clearly, validity is more difficult to assess. The simplest and most straightforward approach to the issue is the following: a quantification that is not theory-oriented is not interesting.[56] Simple as it is, this observation is fundamental and no amount of hand waving or of methodological machismo can cancel it. It also constitutes a very important link between qualitative and quantitative research.

Different methods have been developed to establish reliability and there is a wide literature on this topic (see, for example, Thompson (2003), Snyder and Lawson (1993) and Reinhardt (1996)). The bottom line is that PSPIs should be submitted to two reliability tests:

a) That the aggregated result is stable;

b) That putative data based on expert opinion are credible.

There are many ways of measuring reliability: the so-called 'internal reliability', the prophecy index and temporal stability among others (Aamodt 1991). These are designed to establish that the items in a questionnaire capture the same underlying construct (for example, that IQ really is a test for intelligence and not something else); that the ideas, values, capabilities or preferences of the respondents have been captured correctly (or at least stably); and that the replications of the test in similar settings will yield similar results. The main instruments to establish reliability have been developed in the context of scale and questionnaire construction (see, for example, Schwab 2005), where the main concern is to capture adequately (or at least stably) some subjective state of mind of the respondents. Note that the concern of indexes is different. At first glance, the difference appears to be a small nuance but in reality it is quite fundamental. For example, when submitting a set of countries and scales to a coder or a set of coders, index builders want to find 'what is out there', not what the state of mind of the coders is.

This takes us to the problem of the reliability of expert opinion. Coders can be considered experts of sorts, and this is discussed in Section 2.1. In effect, several third wave indexes and not only PSPIs rely heavily on expert opinion. The majority of PSPIs would not exist without this source. Its advantages are obvious. First, unlike counts, it is rarely affected by the problem of missing data. Second, data can be collected on a single scale, so that issues related to normalisation are circumvented (see Section 1.4). Third, it is much cheaper to mark a country based on expert opinion than by setting up and maintaining counting offices (which, as outlined in Section 1.1, are a key component of state building). Fourth, scales that come from expert grading have a 'natural' interpretation (based upon, for example, the degree of intensity or the presence of a social good or bad). The already enormous literature about expert opinion, however, reveals that it is dangerous to commit to the process of grading and making predictions based upon the opinions of experts without taking adequate precaution. Diverse experiments in many fields and using a variety of designs have yielded the following conclusions: first, expert opinion can be quite unstable (Cooke 1991). It can change depending on circumstance, personality, conjuncture and more. Second, the choice of expert is crucial. Indeed, 'in areas requiring professional judgment the most critical factor lies with the selection of panel members, because the reliability and quality of the results will reflect the quality of the experts' (Tolley et al. 2001: 309. Also see Martino 1983; Preble 1984; Taylor and Judd 1989). Experts are usually selected purposefully rather than randomly, and unfortunately this leads to a certain amount of bias owing to self-selection. In PSPIs, this issue is particularly significant, as 'in-house' coders are often taken to be the de facto experts.

Third, even if the selection of experts were not a significant issue, the use of expert opinion for prediction remains a topic of debate. Vaillant et al. (2008) tested how well experts predicted the price of horse semen being sold at a market. There was no significant correlation between their guesses and real prices. Ashenfelter (with Ashenfelter 2001; with Storchmann 2003) has conducted similar evaluations on the subject of wine. Such experiments have shown that scores that tasters of wine or coffee, for example, attribute to samples differ when the taster is blindfolded and thus not aware of the brands involved in the exercise (Livat and Vaillant 2006).

Teachers and professors are often considered a type of expert, indeed one professionally trained to assign grades. Noting a strong analogy between grading an exam paper and a country, it is interesting to question their reliability in the grading process. This is a question that has attracted the attention of many outstanding educational researchers and has thus been frequently and rigorously assessed. Results indicate that these grades are *not* very reliable. Even in exact disciplines, such as mathematics or engineering, significant variance is found between marks depending on factors such as the time of day that the marking took place, the quality of the immediately preceding exam marked and other such factors (Newmann et al. 1997).[57] Teachers demonstrate deep bias. An often cited example of this is gender, where, as Reeves et al. (2001) found,

> The difference was significant for mathematics across all years, with the teacher assessment consistently under-rating boys more than girls. The same was found for science, but for English the opposite was found, where females were more frequently under-rated by the TA. For students with special educational needs (SEN), teacher assessment levels were more likely to be lower than test results. In many instances the effect was considerable, for example, 24% of students with SEN were awarded lower levels than the test results for science in 1998. For students whose first language was not English, the only effects were in English

55 Every single text on SEM or factorial analysis stresses the fact that searching for latent variables without any theoretical guide is bound to produce pure noise. See Mulaik (2009) and Kline (2005).

56 This is an even more applicable approach to take when it is impossible *not* to find something, as in factor analysis. Here, the challenge for the researcher is not to find 'something', but to find formal constructs that have meaning in the context of some theory.

57 The best situation is to be marked after a very bad exam.

> in 1998 when TA underrated 25% of these students compared with 15% of others (Harlen 2005: 259).

In a study about physical education teachers, Hay and MacDonald (2008: 153) found that,

> The data indicated that the teachers in this study made progressive judgments about students' level of achievement across each unit of work without explicit or overt reference to the criteria and standards represented in the schools' work programs and in the Senior PE syllabus … Determining students' levels of achievement was for the teachers somewhat 'intuitive', being reliant on their memory of students' performances, and influenced by the construct-irrelevant affective characteristics of the students. That such construct-irrelevance compromised the construct validity and possible inter-rated reliability of the decisions made and advantaged some students and marginalised others on the basis of characteristics that were not specifically related to the learning expected from following the syllabus.

Wyatt-Smith concludes that, when working towards consistent assessment based on teacher judgment there is a need to consider how information about aspects of students' behaviour or knowledge of gender, special educational needs, or the general or verbal ability of a student can impact on teachers' judgments of performance in a particular task (Wyatt-Smith and Castleton 2005).

Exams are very poor tools to establish credible and stable cut-off points, or the mark below which a student fails (Bouyssou et al. 2000). In sum, even in a context of highly trained personnel and strong institutional constraints and monitoring, specific routines and processes to avoid subjectivity, it is difficult to interpret what a grade really means.

None of this should be used for or considered as a nihilistic tirade against expert opinion. Actually, another group of experiments finds that it makes sense to take expert opinion into account as an intuitive finding. A study about the population distribution of 'rare species for which insufficient presence/absence data are available for traditional statistical methods to be used' showed through cross validation' that:

> Utilizing the experts' prior knowledge was much better than ignoring it; each of the prior models led to posterior distributions that had better scores than stepwise logistic regression. The experiment also showed that there can be benefit in modifying a prior distribution in order to reduce the effect of systematic bias in an expert's assessments (Al-Awadhi and Garthwaite 2006: 139).

Hence the lesson here is not and cannot be that expert opinion is meaningless. Rather, unless certain prerequisites are fulfilled it is not reliable. That a mark has been produced by experts or in-house coders after a process designed to yield consensus is not enough to establish reliability, let alone validity. According to Scapolo and Miles (2006: 682),

> Experts' contribution is seen as a help in areas of research where an explicit conceptual framework may not exist or where data are very impoverished (ie where formal methodologies which make the use of any existing theory and data are not available, or underdeveloped, or not widely accepted). Those experts in a particular problem area may possess unstated 'mental models' and knowledge of the causal structure of a particular system, and are likely to have reasonably well-grounded appraisals of the state of affairs in the topics of concern.

Many questions must be answered before utilising *reliable* expert opinion. Yu and Park (2000) believe that a plausible use of expert opinion entails answering at least the following questions:

1. How should the uncertain information elicited from experts be represented?

2. How should the uncertain information obtained from different experts be combined?

3. How should this information be propagated through a system? For example, if the information provided by expert opinion is poor or misleading, the error can be multiplied several times in the process of construction of the indicator.

Note that each of these questions is essential for PSPIs. With regard to combination, for example, it is not evident that two codings by experts using a similar scale (for example, 1 to 4) but with different semantics (one is about democracy, the other one about quality of life) can be added and multiplied.[58] As Yu and Park (2000: 714) conclude:

> The fact that the expert opinion elicitation process involves explicit characterization of uncertainties makes it necessary for two major types of uncertainty to be distinguished: a. uncertainty due to stochastic variability and b. uncertainty due to a lack of knowledge.

Hokstada et al. (1998: 66) assert that there are three indispensable steps required to put expert opinion into a usable form:

1. Preparation
   - **a.** Choosing Experts
   - **b.** Defining the Questions
2. Elicitation
   - **a.** Performing the Interview
3. Calculation
   - **a.** Evaluation
   - **b.** Combination (aggregation)

They also offer a list of minimum requirements for the sensible extraction of expert opinion:

58 This is true even if the marks are provided by the same expert, let alone if the marks come from different sources. This is somewhat different from the summated scales of psychologists and opinion polls.

1. Documentation

Thorough documentation is fundamental in order to make the analysis credible. All assumptions and decisions made by the analyst (person(s) that will administer the experts and carry out the analyses) must be stated, including documentation of the elicitation process the calculations (formulae and data) should be documented or referenced, so they can be checked by peers (Hokstada et al. 1998: 67);

2. Objectivity

Expert judgment is by nature subjective and the main criticism is directed toward this. *Honesty* of the expert is essential in order to achieve *objectivity.* Motivational bias (ie, 'personal interest in the results') can be difficult to reveal and control, thus, the analyst should strive for neutrality, ie, not influencing the experts or taking active part in the elicitation process, and should encourage the experts to state their true opinions (be honest)' (Hokstada et al. 1998: 67);

3. Empirical Control

Empirical control (verification/checking through observation) of the expert judgments may take a very long time, hence, the analyst should to some extent mix control questions (*seed variables*) and the actual questions, to secure that the experts have the same attitude towards both type of questions (Hokstada et al. 1998: 67);

4. Completeness

Requirements on completeness are demanding but necessary, in order to achieve credible results. The group of experts should be so composed that all relevant aspects for deciding the question are illuminated, so that the analyst can design a procedure with a main objective of extracting as much relevant knowledge from the expert as possible. In particular, if the use of seed variables reveals that the expert is biased (consistently overestimating the failure rates), he could nevertheless provide valuable information, provided the bias is discovered, and the estimates are calibrated by subtracting this bias from his estimate. Such a calibration should only be carried out when there is clear evidence of bias being present (Hokstada et al. 1998: 68);

5. Simplicity

These requirements are needed for achieving scientific credibility. Simplicity, however, is needed for a practicable approach that can gain widespread use. The elicitation process and calculation model should be kept as simple and inexpensive as possible (but accounting for all relevant facts) (Hokstada et al. 1998: 68).

They conclude that 'data emerging from experts should not replace operational data' (Hokstada et al. 1998: 74).

In sum, expert grading for PSPI databases should evolve towards:

1. Explicit and careful discussion of how the panel of experts was assembled (*especially* for in-house grading);
2. Record of the grading process;
3. Controlled gathering of the grades, instead of impressionistic consensus mechanisms;
4. Reliability tests, as in other areas.

1.6. INTRINSIC AMBIGUITY[59]

Social science concepts are full of modifiers (or linguistic hedges, as they are called by the fuzzy set literature; see section 2.3). These modifiers can speak about the degree to which a state of the world is present, or its form, or its relation to other states, or the circumstances of time and place that give it meaning. The more complex the social scientific concept is, the more modifiers appear, and the larger the distance between conceptualisation and operationalisation. Think about words such as democracy, state, fragility, legitimacy and efficiency. PSPIs often include these and similar words not only in the *definiendum* but also in the *definiens*. That is, the dictionary of PSPIs includes complex and multidimensional terms not only on the left side – terms to be defined – but also on the right side – terms that define. It is very frequent that the definition process does not attenuate ambiguity (or does not attenuate it substantially). This makes PSPIs different from other indexes.

PSPIs must therefore resolve whether: (a) it is possible totally to eliminate ambiguity in the operationalisation of the relevant concept; and, (b) if the answer is yes, how this can be done. If the answer is no, then how can the ambiguity be dealt with? None of these questions have been addressed, let alone dealt with, by those who build PSPIs. It is clear that, where a degree of ambiguity is not controlled, both the validity and the reliability of any exercise deteriorate. There are yet larger implications to the problem of ambiguity. PSPIs produce marks and ranks, but also produce tags based on cut-off points: so, for example, does the country pass the test of state failure? If it is below a certain mark, then it is called a 'failed' state. In all existing PSPIs these cut-offs are considered crisp.[60] Using them is standard procedure.[61] It is probable that the program of total disambiguation is unfeasible. Some amount of ambiguity may be reduced by tinkering with the definitions,[62] but once again a high residue of intrinsic ambiguity will remain. As will be seen in chapter

59 This section is short and schematic, in spite of the importance of its subject. Some of the co-authors here have already considered ambiguity in detail in separate publications (Gutiérrez and González 2009; Gutiérrez 2009). In Chapters 2 and 3 here we will be able to see several types of intrinsic ambiguity in action.

60 It is worth remembering that these tags have significant policy implications.

61 Extensive discussion has taken place about where to set the correct cut-off point, for example what defines whether a country is or is not in civil war. Yet this discussion has not yet noted that any cut-off yields a question of what to do with counts that are on the immediate boundary of that cut-off. So, if 20 is my cut-off, what do we do with a score of 19 or 21? The problem is even more severe when the problematic nature of convenience samples, even in death-count estimates, is taken into account (Ball 1996).

62 Polity has gradually done this.

3, some tools taken from fuzzy set theory have started to be introduced in PSPIs, which is a distinct improvement. However, general practice is still to treat the operationalisation of definition as if it were unambiguous and crisp.

1.7. THIRD WAVE INDEXES ARE MORE DIFFICULT TO BUILD

Now we have all the criteria to understand why third wave indexes are more difficult to build:

- **a.** Their underlying concepts are much more complex;
- **b.** It is not easy to identify and control the many potential biases that can affect them;
- **c.** They have more variables;
- **d.** The variables are more heterogeneous, including counts, representative polls, non-representative polls, expert judgment (both in-house and external) and press monitoring, among others;
- **e.** It is not clear that these entities can be numerically manipulated as if they were real numbers;
- **f.** The data tend to be ambiguous and patchy;
- **g.** The aggregation is not grounded on any obvious counting unit – it is not clear in what sense the exercise is a measurement proper;
- **h.** The aggregation function goes from a partially ordered set to a totally ordered set. This added structure is ad hoc, unless some additional set of criteria is produced; but if the added structure is ad hoc, a substantial number of the conclusions (rankings and grades) are arbitrary. Until now, the challenge has not been resolved;
- **i.** Producing an index is a tough task. Difficulties emerge from the number of variables, the proportion of incomparability in the domain, the degree of deterioration of the data, and more.

Some of the problems can be coped with relatively easily. For example, conceptual stretching can be avoided, and the number of boxes and variables can be narrowed down. Dimensionality might be further reduced, though violation of IIA has to be controlled for and a residue of high level dimensionality will inevitably remain.[63] Thus, aggregation should be the main concern for PSPI builders, and more generally for third wave index builders. Measuring and disambiguating will also become fundamental tasks.

The quality of data also remains something to highlight. We have limited our discussion here to the quality of scales based on expert opinion but counts coming from convenience samples are not invulnerable either.[64] It has been shown that they can be wrong with significant magnitude (Ball 1996; Freedman 2005). It is easy to slip from here to a fully fledged nihilistic position of denying the possibility of doing any kind of quantitative exercise based on fragmentary, imperfect and/or corrupted data. As asserted in section 1.2., this position would be inconsistent.[65] It is also untenable. Many disciplines work on fragmentary and incomplete data (take, for example, archaeology and palaeontology). Note, however, that for PSPI builders the situation is symmetrically inverted. In the above disciplines, researchers suffer from a chronic drought of data. PSPI builders suffer from a chronic flood of data and face the challenge of using it adequately and making sense of it. This overflow of information is typical of third wave indexes owing to enhanced capabilities of data capture, storage, and retrieval.

In the next chapter we will discuss how this debate affects the concrete process of the construction and use of PSPIs.

63 Standard methods of dimensionality reduction (for example, factorial or principal components analysis) do not respect IIA. Aggregation by neural networks does not respect it either.

64 Unlike censuses or controlled samples, convenience samples have a potentially high component of non-random error. The majority of counts that are plugged into PSPI databases come from convenience samples.

65 If this position were to be taken, it should then be extended to all kinds of data, qualitative and quantitative, and would result in the denial of producing social knowledge.

TABLES: CHAPTER 1

Table 1.1: Illustrations of the three waves of index building

Source: Developed by authors, main source Bandura 2008

Index	Year of creation	What does it measure	Variables
Gini coefficient	1912	Economic inequality	Share of people with lowest to highest income.
Stock market index	1923	The performance of portfolios such as mutual funds.	Some mutual funds; exchange-traded funds and other funds such as pension funds.
GDP	1942	Overall economic output	Private consumption; gross investment; government spending; exports; imports.
Reuters-CRB Index	1957	Price of commodities	28 commodities, 26 of which were traded on exchanges in the U.S. and Canada, and two cash markets.
Composite Index of National Capability	1963	National power	Total population of country; urban population of country; iron and steel production of country; primary energy consumption; military expenditure; military personnel.
Citation index	1965	Use of (mainly scientific) literature	Number of citations.
Physical quality of life index	1970	Quality of life in a country	Basic literacy rate; infant mortality; life expectancy at age one. All are equally weighted on a 0 to 100 scale
Global Competitiveness Index	1979	Set of institutions, policies and factors that set the sustainable current and medium-term levels of economic prosperity	More than 50 variables, including quality of the institutions, corporate ethics, etc.
Human Rights	1981	The data set contains measures of government human rights practices, not human rights policies or overall human rights conditions (which may be affected by non-state actors)	Extrajudicial killing; disappearance; torture; political imprisonment; freedom of speech; freedom of movement; and women's economic, political and social rights, among others.
Human Development Index	1990	A specific concept of development. Three dimensions: life expectancy, knowledge and education, and standard of living	Four variables (life expectancy; adult literacy; educational enrolment; and transformed GDP per capita).
Corruption Perception Index	1995	Perceived level of public-sector corruption in 180 countries and territories around the world. The CPI is a 'survey of surveys', based on 13 different expert and business surveys	The CPI 2005 draws on 16 different polls and surveys from 10 independent institutions.
Index of economic freedom	1995	The degree of economic freedom in the world's countries	The index scores nations on 10 broad factors of economic freedom using statistics from organisations such as the World Bank, the IMF and the Economist Intelligence Unit.
Polity IV	2003	The Polity conceptual scheme is unique in that it examines concomitant qualities of democratic and autocratic authority in governing institutions, rather than discreet and mutually exclusive forms of governance	Three conceptual categories: executive recruitment; executive constraints and political competition.
Satisfaction with Life Index	2006	Subjective life satisfaction	Scales (opinion polls).
State Fragility Index	2007	It is an assessment of the fragility of countries	'Effectiveness' and 'Legitimacy'.
Social Institutions and Gender Index	2009	Gender Equality	Boxes: family code; physical integrity; civil liberties; property rights. Each box is composed of several variables.

Table 1.2: Dubious criticisms raised against indexes

Source: Developed by authors

Criticism	Reasons why criticism is dubious	Tenable aspect of criticism
They simplify reality.	They should simplify reality.	Researchers should be aware of the limitations of context-free products but simplifications should not go too far.
They compare apples and oranges.	Apples and oranges have abstract qualities that can be worked with arithmetically.	Indexes should be based on measurable questions, and when these are measurable and vague, such circumstance should be taken into account.
They isolate reality.	Once again, this is what they should do. The criticism actually leads to the wrong practice of conceptual stretching.	Isolation should be sensible and theory driven. Conceptual stretching should be avoided because it precludes establishing the associations that are really interesting.
They make no sense, because of the poor quality of the data.	This is a very important problem, which must be treated with the utmost care. However, many data problems can be reasonably solved.	Indexes should be punctilious in their treatment of data.

Table 1.3: Examples of the zeros of Polity

Source: Polity IV

Code	Country	Year	Democ	autoc	polity	polity2
339	Albania	1996	3	3	0	0
490	Congo, Democratic Republic of (Zaire)	1996	-77	-77	-77	0
450	Liberia	1996	-88	-88	-88	0
490	Congo, Democratic Republic of (Zaire)	1997	-77	-77	-77	0
450	Liberia	1997	3	3	0	0
451	Sierra Leone	1997	-77	-77	-77	0
490	Congo, Democratic Republic of (Zaire)	1998	-77	-77	-77	0
404	Guinea-Bissau	1998	-77	-77	-77	0
450	Liberia	1998	3	3	0	0
451	Sierra Leone	1998	-77	-77	-77	0
490	Congo, Democratic Republic of (Zaire)	1999	-77	-77	-77	0
450	Liberia	1999	3	3	0	0
451	Sierra Leone	1999	-77	-77	-77	0
490	Congo, Democratic Republic of (Zaire)	2000	-77	-77	-77	0
450	Liberia	2000	3	3	0	0
451	Sierra Leone	2000	-77	-77	-77	0
940	Solomon Islands	2000	-77	-77	-77	0
439	Burkina Faso	2001	2	2	0	0
516	Burundi	2001	-88	-88	-88	0
581	Comoros	2001	2	2	0	0
490	Congo, Democratic Republic of (Zaire)	2001	-77	-77	-77	0
450	Liberia	2001	3	3	0	0
940	Solomon Islands	2001	-77	-77	-77	0
439	Burkina Faso	2002	2	2	0	0
490	Congo, Democratic Republic of (Zaire)	2002	-77	-77	-77	0
437	Cote D'Ivoire	2002	-77	-77	-77	0
450	Liberia	2002	3	3	0	0
940	Solomon Islands	2002	-77	-77	-77	0
439	Burkina Faso	2003	2	2	0	0

437	Cote D'Ivoire	2003	-77	-77	-77	0
439	Burkina Faso	2004	2	2	0	0
437	Cote D'Ivoire	2004	-77	-77	-77	0
439	Burkina Faso	2005	2	2	0	0
437	Cote D'Ivoire	2005	-77	-77	-77	0

Table 1.4: The growth of incomparable pairs in the number of variables

Source: Authors' own calculations

# of variables	Average of incomparable pairs	Minimum for all the runs	Maximum	Standard deviation
2	0.5004	0.4190	0.5901	0.0245
3	0.7506	0.6793	0.8153	0.0209
4	0.8743	0.8265	0.9158	0.0149
5	0.9371	0.8978	0.9616	0.0097
10	0.9980	0.9936	0.9996	0.0080
20	0.9999	0.9970	1	0.00001
50	1	1	1	0

Table 1.5: Preferences of the Colombians in the 1998 elections

40%	35%	25%
Horacio Serpa	Andrés Pastrana	Noemí Sanín
Noemí Sanín	Noemí Sanín	Andrés Pastrana
Andrés Pastrana	Horacio Serpa	Horacio Serpa

Table 1.6: Regressing normalised data

Source: Authors' own calculations

Raw Data	Variable 1 – Year 1	Variable 2 – Year 1	Variable 1 – Year 2	Variable 2 – Year 2
C1	1	7	5	8
C2	2	6	5	7
C3	3	6	5	6
C4	4	1	6	5
Normalised Data				
C1	0	1	0	1
C2	0.33	0.83	0	0.66
C3	0.66	0.83	0	0.33
C4	1	0	1	0

Table 1.7: Standard normalisation does not respect IIA

Source: Authors' own

Raw data	V1	V2	Average	Rank
C1	5	1	3	1
C2	1	5	3	1
C3	0	2	1	4
C4	4	1	2.5	3
Normalised data				
C1	1	0	0.0444	2
C2	0.2	1	0.1111	1
C3	0	0.25	0.1	4
C4	0.8	0	0.4	3
Normalised data with a new country				
Raw data country	0	10		
C1	1	0.1	0.5	1
C2	0.2	0.5	0.3222	3
C3	0	0.2	0.0555	4
C4	0.8	0.1	0.4	2
C5	0	1	0.5	

2. A GLANCE AT PSPIs

2.1. VALIDITY AND RELIABILITY

2.1.1. RELIABILITY

The reliability of PSPIs and other third wave indexes such as Polity is as yet a basically unexplored theme. Fabra and Ziaja (2009: 29) make the following reflexive observation:

> How do the results of the fragility indices differ? As most indices rely on similar data sources and apply mostly additive aggregation methods (of similar conceptual attributes), one may ask whether the resulting index scores resemble each other as well. Bivariate correlations are used to determine how similar two indices' scores are. The resulting coefficients between indices imply a large degree of similarity: for the most part, they range between 0.7 and 0.9. This is not unusual, however, for macro-social indicators. There are two possible reasons why the scores of fragility indices are highly similar. First, it is possible that indices actually measure their respective concepts with a high degree of accuracy. High correlations would show that the real-world phenomena that are being measured often occur jointly. Second, it is possible that the indices do not measure the concepts accurately. Then, high correlations could be caused by the fact that most indices use highly similar data sources.

Note that this observation is based upon the stability of the result: it is a criterion of similarity. Other authors have claimed that the stability of PSPI ratings is evidence of their reliability (Carmines and Zeller 1979; Nardo et al. 2005). A similar idea appears implicitly in the procedure described by Kaufmann et al. (2009). The rating of a country by certain variables is carried out by two in-house coders. The codebook reports that consensus is achieved very frequently. However, this interpretation of reliability offers no comfort because the PSPI data are actually mainly a codification of expert opinion (see section 2.2).

The rating exercises we present here are reliable in a very precise sense: they capture the beliefs of the coders well, given certain information. Psychological literature is very clear on whether and how this can be done (Shapiro 1997). The ratings act as tests that capture a snapshot perspective of a set of objects (the countries) by a very specific subset of the population (the coders) in a stable fashion. In principle, such reliability does not convey any information whatsoever about 'what is really happening out there'. Indeed, we can assume that coders have much better information than the average person, which makes their beliefs a relevant datum but, on the other hand, may give them strong bias.[66] Their opinion may not match the opinion of the experts on the given country. Several authors have reported wrong or equivocal ratings of PSPIs that go against the consensus of the academic community (for example, Vreeland 2008). These are outliers, of course, and the normal situation will be different: the experts' opinions will be too divided to provide any sort of usable benchmark. Under these circumstances, it is hard to tell if any convergence in the opinions of the coders speaks to the actual situation of the countries. The issues with expert opinion explored above in section 1.5 are highly relevant here.

There is yet another, somewhat less obvious, relevant observation about the percentages of consensus reached by the indexes. These should be calculated over pair-wise orderings, separating the 'comparable' and the 'non-comparable' pairs. The gist of the problem is that the former are fairly obvious and do not say much about the reliability of an index.[67] You do not need one to know that Germany's state is stronger than Colombia's. Yet do the orderings coincide when speaking about Colombia and Venezuela, or Rwanda and Uganda? The underlying structure here is easy to understand: if there is a dominant country (in the sense that it is better than other in all respects), then any aggregation function that deserves that name will put it above the rest. This coincidence implies no surprise, and it is not very interesting.

Thus we will take the analysis a step further and evaluate six of the main PSPIs, to check what would happen to the coincidences in ratings that they report once we control for comparability. What we did was the following:

a. For each comparison between two indexes, we took the countries that were contained in both;

b. We made all possible comparisons for these countries for each index;

c. We calculated the proportion of common rankings.

The results are shown in Table 2.1. Note that this is a quite straightforward test because we are evaluating the stability of the *rankings* of the indexes, independently of the concrete scores that they attribute to the countries. We would expect the rankings to be very stable – more even than scores – because they evaluate the relative position of the countries. They are not.[68] As shown, rates of agreement are far from spectacular. There are indeed many ranking reversals, some of which are over 20%.[69] We stress that:

a. This is not an R^2. An R2 of 80% should be considered quite high. This is the percentage of stability of the relative ordering of pairs. This means that the capacity of predicting the ordering of one index based on another one is 80%: not high.

b. Our test is the easiest one with respect to ordering. If we had taken the orderings of triples, quadruples … n-tuples, the degree of stability would have decreased dramatically.

We show how these reversals work in practice in Table 2.2. We take eight state and governance indexes and six countries across Africa,

66 Bias may emerge because they belong, more or less to a specific (global) social sector, or because they face organisational constraints, among other reasons.

67 Besides, by definition any index has to be Paretian/monotonous (Beliakov et al. 2007).

68 Indeed, it confirms instead our claim about the essential instability of rankings of incomparable cases for functions that go from posets to partially ordered sets (see section 4.2.4). This is not surprising, as the latter is a formal result, which suggests that the instability of rankings should grow with the number of variables and boxes.

69 20% of ranking reversals should be considered quite high.

Asia and Latin America that appear frequently in the 'state failure' literature. The values in the table are the score that the index attributes the country, and (in parentheses), is the relative position that the country holds within the set of six countries (where **(1)** is the most failed, and **(6)** the least). Two positions are relatively fixed: DRC is here the worst performer and South Africa the best. The rankings relative to those countries are pretty stable, but the remaining four countries vary up and down within the intermediate positions.

Some PSPIs behave well with respect to other basic reliability tests. The most elementary and popular of all is the so-called 'internal reliability', or *Cronbach alpha*, which measures how consistent the construct is. In other words, it asks: are the variables really measuring the same construct? It is defined as:

$$\alpha = \left(k/(k-1)\right)\left(1-\frac{\Sigma\sigma_i^2}{\sigma_y^2}\right) \quad \text{f(2.1.1)}$$

Where:

K is the number of variables;

$\Sigma\sigma^2_i$ is the sum of the variances of the variables;

$\sigma y2$ is the variance of the total sum of the variables.

The results when this is applied to some of the most important indexes are evident in the matrix of correlations of the Index of African Governance (IAG) (Table 2.3).

2.1.2. VALIDITY

The main problem with the validity of indexes is related to the operationalisation of the concept. Does the concept express well what it is supposedly operationalising? There is already a wealth of literature about the validity of third wave indexes in the social sciences (on Polity see, for example, Munck and Verkuilen 2002; on the quantification of internal conflict, see Cramer 2006; specifically on PSPIs validity, see Cammack et al. 2006 and Di John 2008). Given this, we focus here on exploring the sources of lack of validity in PSPIs.

- *At the definitional level*

It is frequent that the concept is not captured by the operationalisation, and that the latter – especially when it is crafted during a multi-step procedure – is internally inconsistent. For example, the CPIA defines fragility thus:

> Fragile states is the term used for countries facing particularly severe development challenges such as weak institutional capacity, poor governance, political instability, and frequently on-going violence or the legacy effects of past severe conflict (World Bank 2007: 2).

Fragility here is understood to be a 'severe developmental challenge'. Now let us see how this definition is transformed into variables. Table 2.4 illustrates the mismatch. According to the detailed definition reproduced above, ongoing or lagged violent conflict is one of the possible contextual components of fragility: but there is no variable to measure this.[70] Instead, structural policies, which do not appear in the definition, appear in the 'operationalisation' – the breakdown of how the concept is to be actually measured. These might be considered synonymous with 'good governance', but if so then this should first be tested empirically. The qualitative assessments of the equation that we have at hand (structural policy = good governance) invites scepticism. Finally, the World Bank, like any other institution, has changed its vision of what a good policy is.

The most striking case of problematic definitional validity, though, is reported by Di John (2007). A sophisticated concept is constructed and then dropped at exactly the moment that it should be made operational. The result is an exercise that is in many respects interesting but marred by suspect validity.

- *Conceptual stretching.*

The definition of state weakness according to the *Index of State Weakness* is the following:

> We define weak states as countries that lack the essential capacity and/or will to fulfil four sets of critical government responsibilities: fostering an environment conducive to *sustainable and equitable* economic growth; establishing and *maintaining* legitimate, transparent, and accountable political institutions; securing their populations from violent conflict and controlling their territory; and meeting of the basic human needs of their population (Rice and Patrick 2008: 3. Emphasis added).

'Equitable and sustainable growth' are very open terms and it is difficult to see where they fit in the theory of the state. They might be correlates of state strength (for a historical period or for a region), but why make them part of the definition? An analogous comment can be made, with even more emphasis, about the demand that strong states be 'transparent' and 'legitimate' (more about this in section 2.3.2).

The set of boxes and of variables of the PSPI designed and conducted by the Fund for Peace is shown in Table 2.5. This is a rather unwieldy set of variables, and the question of what they have in common suggests itself. As Cammack et al. (2006) observe, there are several sources that feed this type of definition and operational indicator:

a. Analytic concerns proper (interventions of other states, or the incapacity of the state to fulfil its putative functions);

b. Hunches about possible causes (which should then be put out of the definition and as a covariable in a

70 We stress: it is not that violence is associated with fragility, but that violence – according to the definition that this indicator is designed to measure – *defines* fragility.

regression). For example: demographic pressures, or uneven economic development along group lines;

c. Normative concerns (for example, violation of human rights or lack of transparency);

d. Policy concerns (for example, massive migration caused by different factors from the South to the North).

All these are cases of conceptual stretching. The meaning of state fragility (or weakness) is broadened so much that it includes putative causes and consequences. This would be equivalent to including in the definition of lung cancer, together with the phenomenon proper (uncontrolled growth of cells in the lung), indicators such as cigarette smoking (an independent variable that explains an important part of its incidence) and family crisis (a possible consequence).

There is yet another issue with validity. The very notion of fragility is, inevitably, grounded on certain prototypes that are observed in the world. These are typically strong (non-fragile) and weak (fragile) states (for example, Switzerland and Afghanistan). Any good index should put Switzerland very near the non-fragile end of the spectrum, and Afghanistan near the fragile end. This, in itself, is a form of validation. However, for the reasons discussed in section 1.3 and at the beginning of this section, we have no equivalent to validate the ranking of the 'lost middle'. Furthermore, if we introduce into our definition the notion of a continuum, as all index builders do, which goes from great strength to the breakdown of the state, then we face a new quandary: there are many very different forms of state disruption or termination. Consider the following situations:

a. The state stops working (for example, Afghanistan, DRC, Albania (1996–1997));

b. The state disappears as a de jure recipient of sovereignty, but without major social disorder (as in the breakdown of Czechoslovakia);

c. The state disappears as a de jure recipient of sovereignty and there is a fair amount of disruption, but some entity continues in control (such as East Germany in the aftermath of the fall of the wall (Diewald et al. 2006));

d. The state disappears as a de jure recipient of sovereignty and in some regions and areas it ceases to work, but the state does not fully break down (such as in the former Soviet Union);

e. A state apparently works very well, but its status as de jure recipient of sovereignty is hotly contested by significant sectors of the population (many examples can be given here, including Canada, Belgium, Spain,[71] Iraq);

f. The state remains as a de jure recipient of sovereignty, but it suffers a major disruption by an external invasion (Iraq, Afghanistan).

From the Darwinian point of view prevalent in the theories of the state (that the main outcome is survival or total unravelling) these situations have strong commonalities. Nice death is better than horrible death, but it is still death after all. The main point is that, despite describing states of the world in which the state – as a juridical entity – vanishes, the above situations *do not* constitute a continuum. As we do throughout this text, we insist on separating two types of orders: total order and partial order. In the latter there is some hierarchy, but not all comparisons can be established only in hierarchical terms. When there is only partial order, classification goes beyond ranking.

- *A special case of stretching: democracy and legitimacy*

As seen in Table 2.6, several indexes incorporate political dimensions into the definition of fragility. Of course, the relationship between a certain type of political regime (normatively desirable) and fragility is a very important empirical question but we will not be able to answer it if we include democracy in the definition of fragility. Introducing legitimacy adds an additional problem. Legitimacy is at least as complicated a concept as state fragility. Nobody has demonstrated that to observe the former is easier than to observe the latter. If we are able to mark legitimacy directly, why not mark fragility directly as well? Why bother with an index? Legitimacy is a typical example of a variable that does not correspond to an observable, and that hence can become a container of any source of bias, or simply of noise.[72]

2.1.3. CONCLUSIONS

Many of the definitions underlying PSPIs are not oriented by theories. The main result of this is conceptual stretching (Sartori 1984). Conceptual stretching is a major flaw because it prevents isolation, which is a key step in large N comparative exercises, and one of the main desirable outcomes of sensible quantification. Hence, conceptual stretching blocks the possibility of establishing the correlates between different social goods and bads. The co-authors of this text, for example, consider that 'democracy' (in a purely formal sense) and 'state strength' are both desirable. But if we subsume them under a common tag we will never be able to discern what the real covariation between one and the other is. We would certainly want to know this. The assumption is that all good things come together (Hirschman 1981), but this inevitably undermines validity (Putzel 1997). In terms of policy, conceptual stretching results in a laundry list of demands, and this has a potentially crippling effect on states of non-developed countries. Another typical erroneous operation is conceptual replacement (where one starts out with one notion, only to replace it with another one in the process of operationalisation). This may coexist with inconsistent operationalisation.

71 Secession, after all, is a very real issue today.

72 It would be interesting also to do reliability tests on the gradings of legitimacy.

The main concern of index builders should be to maintain validity. The PSPI field has recently started to pay attention to reliability issues but it is still not clearly established. Achieving informal consensus between in-house coders is still no proof of genuine reliability. Here, grading acts as a test to elicit beliefs from coders, not as a way of gauging the 'real situation' of the countries. Arriving at similar rankings may have more import, but when one controls for comparability, the stability of the rankings becomes moot. As yet, we have no evidence that collection of ordinal data based on expert opinion is reliable in any meaningful sense.

2.2. THE DATA SETS

In this section, we briefly describe how the PSPI datasets are built.

2.2.1. WHAT CONCEPT IS BEING OPERATIONALISED?

Table 2.7 shows the definition of fragility by the major PSPI databases. As asserted in previous sections, the BTI operationalisation is theory-oriented but in the other datasets putative causes, consequences, correlates and definitional aspects are collapsed into a single catch-all concept. Besides, as was also observed previously, not all of the aspects that are highlighted in the definition have a correlate in the operationalisation indicators, and vice versa.

On the other hand, the datasets are outstanding for the exuberance of the information they contain. The CIFP contains up to 73 variables; the lower bound is of course the BTI, with two. All in all, however, the indexes are supported by very rich information.

2.2.2. THE NATURE OF THE DATA

From the point of view of the data they store, PSPI datasets are also quite heterogeneous. For example, the BTI deals solely with expert opinion. As the Bertelsmann Foundation explain, 'the indicator is carried out in consultation with an interdisciplinary board of experts … The country reports are generally written by external experts for each state and then reviewed by a second expert from each respective state' (Bertelsmann 2010: 5). Nothing is said about the actual selection of experts.

So how does the BTI data-collection process work? The values of the variable 'monopoly on the use of force' come from a poll sent to the set of experts with the following question: 'to what extent does the state's monopoly on the use of force cover the entire territory?' The codification of the answers delivered by the experts is then done according to the following scale of 10–1 (Bertelsmann 2010).

- Grades 10 and 9: There is virtually no competition with the state's monopoly on the use of force throughout the entire territory;
- Grades 8, 7 and 6: The state's monopoly on the use of force is established nationwide in principle but it is threatened (or challenged) by organisations in territorial enclaves (guerrillas, mafias, clans);
- Grades 5, 4 and 3: The state's monopoly on the use of force is established in key parts of the country, but there are organisations (guerrillas, paramilitaries, clans) that are able to usurp the state's monopoly on the use of force in large areas of territory;
- Grades 2 and 1: There is no state monopoly on the use of force. Instead, there is anarchy, civil war, a clan oligopoly or equivalent.

The other variable is constructed similarly. Other databases that are entirely based upon expert opinion and in-house coding are the Country Policy and Institutional Assessment (CPIA) and the State Fragility Index. PSPIs also include count variables, or ordinal variables based on counts. Notable for its innovation is the Conflict Assessment System Tool (CAST), used by the Fund for Peace, which consists of an automated screening of 1,100 sources of the world's media (Baker, 2006).

Table 2.8 (based on Fabra and Ziaja 2009) presents the overall landscape of PSPI sources and data types. Table 2.9 describes the variables used by PSPIs and reveals that the majority of them lean on expert opinion to produce a significant portion of their variables (of six, three are totally expert-based). No discussion whatsoever of the quality of this kind of data is advanced by the host organisations for these PSPIs. It is as if the equivalence of expert assessment and hard data is obvious and uncontroversial. Other sources also escape critical scrutiny. It is almost a sin to criticise developments such as the CAST system, which is a kind of symbol of the powerful way in which third wave indexes can profit from our current informational overflow. But it is necessary to reflect in more depth about what the data mining in 11000 periodicals can return: some countries are over-represented, and others severely under-represented in the media. How should this and other issues be dealt with? What does the system do with contradictory information? If this black box is not opened, it is difficult to use this as a source of valid and reliable information.

2.2.3. MISSING VALUES

One of the advantages of expert-opinion variables is that they avoid the problem of missing values. So the missing data problem does not exist for the BTI, for example. The Index of State Weakness for the Developing World reports 3.5% missing data. It does not impute them, but rather aggregates each country vector with the available data. The State Fragility Index also has a small percentage of missing values, and imputes them by averaging within the same category and the same country. The Failed State Task Force reports:

> We created a version of the data set that imputes missing values so that composite factor scores can be calculated. The imputation was performed according to the following decision rule: for a missing value, assign an imputed value equal to the average score within the same category and the same country. For the 1999 data, 83 missing values were imputed out of 2,226 values; only 3.7 per cent of the information used to create the 1999 capacity index is based on imputation. For the 1990 data the corresponding number was 118 (5.2 per cent) (Goldstone et al. 2000: 70).

The World Bank CPIA imputes missing values, 'based on the relationship of the sum of available data to the total in the year of the previous estimate' (World Bank 2010). It also refrains from, imputing when 'missing data account for more than a third of the total in a given year' (World Bank 2010). Other imputation techniques are based on proxies or on assignment by rules like the following: 'for a missing value, assign an imputed value equal to the average score within the same category and the same country' (Goldstone et al., 2000:70).

2.2.4. CONCLUSIONS

The main problem of PSPI dataset building is validity. Although the use of conceptual substitution, as by the Failed States Task Force – defining a concept, and then using a different operational definition as reported by Di John (2008) – is an outlier, too many things are still put into the stew of fragility. In particular, there appears to be a systematic confusion between correlates and definitional characteristics of a phenomenon or state of the world. Some PSPIs explicitly search for variables that correlate highly with fragility to include them in the definition. We cannot agree with this practice, which replaces theoretical reflection by number crunching. Resorting to an example presented in previous sections, this would be like including 'smoking heavily' in the definition of lung cancer.

Another key issue is the quality of data based on expert assessments. In its present form, we can hardly consider a variable built upon expert assessment as bona fide data. The way in which the opinions are gathered and aggregated is not reported; nor is the choice of the experts (and it seems to be anything but random). What we can glean about the questionnaires is that they are heavily hedged, and there is no instruction as to how to deal with the (potentially very high percentage of) ambiguous situations. The reliability of other sources such as the media is open to discussion. An area that also allows room for improvement is the imputation of missing data (when it is necessary).

2.3. AMBIGUITY

In this third section of Chapter 2 we illustrate three types of ambiguity that can easily be found in PSPIs: linguistic hedges, broad terms and ad hoc cut-off points.

2.3.1. HEDGED DEFINITIONS

Definitions of concepts such as state fragility or democracy are, and have to be, heavily 'hedged'. To be clear, we display in Table 2.10 some of the definitions, highlighting both the 'hedges' and the broad terms. By 'hedges' we mean modifiers that express intensity or modality. By 'broad terms' we mean terms that are as complicated as the one being defined, and where the status of 'observable' is at least suspect.

Hedges are inevitable in the social sciences. The problem is that no tools to deal with them have been designed by extant PSPIs (Gutiérrez 2009). Thus, when, for example, the sentence 'inability to provide reasonable public services' is included in a definition, the decision of what is 'reasonable' or not remains in the hands of the coders.[73]

2.3.2. LEGITIMACY AND FRAGILITY

Unlike hedges, broad terms may be inevitable. In general, no operational definition should contain concepts that are as complicated and as (directly) unobservable as the original. Consider, for example, what happens when we try to operationalise 'legitimacy'. As seen in sections 2.1 and 2.2, legitimacy is integrated as part of the definition of fragility in several indexes. How can we capture it? The following definition gives some tips:

> Legitimacy refers to the ability of a state to command public loyalty to the governing regime, and to generate domestic support for that government's legislation and policy. Such support must be created through a voluntary and reciprocal arrangement of effective governance and citizenship founded upon principles of government selection and succession that are recognized both locally and internationally. States in which the ruling regime lacks either broad and voluntary domestic support or general international recognition suffer a lack of legitimacy. Such states face significant difficulties in maintaining peaceful relations between and among various communities within the state; any security that exists is likely the result of coercion rather than popular consent. As a result, such states are inherently vulnerable to internal upheaval and are likely to remain fragile so long as legitimacy remains wanting (Carment et al. 2006: 7).

This definition introduces new sources of ambiguity. It collapses two types of legitimacy – internal and external – which, in principle, should not be done. If a state is able to gather broad voluntary internal support but not international recognition, or vice versa,[74] then what mark does it deserve? Neither the coder, to our knowledge, nor the user of the index really knows.

73 Symptomatically, the hedges of the State Failure Task Force can be easily operationalised. The Task Force has put a lot of stress on formal rigour, but at the same time has replaced its original concept by a different, and much simpler, one.

74 There are hundreds of historical cases where this happened. Additionally, domestic or international support can be highly fractured. How will these cases be coded?

Thus, the CIFP legitimacy marks are rather difficult to interpret. A sample is shown in Table 2.11. The marks hardly correspond to observables. For example, DRC has a score of 6.15 legitimacy, while North Korea has a score of 7.14 (here, higher is worse). Yet the DRC state has an ongoing rebellion against it. So in what sense is the North Korea state more illegitimate than the Congolese? Something similar can be said about Somalia. Iran is a typical case of fragmented support. The regime has a broad mass of staunch supporters but at the same time confronts widespread opposition that has not become an open rebellion. In what sense does it have a high level of illegitimacy?

Let us now consider the way in which the State Fragility Index conceptualises legitimacy (or a lack thereof). Within the State Fragility Index there are four types of legitimacy (security, political, economic and social). Political legitimacy includes leadership characteristics and factionalism, salience of elite ethnicity and polity fragmentation, and exclusionary ideology of the political leadership. It is not obvious how these items are to be assessed: the 'exclusionary' nature of an 'ideology' is something open to debate, even among experts. How are these scales built and marked?

The Fund for Peace uses legitimacy in a dual characterisation of states (criminalisation and/or delegitimisation). Once again, marking two distinct characteristics (for example, making a numerical characterisation of two different states of the world at the same time) introduces inevitable ambiguity. Since these notions, criminalisation and delegitimisation, are not coterminous and a state can indeed have high levels of criminalisation without being illegitimate and vice versa, it is difficult to know what the mark really means. Table 2.12 displays the grading for North Korea for the years 2007, 2008 and 2009. Some grades seem fairly sensible. Indeed, according to the specialised literature, the security apparatus in North Korea constitutes a state within the state (Kihl 2006). Human rights are violated massively, and there is no liberty of expression. However, it is difficult to interpret the high mark that the 'rise of factionalized elites' receives. 'Seeking group grievance or group paranoia' is marked with one number despite including two distinct ideas, both of which are rather hazy. We simply do not know the evidence that coders can use to give a high or low mark.

2.3.3. AD HOC CUT-OFF POINTS

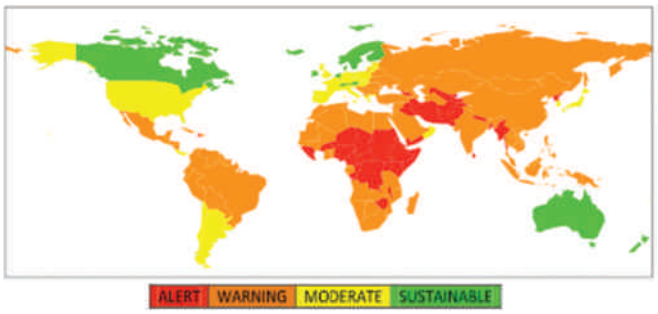
In addition to ranks and marks, some PSPIs produce a list of countries that perform particularly poorly. Some establish ad hoc thresholds below which a state should be considered fragile or failing. Others base their threshold on the database itself, which seems much better (see Table 2.13).

This is an example of the problem of cut-off points, which is widespread. The majority of PSPIs rely heavily on expert grading based on ordinal scales. But what is the difference between a 3 and a 4 according to expert A?[75] What will he/she do when scores are close to the boundaries or at the limit? Will his/her grading criteria coincide with those of expert B?

A point that should be flagged here is that very small changes in the cut-off points produce significant changes in classification. The Fund for Peace, for example, classifies states in four broad categories, depending on their final score (which is the sum of its four boxes, see Table 2.14).

Figure 2.1: Original cut-offs for Fund For Peace (2009)

Source: Authors' own based upon Fund for Peace 2010



Suppose we choose instead to place the cut-offs at the quartiles of the data (which in fact seems more logical). Then, based upon 2007 Fund for Peace data, we would have:

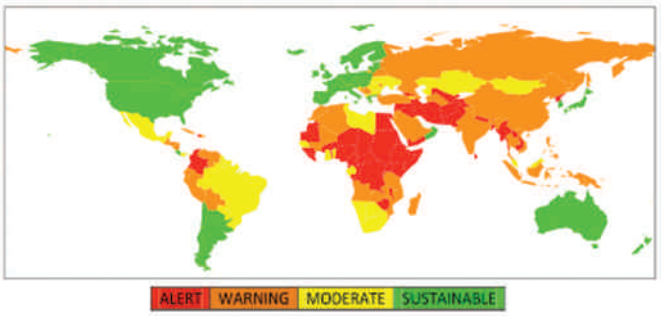
	Fund for Peace Index numerical value
Percentile 25	57.2
Percentile 50	75.9
Percentile 75	86.9

We show the new map in Figure 2.2. The differences are significant. Brazil and South Africa improve from 'warning states' to moderate ones. All of Western Europe becomes 'sustainable', whereas before large areas of it were only 'moderate'. Colombia, Laos and Cambodia are degraded, and get an 'alert'.

Maps are powerful heuristic devices, but here a substantial portion of the message was an artefact of ad hoc chosen cut-offs.

Figure 2.2: Indicator Fund for Peace (2009)

Source: Authors' own based upon Fund for Peace 2010



75 Note that some ordinal scales are invulnerable to these questions. For example, the World Bank ordinal GDP is based on intervals, not on marks by experts.

2.4. NORMALISATION AND AGGREGATION

In this section we will discuss some of the problems related to normalisation and aggregation. We show what would happen to marks and ranks if small changes in the parameters were chosen in an ad hoc way.

2.4.1. NORMALISATION

Table 2.15 describes the normalisation procedures of a subset of major PSPIs. Some of them do not need normalisation because their data is introduced into the database in a directly usable form. The BTI and the CPIA, where data is described by a number in the same ordinal scale (1-10 and 1-6 respectively), pertain to this category. Others do not explain how normalisation is undertaken.

We discuss here in some detail the procedure of an index that does normalise, and explain how this is done. The Index of State Weakness in the Developing World normalises and aggregates its data in the following way:

> The Index is based on 5 indicators in each basket. Taken together, the 20 indicators yield a balanced picture of how developing countries perform or fail to perform along multiple dimensions. Within each basket, the indicator scores are standardized and aggregated, creating individual indicator and basket scores ranging from 0.0 (worst) to 10.0 (best). The 4 basket scores are then averaged to obtain an overall score for state weakness, ranging from just above 0 to just short of a perfect 10, to produce a ranking of states on the basis of their relative weakness (Rice and Patrick 2008: 8).

The variables are re-scaled using the following formula:

$$\frac{x - \min(Vi)}{\max(Vi) - \min(Vi)} \qquad \text{(f.2.4.1),}$$

where *min(vi)* is the minimum value for variable *i, max(vi)* is the maximum value for variable *i*, and *x* is the value of the given country for variable *i*. As seen in section 1.4, this is the re-scaling normalisation that may violate IIA and can eventually change the sign of the regression (with respect to the raw data). Does this actually happen? PSPI builders have not cared to evaluate this with their own data, or experimentally.

Following this re-scaling, the variables are averaged within each basket and this output is re-scaled once again in the following way:

$$10 * \left[\frac{(x_{ki1}) - \min\{x_{ki1}\}}{\max\{x_{ki1}\} - \min\{x_{ki1}\}}\right] \qquad \text{(f.2.4.2)}$$

See Table 2.16 for examples of the operation for a subset of the countries evaluated by the index. The five variables that compose the economic basket are averaged. The result is shown in the column 'Average of the variables'. Then the minimum and maximum figures in this column are found and the equation *f.2.4.2* is applied. After this has been done, the four baskets are averaged once more.

It is observed in Table 2.16 that each of the baskets has different minimum and maximum figures. This was to be expected. Thus, the normalisation process inadvertently changes the relative weights of each. Additionally, the basket procedure does not respect IIA. This is not just a theoretical issue. Suppose we have a subset of countries and we improve the performance of the best one so that the maximum of the averaged column (the aggregation of the basket) changes. As is shown in the table, the hierarchy then changes. Croatia improves while Chile falls. Changing the maximum value of the Economic Basket, where Hungary is still the best in this dimension (preserving order), but modifying the maximum value from 7.94 to 10, changes the final order of hierarchy. This is not surprising. As seen in Chapter 1, we know that this may happen – and not only between countries but also between years.[76]

2.4.2. AGGREGATION

Table 2.19 describes the aggregation function of a selection of relevant PSPIs. We find a major flaw: all of the functions they utilise are compensatory but only one index explains the way that the weights used to aggregate were obtained (this is the Political Instability Ledger and the explanation is provided not in the process of aggregation but during the explanation of a statistical exercise (see Hewitt et al. 2010)). So it seems that the imputation of these weights is very similar to an act of magic.

We can only conjecture what these weights mean. First, since in some of the indexes all the variables and/or baskets have the same weight,[77] then the weighting can be interpreted as an implicit appeal to Laplace's principle of insufficient reason. In reality, however, in this context there are no grounds to make such an appeal, and in no codebook is any made. Actually, we have every reason to believe that the PSPI variables – if they are commensurable, ie, if weights that express their relative importance with respect to state fragility can be attributed to them – should be weighted differentially. Women's rights and the degree of state repression act in a very different manner as regards fragility but we do not have a theory or empirical evidence that would enable us to say which variable is more important, let alone how much so. The alternative is that these weights are simply ad hoc parameters. Given that the codebooks do not explain the weights, and do not

76 The State Fragility Index is quite robust technically but not so clear conceptually. By suddenly introducing an income criterion in the definition of fragility, it seems to stretch the concept (making it impossible in the process to regress income on fragility). See Table 2.4.

77 Though imputing the same weights to baskets may change the weights of the variables. Thus, verbally it may be asserted that the variables are equally important, but operationally they work differently. See below.

and could not appeal to Laplace's principle (Dembo and Zeitouni 1998), we have strong reason to believe that this is the case.

Ad hoc weights have no possibility of altering the ranking of pairs of comparable cases because all of the aggregation functions used by PSPIs are Pareto and monotonous. In particular, simple averages fulfil these extremely desirable conditions. However, ad hoc weights can affect the rankings of non-comparable countries. The strength of the effect of this might vary. In a database such as BTI, where there are only two ordinal variables, one would not expect to find too many of these. We can see in Table 2.18 that, in 2008, 90.17% of the pairs in the BTI were comparable. Having ad hoc weights for the two variables will affect only 10% of the rankings, which is not negligible but not overwhelming either. However, as seen in section 1.3, the problem of order grows 'naturally' and very fast in the number of variables. For an index such as the Fund for Peace, nearly two thirds of the paired comparisons are not comparable in the domain (see Table 2.18). The CPIA is far from the baroque end of the spectrum of PSPIs, but two thirds of its paired comparisons also lack an 'obvious order' in the domain. This, indeed, corresponds to the general case.[78] This means that a substantial number of the rankings of these indexes are an artefact of ad hoc parameters.

Let us now illustrate how all this works. Take the CPIA, which in many regards is a very carefully built index (Table 2.20). The substantive interpretation of its weights is not evident. A critic observes that:

> The CPIA gives equal weight to all its constituent elements, although some of them may have much more bearing on state building and peace building than others. It could be argued, for example, that improvements in the efficiency of resource mobilization or in the equity of public resource use should take precedence over some macroeconomic indicators in the CPIA if state building is a key objective (World Bank 2006: 45).

The CPIA aggregation function is the following:

$$\frac{\frac{x_{1i}+x_{2i}+x_{3i}}{3}+\frac{x_{4i}+x_{5i}+x_{6i}}{3}+\frac{x_{7i}+x_{8i}+x_{9i}+x_{10\,i}+x_{11\,i}}{5}+\frac{x_{12\,i}+x_{13\,i}+x_{14\,i}+x_{15\,i}+x_{16\,i}}{5}}{4} \qquad \text{(f.2.4.3)},$$

Where the *x*'s are the variables and the *'4'* corresponds to the number of boxes. The CPIA, in effect, is an average of averages, so that boxes have equal weights but variables do not (because each box has a different number of variables).[79] Is there some substantive interpretation of this distribution of weights, where variables are attributed unequal values and boxes identical ones?

Expanding *f.2.4.3.* we get the following aggregation function:

F1(xi)=

$$\frac{1}{12}x_1+\frac{1}{12}x_2+\frac{1}{12}x_3+\frac{1}{12}x_4+\frac{1}{12}x_5+\frac{1}{12}x_6+\frac{1}{20}x_7+\frac{1}{20}x_8+\frac{1}{20}x_9+\frac{1}{20}x_{10}+\frac{1}{20}x_{11}+\frac{1}{20}x_{12}+\frac{1}{20}x_{13}+\frac{1}{20}x_{14}+\frac{1}{20}x_{15}+\frac{1}{20} \qquad \text{(f.2.4.4)},$$

Suppose somebody feels quizzical about giving different weights to the variables and equal weights to the boxes, and proposes an aggregation function that works the other way around. In this case, variables would have the same values and boxes would be allowed to vary. Is one of these functions ostensibly superior to the other? No, although if custom has any value, the second should be preferred. Several other indexes simply add or average the variables, not over the boxes. For example, BTI and the State Weakness Index do this. The new aggregation function will look like this:

78 We have made the analysis for every single database, including ours, but do not include all the results to avoid becoming repetitious.

79 The variables placed in boxes with a small number of variables are worth more.

$$F2(x_i) = \frac{1}{16}x_1 + \frac{1}{16}x_2 + \frac{1}{16}x_3 + \frac{1}{16}x_4 + \frac{1}{16}x_5 + \frac{1}{16}x_6 + \frac{1}{16}x_7 + \frac{1}{16}x_8 + \frac{1}{16}x_9 + \frac{1}{16}x_{10} + \frac{1}{16}x_{11} + \frac{1}{16}x_{12} + \frac{1}{16}x_{13} + \frac{1}{16}x_{14} + \frac{1}{16}x_{15} + \frac{1}{16}x_{16} \quad \text{(f.2.4.5)}$$

This innocent looking change alters the CPIA rankings. The change is not marginal. An example is given in Table 2.21.

The relative positions of Ivory Coast, Togo, Eritrea and Sudan have changed. Thus, the fact that changes in the weights can alter the rankings is not purely 'theoretical'. 1.92% of the rankings of the pairs are reversed in 2005 when the weights are changed from 1/12 and 1/20 to 1/16. Note that the fact that the rank (and *a fortiori* the mark) changes in a non-negligible manner with small and 'natural' perturbations of the weights is completely independent from the rank reversals between indexes discussed in section 2.1. Note as well that some indexes give different values to their variables, without explaining where these came from.

The 'official' weights attributed to variables and boxes in the indexes are not the real weights. These come from the composition between the correlation between variables (and boxes, when these are a unit of aggregation) and the official weights. As we saw in section 2.1., the correlation between variables in some PSPIs is high, a characteristic that is actually positive, but which should be made explicit during the aggregation process. The non-orthogonality of all the PSPI variables is consequential for the aggregation process. This is yet another key issue that has not been acknowledged explicitly. Since correlations between variables vary, they act as 'hidden weights' that have to be composed with the 'official' ones. Table 2.22 shows the correlations between variables in the State Fragility Index. As is evident, they vary a lot. There are variables that are almost orthogonal (social legitimacy and social effectiveness), some that have a decent correlation (of above 0.6), and yet others that have a negative correlation. All this changes the weights and also the rankings.

Our final observation as regards weights relates to their arithmetic means, which are not only compensatory but *fully* compensatory. This means that a complete catastrophe in one variable (box) can be outweighed by improvements in the other boxes and,

> As a consequence, all indices allow their conceptual components to partially compensate for each other. Assuming an index of two equally weighted dimensions (eg, security and political), absolute failure in the first dimension would still allow a country to reach 50 per cent on the overall score if it performs optimally in the second dimension. In other words, no index assumes any function of the state to be a necessary condition – considering the strong theoretical focus on the monopoly of violence, this finding is rather surprising (Fabra and Ziaja 2009: 29).

This is indeed surprising. It implies that, if monopoly of force is entirely lost, a good enough economic opportunity might make up for this loss. This, of course, is not a finding but an assumption, and a heroic one at that. Once more, this assumption is neither discussed nor made explicit at any stage.

To summarise this chapter: the majority of PSPIs do not need to normalise because they aggregate over identical ordinal scales. Aggregation, in turn, is full of problems. We have focused here on the main one, that all of them use compensatory functions despite the fact that only one (the Political Instability Ledger) grounds the weights on the data. In general, there seems to be no awareness at all about the crucial nature of the following decisions: (a) choosing a compensatory aggregation function; and (b) attributing specific weights to variables or boxes.

We showed then that:

a. There is no substantive interpretation of the weights. Their meaning remains one of the PSPI unsolved mysteries;

b. Relatively small changes in the weights produce significant rank reversals (and of course, very substantial alterations of the grading);

c. It is not clear what criteria are utilised to aggregate within and between boxes;

d. The non-orthogonality between the variables of the index has not been taken into account. We can ask: which variable weights are correct – the official ones, or those that we get after composing them with the weights?

TABLES: CHAPTER 2

Table 2.1: Proportion of coincident pair-wise rankings of non-comparable countries

Source: Authors' own calculations

Indicator	Bertelsmann	CIFP	CPIA	Failed State Index	State Weakness	State Fragility Index
Bertelsmann	1.000	0.667	0.708	0.762	0.775	0.781
CIFP	0.667	1.000	0.710	0.756	0.887	0.873
CPIA	0.708	0.710	1.000	0.680	0.760	0.693
Failed State Index	0.762	0.756	0.680	1.000	0.7804	0.815
State Weakness	0.775	0.887	0.760	0.7804	1.000	0.869
State Fragility Index	0.781	0.873	0.693	0.815	0.869	1.000

Table 2.2: Definitional Level

Source: Authors' own calculations

	State Fragility Index(2007)	BTI (2008)	CIFP (2007)	CPIA (2008)	Failed States Index (2007)	Index of African Governance (2007)	Index of State Weakness (2008)	Political Instability Index (2007)
Democratic Republic of Congo	20 (1)	2.5 (1)	6.5 (1)	2.73 (1)	105.5 (2)	37.3 (1)	1.67 (1)	7.2 (2)
Colombia	11 (3)	4.5 (2)	5.24 (4)		89.7 (4)		5.63 (4)	6 (3)
North Korea	8 (4)	9 (6)	5.88 (3)		97.7 (3)		3.87 (3)	3.7 (6)
Venezuela	10 (5)	7.5 (4)	5.13 (5)		79.8 (5)		6.33 (5)	4.3 (4)
Zimbabwe	16 (2)	5 (3)	5.92 (2)	1.4 (2)	110.1 (1)	47.3 (2)	3.44 (2)	8.8 (1)
South Africa	9 (5)	8.5 (5)	4.84 (6)		57.4 (6)	68.4 (3)	7.5 (6)	4 (5)

Table 2.3: Matrix of correlations of the Index of African Governance (IAG)

Source: Authors' own calculations

	Safety & Security	Rule of Law	Participation & HR	Sust.Econ.Opp	Human Development
Safety & Security	1				
Rule of Law	0.527	1			
Participation & HR	0.291	0.537	1		
Sust.Econ.Opp	0.449	0.598	0.177	1	
Human Development	0.451	0.586	0.098	0.755	1

Table 2.4: The CPIA operationalisation

Source: World Bank (2007)

Definitional dimensions of fragility		Operational dimensions of fragility
• Weak Institutional capacity • Poor governance • Political instability • Ongoing violence or the legacy effects of past severe conflict	versus	• Economic management • Structural policies • Policies for social inclusion/equity • Public sector management and institutions

Table 2.5: The boxes and the variables of The Fund For Peace

Source: Fund For Peace (2010)

Social indicators	Economic indicators	Political indicators
1. Demographic pressures. 2. Massive movement of refugees and internally displaced peoples. 3. Legacy of vengeance – seeking group grievance 4. Chronic and sustained human flight	5. Uneven economic development along group lines. 6. Sharp and/or severe economic decline	7. Criminalisation and/or delegitimisation of the state 8. Progressive deterioration of public services 9. Widespread violation of human rights 10. Security apparatus as state within a state 11. Rise of factionalised elites 12. Intervention of other states or external factors

Table 2.6: Some indexes incorporate democracy or legitimacy in their definition of state fragility

Source: Authors' own data

Index	Definition of State	Includes democracy/ legitimacy?
Bertelsmann Transformation Index (BTI) State Weakness Index	'Successful transformation requires that a state have functioning administration structures and that it secure its monopoly on the use of force. Without these two in place, a state cannot guarantee and provide for the rule of law or the security of its population' (Bertelsmann 2008: 6)	No
Country Indicators for Foreign Policy	'Fragile states lack the functional authority to provide basic security within their borders, the institutional capacity to provide basic social needs for their populations, and/ or the political legitimacy to effectively represent their citizens at home and abroad. ... Failed States [are] characterized by conflict, humanitarian crises, and economic collapse. Government authority, legitimacy, and capacity no longer extend throughout the state, but instead are limited either to specific regions or groups' (Carleton 2010).	Yes
Country Policy and Institutional Assessment (CPIA)	'Fragile states is the term used for countries facing particularly severe development challenges such as weak institutional capacity, poor governance, political instability, and frequently on-going violence or the legacy effects of past severe conflict' (World Bank 2007: 3).	No
Failed States Index	'A state that is failing has several attributes. One of the most common is the loss of physical control of its territory or a monopoly on the legitimate use of force. Other attributes of state failure include the erosion of legitimate authority to make collective decisions, an inability to provide reasonable public services, and the inability to interact with other states as a full member of the international community' (Fabra and Ziaja 2009: 53)	Yes
Index of State Weakness in the Developing World	'We define weak states as countries that lack the essential capacity and/or will to fulfil four sets of critical government responsibilities: fostering an environment conducive to sustainable and equitable economic growth; establishing and maintaining legitimate, transparent, and accountable political institutions; securing their populations from violent conflict and controlling their territory; and meeting the basic human needs of their population' (Rice and Patrick 2008: 3)	Yes
State Fragility Index	'A state may remain in a condition of fragile instability if it lacks effectiveness or legitimacy in a number of dimensions; however a state is likely to fail, or to already be a failed state, if it has lost both' (Marshall 2008: 13).	Yes

Table 2.7: The concept

Source: Developed by the authors, based on Fabra and Ziaja 2009

Index	Basic aspects of the definition	Boxes that operationalise the concept
Bertelsmann Transformation Index (BTI) State Weakness Index	Administration structures and monopoly on the use of force.	'Monopoly on use of force' and 'basic administration'
Country Indicators for Foreign Policy	Security, institutional capacity, political legitimacy.	Governance, economics, security & crime, human development, demography, environment
Country Policy and Institutional Assessment (CPIA)	Institutional capacity, governance, political instability, violence.	Economic management, structural policies, policies for social inclusion/equity, public sector management and institutions
Failed States Index	Control of its territory, monopoly on the legitimate use of force.	Social indicators, economic indicators, political indicators
Index of State Weakness in the Developing World	Environment to sustainable and equitable economic growth, political institutions, security and control of its territory	Security, political, economic, social welfare
State Fragility Index	Effectiveness and legitimacy	Effectiveness and legitimacy

Table 2.8: Type of data used by PSPIs

Source: Fabra and Ziaja 2009

Indicator	Data type	Sources
Bertelsmann Transformation Index (BTI) State Weakness Index	Expert survey	Uses two out of forty-nine questions from the BTI Country Assessments, which employ one primary researcher per country, one peer reviewer and two calibration rounds by regional and global coordinators. Information on how much calibration has impacted on original expert judgments is not reported.
Country Indicators for Foreign Policy	Expert data/ public statistics	Central Intelligence Agency, Centre for Systemic Peace, CLRI Human Rights Data Project, Food and Agriculture Organization of the United Nations, Freedom House, The Fund for Peace, Global Footprint Network, Heritage Foundation, Minorities at Risk, The Office of the United Nations High Commissioner for Refugees, Transparency International, UNDP Uppsala Conflict Database, US National Counterterrorism Centre, World Bank, Worldwide Governance Indicators.
Country Policy and Institutional Assessment (CPIA)	Expert survey	Country ratings are conducted by World Bank staff. They are preceded by an intensive benchmarking study on a smaller sample of countries and accompanied by consultation with country authorities.
Failed States Index	Content analysis / expert survey / public statistics	The Fund for Peace collects its own data. The core of data generation is a tool for content analysis of electronically available documents, termed 'Conflict Assessment System Tool' (CAST). It is accompanied by a ranking of countries based on public statistics (provided by the World Health Organization, the World Bank, The Office of the United Nations High Commissioner for Refugees, UNDP and others), and by calibration carried out by Fund for Peace experts. With regard to their database, the producers state: 'We receive our full text documentation from Meltwater, a news feed organization which provides us with links to over 11,000 sources originating from 110 countries in 50 languages' (Fund For Peace 2010).
Index of State Weakness in the Developing World	Expert data / opinion polls / public statistics	Archives, Centre for Systemic Peace, Economist Intelligence Unit, FAO, Freedom House, International Monetary Fund, Political Instability Task Force, Political Terror Scale, the UN, UNICEF, World Bank, Worldwide Governance Indicators.
State Fragility Index	Expert data / public statistics	Centre for Systemic Peace, Elite Leadership (Gurr / Harff), Leadership. Duration (Bienen / van de Walle), Minorities at Risk, Political Terror Scale, UNDP, US Census Bureau, World Bank

Table 2.9: Variables and use of expert opinion

Source: Authors' own data

Indicator	Type of variable	# of variables	Count variables	Expert opinion variables
BTI	Ordinal	2	None	2 Variables
CIFP	Continuous	83	Several: more than 20 variables	Several: more than 15 variables
CPIA	Ordinal	1-6	None	Several: more than 10 variables

Table 2.10: Hedged definitions of existing PSPIs

Source: Authors' own data, based upon Fabra and Ziaja 2009

Index	Definition	Linguistic hedges and broad terms
Country Policy and Institutional Assessment (CPIA) / International Development Association (IDA) Resource Allocation Index (IRAI)	'Fragile states is the term used for countries facing particularly severe development challenges such as weak institutional capacity, poor governance, political instability, and frequently on-going violence or the legacy effects of past severe conflict' (World Bank 2007: 2)	Severe, weak, poor, frequently
Brookings Institution	'We define weak states as countries that lack the essential capacity and/or will to fulfil four sets of critical government responsibilities: fostering an environment conducive to sustainable and equitable economic growth; establishing and maintaining legitimate, transparent, and accountable political institutions; securing their populations from violent conflict and controlling their territory; and meeting the basic human needs of their population' (Rice and Patrick 2008: 3)	Essential, sustainable, equitable, legitimate, transparent, accountable
Fund For Peace	'A state that is failing has several attributes. One of the most common is the loss of physical control of its territory or a monopoly on the legitimate use of force. Other attributes of state failure include the erosion of legitimate authority to make collective decisions, an inability to provide reasonable public services, and the inability to interact with other states as a full member of the international community'. (Fabra and Ziaja, 2009: 53)	Legitimate, reasonable full-member.
USAID	'Crisis states are those where the central government does not exert effective control over its own territory or is unable or unwilling to assure the provision of vital services to significant parts of its territory, where legitimacy of the government is weak or nonexistent, and where violent conflict is a reality or a great risk (Warren 2006: 4)	Effective, significant
CIA	'State failure was defined to include four categories of events: Revolutionary wars (Episodes of sustained violent conflict between governments and politically organized challengers that seek to overthrow the central government, to replace its leaders, or to seize power in one region), Ethnic wars (Episodes of sustained violent conflict in which national, ethnic, religious, or other communal minorities challenge governments to seek major changes in status), Adverse regime changes (Major, abrupt shifts in patterns of governance, including state collapse, periods of severe elite or regime instability, and shifts away from democracy toward authoritarian rule), Genocides and politicides (Sustained policies by states or their agents, or, in civil wars, by either of the contending authorities that result in the deaths of a substantial portion of a communal or political group)' (Goldstone et al. 2000: 9)	Sustained, substantial major, severe

Table 2.11: Countries with high levels of illegitimacy according to Foreign Policy (2007)

Source: CIFP Country Ranking Table 2007

Legitimacy Rank	Country	Legitimacy
1	Saudi Arabia	7.41
2	Libya	7.17
3	Korea, North	7.14
4	Yemen, Rep.	7.06
5	Somalia	7
6	Iraq	6.96
7	United Arab Emirates	6.92
8	Turkmenistan	6.89
9	Equatorial Guinea	6.79
10	Iran	6.76
40	Congo, Dem. Rep.	6.15

Table 2.12: Fund for Peace marks for Republic of Congo

Source: Fund for Peace 2010

Variable		Years		
		2007	2008	2009
Mounting Demographic Pressures	I-1	8.7	8.7	8.9
Massive Movement of Refugees or Internally Displaced Persons creating Complex Humanitarian Emergencies	I-2	7.3	7.7	7.8
Legacy of Vengeance-Seeking Group Grievance or Group Paranoia	I-3	6.8	6.8	6.5
Chronic and Sustained Human Flight	I-4	6.1	6.1	6.1
Uneven Economic Development along Group Lines	I-5	8.1	8.1	8
Sharp and/or Severe Economic Decline	I-6	8.3	8	8
Criminalization and/or Delegitimisation of the State	I-7	8.5	8.8	8.6
Progressive Deterioration of Public Services	I-8	8.8	8.8	8.8
Suspension or Arbitrary Application of the Rule of Law and Widespread Violation of Human Rights	I-9	7.9	7.9	7.9
Security Apparatus Operates as a 'State Within a State'	I-10	7.9	7.9	7.8
Rise of Factionalised Elites	I-11	7.2	7.2	7.1
Intervention of Other States or External Political Actors	I-12	7.4	7.4	7.6
Total	Total	**93**	**93.4**	**93.1**
Rank		7	6	5

Table 2.13: Cut-off point and classification

Source: Authors' own data. based on Fabra and Ziaja 2009

Index	Cut-off point and Classification
Bertelsmann Transformation Index (BTI) State Weakness Index	The BTI (2007:8) maps 'failed states' (scores of 1.0–2.5), 'very fragile states' (3.0–4.0) and 'fragile states' (4.5–5.5); the remaining countries are not classified.
Country Indicators for Foreign Policy	Score thresholds and rank fraction: performing well relative to others (scores below 3.50); performing at or around the median (3.50 to 6.50); performing poorly (above 6.50); worst global performers (5% worst ranking)
Country Policy and Institutional Assessment (CPIA) / International Development Association (IDA) Resource Allocation Index (IRAI)	Countries scoring 3.2 and below are termed fragile states
Failed States Index	Countries are categorised by score quartiles: alert (scores of 90–120), warning (60–90), moderate (30–60), sustainable (0–30)
Political Instability Index	Thresholds (determination not explained): very high risk (above 7.4), high risk (5.8–7.4), moderate risk (4.0–5.7), low risk (below 4.0).

Table 2.14: The risk categories of the Fund for Peace

Source: The Fund for Peace (Fabra and Ziaja 2009: 54)

The risk categories			Cut offs at quintiles		
If	$(\sum_{i=1}^{12} Indicador_i) \geq 90$	then 'Alert'	If	$(\sum_{i=1}^{12} Indicador_i) \geq$ Percentile 75	then 'Alert'
If	$60 \leq (\sum_{i=1}^{12} Indicador_i) < 90$	then 'Warning'	If	Percentil 50 $\leq (\sum_{i=1}^{12} Indicador_i) <$ Percentile 75	then 'Warning'
If	$30 \leq (\sum_{i=1}^{12} Indicador_i) < 60$	then 'Moderate [danger]'	If	Percentil25 $\leq (\sum_{i=1}^{12} Indicador_i) <$ Percentile 50	then 'Moderate'
If	$(\sum_{i=1}^{12} Indicador_i) < 30$	then 'Sustainable state'	If	$(\sum_{i=1}^{12} Indicador_i) <$ Percentile 25	then 'Sustainable'

Table 2.15: Normalisation

Source: Fabra and Ziaja 2009 with authors' additions

Index	Index Scores	Normalisation
BTI	Coding applies a 1–10 (worst to best) score which is not transformed before aggregation.	Categorical variables in a 1–10 scale. They are not normalised
CIFP	Indicators are rescaled to a range of 1–9 (Best to worst)	No normalisation
CPIA	Coding applies a 1 to 6 scale (worst to best)	No normalisation, all are categorical variables under the same scale
Failed States Index	Indicators are standardised to a 0.0–10.0 scale (best to worst)	No normalisation, all are categorical variables
Index of State Weakness in the Developing World	Indicator values are converted to a range between 0 and 10	All the variables belong to a 0–10 scale. The variables are averaged within each box, and this average is restored to a 0–10 scale
State Fragility Index	Sub-categories are transformed to a four-point scale (0–3) by thresholds: 0 'no fragility', 1 'low fragility', 2 'medium fragility', 3 'high fragility'	No normalisation, categorical variables

Table 2.16: Examples of the normalisation of the State Weakness in the Developing World Index

Source: Authors' own calculations

	Economic Basket					Average of the variables of the basket	Average normalisation
Country	Per capita GNI	GDP Growth	Income Inequality	Inflation	Regulatory Quality		
Somalia	0.12	--	--	--	0	0.06	0
North Korea	--	--	--	--	0.47	0.47	0.52
Zimbabwe	0.23	0	4.99	0	1.2	1.284	1.55
Iraq	0.95	2.98	--	--	3.02	2.31	2.86
Eritrea	0.09	2.63	--	5.25	2.02	2.4975	3.09
Liberia	0.04	1.81	--	6.48	2.59	2.73	3.39

Table 2.17: Baskets of the State Weakness in the Developing World Index

Source: Rice and Patrick 2008

Economic	Political	Security	Social Welfare
• Gini per capita, 2006 • GDP Growth, 2002–2006 • Income Inequality, 2006 • Inflation, 2002–2006 • Regulatory Quality, 2006	• Government effectiveness, 2006 • Rule of law, 2006 • Voice and Accountability, 2006 • Control of corruption, 2006 • Freedom Ratings,2006	• Conflict Intensity • Gross Human • Rights Abuses • Territory Affected by Conflict • Incidence of Coups • Political Stability and Absence of Violence	• Child Mortality • Access to Improved Water and Sanitation • Undernourishment • Primary School Completion • Life Expectancy

Table 2.18: Pair-wise comparisons and incomparable cases

Source: Authors' own calculations

	BTI, 2008	The Fund for Peace, 2006	CPIA, 2007
Total	15625	21316	5625
Feasible pair wise comparisons	14089	8334	1861
Non-comparable	1536	12982	3764
% comparables	90.17%	39.10%	33%
% non-comparables	9.83%	60.90%	67%

Table 2.19: Some major indexes and their aggregation procedures
Source: Based upon Fabra and Ziaja 2009

Index/ Aggregation function	Aggregation levels	Aggregation procedure	Number of indicators	Range of weights per indicator	Compensatory?	If yes, explanation of the weights
BTI	1	(Monopoly of Violence + Basic Administration) / 2	2	0.5	Yes	None
CIFP	2	Governance + Economics + Security & Crime + Human Development + Demography + Environment Categories are calculated by arithmetic means of their indicators.	83	0.007–0.019	Yes	None
CPIA	2	(Economic Management + Structural Policies + Policies for Social Inclusion/Equity + Public Sector Management and Institutions) / 4. All categories consist of 4 indicators each.	16	Depends on the box where the variable is. The four boxes of the CPIA have the same weight each (0.25).	Yes	No
Failed States Index	1	I-1 + I-2 + … + I-12	12	0.83	Yes	No
Global Peace Index	2	0.4 *External Peace + 0.6 * Internal Peace Weights between 1 and 5 assigned to individual indicators comprising the categories (calculated by weighted means).	23	0.012–0.061	Yes	No
Harvard Kennedy School Index of African Governance	3	(Safety and Security + Rule of Law, Transparency, and Corruption + Participation and Human Rights + Sustainable Economic Opportunity + Human Development) / 5 All categories consist of 2–4 subcategories which consist of 1–11 indicators. All are weighted equally (calculated by arithmetic mean) except for Safety and Security = (2 * National Security + Public Safety) / 3.	55	0.006–0.067	Yes	No
Index of State Weakness in the Developing World	2	(Economic Basket + Political Basket + Security Basket + Social Basket) / 4. Categories are arithmetic averages of 5 indicators each.	20	0.05	Yes	No

Peace and Conflict Instability Ledger	1	Model driven (Logistic Regression Estimates); employed variables: Inconsistency of the governing regime, high infant mortality rates, lack of integration with the global economy, the militarisation of society, and the presence of armed conflict in neighbouring states.	5	Vary with country and year (based on logistic regression)	Yes	Yes (weights for each factor reflect the relative influence that each has on explaining future instability)
State Fragility Index	3	Effectiveness score + Legitimacy score. Effectiveness Score = Security Effectiveness + Political Effectiveness + Economic Effectiveness + Social Effectiveness. Legitimacy score = Security Legitimacy + Political Legitimacy + Economic Legitimacy + Social Legitimacy. Sub-categories consist of 1 to 3 indicators each.	14	0.031–0.125	Yes	No

Table 2.20: The substitution rates of the CPIA

Source: Authors' own calculations from World Bank. (2009)

	Macro. Mgt.	Fiscal Policy	Debt Policy	Trade	Financial Sector	Business Regulatory	Gender Equality	Equity of Public Resource Use	Building Human Resources	Social Protection & Labour	Pol. & Instit. for Environ. Sustain.	Property Rights	Quality of Budget. & Finan. Mgt.	Effic. of Revenue Mobil.	Quality of Public Admin.	Transpar., Account. & Corrup. in Pub. Sec.
Macro. Mgt.	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Fiscal Policy	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Debt Policy	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Trade	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Financial Sector	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Business Regulatory Environ.	1	1	1	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Gender Equality	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Equity of Public Resource Use	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Building Human Resources	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Social Protection and Labour	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Pol. & Instit. for Environ. Sustain.	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Property Rights and Rule-based Govern.	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Quality of Budget. and Finan. Mgt.	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Effic. of Revenue Mobil.	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Quality of Public Admin.	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1
Transpar., Account. and Corrup. in Pub. Sec.	1.67	1.67	1.67	1.67	1.67	1.67	1	1	1	1	1	1	1	1	1	1

Table 2.21: Rank changes after modifications in the weights of the CPIA variables

Source: Authors' own calculations

Country	Rank F1	CPIA Score	Country	Rank F2	CPIA F2 (*) Score
ZIMBABWE	1	1.82	ZIMBABWE	1	1.88
CENTRAL AFR. REP.	2	2.39	CENTRAL AFR. REP.	2	2.34
COMOROS	3	2.42	COMOROS	3	2.44
TOGO	4	2.49	COTE D'IVOIRE	4	2.47
COTE D'IVOIRE	5	2.49	TOGO	5	2.47
ERITREA	6	2.50	SUDAN	6	2.53
ANGOLA	7	2.58	ANGOLA	7	2.56
SUDAN	8	2.59	ERITREA	8	2.63
GUINEA-BISSAU	9	2.68	GUINEA-BISSAU	9	2.69

Table 2.22: Correlations between variables in the SFI

Source: Authors' own calculations

	Security Effectiveness 2007	Security Legitimacy 2007	Political Effectiveness 2007	Political Legitimacy 2007	Economic Effectiveness 2007	Economic Legitimacy 2007	Social Effectiveness 2007	Social Legitimacy 2007
Security Effectiveness 2007	1	0.632658	-0.01426	0.223362	-0.10854	-0.24898	-0.09195	-0.15871
Security Legitimacy 2007	0.632658	1	0.156003	0.421232	0.043524	0.015576	0.091641	-0.0267
Political Effectiveness 2007	-0.01426	0.156003	1	0.180326	0.276119	0.215027	0.183501	0.076795
Political Legitimacy 2007	0.223362	0.421232	0.180326	1	0.036738	0.10317	0.027973	-0.04069
Economic Effectiveness 2007	-0.10854	0.043524	0.276119	0.036738	1	0.104622	0.692208	0.277359
Economic Legitimacy 2007	-0.24898	0.015576	0.215027	0.10317	0.104622	1	0.137976	0.180318
Social Effectiveness 2007	-0.09195	0.091641	0.183501	0.027973	0.692208	0.137976	1	0.751683
Social Legitimacy 2007	-0.15871	-0.0267	0.076795	-0.04069	0.277359	0.180318	0.751683	1

3. ALTERNATIVES AND PERSPECTIVES

3.1. TOOLS FOR PARTIAL IMPROVEMENTS

3.1.2. FUZZY SETS AND OTHER INPUTS: THE TREATMENT OF NON-PROBABILISTIC UNCERTAINTY

We draw heavily on the literature that explores formal representations of poorly defined situations. This is a very rich and bulging area of research. The data is often abundant but ambiguous, possibly poor and corrupted, and sometimes difficult to interpret. Analysis is usually made possible through the 'formalisation' of social and managerial problems. This is the focus of our discussion in this section, where we put forth 'fuzzy set theory' as a useful tool.

Fuzzy set theory is a branch of mathematics developed in the 1960s by Lofti Zadeh, whose basic argument was that in many situations objects can belong only partially to a set, but that in this partial belonging they can belong to many sets at the same time (Zadeh 1975a). As with many other tools for 'approximate reasoning', fuzzy mathematics is 'subsidiary' to crisp set theory (Peters et al. 2007). It is a well-established area and the solidity of its usefulness is beyond reasonable doubt, having been successfully used and tested in engineering, robotics, transportation and basic research (Lee 1990). In academic fields directly concerned with social indicators, fuzzy mathematics has become a standard and powerful tool for pattern recognition (Bezdek et al. 1999), clustering and other forms of classification (Yao et al. 2001), the building of aggregation functions (Beliakov et al. 2007), the evaluation of similarity functions (Zadeh 1971), the analysis of preference orderings (including Arrovian themes; Taylor and Pacelli 2008; Baliamoune-Lutz and McGillivray 2008), multi-objective and multi-analysis decision making (Chen and Hwang 1992) and diagnostic decision making (Kuncheva 1991), among other tasks. In many of these fields it has become an indispensable tool.

During this period a debate emerged regarding the comparative usefulness of fuzzy ('possibility') approaches and statistical ('probability') approaches (Kim et al. 1996). This was replaced by a much more constructive dialogue on the possibility of combining these two approaches (Zadeh 1978). This argument proposed coordinating the strengths of fuzzy and statistical approaches, and overcoming their limitations. We will here highlight two criticisms levelled against fuzzy set indicators and aggregation functions.

First, in fuzzy operations, objects are not forced into belonging or not belonging to a set. They can belong partially. Said in other terms, while in standard operations the membership function of an object to a set can be only zero or one (zero if it does not belong, one if it does), in fuzzy operations it can be *any* number between zero and one. But this implies that the membership function of each object to each set has to be created. The form and parameter values of this new membership function are often chosen idiosyncratically (Bouyssou et al. 2000). There is no clear set of instructions on how to choose one form or another alternative. Through interaction with the data the researcher can calibrate some of the parameters but this remains more of an art than a rigorous and replicable algorithm. For an analysis of the 'calibration problem' in the social sciences, see Ragin (2008), and for a good analysis of fuzzy calibration more generally see Jawahar (2002).

Second, syntactically fuzzy sets pass any test with flying colours. They have also demonstrated repeatedly that they are powerful tools 'for doing things in the world'. However, in some contexts their semantic interpretation is not fully developed (Lakoff 1973).

Both of these criticisms have to be taken into account, but are they not insurmountable.

While fuzzy set theory is an area that has already reached maturity, others of its 'soft computing' siblings are in a process of development. Together with fuzzy sets they constitute the field of 'approximate reasoning' (Zadeh 1975b).[80] Approximate reasoning approaches:

- Take on board explicitly non-probabilistic forms of uncertainty (fuzzyness, vagueness);
- Relax some of the key axioms of classical logic (such as the axiom of the excluded middle);
- Are designed to 'compute with words' (Zadeh 1975b);
- Are designed to deal with an overflow of (possibly inferior, poorly defined or deteriorated) data, and make sense of it. They behave well in front of data corruption or incompleteness;
- Formalise ill-defined problems and complex concepts that are heavily hedged.

Typically, formal approximate reasoning is a product of the same wave of technological change that led to the development of third wave social indicators. One of its main practical concerns is the optimisation of (perhaps not very well defined) queries on the internet or in very large databases (Jarke and Koch 1984; Kooi 1980).

We believe that large N social science studies – especially those related to comparative politics – cannot afford to ignore this modality of formalisation, since it able to address some of the typical issues that third wave indexes face. Some researchers within the approximate reasoning area have proposed a 'non-invasive' data treatment method (Düntsch and Gediga 1999). Non-invasiveness is characterised by

80 There is a rather baroque naming activity in related fields of research. The generic tags of 'soft computing' and 'granular computing' are also used. All these fields contain relevant and valuable developments for social researchers interested in dealing formally with ill defined, incomplete and deteriorated data. Purposefully we do not discuss neural networks, which are frequently used in conjunction with fuzzy sets and other tools. Neural networks are an extremely potent classificatory tool, but semantically they are black boxes.

the minimisation of ad hoc parameters in data analysis; the fewer ad hoc parameters you have, the better.

We draw on the perspectives and results of approximate reasoning and non-invasiveness here to put forth our argument. We believe that to treat multidimensional and fuzzily or vaguely defined concepts as if they were crisp and exact is a travesty, and one that ultimately deforms the entire analysis. Imposing total order on specific but unexplained forms of data organisation (that are as good as any other) and ad hoc parameters produces results that are hard to read into. Social indicators have become more ambitious and today seek to measure multidimensional and 'subjective' issues. The tools to do so should be adapted so that a comprehensible language is at hand to sensibly capture the meaning and the specifics of the data that are being used. Introducing more approximate reasoning and non-invasive tools into social scientific formal analysis promises a move from spurious precision to systematically treated and 'tamed' ambiguity.

3.1.2. THE CONTRIBUTIONS OF CHARLES RAGIN AND A TRIBE OF ENGINEERS

A similar line of reasoning was used by Charles Ragin in his pioneering methodological books (Ragin 2000). Ragin (2008: 1) criticises King, Keohane and Verba's controversial view that:

> The proper template [for all social research] is provided by large-N quantitative research, with its well-defined and seemingly limitless populations and its focus on calculating the net effects of 'independent' variables in properly specified linear models.

He argues:

> It is this template for conducting research that is at issue … the problem is not that it is a bad template. It is a wonderful, well-articulated template. The problem is that it is too often promoted as the best template or even the only template.

The focus of Ragin's attention has been qualitative research, and in particular the quest for systematic ways to treat data in small N studies. However, many of his observations are relevant for quantitative research also. The tool that he appropriately describes as 'wonderful' is, in effect, wonderful because the models are 'properly specified' and 'well defined'. If these conditions are missing, and as we have demonstrated here they are egregiously missing in PSPIs, then new ways of accommodating the data should be a priority for people interested in making sense of large N political datasets.

In subjects other than the social sciences, an acute sensitivity to the subjectivity and elusiveness of multi-attribute decision processes has developed. After working for a relatively long period on a critical analysis of PSPIs, we came across the brilliant works of Bouyssou and Perny (1992), a group of engineers and operation researchers exploring indicators and their limits. This work soon became a reference point for us. We found far more affinity with the arguments and ideas of these engineers than we did with some economists and political scientists, who often treat the fourth decimal place of an average over expert data as if it were an alien object. As an aside, it is possible to explain why students of multi-attribute decision making remember and implement lessons that social scientists tend to forget, despite the fact that the latter are professionally trained to understand these issues and the former not. Since engineers in the area of multi-attribute decision making face the task of expressing the preferences of concrete human beings – entrepreneurs, for example – they rapidly understand that this can be a difficult task, and its results hazy and unstable.

3.1.3. FUZZY TOOLS IN PSPIs

3.1.3.1. FUZZIFYING THE CPIA

Baliamoune-Lutz and McGillivray (2008) identify three main sources of ambiguity in the CPIA, a World Bank Index:

1. Ambiguity may arise simply from the lack of robust and conclusive evidence on aid effectiveness and the effect of some economic policies and institutional reforms on growth and poverty reduction;[81]
2. Ambiguity could result from the questions in the World Bank questionnaire used in collecting information on the clusters;
3. The respondent's perception of what he/she thinks to be a true picture could also be a source of ambiguity.

In order to address this, they decided to 'fuzzify' the classificatory boundaries of the CPIA and in particular the critical cut-off (3).[82] They transformed the CPIA scores to 'examine changes in country ranking depending on the degree of vagueness and the level (threshold) at which performance changes from disastrous to average or good (depending on how we define the cut-off point)' (Baliamoune-Lutz and McGillivray 2008: 3). The transformation is performed, as happens in fuzzy formalisation, by defining a membership function. The concrete form of this membership function is:

$$\mu(x_i) = \frac{1}{1+e^{-\alpha(x_i-\beta)}} \qquad \text{(f.3.1.1)}$$

The parameters α and β represent the degree of ambiguity and threshold of identification, respectively. These parameters are identified in the following way:

81 Remember that the CPIA includes 'correct policies' as part of the definition of fragility (fragile countries do not follow correct policies). What Baliamoune-Lutz and McGillivray are claiming is that in some cases there is not enough evidence to claim that some policies are correct and others not.

82 For the CPIA, a country below 3 is failed.

$$\alpha = \frac{\ln\left(\frac{\mu_h}{1-\mu_h}\right) - \ln\left(\frac{\mu_l}{1-\mu_l}\right)}{x_h - x_l}, \quad \beta = \frac{\ln\left(\frac{\mu_h}{1-\mu_h}\right) - \ln\left(\frac{\mu_l}{1-\mu_l}\right)}{\ln\left(\frac{\mu_h}{1-\mu_h}\right) - \ln\left(\frac{\mu_l}{1-\mu_l}\right)}$$

(f.3.1.2)

'where *μh* represents the membership degree of the highest achievement (*xh*) of the goal, and μl represents the membership degree of the lowest achievement (*xl*) of the goal' (Baliamoune-Lutz and McGillivray 2008: 3).

The transformation is implemented as seen in Table 3.1.[83] Once the transformed CPIA values are obtained, the countries are clustered by quintiles. The output is compared with the crisp CPIA marks and ranks. The conclusion is very relevant for us: 'It is clear that once we take into account the ambiguity of the outcomes we get different scores from the World Bank's CPIA' (Baliamoune-Lutz and McGillivray 2008: 6).

This work of Baliamoune-Lutz and McGillivray is a building block for any attempt to improve PSPIs. A problem that remains unsolved here, though, is the specific functional form that the membership function takes. Why should one form be chosen over another? Why use the *logarithm* when other functional forms would have worked equally well, and might have produced substantially different results? From a 'non-invasive' treatment of the data perspective, it would be ideal to uphold a unique membership function (perhaps up to linear or affine transformations) based on the data.

3.1.3.2. WORKING WITH NON-NUMERICAL OBJECTS

Carment et al. (2006) developed a method of visualisation of triangular 'country profiles', to take into account the multidimensional nature of the country characterisation typical of PSPIs. The method stems from a simple, attractive and powerful insight. Each of the three boxes of the index (authority, legitimacy and capacity) represents a vertex of the triangle (see Carment et al. 2006: 19 for an example). Fragility profiles observed in the figure come from the recognition that:

> Authority, legitimacy, and capacity are analytical constructs, reflecting the functions of a state and its component parts. The three dimensions are inextricably interlinked; for instance, authority correlates with legitimacy at 0.58 and with capacity at 0.62, while legitimacy and capacity correlate at 0.75. As a result, shortfalls in any one dimension have implications for a given state's functionality along the other two, thus providing additional insight into the overall fragility of the state (Carment et al. 2006: 5).

This type of representation is used to identify which of the three dimensions – authority, legitimacy and capacity – is the main source of weakness and to provide 'additional insight into the overall fragility of the state' (Carment et al. 2006: 6). In Carment et al. 2006: 19 (Table 8. Detailed Country Fragility Profile for Sri Lanka), the triangular representation of Sri Lanka is shown. According to Carment:

> Sri Lanka exhibits weak authority, largely as a result of the decades-long confrontation between the Liberation Tigers of Tamil Eelam (LTTE) and the Sri Lankan government. Indicators related to political violence, organized crime, number of refugees produced, and other measures of security tend to reflect various destabilizing aspects of the conflict. Human development indicators suggest that the country is performing relatively well when compared to regional averages, with moderate levels of literacy, infant mortality, and HIV/AIDS infection, given the state's overall level of economic development (Carment et al. 2006: 19).

Carment's insight suggests that we should move towards more complex representations than (cardinal or ordinal) numbers (as we have argued in the introduction here) This is a step in precisely such a direction. However, the triangular representation has not yet been put in operational form. It is not possible to perform comparisons or additions, let alone regressions, over it.

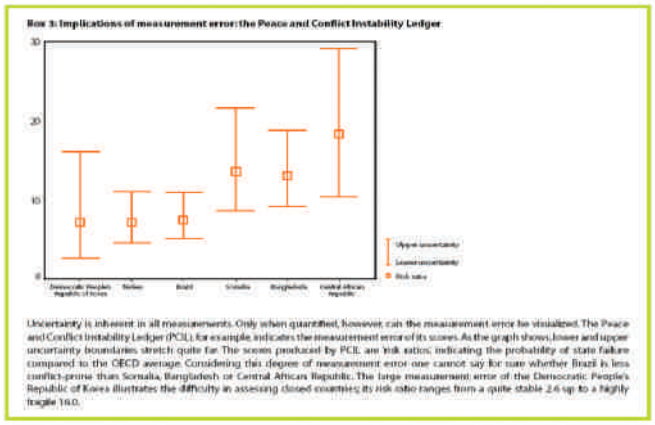
3.1.3.3. RISK RATIOS

Fabra and Ziaja (2009) cite a measure of error developed by the Peace and Conflict Instability Ledger to capture different forms of uncertainty (see Figure 3.1). They argue that:

> Uncertainty is inherent in all measurements. Only when quantified, however, can the measurement error be visualized. The Peace and Conflict Instability Ledger (PCIL), for example, indicates the measurement error of its scores. As the graph shows, lower and upper uncertainty boundaries stretch quite far. The scores produced by PCIL are 'risk ratios', indicating the probability of state failure compared to the OECD average. Considering this degree of measurement error one cannot say for sure whether Brazil is less conflict-prone than Somalia, Bangladesh or Central African Republic. The large measurement error of the Democratic People's Republic of Korea illustrates the difficulty in assessing closed countries; its risk ratio ranges from a quite stable 2.6 up to a highly fragile 16.0 (Fabra and Ziaja 2009: 13).

83 Note that the authors call 'clusters' what we call here 'boxes'.

Figure 3.1: A way of representing measurement error: The Peace and Conflict Instability Ledger
Source: Fabra and Ziaja 2009: 13.



They also noted that there was a problem with weights: 'Only the peace and Conflict Instability Ledger and WGI Political Stability and absence of Violence use model driven approaches in which weights are extracted from the data by mathematical algorithms' (Fabra and Ziaja 2009: 28). The methods used to select weights in the Ledger are factorial analysis and principal components.[84]

This contribution cannot be overestimated. They come close to representing countries by intervals, and find that PSPIs databases are affected by intrinsic uncertainty, not simply by probabilistic uncertainty. The question that emerges is how to work with these new objects. Until now, non-numerical representations of countries have not been put in operational form; they are only heuristic tools. In the forthcoming sections here, we will suggest some answers.

3.2. THE MONOPOLY-ADMINISTRATION-TERRITORY (MAT) DATABASE

3.2.1. THE SUBJACENT THEORY: THE THREE DIMENSIONS OF FRAGILITY (STRENGTH)

In this presentation of our database we use the approach and results of the Crisis States Research Centre (CSRC). The insights of Di John and Putzel (2009) are particularly useful. The notions that guide the structure of the MAT database and the recollection of the data draw heavily upon this work, though the database is not a literal transcription of them. This section is not intended to replace in any form a codebook. Rather, it is an introduction and an explanation. We will simply sketch the main direction of the argument, with some of the problems we have found and address specific theoretical concerns.

The CSRC has advanced and strengthened the understanding of statehood as based, at core, on three elements: a monopoly of violence, bureaucratic strength and the territorial reach of the state. According to Tilly (1978; 1990), statehood is defined by a continuum that goes from oligopoly to the monopoly of violence and coercion, and states 'specialize in the control and use of coercive means – surveillance, detention and armed force (Tilly 1989: 63).[85] Bureaucracy is the signature of modern states, and there is also a broad consensus in the social sciences that it is both a key characteristic of statehood in its own right and a proxy of the capacities of the state (Weber 1922; Mann 1984). Indeed, what matters most here is not the size of the bureaucracy, though if extreme this can be decisive,[86] but its efficiency. Finally, the ability to control a contiguous territory and to operate on it is at the very heart of the definition of modern sovereignty (North et al. 2009; Blanton and Fargher 2008; Jackson 1990). The *Social Science Encyclopaedia* defines a state as 'a territorial unit ordered by a sovereign power, and involved officeholders, a home territory, soldiers distinctively equipped to distinguish them from others, ambassadors, flags, and so on' (Minogue 1994: 1435). Practically any working definition of the state includes officeholders (bureaucracy), a home territory (territorial reach) and coercion (soldiers), in any order. These are the invariants of the state. The variables that cause the strength or weakness of these dimensions or that are influenced by it are multifarious, but the core notion of statehood should not be confounded with causes and consequences.

This captures the essentials of Weber's classical definition of state, where 'monopoly on the legitimate use of physical force within a given territory' (Weber 1968) is key. What, we might ask, is 'legitimate'? Weber explicitly considered legitimacy to be subjective submission to instituted authority (Weber 1922). In this use of the word, legitimacy cannot be treated as a synonym of 'morally acceptable'. Note, however, that both the Weberian and the moral definition of legitimacy require clarification of to whom the term should apply. Lack of legitimacy should be observable. This poses a challenge to empirical investigators, as well captured by the *Social Science Encyclopaedia*: 'a government or state is considered "legitimate" if it possesses the "right to rule"' (Kuper (eds) 2003: 794). Unfortunately, this definition begs the most crucial question: what is a 'right' to rule, and how can its existence and meaning be determined?[87] Generally speaking, this question has been answered in two ways. One school of thought has argued with Weber (1968) that, 'it is only the probability of orientation to the subjective *belief* in the validity of an order which constitutes the valid order itself' (Weber and Parsons 1997: 126). According to this view, 'right' is reduced to belief in the

84 These, of course, are well-established statistical methods that already come with every single commercial statistical package (also with open source packages, as R). As noted in section 1.3, however, they are unstable and depend on the set of cases that are chosen. This can be consequential for the rankings that are produced.

85 Note that here Tilly is including the functions of the army and the police.

86 In other words, if there is only one employer, the state, this is a significant datum. On the other hand, if the state has almost no employees, that is a significant datum as well.

87 There's another crucial question left out here: for whom? A government can be highly legitimate for some sectors of the population, and illegitimate for others.

appropriateness of an existing order and the 'right to rule'. The presence of objective, external or universal standards for judging rightness grounded in natural law, reason or some other trans-historical principle is typically rejected as philosophically impossible and sociologically naïve. In his sociology of legitimacy, Weber attempted to guard against the relativistic consequences of such a conception by identifying four reasons for ascribing legitimacy to any social order: tradition, affect, value-rationality and legality. We fully ascribe to this Weberian position, with its explicit rejection of some sort of trans-historical morality. This, in our view, can only impoverish empirical, and particularly large N, studies. We are aware as well of the criticisms directed at the Weberian notions of legitimacy by authors such as Habermas (1975). The problem of these alternative views is that they have not been able to produce any kind of operational tool to solve the problem of 'observation' in a satisfactory manner. The *Encyclopaedia* concludes that:

> In light of such an impasse in philosophy, since the mid-1970s work on legitimacy in the social sciences has proceeded generally in three directions. First, social scientists attracted to empirical investigation have either worked towards testing hypotheses about legitimation in experimental settings, or they have dropped the term legitimacy altogether, hoping to avoid troubling normative issues while searching for measurable levels of 'regime support'. Second, some have moved towards developing theories about *il*legitimacy or *delegitimation*, arguing that the real problems of the modern state lie with its essential lack of legitimacy, as illustrated most dramatically by the collapse of the former Soviet Union and events in Europe after 1989. Third, in a related move, others have focused attention on state structure and policy or the relationship between state and civil society in an effort to understand the factors conditioning legitimacy (Kuper (eds) 2003: 795).

Since large N studies are based on simplification and isolation (section 1.2), a fatal error would be to introduce these unsolved complexities into a numerical database. A significant deterioration of subjective belief should produce observable phenomena (such as the deterioration of monopoly).

Apparently bureaucracy is taken for granted in this discussion, but it should not be. It plays a very prominent role in Weber's theory (Swedberg 1999 and Giddens 1984). There are three reasons that link governmental efforts to maintain the monopoly of violence and to control a territory with bureaucratic build up. First, to build working apparatuses of coercion, rulers have to engage in massive rent extraction, which cannot be implemented without bureaucratic organisations (Olso 1993; Tilly 1990; Levi 1988; Schumpeter 1942). Second, the state unfolds in the space left by violence, so it has to find a way to establish a chain of command and a network of communications. It must coordinate actions between the decision-making centre and the periphery. Third, rent extraction only becomes stable when 'roving bandits' become 'stationary bandits' (Olso 1993), so the state has to offer a *quid pro quo* to the populations it aspires to rule (Tilly 1990; Blanton and Fargher 2008; Mann 1984; Przeworski et al. 2003).

We use these three core components of statehood as the basis of the MAT database. We deliberately use a lean definition of statehood and thus of fragility so that we can correlate it with income, democracy and other phenomena and indexes. This has two implications:

a. Our boxes (dimensions) have non-negligible correlations.[88] Indeed, they try to capture distinct aspects of each construct (for example, homicides and deaths in combat refer to the 'police' and 'army' dimensions of security respectively), but even then there is an inevitable interaction between them. Thus, they are not orthogonal, and any aggregation function will have to take this into account.

b. We differentiate state performance (measured on monopoly of violence, bureaucratic build up and territorial reach of the state) from levels of democracy and development. Naturally, this does not mean that we believe that democracy or well - being are irrelevant or uninteresting for statehood in some esoteric sense. They are very transcendental political realities and social goods. As empirical researchers, we want to see how strongly associated state performance, democracy and development are.[89] The only way to do this is to treat them as distinct concepts, as the overwhelming majority of political theorists do (see Przeworski et al. 2003).[90] Here, we follow Przeworski's argument in favour of avoiding conceptual stretching, precisely to be able to capture the interactions of different aspects of social reality.

3.2.2. GUIDING PRINCIPLES FOR DATABASE CONSTRUCTION

We followed five principles for the construction of the MAT database:

a. We should never replace unobservables with unobservables and pretend that this is a genuine operationalisation;

b. Ordinal scales that capture information from counts are one of the most solid quantifications available;

c. Counts of relatively rare and public, and thus highly observable and traceable, phenomena (for example invasions, or coups) are also credible;

88 In some operationalisations of reliability, this is an indicator of the fact that in reality a single construct is being measured.

89 We do this in Gutiérrez et al. 2010. We find that democracy and income levels (or other developmental indicators) explain, but only partially, the variance of fragility.

90 The study of the state and the study of political regimes are considered to be related but clearly distinct fields.

d. Massive counts (such as deaths) should be used with prudence and after being properly treated and/or transformed (Bond et al. 2003);[91]

e. Some state and/or NGO assessments can be transformed into ordinal scales.

Missing data can be imputed, if the method is appropriate and the assumptions for the imputation are met. We imputed missing data by ANCOVA and hotdeck methods[92] (see section 3.2.4 for further explanation of these):

- Expert and in-house coding of public, relatively rare, and traceable events (for example, military interventions)[93] is plausible and can be utilised;

- Expert and in-house coding of ill-defined, not directly observable, frequent or chronic phenomena and states of the world can only be used with extreme prudence because it can be highly biased and we are as yet ignorant of the conditions in which is it produced. This kind of 'data' can only become credible if produced and elicited in rigorous and replicable conditions.

Thus, most of our variables are directly observable and are counts and ordinal scales, both based and not based on counts. When possible, we quantify highly observable and traceable phenomena. After all, there is no such thing as a clandestine loss of the monopoly of violence by the state!

Note that there is anyway a difference between the boxes. In the 'monopoly' box we have a 'natural' 'absolute zero'. Nobody is assassinated in the streets; there is no rebellion; the country has not been invaded: this is 'perfect' monopoly. There is no full analogy for the other two boxes. Bureaucratic quality is particularly difficult to operationalise, as the size of the payroll obviously counts but it is not clear how. Beyond a certain threshold, a high number of bureaucrats may become a nuisance and a cost. But where should the threshold be established?[94] We thus avoid including the variable 'number of public employees in the bureaucracy' box.[95]

3.2.3. THE BOXES AND THE VARIABLES

3.2.3.1. BOX 1: MONOPOLY

3.2.3.1.1. MILITARY INTERVENTIONS

We define a military intervention as a stable incursion within the territory, with the intention of producing a substantial alteration of the political regimen of a sovereign country, using forces directly or through a third party (Gutiérrez and González 2009). This is a 'fuzzification' of the major existing military interventions databases plus additional documentation. It is a membership function that goes between 0 and 1 and reflects how large and authority-oriented the event is (following Rosenau's canonical definition (1969)). If the value is bigger than zero this means that the country suffered some kind of military intervention. Values near one correspond to 'canonical' interventions.

Sources: Regan 2002; Pearson and Baumann 1993; Tilema 1991; *New York Times* from 1960 until 2009.

3.2.3.1.2. DEATHS CAUSED BY ARMED CONFLICT

This variable is conceived as a proxy of the intensity of the conflict. If a state hosts an inactive or small guerrilla force, the deterioration of the monopoly of violence is less than if the guerrilla force has the clout to fight actively. We use here 'the PRIO Battle Deaths Dataset', which captures 'deaths resulting directly from violence inflicted through the use of armed force by a party to an armed conflict during contested combat'[96] (Bethany and Gleditsch 2005). We used the variable 'best estimate of annual battle fatalities' (*bdeadbest*). When not available, we used the dataset 'estimates of a floor and a ceiling of deaths' (variable *bdeadlow*[97] and *bdeadhigh*[98]). We averaged both and then divided the result over the total population.

Sources: The Battle Deaths Dataset version 3.00, released October 2009. Centre for the Study of Civil War – PRIO.

3.2.3.1.3. HOMICIDE RATES

We include here only violent homicide.[99] Note that while variables related to invasions and armed conflict fall into the province of the army, this one falls into the province of the police. As is standard, these are counted in rates per 100,000 inhabitants. Missing data were imputed in Mathematica®, using the ANCOVA method (Little and Rubin 2002).

Sources: WHO, Eurostat, United Nations Survey of Crime Trends and Operations of Criminal Justice Systems, some statistical agencies.

91 This point may be adjusted in time.

92 Programmed by us in Mathematica®. We used ANCOVA for all the continuous variables and Hotdeck for the categorical variable Quality of bureaucracy.

93 In Gutiérrez and González (2009) it is shown that these counts can benefit from a process of fuzzification. As seen below, we use the fuzzy version in our database.

94 This, of course, links with theoretical and empirical themes that have been debated endlessly.

95 It also has a high percentage of missing data.

96 'Contested combat is use of armed force by a party to an armed conflict against any person or target during which the perpetrator faces the immediate threat of lethal force being used by another party to the conflict against him/her and/or allied fighters. Contested combat excludes the sustained destruction of soldiers or civilians outside of the context of any reciprocal threat of lethal force (e.g. execution of prisoners of war)' (Bethany and Gleditsch 2005:3).

97 'Low estimate of annual battle fatalities' (Bethany and Gleditsch 2005).

98 'High estimate of annual battle fatalities' (Bethany and Gleditsch 2005).

99 Excluding, for example, deaths caused by transit accidents.

In summary, the monopoly basket has three variables. Of these, one is a count of rare and public events: an invasion. We will not expect here to see a high degree of over– or under estimation. Two are cardinal counts coming from convenience samples (see Tables 3.2 and 3.3).

3.2.3.2. BOX 2: TERRITORIAL CONTROL

3.2.3.2.1. ILLEGAL ECONOMIES

Both theory and many empirical studies tell us that if a state cannot tax and control a significant crop, this will be the source of continuous fractures of statehood (Rotberg 2004; Sung 2004). We count here *only* production.[100] The variable is an ordinal scale, between 0 and 1. It is zero when no illegal production is reported; 0.25, when the country produces an illegal substance but is not classified as a major producer by the United Nations Office of Drugs and Crime; and a mark between 0.5 and 1 is given to large-scale drug producers.[101]

Source: UNODC.

3.2.3.2.2. ROADS

This, of course, is a 'natural' territorial variable. 'Paved roads are roads surfaced with crushed stone (macadam) and hydrocarbon binder or bituminized agents, with concrete, or with cobblestones, as a percentage of all the country's roads, measured in length' (International Road Federation). Missing values were imputed with ANCOVA method.

Source: World Bank (World Development Indicators) and CIA, The World Factbook.

3.2.3.2.3. CONNECTIVITY – COMMUNICATIONS

This is the simple average of telephone lines and number of postal offices per 100,000 inhabitants. The variable 'telephone lines' is defined in the following way:

> Main (fixed) telephone lines refer to telephone lines connecting a customer's equipment (e.g., telephone set, facsimile machine) to the Public Switched Telephone Network (PSTN) and which have a dedicated port on a telephone exchange. Note that for most countries, main lines also include public payphones. Many countries also include ISDN channels in main (fixed) lines (see below ISDN and ADSL). Main (fixed) telephone lines per 100 inhabitants are calculated by dividing the number of main lines by the population and multiplying by 100 (World Bank 2010).

The variable 'total number of permanent post offices' is defined in the following way:

> It represents all offices open to the public and operating on fixed premises. Permanent post offices include offices staffed by officials of the designated operator and offices staffed by persons not connected with the designated operator. Offices staffed by designated operator officials (heading 3.2) may be full-service offices or secondary offices. Full-service post offices are post offices to which, in principle, customers may go for all postal services. This category also includes sections of exchange offices or sorting offices offering similar services. Secondary offices generally have reduced services and, in principle, come under a main post office (Universal Postal Union 2010).

Missing values were imputed with ANCOVA method.

Source: World Bank World Development Indicators and CIA, The World Factbook (telephone lines); Universal Postal Union (post offices).

In summary, territorial control includes four variables. Of these, one is a count of a rare and public event or state of the world, the others are cardinal counts that come from convenience samples and have plenty of missing values that were imputed (see Tables 3.4 and 3.5).

3.2.3.3. BOX 3: BUREAUCRACY

3.2.3.3.1. MISSING VALUES

This variable is simply a counter for the missing data in the rest of the variables over the total by year. It is supposed to be a proxy of the capacity of the state to produce information, which is a fundamental tenet of organisational quality. Indeed, it involves a clear bias against centrally planned economies, but since the bias is systematic it can eventually be controlled for.

Source: In-house count.

3.2.3.3.2. QUALITY OF THE BUREAUCRACY

This is a mark granted to countries by the PRS Group. It is defined in the following way:

> The institutional strength and quality of the bureaucracy is another shock absorber that tends to minimize revisions of policy when governments change. Therefore, high points are given to countries where the bureaucracy has the strength and expertise to govern without drastic changes in policy or interruptions in government services. In these low-risk countries, the bureaucracy tends to be somewhat autonomous from political pressure and to have an established mechanism for recruitment and training. Countries that lack the cushioning effect of a strong bureaucracy receive low points because a change in government tends to be traumatic in terms of policy formulation and day-to-day administrative functions (Mckenzie 2002: 33).

100 The potential effect of the existence of other modalities of illegal economies is probably captured by other variables, in other boxes.

101 According to the following formula: [(1-0.5)/(Max(X)-Min(X))].(x_i-Min(X))+ 0.5.

Note that this is a heavily hedged definition (no 'drastic changes', 'somewhat autonomous'). Missing values were imputed with the hotdeck method using Manhattan distance (Little and Rubin 2002). See Tables 3.6 and 3.7.

Source: The PRS Group - International country risk guide.

3.2.3.3.3. TAXES

Taxation is both historically and conceptually one of the *sine qua non* dimensions of statehood (Brautigam and Fjeldstad 2008). The ability of a state to collect tax is intimately related to the regulatory capacity of the state. Of course, taxation does not behave in a linear manner as the state can overtax and, by this and other forms, engage in predatory dynamics. Even then, below a threshold, taxation is a key proxy of bureaucratic power.

We built this variable from the World Bank 'tax revenue (% of GDP)':

> Tax revenue refers to compulsory transfers to the central government for public purposes. Certain compulsory transfers such as fines, penalties, and most social security contributions are excluded. Refunds and corrections of erroneously collected tax revenue are treated as negative revenue (World Bank 2010).

To fulfil informational gaps in the World Bank variable, we used other sources.[102] We divided tax revenue by GDP. Missing values were imputed with the ANCOVA method.

3.2.4. MISSING DATA

Our database has ten variables and three boxes. Of these, four did not need imputation because they did not have missing values (they were based on counts of rare and public events). Four other variables had a marginal percentage of missing values.

The main method of imputation we used was ANCOVA (programmed in Mathematica®). The ANCOVA has some properties that are highly desirable in our context. It is not iterative and hence does not have convergence problems; if there is a single pattern of missing values, as when all of the data of one year are missing, the method discerns it, something that iterative ones do not do; the method produces correct estimations for the sum of squares, standard errors and F tests.

To evaluate the performance of the imputation method, we performed the following experiment: we deleted randomly from a data table with 50 individual and 12 years (between 1996 and 2005) a certain percentage of the data and we replaced it by ANCOVA-imputed data. We calculated the NRMSE (normalized root mean squared error):

$$\frac{\sqrt{mean[(\hat{y}-y)^2]}}{std(y)} \quad \text{f(3.2.4),}$$

where $\hat{y}$ is a vector containing the imputations, and y is the vector with the real values that were deleted. The media and the standard deviation are calculated over these values. In the experiment, we increased the percentage of deleted data (10%, 20%,…,90%). For each level of deletion we ran the experiment 100 times. As can be seen in Table 3.8., ANCOVA performs quite well, even in the most demanding scenario.

3.2.5. CONCLUSIONS

As did other PSPI database builders, we faced many problems related to having access to a wealth of data, much of which is incomplete, deteriorated, noisy and/or not highly reliable. Only some of these problems were solved. What we did achieve was the following:

- We avoided conceptual stretching, and organised the capture of data from a clearly defined, and parsimonious, theoretical perspective, based on a long-term research programme. We left out developmental and political regime variables. This allowed us to produce regressions that empirically evaluate the relation between them and the conception of the state we are operationalising.
- We did not define unobservables using unobservables.
- We maximised the number of variables that are based on counts of public and rare events, that are near the 'census' end of the spectrum of counts, that can be controlled and corrected, and that have small probability of resulting in gross under– or overestimations (military interventions, illegal economies, missing data).
- We minimised the use of variables based on ordinal scales created by experts (or in-house coders). We include two externally created (thus, not controlled by us) scales of this type in the bureaucracy box because they appeared too central to the dimension of statehood we were trying to capture to eschew them lightly. We also include two in-house-built scales. The first one is military interventions, which was based upon the main existing relevant databases (plus additional documentation) and coincides with them 100% in the major events (for the details, see Gutiérrez and González 2009). The second one is missing values, which is a simple, public and completely replicable variable. We have full control over none of them.
- We also minimised avoidable ambiguity. Intrinsic ambiguity remains, but we can live with that (see below).

102 Regional Banks (CEPAL, EUROSTAT, African Bank, Asian Development Bank) and statistical offices of each country. We used the source that provided more information, and if all sources were equivalent we used the World Bank data. For details, see MAT Codebook.

3.3. NORMALISATION AND AGGREGATION FUNCTIONS

3.3.1. INTRODUCTION

Our database intends to capture the core dimensions of statehood. It gathers quantitative data about the level of deterioration of each of these dimensions (our baskets). This is the operational meaning of fragility. Hence, in MAT more is less (a high score means more deterioration). As with other PSPIs, we have to deal with heterogeneous data[103] and a three-process transformation:

a. Normalisation (all the variables have to be put on a unique scale);

b. Variable aggregation;

c. Box aggregation.

We used one normalisation procedure, and three aggregation procedures. Of these, two produce numbers and orderings. They can be seen as 'competitive' and 'redundant'. The third produces intervals and orderings. All the aggregation rules that we present here are based on the following assumption: the key data for characterising fragility are extreme values. This is our basic assumption, which is reflected in the choice of weights when building a compensatory method and in the structure of comparisons for non-compensatory methods. The underlying intuition is the following: very good or poor performance in specific areas has significant spill over effects, many of which are not captured by the quantification. Actually, things can go the other way round. The obvious example is that weak statistical offices may be too benevolent (not out of malice but out of simple incapacity) in the count of homicides, thus increasing artificially the 'strength' of the state. In such cases, we would like to remain with the extreme datum (poor behaviour of the information agencies, quantified for example by the amount of missing data), and not with the intermediate one (a mediocre but not outright bad violence record, which exists only in the books).[104] We would like a country that has an excellent performance in one dimension and a disastrous one in the other to get a different mark from one that has a mediocre performance in all three – and we would also want them to fall in different clusters in a good classificatory exercise.[105]

The assumption is strengthened by two key sources. The first one is social history on state construction, which stresses the critical role played by 'virtuous' or 'vicious' circles. For Tilly (1993), for example, strong armies 'pulled' ahead sequentially other dimensions of state strength (bureaucracy and taxation). Similar visions have been proposed repeatedly with a wealth of historical evidence (Tilly and Stinchcombe 1997). The second source strengthening our assumption is data. We know that in PSPIs: (a) there is a high correlation between the variables, which is desirable *if* it is explicitly taken into account; and (b) part of the data is hazy, non-credible, difficult to interpret or corrupted. Putting together (a) and (b), *credible* extreme values should have a special influence on the grading and ranking of the countries.

We utilised two normalisation procedures and developed three aggregation functions (compensatory numerical, non-compensatory numerical, non-compensatory interval). We briefly describe them below.

3.3.2. THE NORMALISATION FUNCTION

All of the values are normalised so that the values of variables and boxes fall within a scale of 0–1. Using this scale means that some variables (including homicide rates) do not have to be normalised.

We normalised the variables in the following way:

1. Monopoly:

- **a.** Conflict intensity and homicides were transformed as rates;

2. Territorial control:

- **a.** We logarithmically transformed the variables 'postal offices', 'telephone lines' and 'roads' because their distribution was skewed to the left;
- **b.** We built 'illicit economies' as a categorical variable, according to the size of coca and opium production as evaluated by the relevant sources (UNODC Crime Trends 2010);

3. Bureaucracy:

- **a.** Quality of bureaucracy is a monthly evaluation, so we averaged the values over a twelve-month period.

We logarithmically transformed the 'taxes' variable.

3.3.3. AGGREGATION FUNCTIONS

3.3.3.1. THE COMPENSATORY PROCEDURE

We aggregated the variables by an Ordered Weighted Average (OWA) and then aggregated the boxes by the Choquet integral.

In the OWA operator, the data to be aggregated (in this case, the variable values in each country vector) have to be sorted in descending order: $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(p)}$

103 We reduce heterogeneity with respect to the majority, with the exception of BTI (which has only two variables, ordinal scales constructed by in-house coders). See section 3.2.

104 What happens when the wrong data appears at the extremes? This seldom happens, because for example, in regard to violence counts, the endemic problem is under– not over-estimation. Besides, obviously erroneous outliers are easier to identify, correct or prune.

105 Actually, one of the criteria to evaluate whether a classificatory exercise is good or bad is to see if it can tell apart these two different types of case.

Then, the variables are multiplied by the weights:

$$OWA(x_1, x_2, \ldots, x_p) = \sum_{i=1}^{p} w_i x_{(i)} \qquad \text{(f.3.3.1.)}$$

As seen in section 1.3, one of the reasons that compensatory methods are vulnerable relates to the choice of weights. Ad hoc weights return ad hoc marks and rankings. Here, we imputed weights using the method of 'minimum variance', which – as discussed above – avoids many of the potential problems in this area.[106] For this, we minimised the following function:

$$min \sum_{i=1}^{p} w_i^2 \quad \text{(f.3.3.2)}$$

$$\text{Subject to } \sum_{i=1}^{p} w_i = 1,\ \sum_{i=1}^{p} w_i \frac{p-i}{p-1} = \alpha,\ w_i \geq 0.$$

The operation yields the following weights for the variables in the three dimensions:

Monopoly	Territorial control	Bureaucratic capacity
0.533333	0.533333	0.533333
0.333333	0.333333	0.333333
0.133333	0.133333	0.133333

It is very important to note that the weights are *not* assigned to the variables but to the values. It may happen that for a country *A* the biggest weight is assigned to variable *x1*, while for country *B* the most important variable is *x2*. For example, for Haiti the order in the bureaucratic box (see Tables 3.9, 3.10 and 3.11) is quality of bureaucracy > taxes > missing values, so that the calculation of its bureaucracy value is ***(1 x 0.5333) + (0.3382 x 0.3333) + (0.1875 x 0.1333) = 0.6711***, while the order for Japan is taxes > quality of bureaucracy = missing values, so its bureaucracy value is ***(0.327 x 0.5333) + (0 x 0.3333) + (0 x 0.1333) = 0.1744***.

Using this method ensures that weights do not appear as substitution rates but as indications of 'relative importance with respect to extremeness'. Even if they were interpreted as substitution rates, they have a non-linear relationship to each other (as would be expected), they do not allow for full compensation, and they are grounded on the data of each country (for the details see Annex 1).

Having attained the values for each box, we aggregated them by a Choquet integral. This fuzzy aggregation function takes into account not only single variables but every subset of variables. By definition, the Choquet measure of the empty set is 0; the Choquet measure of the universe is 1. In order to attribute the weights to the other subsets of boxes we used the correlations between them (for each year). Note that combining both procedures – for variables and boxes – we take non-orthogonality fully into account. This means that the 'official' weights are not distorted by unaccounted-for correlations. For example, for the year 2005 the correlation between the boxes is shown in Table 3.12. The inverse of these figures is shown in Table 3.13. Then we divided each inverse by the sum of all three, which yields the results shown in Table 3.14.

The subsets of cardinality one are given the least fuzzy measure for sets of cardinality two. Thus, the fuzzy measure (weights) of each subset are imputed for this example in Table 3.15. Once this has been done, the algorithm for the Choquet integral is applied (see Table 3.16).[107] (For a relevant bibliography on the Choquet integral see Grabisch 2000 and Mufofushi and Sugeno 2000).

3.3.3.2. DOWNSETS

Downsets are one of the most important order structures in posets (Roman 2008). In lattices, which are an important type of poset,[108] downsets are called ideals (and the downset of a single element is a principal ideal). An idea that naturally comes to mind when trying to build a function from $\mathbb{R}^n$ to $\mathbb{R}$ that sends order structures into order structures in the best possible way is to count the cardinality of the downsets of each case (for us, country) and divide it by the total number of cases. Naturally, an analogous operation can be done with up-sets (filters, in the context of lattices). If the poset has a top, then it will be attributed a 1, and if it has a bottom then it will get 0. Otherwise, $\downarrow x = \{y \mid y \leq x\}$. The indicator tells how many countries are distinctly (e.g., in all dimensions) worse or better than the given case x. It is quite obvious that down- and up-sets can be used to aggregate variables and boxes.

It turns out that in many respects this is a quite well-behaved non-compensatory function. No parameter has to be taken out of the hat to produce a working characterisation of the country.[109] This is a fully non-invasive aggregation function. In the spirit of section 1.3, we would expect it to have its downsides. In effect, the method has two problems. The first one is that it only captures hierarchies, not values. Any regression utilising this indicator only 'sees' how the country has evolved in time with respect to other countries. It captures the 'relative' not the 'absolute' trajectory. If the data have deteriorated enough, this may actually be an advantage. The second problem is that it does not respect IIA.[110] This is potentially very serious. However, an experiment with computer-generated data suggests that the violation is basically marginal (see Table 3.17).[111] The theory says that the problem will appear, but the experiment suggests that it will only very seldom appear.

106 Subject to an ority of $\alpha = 0.7$. For the details, meaning of ority, etc., see Beliakov, Pradera and Calvo 2007. We chose a relatively high ority in order to give high weights to extreme values, without wholly discounting the rest. This is in line with our basic assumption.

107 For a relevant bibliography on the Choquet integral see Grabisch 2000 and Murofushi and Sugeno 2000.

108 Where the operations of finding the lowest upper bound and biggest lower bound can be performed over every pair of elements.

109 It can also produce clustering algorithms.

110 There is probably no function that is not the leximin that is non-compensatory, anonymous, symmetric, Pareto and IIA.

111 See an explanation of the experiment in Annex 1.

The experiment reveals that when ten countries are taken out randomly of the data base, 1% of the rankings are reversed. This seems tolerable. Note as well that downset aggregation allows us to deal well with deteriorated information.[112] The correlation between the Choquet and the downsets marks are shown in Table 3.18. However, there are many ranking reversals between them. The percentage of reversals for year 2005 is 22.95%.

3.3.3.3. MIN–MAX INTERVAL REPRESENTATION

If the assumption is that extreme values have a special status and should be flagged in some way or another, the simplest possible aggregation is to take the minimum and the maximum of the given vector of values and represent each case xi by the interval $\{x_{imin}, x_{imax}\}$. Of course, this rule can also be applied to variables and boxes (see example in Table 3.19).

In principle, since the *range* of the *minmax* function is only a poset $(R^2\ \mathbb{R}^2)$ it does not respect the universality of domain. As discussed in section 1.3, this axiom does not seem fundamental for PSPIs. For some pairs, it may be desirable to acknowledge that it is not possible to know which of the two is better. However, since the number of non-comparable pairs is relatively high after applying the function, we complete the min–max aggregation with the fuzzy plausibility function:

$$A(A < B) = \frac{m(B) - m(A)}{w(B) + w(A)}$$

where $m(A)$, $m(B)$ are the centres of A and B, and $w(A)$, $w(B)$ their half-width (Sengupta and Pal 2009).

This returns a value that decides how plausible the statement 'interval A is bigger than interval B' is. In our context, this would mean 'country A is more fragile than country B'. Thus, our complete *minmax* function operates in the following way:

a. As a function that produces values, simply take the min and max of the vector of values of each country, and form the interval {min, max} using them;

b. As a function that produces *ranks*, take the min and the max as before, and then apply the plausibility to the unsolved comparisons. It is easy to see that this produces a consistent hierarchy.

Formulated in these terms, min–max works rather nicely. It fulfils all the desirable conditions (axioms) discussed in section 1.3. In particular, it is:

a. Non–compensatory;

b. Pareto, monotonic, anonymous, symmetric, independent of irrelevant alternatives and universal.

This small miracle was obtained at a cost: added complexity. Intervals give more information than numbers, so they are more complex to read. Importantly, an interval is *not a* number. It has its own arithmetic (Sengupta and Pal 2009). Have we produced, then, a purely heuristic device, usable only to represent graphically our intuition of where a country stands? As will be shown in the next section, our method holds far greater potential: these intervals can be operated on, and – crucially – regressed on.

3.3.4. CONCLUSIONS

This section was dedicated to the discussion of the aggregation functions that we developed to operate with the data contained in the MAT database. Their characteristics are synthesised in Table 3.20. The Choquet aggregation works well, but the problem is that it is compensatory. Yet all of the weights are taken from the data, and have a plausible interpretation as importance in explaining variance, not as substitution rates. If they are treated as substitution rates, at least they behave non-linearly, and are different for each country. A key trait of the Choquet integral is that it is *not* fully compensatory: above a threshold an increase of fragility in one dimension *cannot* be compensated by improvements in the others (for illustrations, see Annex 1). All of this seems to tell a more reasonable story than simply averaging. Indeed, we feel that it represents a distinct progress with respect to ad hoc, unexplained and fully compensatory weighting. At the same time, two issues remain: one varying parameter[113] and the fact that, within bounds, dimensions can be represented numerically in terms of others (though this representation changes for each country).

The aggregation by the cardinality of down- and up-sets captures the implicit hierarchy in the domain of the function, and is neatly non-compensatory. It does not respect the independence of irrelevant alternatives, which is not that serious because violations appear to be marginal. The fact that it only captures relative position, and not hard data, is indeed a limitation. However, in contexts where the PSPI data seems uncertain or hazy, the best way of doing any kind of serious quantitative exercise is to shed spurious precision and stick to the ranks.

Minmax is also non-compensatory. Note that both *minmax* and downsets are also non-invasive. The researcher does not have to invent things that might (and, as shown in chapters 1 and 2, do) abruptly change the results of the analysis.[114] The additional information provided by the interval representation can be put in operational form, though there is still a long road ahead in at least two senses: translating key operations into interval language is not a trivial task, and after doing this there is still much to be achieved.

112 You need only compare on the existing data, though we did not test this characteristic.

113 Ority. However, we did not establish it on an ad hoc basis,. We calibrated it with the data so that it fitted our basic assumption without discounting the other variables too heavily (Ragin 2008). Let us say that this is a 'moderately pessimistic' aggregation (it gives more importance to the most fragile dimension, without discounting the rest).

114 We also used the Choquet integral in a non-invasive way, minimising the parameters established exogenously.

3.4. FUZZY TOOLKIT

3.4.1. INTRODUCTION

This section constitutes a brief presentation and illustration of the fuzzy (and more generally 'approximate') tools that we have programmed or developed.[115] We used the data as our guide and only resorted to new tools when:

a. The normally used tools were found wanting in some crucial sense for our purposes;

b. The new tools were important complements to the traditional tools;

c. Our formalisation produced, or demanded, objects that could not be manipulated in the traditional fashion.

We present here four tools:

a. Multivariate regression over fuzzy triangular numbers. After testing several versions, we chose the Hojati et. al. regression (Hojati et al. 2005). There is no commercial version of this, so we programmed it in Mathematica®, and validated it with several examples, including those that appear in Hojati's paper;

b. A multivariate interval regression, analogous to the Tanaka one but tailored to deal with intervals. We created this in the context of our research programme. It was also submitted to diverse forms of validation (Gutiérrez et al. 2010). This was programmed in Mathematica® as well;

c. A multilevel regression, that can admit the triangular or the interval one as a building block, and that captures various levels of variation (in our case, country and time period). This was developed by us and programmed in Mathematica®;

d. Several forms of fuzzy clustering, based on Miyamoto et al. 2008.

Other developments, conceived to enhance the power and flexibility of the fuzzy/approximate toolkit, are in progress. We will not indulge here in the exposition of the technical details, but rather sketch a contrast with the crisp/probabilistic regression, and discuss some of the implications of the differences. Then we present a concrete example of the tools discussed here: a fuzzy clustering of the cases contained in the MAT database (year 2005), which separates countries 'horizontally' and 'vertically', producing a much neater classification of the world states than the one that PSPIs normally offer – e.g., only 'vertical' (hierarchical). The use of *minmax* aggregation in fuzzy regressions proper is circulated separately. We end by flagging some open questions.

115 Others are in progress.

3.4.2. CRISP AND FUZZY REGRESSIONS

It would be extremely naïve to claim that fuzzy regressions are 'superior' in any reasonable sense to crisp ones; rather the contrary is true. Probabilistic models are backed by a very powerful mathematical theory, which has persistently demonstrated its power to orient humans in the world (and to help them transform it). Hard work has given origin to a rich interface between probability theory and 'possibility theory', one of the aliases of fuzzy sets, but all this is still relatively recent (see for example Ross et al. 2002 and Manton et al. 1994).

The problem with the impressive machinery behind probabilistic models is that it demands that the data meet certain rather severe conditions. First of all, there must be data proper, such as numbers, over which you can operate arithmetically. As was seen in the previous two chapters, this is not a truism. Second, probabilistic modelling is based on the assumption of the existence of an underlying distribution. This assumption is stronger than what is normally admitted (Freedman 2005). Third, as Freedman (2005) and several other authors show, when the preconditions are not met many of the statistical operations in the social sciences and other disciplines end up not making sense (Bouyssou et al. 2000). Examples include:

a. Operations on non-numbers;

b. Statistical inferences over very deteriorated data;

c. Statistical inferences over apparently well-behaved data, but which has not been appropriately collected. A very important instance of this is the so-called convenience samples, which are a very fundamental input of PSPIs and third wave social indicators in general. Convenience samples contain a potentially huge mass of information but, since they are not censuses nor have they been collected randomly, they are *not* fit to produce statistical inferences. For example, Ball reports that the difference between the 'true' number of conflict-caused homicides and those produced by the convenience samples in Peru was nearly 1:2 (Truth Commission 2003).[116] Also, the proportions changed: in the convenience samples, the state was the main victimiser, in the corrected version it was the guerrilla force. Ball and others have further shown that an increase in convenience counts can simply be a product of an improved efficiency of the given office (or of more attention by the public on the phenomenon that is being counted) (Ball 2001).[117] Note that the aggregation functions we have chosen diminish this problem (see section 3.3).

For these and other reasons, there are many situations where the use of probabilistic models is dangerous and likely to produce

116 In extreme cases, the difference can be of one order of magnitude.

117 Anybody who has studied qualitatively the workings of relatively weak states must be aware of these types of effect. Of course, as stressed in chapter 1, this does not mean that anything goes. Certainly the figure of homicides in DRC or Colombia is imprecise, but it is higher than Norway's or Japan's.

mainly noise. Freedman (2005) has made an extraordinary job of explaining why many of the statistical inferences drawn in the social sciences are spurious. As seen is previous chapters and sections, PSPIs are plagued by several of the problems reported here. They do not deal with numbers proper (for a rich discussion of this in another context, see Bouyssou et al. 2000). They capture expert information about very vague and poorly defined questions. They deal with intrinsic ambiguity, and besides add new layers of inevitable ambiguity in the process of database construction and aggregation. Their aggregation functions are built upon a fiction (the existence of substitution rates) that has no support whatsoever, theoretical or empirical. Additionally, the concrete rates are produced by ad hoc decisions that are not even discussed.[118] Hence in some cases, hence, the construction of PSPI-based probabilistic models and statistical inferences may yield over-optimistic results.

The question is therefore if there are formalisms that produce coarser outputs but that are based on less stringent assumptions. 'Approximate' (fuzzy, rough, case by case) models intend to do precisely that. Thus, a fuzzy regression is similar to a crisp one in the sense that it tries to quantify the associations between several variables, but has two fundamental differences:

1. It manipulates other types of objects, which express other types of uncertainty (non-probabilistic uncertainty). For example, degrees of belief can be expressed via a triangular fuzzy number, where the most plausible value gets a 1 while any other values to the right and to the left gradually decrease in plausibility (not necessarily symmetrically).

2. It produces other kinds of outputs, which are analogous (for example, coefficients) but have a different interpretation.

These unfold in many other differences, which are both interpretative and operational. The main ones are detailed in Table 3.21. Crisp regressions study variability, fuzzy ones variability in the face of ambiguity. Any self-respecting crisp regression has to meet some crucial conditions. Indeed, much basic statistical research has been able to weaken some of these, but there is a core method that will inevitably remain (Freedman 2005). Fuzzy regressions are not tied by these assumptions.

The interpretative antithesis has a clear operational correlate, which, as said above, consists basically of the fuzzy regression producing coarser but much more 'ambiguity-friendly' quantitative results. Note that confidence intervals have little to do with the fuzzy estimations of the dependent variable. An extremely crucial operational difference is the following: in crisp regressions, the more data you have the better, because you are able to decrease gradually the inexactitude of your estimation. Standard deviations fall with the number of data. In fuzzy regressions, exactly the contrary takes place. The more data you have, the bigger ('coarser') are the estimates (be they fuzzy numbers or intervals) that you need. This is caused by the fact that in fuzzy regressions you 'cover' the data with your estimations, and thus, as the space filled by the data increases, so does the cover. We believe that this captures neatly the contrast between both regressions, and the big potential of fuzzy regression when dealing with ambiguous and faulty data. By manipulating objects that are not crisp numbers, fuzzy regressions are able to take ambiguity on board and at the same time establish a penalty for bringing in too much noise. This depends upon an optimisation operation by the researcher: he/she should use all the data available until it generates excessive (intractable) ambiguity.

There is a whole family of fuzzy regressions. This is not worrying in the least, and is paralleled in the crisp regression world. The Hojati regression has several advantages, the main one being that in contrast to others it is very easy to interpret. For example, its estimates are based on notions of area (Hojati et al. 2005: 175). On the other hand, the crucial advantage of our own interval regression is that it is fully non-invasive. While for fuzzy numbers a series of parameter decisions have to be taken – the form of the number (whether triangular, Gaussian, etc.), the functional form of the membership function – intervals are flat and non-invasive. No number or parameter is taken from the hat and plugged more or less arbitrarily into the model. There is a mathematical theory and arithmetic of intervals,[119] which we used to develop this regression (Moore and Bierbaum 1979).

In summary, fuzzy regressions are powerful tools for working with ambiguous, incomplete or severely deteriorated data. In this context, they produce coarser but more credible results (see Kim et al. 1996). There are several types of fuzzy regression. We programmed a very elegant variant found in the literature but developed in parallel an interval version that is non-invasive. Both can serve as the core machinery for any type of multilevel, hierarchical model. Hence, interval aggregation produces objects that are not only heuristic, but we can operate on them, and incorporate them in regressions.[120]

3.4.3. FUZZY CLUSTERING

We do not have to go that far to gain new insights from the data from fuzzy tools. We illustrate this point here with an exercise that unveils classificatory aspects that are generally lost in PSPI outputs. The point is that state typologies have to take into account not only 'hierarchical' but also 'horizontal' criteria. This means that we are interested in differentiating countries according to their 'degree of fragility', but also according to their differential performance across the three main dimensions of statehood measured in the database. All clustering exercises have the following characteristics:

118 And besides, the 'official' weights are not the true ones, because they are perturbed by non-orthogonality.

119 From another point of view: intervals are not fuzzy numbers. Thus, we have incurred in a certain notational abuse calling the interval regression a fuzzy one. However, all the discussion about the specifics of the interpretation of the regression results stand.

120 Both Hojati and interval regressions can operate with mixed data (in the Hojati case: real, ordinal and fuzzy variables; in the other one real, ordinal and interval variables).

a. They extract and reorganise data from a database, according to a notion of distance;
b. They separate countries in categories, which is one of the specific objectives of PSPIs.

The main difference between crisp and fuzzy clustering is that, while the former attributes one object to one and only one cluster, the latter imputes memberships (from 0 to 1) of all objects to all clusters. In fuzzy clustering, a country can partially belong to several categories. This has a very natural and intuitive interpretation in the social sciences in general, and in PSPIs in particular. A country such as the Philippines can be relatively close to different prototypes of states. Vectors of real numbers or integers can be fuzzily clustered. To perform our exercise, we used our package FuzzyClust programmed in Mathematica®.[121]

We ran the 2005 MAT data[122] with the F c-means algorithm (Miyamoto et al. 2008). The results, shown in Table 3.22, are rather striking. Note that the clusters are hierarchically ordered, but only partially. There are nine clusters. The best performers – the developed countries – all belong to cluster 8. The ones that fare worst appear in cluster 3. The countries that have failed in the crucial terrain of the monopoly of violence, but that otherwise have fairly acceptable behaviour for their level of development (Colombia, Sudan) appear in a separate cluster (2). Note that each cluster, save at the extremes, where everything or nothing works well, contains countries at different levels of development. Even after its economic miracle, China has a low GDP per capita, but its state appears as a quite good performer, though with some problems (related to quality of bureaucracy), which it symptomatically shares with Chile. Note also that each cluster is marked by its own combination of strengths and weaknesses. Differential performance here is the focus of attention, overriding concerns about ranking. Since the exercise is fuzzy, several countries belong partially to several clusters. For example, Nigeria shares traits of type 1 and type 5 countries.

Naturally, this has been a purely heuristic presentation, oriented towards showing how other foci of attention and forms of organising the information can offer interesting insights of key theoretical problems (here, differential performance). In the technical annex we will walk the reader through the details of the exercise. As in other formal models, the choice of parameters is of paramount importance; these are explained in the Annex.

121 Different from the commercial package Fuzzy Logic® by Marian Stachowicz and Lance Beall, which also runs over Mathematica® and has a good clustering routine. But we needed the flexibility to change rules, parameters and algorithms, so we ultimately decided to make our own package. Fuzzy Logic was very instrumental for validating our program (where the clustering algorithms coincided).

122 Aggregated by OWA in each dimension. Thus, in this exercise each country is represented by a vector of three numbers.

TABLES: CHAPTER 3

Table 3.1: Fuzzy transformations of CPIA scores

Source: Baliamoune-Lutz and McGillivray 2008: 11

Cluster	Best outcome	μ_h	Worst outcome	μ_l	A	B
A: Economic management	5	0.833	2	0.334	4.6049	0.4838
B: Structural policies	6	0.999	2	0.334	11.4135	0.3939
C: Policies for social inclusion/equity	5	0.833	1	0.167	4.8259	0.500
D (1): Public sector management and institutions	6	0.999	3	0.500	13.8412	0.500
D (2): Public sector management and institutions	6	0.999	4	0.667	18.6969	0.6296

Table 3.2: Correlation – monopoly

	Military interventions	Deaths caused by the armed conflict	Homicide rates
Military interventions	1		
Deaths caused by the armed conflict	0.25737519	1	
Homicide rates	0.06351079	0.066329335	1

Table 3.3: Variables and % of missing data – monopoly

Year	Number of countries	MONOPOLY			
		Military interventions	Deaths caused by the armed conflict	Homicide rates	
				Original	Imputed
2001	179	0	0	46.9%	1.7%
2002	179	0	0	48.0%	1.1%
2003	179	0	0	31.8%	2.2%
2004	179	0	0	2.2%	1.1%
2005	179	0	0	34.6%	1.7%
2006	181	0	0	35.4%	1.1%
2007	180	0	0	41.1%	1.7%
2008	180	0	0	52.2%	1.7%
Total	1436	0	0	36.6%	1.5%

Table 3.4: Correlation – territorial control

	Connectivity	Roads	Illegal economies
Connectivity	1		
Roads	0.75187533	1	
Illegal economies	0.0530346	0.008723562	1

Table 3.5: Variables and % of missing data – territorial control

Year	Number of countries	TERRITORIAL CONTROL						
		Conectivity				Roads		Illegal economies
		Telephone lines		Post offices				
		Original	Imputed	Original	Imputed	Original	Imputed	
2001	179	3.4%	1.1%	21.8%	4.5%	52.5%	5.0%	2.8%
2002	179	2.8%	0.6%	22.9%	4.5%	54.2%	5.0%	2.2%
2003	179	2.2%	1.1%	16.2%	4.5%	60.3%	5.0%	2.2%
2004	179	3.9%	0.6%	16.8%	4.5%	59.8%	5.0%	2.2%
2005	179	2.2%	0.6%	17.3%	4.5%	70.4%	5.0%	1.7%
2006	181	2.2%	0.6%	23.8%	5.5%	70.7%	6.1%	1.7%
2007	180	2.8%	0.6%	17.8%	4.4%	74.4%	6.1%	1.7%
2008	180	0.6%	0.6%	23.3%	4.4%	86.7%	6.1%	2.2%
Total	1436	2.5%	0.7%	20.0%	4.6%	66.2%	5.4%	2.1%

Table 3.6: Correlation – bureaucracy

	Quality of the bureaucracy	Taxes	Missing values
Quality of the bureaucracy	1		
Taxes	0.26382985	1	
Missing values	0.27038956	0.092264968	1

Table 3.7: Variables and % of missing data – bureaucracy

Year	Total countries	BUREAUCRACY			
		Missing values	Quality of the bureaucracy	Taxes	
				Original	Imputed
2001	179	0	22.3%	12.8%	2.2%
2002	179	0	22.3%	11.2%	1.7%
2003	179	0	22.3%	7.3%	1.7%
2004	179	0	22.3%	6.7%	1.7%
2005	179	0	22.3%	7.8%	1.1%
2006	181	0	23.8%	7.2%	1.1%
2007	180	0	22.8%	8.9%	1.1%
2008	180	0	22.8%	22.8%	1.1%
Total	1436	0	22.6%	10.6%	1.5%

Table 3.8: Mean NRMSE for ANCOVA method

%deleted	10	20	30	40	50	60	70	80	90
NRMSE	0.219	0.202	0.234	0.201	0.221	0.219	0.225	0.228	0.234

Table 3.9: Examples of the OWA aggregation for the monopoly box

Country	Military intervention	Conflict deaths	Homicide rate	OWA
United States of America	0	0.0048	0.0963	0.0530
Haiti	0.8	0	0.1041	0.4614
Colombia	0.4458	1	0.7233	0.8339
Switzerland	0	0	0.0136	0.0073
Burkina Faso	0	0	0.0073	0.0039
Iraq	1	1	0.0951	0.8793
Japan	0	0	0.0075	0.0040

Table 3.10: Examples of the OWA aggregation for the territorial control box

Country	Connectivity	Roads	Illegal economies	OWA
United States of America	0.1809	0.1074	0	0.1323
Haiti	0.7425	0.3522	0	0.5134
Colombia	0.3112	0.4730	0.5039	0.4679
Switzerland	0.1061	0.0000	0	0.0566
Burkina Faso	0.6496	0.7563	0	0.6199
Iraq	0.4801	0.0447	0	0.2710
Japan	0.1674	0.0595	0	0.1091

Table 3.11: Examples of the OWA aggregation for the bureaucracy box

Country	Bureaucratic quality	Missing values	Taxes	OWA
United States of America	0	0	0.3082	0.1644
Haiti	1	0.1875	0.3382	0.6711
Colombia	0.5	0.0625	0.2032	0.3427
Switzerland	0	0	0.3262	0.1740
Burkina Faso	0.75	0.0625	0.3061	0.5104
Iraq	1	0.125	0.3788	0.6763
Japan	0	0	0.3270	0.1744

Table 3.12: Correlation between the boxes

	2001	2002	2003	2004	2005	2006	2007	2008
M – CT	0.3394	0.3509	0.3424	0.4422	0.3128	0.2929	0.3175	0.3079
M – B	0.3248	0.3553	0.3560	0.3752	0.2967	0.3330	0.3868	0.3080
CT – B	0.5887	0.5763	0.5908	0.5665	0.5454	0.5165	0.5445	0.4875

Table 3.13: Inverse correlation between the boxes

	2001	2002	2003	2004	2005	2006	2007	2008
M – CT	2.9457	2.8469	2.9199	2.2611	3.1966	3.4134	3.1496	3.2476
M – B	3.0784	2.8137	2.8089	2.6647	3.3698	3.0028	2.5846	3.2463
CT – B	1.6984	1.7350	1.6925	1.7651	1.8331	1.9358	1.8363	2.0512
Sum	7.7225	7.3984	7.4214	6.6910	8.3996	8.3520	7.5706	8.5451

Table 3.14: Weights of subsets of cardinality two

	2001	2002	2003	2004	2005	2006	2007	2008
M – CT	0.3814	0.3851	0.3934	0.3379	0.3805	0.4086	0.4160	0.3800
M – B	0.3986	0.3803	0.3784	0.3982	0.4011	0.3595	0.3414	0.3799
CT – B	0.2199	0.2345	0.2280	0.2638	0.2182	0.2317	0.2425	0.2400

Table 3.15: Fuzzy measure (weights) of subsets

	2001	2002	2003	2004	2005	2006	2007	2008
Measure	0.2199	0.2345	0.2280	0.2638	0.2182	0.2317	0.2425	0.2400

Table 3.16: Weights imputed to the Choquet integral for 2005

	{}	{M}	{C}	{B}	{M, C}	{M, B}	{C, B}	{M, C, B}
2001	0	0.2199	0.2199	0.2199	0.3814	0.3986	0.2199	1
2002	0	0.2345	0.2345	0.2345	0.3851	0.3803	0.2345	1
2003	0	0.2280	0.2280	0.2280	0.3934	0.3784	0.2280	1
2004	0	0.2638	0.2638	0.2638	0.3379	0.2638	0.2638	1
2005	0	0.2182	0.2182	0.2182	0.3805	0.4011	0.2182	1
2006	0	0.2317	0.2317	0.2317	0.4086	0.3595	0.2317	1
2007	0	0.2425	0.2425	0.2425	0.4160	0.3414	0.2425	1
2008	0	0.2400	0.2400	0.2400	0.3800	0.3799	0.2400	1

Table 3.17: Violation of IIA by downset aggregation

#Countries taken out	#Samples	Total pairwise rankings	#Average of reversals in the total	%Reversals
1	174 (all)	14878	18	0.15%
2	100	14706	31	0.31%
3	100	14535	46	0.42%
4	100	14365	61	0.43%
5	100	14196	72	0.51%
10	100	13366	112	0.84%
20	100	11781	176	1.49%
30	100	10296	194	1.88%

Table 3.19: Ranking Reversal (Year 2008)

Country	Monopoly	Territorial control	Bureaucracy	Min	Max
United States of America	0.0472	0.1370	0.2444	0.0472	0.2444
Haiti	0.4629	0.4516	0.6771	0.4516	0.6771
Colombia	0.6801	0.4768	0.3755	0.3755	0.6801
Switzerland	0.0064	0.0540	0.2118	0.0064	0.2118
Burkina Faso	0.0510	0.6216	0.5095	0.0510	0.6216
Iraq	0.8799	0.2570	0.5143	0.2570	0.8799
Japan	0.0040	0.1070	0.2144	0.0040	0.2144

Table 3.20: Characterisation of our three aggregation functions

Characteristic/function	OWA + Choquet integral	Ideals	Minmax interval
Pareto	Yes	Yes	Yes
Monotonicity	Yes	Yes	Yes
Transitivity	Yes	Yes	Yes
Non-compensatory	No (but it is not fully compensatory, see Annex 1)	Yes	Yes
Symmetric	Yes	Yes	Yes
Anonymous	Yes	Yes	Yes
Independence of irrelevant alternatives	Yes	No	Yes
Universality of domain	Yes	Yes	Yes (adjusted version, with plausibility deciding the 'ties')

Table 3.21: Crisp vs fuzzy

CRISP	FUZZY
The difference between observed and estimated data is interpreted as an observational error.	Here it is interpreted as the degree of 'blurredness' because the system is undefined.
Probabilistic models evaluate stochastic variability of phenomena.	Possibilistic models evaluate uncertainty due to some form of human influence.
There is a crisp relation between independent and dependent variables.	The relation is not crisp.
It is based on severe assumptions about the properties of the model (normality, heteroscedasticity, no autocorrelation). When these conditions are not met, this can affect the validity and performance of the model.	Does not demand these assumptions, and can be used when they are not met; it is designed to quantify human judgement, ambiguous processes, or where crisp data is poor, deteriorated or unavailable (Bardossy 1987; 1990; Gharpuray 1986).
The estimation is based on the minimization of differences between the observed data and the expected data produced by a concrete probabilistic model.	The estimation tries to find the parameters that make the membership value of the observed datum with respect to the expected one at least H. It minimizes the sum of the amplitudes of the objects (fuzzy numbers or intervals).
It demands only the specification of a model (select the independent variables and the functional form that links them to the dependent one) before implementing the estimation.	Besides the specification of the model, the researcher has to establish a level of plausibility H.
The estimations of the parameters are crisp numbers. Confidence intervals can be built for the parameters if the assumptions are met. These intervals are interpreted for example as: for a level of significance of 0.95, if we take many random samples of the same size 95% of them will fall within the confidence interval.	The estimations of the parameters are fuzzy numbers or intervals. These are fuzzy estimations of the dependent variable F. A value of H=0.95, for example, refers to the narrowest interval whose 0.95 alpha cut contains the whole observation in the given sample.
It includes tests of specification and significance of the model with a strong underlying statistical theory. But they depend on the credibility of the assumptions.	Fuzzy regression tests only establish how plausible the models are and how adequate they are to estimate the dependent variable.
The bigger the quantity of data, the smaller the standard deviation.	The bigger the number quantity of data, the bigger the amplitude of the estimations (Kwang Jae Kim, 1996).

Table 3.22: Nominal classes – classification of fragility in nominal classes (some order, but total order has been sacrificed)

Cluster	Countries – exemplars	Interpretation
1	Bolivia, Brazil, Cameroon, Peru, Uganda	
2	Colombia, Iraq, Moldova, Sudan	Poor monopoly, acceptable bureaucracy
3	Afghanistan, Burundi, Central African Republic, Chad, Congo Democratic Republic of (Zaire), Cote D'Ivoire, Haiti, Liberia, Somalia	Problems in all dimensions
4	Algeria, Argentina, Botswana, Bulgaria, Chile, China, Costa Rica, Cuba, Mexico, Uruguay	Monopoly problems are basically solved, also territorial reach, but instead quality of bureaucracy is still problematic (though not deteriorated)
5	Burkina Faso, Cambodia, Eritrea, Ethiopia, Gabon, Mozambique, Nicaragua, Zambia	The dimension of territorial control is deteriorated, other dimensions work relatively well
6	Ecuador, Egypt, El Salvador, Guatemala, Indonesia, Lebanon, Lesotho, Pakistan, South Africa, Syria, Venezuela, Vietnam	No dimension wholly consolidated, but no one is completely deteriorated either
7	Iran, Kuwait, Libya, Romania, Russia, San Marino, Seychelles, Ukraine	Problems with respect to the quality of bureaucracy
8	Canada, Croatia, Czech Republic, Denmark, Finland, France, German Federal Republic, Greece, Japan, South Korea, Spain, Sweden, Switzerland, United Kingdom, United States	OK in all dimensions
9	Cyprus, Georgia, India, Israel, Jamaica, Jordan, Serbia, Sri Lanka, Tajikistan, Trinidad and Tobago	OK in all dimensions, but worse than cluster 8

4: A CURSE OF EXCESS

ORDER PRESERVING FUNCTIONS FROM FINITE POSETS TO $\mathbb{R}$: HOW MANY ARE THERE?

FRANCISCO GUTIERREZ AND CAMILO ARGOTY

It is important for the study of social indicators to know how many ordered preserving functions can be sent from a partially ordered set (from now on, poset) to a totally ordered one, such as $\mathbb{R}$.

Social indicators are built through aggregation functions that typically take an Rn vector of data and assign a real number to it. Aggregation functions must fulfill the following two conditions (Beliakov et. al. 2007):

a. Boundary conditions. If poset $\mathcal{L}$ has a top, which we will denote, following convention, by a 1, and the codomain is a compact (closed and bounded) subset of $\mathbb{R}$ $f(1) = 1_{cod}$, where 1_{cod} is the maximum value that the function can achieve in the codomain. If $\mathcal{L}$ has a bottom, which we will conventionally denote by a 0, then $f(0) = 0_{cod}$, where 0_{cod} is the minimum value that f can achieve in the codomain.

b. Monotonicity. f preserves the order of the domain (and if the domain is a poset then f is a poset homomorphism). Hence, if $x \geq y$ in $\mathcal{L}$, then $f(x) \geq f(y)$ in the codomain ($\mathbb{R}$ or a proper subset of it).

We concentrate on functions from $\mathcal{L}$ to $\mathbb{R}$, thus ignoring property a[138]. From now on, $\mathcal{L}$ will always be finite. The substantive interest of the questions discussed in this paper is the following. We want to evaluate how stable are the rankings established by an aggregation (in our context, an order preserving) function f over $\mathbb{R}$, and establish some bounds to such variation. This will allow us to understand the meaning of the rankings of genuinely multidimensional social indicators – eg social indicators whose domain cannot be transformed into a totally ordered set and that potentially contains a high percentage of incomparable pairs.

The exposition proceeds in the following order. In the first section, we lay out some elementary facts about posets. In the second, we present our main results, which show that the order established by any well-formed function f can always be reversed by another well-formed function g. Furthermore, the number of alternative order-preserving functions – and thus, of alternative rankings – grows very fast in the width of the poset (the cardinality of the biggest antichain). In the conclusions, we synthesise and offer some interpretations.

4.1 PARTIALLY ORDERED SETS: PRELIMINARIES

Definition 1.1. A *partially* ordered set $(\mathcal{L}, \geq)$ is a set endowed with an order relation $\geq$. When the order relation is understood, the poset can be denoted by $\mathcal{L}$.

Definition 1.2. An *order* is a binary relation that is reflexive, antisymmetric and transitive. An order relation $\geq$ gives origin to strict inequality when $x > y$ if $x \geq y$ and $x \neq y$.

Definition 1.3. Two elements x, y of a poset can be comparable or incomparable (independent). They are *comparable* if $x \geq y$ or $y > x$.

Definition 1.4. When two elements a and b of a poset are incomparable, this relation is denoted as $a \parallel b$.

Definition 1.5. A poset $\mathcal{L}$ is *bounded* if it has a least element, denoted by 0, and a greatest element, denoted by 1.

Definition 1.6. Let $\mathcal{L}$ be a poset. *A chain* of $\mathcal{L}$ is a non-empty subset that is linearly ordered. A non empty set A ? $\mathcal{L}$ is said to be an antichain if for every $a, b \in A$, $a \parallel b$.

Definition 1.7. Let $\mathcal{L}$ be a finite poset. The maximum cardinality of a chain in $\mathcal{L}$ is called the *length* in $\mathcal{L}$. Similarly, the maximum cardinality of an antichain in $\mathcal{L}$ is called the *width* of $\mathcal{L}$.

138 Because $\mathbb{R}$ has neither top nor bottom. This is not consequential for the discussion below.

4.2 RESULTS: ORDER PRESERVING FUNCTIONS FROM $\mathcal{L}$ TO $\mathbb{R}$

Definition 2.1. Let $\mathcal{L}$ be a poset and let $a \in \mathcal{L}$. Then the *up-set* of a in $\mathcal{L}$ is the set

$$Up(a)=\{b \in \mathcal{L} | b > a\}$$

Similarly, the down-set of a is

$$Lo(a)=\{b \in \mathcal{L} | b < a\}$$

Definition 2.2. Let $\mathcal{L}$ be a poset. We call a function f from $\mathcal{L}$ to $\mathbb{R}$ *order preserving* if whenever $a > b$, then $f(a) > f(b)$.[139]

Theorem 2.3. Let $\mathcal{L}$ be a finite poset. Then there exists at least one order preserving function from $\mathcal{L} \longrightarrow \mathbb{R}$.

Proof. We define the function by induction. Let $\mathcal{L}$ be finite. Let $\mathcal{L} = \{a_1, a_2,..., a_\ell\}$, where $\ell = |\mathcal{L}|$. We choose $f(a_1) = 0$. If $f(a_1),..., f(a_j)$ are already defined, we choose $f(a_{j+1})$ to be a real number such that $f(a_{j+1}) < f(a_i)$ if and only if (iff) $a_{j+1} < a_i$ and $f(a_{j+1}) > f(a_i)$ iff $a_{j+1} > a_i$, for $i = 1,\ldots, \ell$. At some moment the procedure terminates, because $\mathcal{L}$ is finite.

Definition 2.4. Let $\mathcal{L}$ be a poset, $f: \mathcal{L} \longrightarrow \mathbb{R}$ a well formed function and $a \in \mathcal{L}$. The *pseudo-up-set* of a determined by f is the subset of $\mathcal{L}$

$$Psup_f(a) = \{b \in \mathcal{L} \mid f(b) > f(a)\}$$

Analogously, the pseudo-downset of a is the subset of $\mathcal{L}$

$$Plow_f(a) = \{b \in \mathcal{L} \mid f(b) < f(a)\}$$

Definition 2.5. Let $\mathcal{L}$ be a lattice and $A \subset \mathcal{L}$. The *up-set* of A in $\mathcal{L}$ is $Up(A) = \{b \in \mathcal{L} | b > a$ for all $a \in A\}$

In a similar way, the downset of A is

$$Low(A) = \{b \in \mathcal{L} | b<a \text{ for all } a \in A\}$$

Definition 2.6. Let $\mathcal{L}$ be a lattice, $f: \mathcal{L} \longrightarrow \mathbb{R}$ an order preserving function, and A $\subset \mathcal{L}$. Then, the pseudo up-set of A is

$$Psup_f(A) = \{b \in \mathcal{L} \mid f(b) > f(a) \text{ for all } a \in A\}$$

And, analogously,

$$Plow_f(A) = \{b \in \mathcal{L} \mid f(b) < f(a) \text{ for all } a \in A\}$$

Theorem 2.7. Let $\mathcal{L}$ be a lattice and $f: \mathcal{L} \longrightarrow \mathbb{R}$ an order-preserving function. Let $a, b \in \mathcal{L}$ such that $a \| b$. Whenever $f(a) > f(b)$, there exists another order-preserving function g such that $g(a) = f(b)$ and $g(b) = f(a)$ (so that g reverses the ordering between a and b established by f).

Proof. Let k be a real number such that $k > f(a) - f(b)$. Now let h be the function $f: \mathcal{L} \longrightarrow \mathbb{R}$ defined in the following way:

$$h(x) = f(x) + k \text{ when } x \in Up(b)$$

$$h(x) = f(x) - k \text{ when } x \in Lo(a)$$

$$h(x) = f(x) \text{ otherwise}$$

139 Note that we have dropped the boundary conditions, because we are working over $\mathbb{R}$. This has no implication on subsequent discussions.

We start proving the following claim: h is order preserving. In effect, let $x, y \in \mathcal{L}$ and suppose without loss of generality that $x < y$. Then we have five possible cases:

a. $x \in Up(b)$. Then by transitivity $y \in Up(b)$ too, which implies $f(x) < f(y)$ [by hypothesis f is order preserving]. This implies that $f(x) + k < f(y) + k$, hence $h(x) < h(y)$

b. $y \in Lo(a)$. Then also $x \in Lo(a)$, and we proceed as in the previous case

c. $x \in Lo(a)$ and $y \in Up(b)$. Then $h(y) = f(y) + k > f(y) > f(x) - k = h(x)$

d. $x \in Lo(a)$ and $y \notin Up(b)$, $y \notin Lo(a)$. Then $h(y) = f(y) > f(x) > f(x) - k = h(x)$

e. $x \notin Lo(a)$, $x \notin Up(b)$, and $y \in Up(b)$. $h(y) = f(y) + k > f(y) > f(x) = h(x)$

Now let us define the function $g: \mathcal{L} \longrightarrow \mathbb{R}$ in the following way:

$g(x) = f(b)$ if $x = a$

$g(x) = f(a)$ if $x = b$

$g(x) = h(x)$ otherwise

We now show that g is order preserving and we are done. Take two distinct $x, y \in \mathcal{L}$, and assume without loss of generality that $x < y$. Then there are five possible cases.

a. $x, y \neq a, b$. By the previous claim about h, $g(x) = h(x) < h(y) = g(y)$

b. $x = a$. Then, $y \in Up(a)$ and $g(y) = h(y) = f(y) + k > f(y) > f(a) > f(b) = g(a)$

c. $x = b$. Then $y \in Up(b)$ and $g(y) = f(y) + k > f(a) = g(b)$

d. $y = a$. In this case, $x \in Lo(a)$, and $g(x) = f(x) - k < f(b) = g(a) = g(y)$

e. $y = b$. Here, $x \in Lo(b)$, and $g(x) = h(x) \leq f(x) < f(b) < f(a) = g(b) = g(y)$

Theorem 2.8. Let $\mathcal{L}$ be a lattice and $f: \mathcal{L} \longrightarrow \mathbb{R}$ an order-preserving function. Let $a, b \in \mathcal{L}$ such that $a \parallel b$ and $f(a) > f(b)$. Then there exists an order-preserving function g such that $g(a) = f(b)$, $g(b) = f(a)$ and $g(c) < g(d)$ whenever $f(c) < f(d)$ with $c, d \in Psup_f(a) \cup Plow_f(b)$

Proof. Let k be a real number such that $k > f(a) - f(b)$. Let h be the function $h: \mathcal{L} \longrightarrow \mathbb{R}$ defined in the following way:

$h(x) = f(x) + k$ when $x \in Up(b) \cup Psup_f(a)$

$f(x) - k$ when $x \in Lo(a) \cup Plow_f(b)$

$f(x)$ otherwise

Claim 1. h is order preserving.

Proof. Let $x, y \in \mathcal{L}$ and suppose, without loss of generality, that $x < y$. Then we have the following five cases:

a. $x \in Up(b) \cup Psup_f(a)$. Then, y $\in Up(b) \cup Psup_f(a)$ too, and $f(x) < f(y)$. Therefore, $f(x) + k < f(y) + ksoh(x) < h(y)$

b. $y \in Lo(a) \cup Plow_f(b)$. Analogous to the previous case

c. $x \in Lo(a) \cup Plow_f(b)$ and $y \in Up(b) \cup Psup_f(a)$. Hence, $h(y) = f(y) + k > f(y) > f(x) > f(x) - k = h(x)$

d. $x \in Lo(a) \cup Plow_f(b)$ and $y \notin Up(b) \cup Psup_f(a)$. Then $h(y) = f(y) > f(x) > f(x) - k = h(x)$

e. $x \notin Lo(a) \cup Plow_f(b)$ and $y \in Up(b) \cup Psup_f(a)$. Then, $h(y) = f(y) + k > f(y) > f(x) = h(x)$

QED for claim

Claim 2. Whenever $f(c) = f(d)$ and $c, d \in Psup_f(a) \cup Plow_f(b)$, we have that $h(c) < h(d)$

Proof. Let $c, d \in Psup_f(a) \cup Plow_f(b)$ be such that $f(c) < f(d)$. Then we have the following three cases:

a. $c \in Psup_f(a)$. Then, $d \in Psup_f(a)$ too. Therefore, $f(c) + k < f(d) + ksoh(c) < h(d)$.

b. $d\ Plow_f(b)$. Analogous to the previous case

c. $d \in Plow_f(b)$ and $c \in Psup_f(a)$. Hence, $h(d) = f(d) + k > f(d) > f(c) > f(c) - k = h(c)$

Now let us define the function g in the following way:

$g(x) = f(b)$ if $x = a$

$g(x) = f(a)$ if $x = b$

$h(x)$ otherwise

QED for Claim

Claim 3. g is order preserving.

Proof. Let $x, y \in \mathcal{L}$ and suppose $x < y$. The we have these five cases:

a. $x, y \neq a, b$. By the previous claim, $g(x) = h(x) < h(y) = g(y)$

b. $x = a$. Then $y \in Up(a)$ and $g(y) = h(y) \geq f(y) > f(a) > f(b) = g(a)$

c. $x = b$. Then $y \in Up(b)$, and $h(y) = f(y) + k > f(a) = g(b)$

d. $y = a$. In this case, $x \in Lo(a)$ and $g(x) = f(x) - k < f(b) = g(a) = g(y)$

e. $y = b$. Here, $x \in Lo(b)$ and $g(x) = h(x) \leq f(x) < f(b) < f(a) = g(b) = g(y)$

QED for Claim

Claim 4. Whenever $f(c) < f(d)$ and $c,d \in Psup_f(a) \cup Plow_f(b)$, we have that $g(c) < g(d)$

Proof. Analogous to Claim 2.

Q.E.D. for Theorem

Lemma 2.9. Let $\mathcal{L}$ be a finite poset and $A \subseteq \mathcal{L}$ be an antichain in $\mathcal{L}$. Let $f: \mathcal{L} \longrightarrow \mathbb{R}$ be an order-preserving function and $g: A \longrightarrow \mathbb{R}$ any function. Then, there exists an order-preserving function h such that:

a. $h(a) \leq h(b)$ whenever $a, b \in A$ and $g(a) \leq g(b)$

b. $h(c) < h(d)$ whenever $f(c) < f(d)$ and $c,d \in Psup_f(A) \cup Plow_f(B)$

Proof. Let us denote by $r_M = max\,(f(a))$. Be $r_m = min\,(f(a))$. Let us denote by a_M an element in A such that $g(a_M) = max(g(a))$, and call a_m an element of a such that $g(a_m) = min(g(A))$. By procedures similar to those used in Theorems 2.7 and 2.8 we can build an order-preserving function f_1 such that condition b holds and $f_1(a_M) = r_M$ and $f_1(a_m) = r_m$. Let $A_1 = A \setminus \{a_m, a_M\}$. Starting again with A_1 and continuing in this way we get the desired function.

Lemma 2.10. Let $\mathcal{L}$ be a finite poset and let n be the width of $\mathcal{L}$. The number up to order isomorphism of order-preserving functions $f: \mathcal{L} \longrightarrow \mathbb{R}$ is greater than the number

$$\sum_{k_1+k_2+\cdots+k_m=n-1} (n-1)!$$

Proof.

Claim. Let A be an antichain in $\mathcal{L}$ with width n, let $a_0 \in A$ and let B a chain of length $m-1 \geq 1$ such that for every $b \in B$ and $a \in A \setminus a_0$, $b \parallel a$. Then, the number of 1-1 order-preserving functions $f: A \cup B \longrightarrow \mathbb{R}$ up to order isomorphism is

$$\sum_{k_1+k_2+\cdots+k_m=n-1} (n-1)!$$

Proof of claim: To get a 1-1 order-preserving function $f: A \cup B \longrightarrow \mathbb{R}$, we have to place the $n-1$ of $A \setminus \{a_0\}$ between the $m-1$ points of B; that is, place $n-1$ elements in m cases. So, we have to choose a number k_i of elements to be placed in the case i, under the restriction $k_1 + k_2 + \ldots + k_m = n-1$. Once these $k_1 + k_2 + \ldots + k_m$ are selected, we have $\binom{n-1}{k_1 \cdots k_m}$ possibilities for placing $n-1$ elements into m cases. On the other hand, once we have chosen k_i elements for the case i, you have to decide one order between them; this leaves $k_i!$ possibilities of orders. Putting all this together, we get that the number of 1-1 order-preserving functions $f: A \cup B \longrightarrow \mathbb{R}$ up to order isomorphism is

$$\sum_{k_1+k_2+\cdots+k_m=n-1} \binom{n-1}{k_1 \cdots k_m} k_1! \cdots k_m! = \sum_{k_1+k_2+\cdots+k_m=n-1} (n-1)!$$

Q.E.D. for Claim.

Now, by Lemma 2.9 every order-preserving function $f: A \cup B \longrightarrow \mathbb{R}$ can be extended to an order-preserving function $f: \mathcal{L} \longrightarrow \mathbb{R}$. This implies that the number up to order isomorphism of order-preserving functions $f: \mathcal{L} \longrightarrow \mathbb{R}$ is greater than the number

$$\sum_{k_1+k_2+\cdots+k_m=n-1} (n-1)!$$

Q.E.D. for Lemma

Remark 2.11.

Given that $m \geq 2$, we have that

$$\sum_{k_1+k_2+\cdots+k_m=n-1} (n-1)! \geq (n-1)!\binom{n-1}{2} \geq (n-1)!$$

so the number of order-preserving functions from $\mathcal{L} \longrightarrow \mathbb{R}$ is bigger than $(n-1)!$

Fact 2.12. Let $\mathcal{L}$ be a subposet of $\mathbb{R}^p$ for some p, and let n be the least such p.
Then $\mathcal{L}$ has an antichain whose width is greater than or equal to n.

Remark 2.13. The previous fact tells us that if we have a finite poset that is naturally embedded in $\mathbb{R}^n$, then an antichain of size n will appear. So the amount of order-preserving functions from $\mathcal{L} \longrightarrow \mathbb{R}$ increases in n at least as $(n-1)!$ increases with respect to n.

Remark 2.14. We have not utilized the uncountability of $\mathbb{R}$, nor any operation different from addition. Thus, the results apply also to functions f: $\mathcal{L} \longrightarrow \mathcal{I}$, the set of integers.

CONCLUSIONS

Suppose we have an array of rectangular data, where the rows are objects and the columns are variables. The above propositions are telling us that, for this data set:

a. A ranking over incomparable cases produced by any aggregation function can be reversed by another equally acceptable order-preserving function;

b. The number of alternative functions, and thus of potential raking reversals, grows factorially in the cardinality of the biggest antichain.

Thus, ranks predicated over independent pairs are basically a result of the ad hoc choice of that particular function, unless additional conditions for the function choice are established.

5. CONCLUSIONS

5.1. DISCUSSING PSPIs

We come to the end of our journey. We started by placing PSPIs in a context of globalisation and technological change, which created simultaneously a demand for new forms of quantification and the technical means to gather, store and manage huge amounts of data on a scale unprecedented in the history of humanity. This is good news, at least in some cases. We are better off if some kind of agency can evaluate whether human rights are upheld, where, and how much (and all of this, of course entails some type of formalisation!).[123] Furthermore, we have argued that we cannot ask of indexes what they cannot deliver. An index cannot be thick, because its thickness would be a logical inconsistency. Quantitative research is based on isolation and simplification, and thus it will never replace good qualitative social science. Additionally, many of the major data problems that are used to denounce quantitative models appear also in qualitative studies.[124]

This is the setting. Then we considered index building, and suggested that there are 'objective' limits to this, in the sense that it will not be possible for indexes to fulfil a relatively small set of axiomatically desirable conditions (at least if we are considering only aggregation functions whose range are real numbers or integers, or subsets of them). Immediately afterwards we discussed the problems of reliability, validity and ambiguity, and came to the conclusion that PSPIs are more difficult to build than other social indicators. This is so because they operationalise more complex concepts, take into account more variables and more levels of aggregation, have to deal with intrinsic ambiguity, and manage much more heterogeneous and speculative data (in compensation, they have much more of it). The main issue here is that PSPIs have to deal with the problem of order, which has not been acknowledged, let alone solved.

In the following two chapters, we:

- Illustrated these themes, showing how they affected extant PSPIs, and argued that: (a) often the problem simply remained unacknowledged; and (b) when a way out was actually presented it did not necessarily work.

- Proposed some partial solutions, which we believe can be considered genuine advances. For example: theory-driven database building, minimisation of ambiguity, explicit listing of the type of data that the dataset uses, minimisation of ad hoc decisions with respect to parameters, creation of non-invasive aggregation functions that are not interpreted as if substitution rates existed, creation of a fuzzy/approximate toolkit to deal with intrinsic ambiguity and new objects (different from real numbers and integers).

123 It is probably the case that every single state violates human rights in some measure. But we have to differentiate gross, massive violations from marginal ones, as well as between each category. Here, after only two sentences, we already have an ordinal model.

124 On the other hand, the line of defence according to which quantitative models at least make explicit their assumptions is erroneous, or at least it has been oversold. It is impossible to make explicit all the important assumptions. For example, PSPI builders have failed to declare that their quantifications are based on the assumption that there are substitution rates between dimensions or variables of state fragility.

We now pin down some of the claims of these chapters, so as to minimise and manage ambiguity.

5.2. INTERPRETING PSPIs

We identified a number of crucial problems in the production and use of PSPIs. Though we have been quite explicit throughout this report, it is worthwhile insisting once more that we are not fond of technological conservatism, nor do we believe that eschewing the effort of quantification is correct. Statistics might be 'frightfully inadequate' (Keynes 1973), but it is a 'wonderful tool' (Ragin 2008), which historically is intimately associated with state building. The discussion is not, and should not, be posed in terms of creating or restoring a Chinese wall between the quantitative and qualitative. The problems lie elsewhere.

The first problem sounds quite simple, but in reality is very involved. When engaging in a quantification related to PSPI, what are we doing? Is this truly a measurement? In what sense? What precisely is being measured? As yet, this is pretty much an open question in several social disciplines. Psychologists have discussed it a lot. On the other hand, when introducing us to his 2007 book about measurement in economics, Marcel Boumans claims that it is 'the first book that takes measurement in economics as its central focus' (Boumans 2007: 3). He notes that measurement has been developed intensely, sometimes frantically, in many economic fields, but that a general view of what measuring means in economic contexts has not been developed. Some of the authors in his book put forward criteria to evaluate the quality and plausibility of different measurements. What does a unit of fragility mean in the PSPI context?[125]

The second problem is also related to the nature of quantification, this time to the preconditions for making statistical inference. We find no better alternative here than to quote the following fantastic reflections by Collier, Sekhon and Stark (2010: xiv), on how mechanical number crunching works:[126]

> Put in the data, turn the crank, out come quantitative causal relationships, no knowledge of the subject is required. This is tantamount to pulling a rabbit from the hat. Freedman's conservation of rabbits principle says 'to pull a rabbit from a hat, a rabbit must first be placed in the hat'. In statistical modeling, assumptions put the rabbit in the hat … Modeling assumptions are primarily made for mathematical convenience, not for verisimilitude … Can the assumptions be tested empirically? Do they violate common sense? … scientific problems cannot be solved by 'one size fits all' methods. Rather, they require shoe leather: careful empirical work tailored to the subject-matter knowledge and statistical principles.

125 As throughout this text, we do not indulge in discussing the theme here. It deserves a separate treatment. Note, however, that since measuring has been a state activity par excellence, in principle it should not be banished from the field of politics.

126 Which pretend to synthesise Freedman's work.

Or, as Freedman expressed it (2005: xiv):

> Many convergent lines of evidence must be developed … Before anything else, the right question needs to be framed … Naturally, there is a desire to substitute intellectual capital for labor. This is why investigators try to base causal inference on statistical models. The technology is relatively easy to use, and promises to open a wide variety of questions to the research effort. However, the appearance of methodological rigor can be deceptive. The models themselves demand critical scrutiny. Mathematical equations are used to adjust for confounding and other sources of bias. These equations may appear formidably precise, but they typically derive from many somewhat arbitrary choices. What variables to enter the regression? What functional forms to use? What assumptions to make about parameters and error terms? These choices are seldom dictated either by data or prior scientific knowledge. That is why judgment is so critical, the opportunity for error so large, and the number of successful applications so limited.

These are severe words, and perhaps unjust in some points. The fact remains, though, that in these quantifications it is *not* enough to present a working model. If they are based on 'formidable' but spurious precision and unsubstantiated crucial choices, or even violate common sense, then they are equivalent to pulling a rabbit from a hat. Throughout our discussion we found the following:

5.2.1. CONCEPTUALISATION

With the notable exceptions of the BTI and the PIL, PSPIs include in their definitions of fragility: aspects related to statehood proper, putative causes of state fragility, putative consequences of state fragility and putative correlates (lsuch as types of regime or levels of development) of state fragility. There does not appear to be a theoretical discussion about the meaning of the inclusion of one or another dimension in the database.

5.2.2. AMBIGUITY

The definitions of the dimensions and of the variables are heavily hedged, which in the majority of cases is inevitable. There are additional layers of ambiguity, created by the operationalisation of unobservables using unobservables and the uncritical resort to expert opinion or in-house coding in conditions that are not carefully controlled, and responding to hazily formulated questions.

5.2.3. QUALITY OF THE DATA

The data sets of PSPIs are composed mainly of ordinal scales marked by experts in uncontrolled conditions, and averaged in ways that are not completely clear for a third party.[127]

5.2.4. ORDER AND AGGREGATION

None of the indexes acknowledge the problem of multidimensionality, and the multi-attribute character of the aggregation of diverse dimensions of state fragility. The modal aggregation function is the simple average. Here three absolutely fundamental problems pop up. First, the existence of substitution rates is assumed. Second, the weights are imputed in an ad hoc fashion, and simply not discussed. Third, since neither variables nor boxes are orthogonal, the 'official' weights differ from the true ones. Which are theoretically sound: the official or the perturbed ones? A fourth problem is that full compensation is also assumed, which is outright implausible.

Pair-wise comparisons – more generally, rankings – in multidimensional datasets can be of two types: either case A is superior or equal in all variables to case B (or vice versa); or A is superior to B in some variables and inferior in others. Call the first type of situation 'comparable' and the second 'incomparable'. For comparable pairs, all aggregation functions (which by definition are monotonic) behave well. We illustrated here, as was demonstrated elsewhere (Gutiérrez and Argoty 2010), that for incomparable cases any ranking produced by a well-formed function can be reversed by an equally well-formed function. The quantity of possible functions that produce different rankings grows very fast in the number of variables. This means that unless the researcher demonstrates that his/her aggregation function is superior to all the others (perhaps up to linear transformations) in some specific sense, the rankings of incomparable cases are basically an artefact of the ad hoc choice of the function (weights, functional form, etc.).[128] The percentage of incomparable pairs of the overwhelming majority of PSPIs (including ours) is very high. Thus, a substantial number of their rankings should be considered an artifice. They will say – and this is not an artifice – that Norway is better off than Haiti, and Germany than Colombia. This we will have to believe. But the way that they rank Colombia and Venezuela, or Rwanda and Uganda, or China and the Philippines, is only a product of a series of methodological decisions whose underlying rationale we ignore.

One index is, in this regard, in a better position than the rest. The BTI, which is built upon a parsimonious and theoretically oriented definition of fragility, has only two variables, which are highly correlated. It is hardly a multidimensional database; hence, it hosts very few incomparable pairs. More or less any aggregation function would arrive at the same result given this dataset.

5.2.5. USE IN PROBABILISTIC MODELS

The easy reaction to these problems is to condemn quantification in these areas, or in general, and set the issue to rest. Unfortunately, this solves nothing. It is a typical instantiation of technological conservatism. More fundamentally, it leaves open all the problems that the PSPIs have left unsolved. Suppose no quantitative exercises about state fragility are

127 We are not speaking here about the averages between variables, but about the averaging of the marks of the experts to produce the one that is ultimately attributed to the country.

128 Of course, the situation is much worse with the *grading*.

attempted in the next ten years, and we rely only on country monographs. How will we generalise the knowledge they will convey to be able to produce or to criticise, or to evaluate global decisions? How will we aggregate the impacts of these decisions or policies to characterise them? How do we prospect their potential impacts? How do we check the reliability of the data and information? We will end up back at square one.

We believe the correct reaction is to understand the deep flaws and the severe constraints that this type of index building faces, and place them in their proper historical context and technological basis. If we identify these, we will be able to see that the blessing and curse of PSPIs, and the factor that gives origin to the exercise itself, is an overflow of relatively easily accessible and cheap information that is ambiguous, deteriorated, corrupted and noisy. This is why the analogy of third wave indicators with disciplines that have to make big reconstructions based on little bits of evidence is attractive but unsound. The curse of these ideal-type researchers is having very scarce data, but they have good theories and sophisticated measurement devices. The curse of third wave index builders is symmetrically inverted. They have a lot of data (too much, actually), but it is noisy and ambiguous. They frequently also have unclear theories, and do not even know if they are engaging in measurement proper.

5.3. PARTIAL SOLUTIONS

From a purely aesthetic point of view, it would be a pity to throw away these appallingly large masses of data with which a new technological base regales us. From a more prosaic perspective, it is clear that these datasets frequently say something – and sometimes a lot. Yet we still do not know what they say, and how much they say. The gist of the matter is to separate information proper from noise, and take the specifics of the exercise (eg, non-probabilistic uncertainty, multidimensionality, problematic compensation between variables) on board. It is rather surprising that, while outstanding statisticians (like Freedman), engineers and operation researchers (like Bouyssou et al.), and mathematicians (like Zadeh and Pawlak) have been able to develop a very refined understanding of these characteristics of formalisation/quantification in many domains of human activity, an army of economists and social scientists unyieldingly persist in tinkering with heterogeneous and noisy data – including those uncontrolled for expert assessments as if they were equivalent to ultra-exact physical measurements.

They are not. But this does not mean that they are unusable. Ambiguity does not mean that anything goes. The answer to the question 'which GDP is bigger, Sweden's or Afghanistan's?' is *not* 'who knows?' Bounded ambiguity is the framework of our database, aggregation function and fuzzy toolkit. They are a work in progress, certainly constitute no panaceas and entail many thorny issues, of which we are aware. However, we believe that we have achieved some partial solutions. We present a listing of what we believe are genuine – small though they may be – advances.

5.3.1. CONCEPTUALISATION

Following strong traditions in social science and good practices (BTIs, for example), we avoid conceptual stretching. Based on a decade-long participation/interaction within a research programme about fragility and the dimensions of fragility (Crisis States Programme), we defined these and grounded them in the relevant literature.

Since we separated the operational definition of fragility and statehood from potential causes, consequences and correlates, we are in a position to make tolerably good regressions between independent variables; and have aggregated fragility and each of its dimensions separately. PSPIs that incur conceptual stretching cannot regress well on anything, because they try to include everything.

5.3.2. MINIMISING AND MAKING EXPLICIT DIVERSE FORMS OF AMBIGUITY

We arrived at a relatively clear preference ordering with respect to the type of data we wanted. From this point of view, the best data are counts of public, relatively rare events and states of the world (such as invasions, elections, coups, existence of armed groups) that fall near the census end of the spectrum, and that if wrong can be relatively easily and cheaply corrected.[129] Then follow ordinal scales taken from counts (they lose information with respect to counts, but in general become more reliable). At the other extreme, we have (potentially very biased) marks from experts elicited by uncontrolled (and unreported) means and in uncontrolled settings. Our database attempts to achieve equilibrium between trying to have enough proxies for the different relevant aspects of each dimension and minimising the use of suspect data. Even then a fair amount of ambiguity remains. We claim that this is inevitable.

5.3.3. NEW AGGREGATION FUNCTIONS

We developed three aggregation functions. Two produce marks and ranks. The third one produces intervals and ranks. We introduce no ad hoc weights, nor choices that do not take into account the data. The first two functions have downsides (which in our theoretical discussion we consider inevitable anyway) but they do not take the relevant parameters out of the blue. We have not tried to replace a crisp rabbit with a fuzzy rabbit! The OWA-Choquet integral is a compensatory function, but it is not fully compensatory. It takes into account the interactions between variables and dimensions, grounds the weights on the data, and allows an interpretation of them not as substitution rates but as 'importance with respect to variance explained'. The downset-cardinality function is non-compensatory, and its violation of

129 Compare adjusting the number of armed groups that participate in a conflict and the number of homicides produced by it. The former is a task that any careful researcher can undertake reasonably well; the latter is a huge enterprise that requires sophisticated data recording and specialised knowledge (see, for example, Ball 2001).

independence of irrelevant alternatives is marginal.[130] Since the *minmax* function introduces richer information in the aggregation process, it is not so surprising that it passes all the tests with flying colours. But it also introduces added complexity and produces a non-numerical object (an interval). Is this intractable?

5.3.4. NEW FORMAL TOOLS AND USES OF THE DATA

This data is not intractable: it can be regressed on. We have programmed and/or developed in the context of this investigation several fuzzy/approximate tools, able to operate on several types of objects, including intervals,[131] making linguistic queries and classifying noisy objects. We have drawn on burgeoning areas of research (fuzzy sets, pattern matching) that seem tailored to suit the needs of social scientists. Yet more are being developed.

5.3.5. LINGERING PROBLEMS

We face a large number of unsolved problems:

- What does measuring political variables mean?[132]
- What should we do with counts based on convenience samples? They should not be lightly discarded – in the first place they should allow for some kind of Bayesian updating – but they can be grossly wrong (both in the concrete numbers and in tendency).
- How can we capture genuinely global variables? They might have a strong incidence in the performance of states, and if this is the case models that do not take them into account are mis-specified. Much tells us that the nature of the state has changed globally. Some of the best achievements in social theory and history tell us that state strength is intimately related to the international context (Skocpol 2007). As yet, PSPIs persist in observing state by state, without introducing in the definition or in the covariables global trends and situations.[133] Even from a purely policy point of view, we would want to know if the current international environment is more state-building friendly than in the past. With the existing constructs we have not the least possibility of trying to address such a question, which seems pretty important. How can we solve this?

We have to investigate further the properties of the OWA/Choquet aggregation and the downset-cardinality aggregation and compare them. OWA/Choquet is compensatory, downsets do not respect IIA, and there is a non-negligible area of divergence between the two. Perhaps there is a third, superior alternative?[134] There are a lot of interpretative problems related to the parameters of fuzzy regressions and clusterings (pseudo Rs and pseudo Ps, for example). The translation of extant very powerful tools from the crisp to the fuzzy/approximate world is still incipient. The coordination between fuzzy and crisp results is also an open area.

But our assumption is that partial improvements are better than nothing.

130 This last observation applies to our *normalisation*. The sigmoidal normalisation respects IIA, and the standard normalisation – given the behaviour of the data over which we used it – was completely harmless.

131 As noted in the previous section, we use intervals to avoid the 'fuzzy rabbit effect'. See Freedman (2010) and Bouyssou and Vasnick (1986), who flag some downsides of fuzzy aggregations.

132 Analysis is in progress.

133 This might not be too serious when the analysis takes into account only short periods (five, ten years). For longer ones, it can induce severe distortions.

134 Actually we know that there is a third alternative, that is non-compensatory and respects IIA. It is the leximin (Dubois et al. 2001)). Unfortunately, it has too many undesirable properties for our purposes, so it cannot be considered superior.

ANNEX 1:

THE CHOQUET INTEGRAL AND SUBSTITUTION RATES

Source: Developed by the authors[135]

Let us try for example to establish the substitution rate between bureaucratic capacity and monopoly. How much bureaucracy is worth an increment (decrement) of monopoly.

Country	Year	Monopoly variation	Substitution rate	Monopoly	Territorial control	Bureaucratic capacity	Choquet
Afghanistan	2003	Initial value	..	0.66843	0.65111	0.73105	0.67170
		Variation 0.001	-0.00065961	0.66943	0.65111	0.73039	0.67170
Colombia	2003	Initial value	..	0.76412	0.48859	0.37341	0.48157
		Variation 0.001	-0.00037599	0.76512	0.48859	0.37304	0.48157

As above, the calculations are based on the fuzzy measure:

	{}	{M}	{C}	{B}	{M, C}	{M, B}	{C, B}	{M, C, B}
2003	0	0.2280	0.2280	0.2280	0.3934	0.3784	0.2280	1

In both cases, monopoly values are incremented in 0.001 (remember that in our database 'more is less': variables and dimensions grow in fragility), and we seek to establish how big the compensation in terms of bureaucratic capacity needs to be to maintain the original aggregated fragility value. For Afghanistan, the answer is -0.00065961 (we have to diminish the bureaucratic fragility in this amount), while for Colombia it is -0.00037599. The difference comes from the different ordering of the dimensions. For Afghanistan, territory < monopoly < bureaucracy, so its fragility value is calculated in the following way:

$$(1 \times 0.65111) + 0.3784(0.66843 - 0.65111) + 0.2280(0.73105 - 0.66843)$$

After which this transformation is applied:

$$(1 \times 0.65111) + 0.3784(0.66943 - 0.65111) + 0.2280(0.73039 - 0.66943)$$

In Colombia, the dimensions are ordered differently, bureaucracy < territorial control < monopoly, so we have:

$$(1 \times 0.37341) + 0.3934(0.48859 - 0.37341) + 0.2280(0.76412 - 0.48859)$$

which is transformed into:

$$(1 \times 0.37304) + 0.3934(0.48859 - 0.37304) + 0.2280(0.76512 - 0.48859)$$

Note that this already cannot be easily interpreted as a substitution rate, because within the same individual (country) a small increment in a dimension will change the weight it gets (because it might change its place in the ordering of variables).

Take the following example:

Country	Year	Monopoly values	Substitution rates?	Monopoly	Territorial control	Bureaucracy	Choquet
Jamaica	2008	Initial	..	0.53333333	0.22384787	0.21158951	0.29054
		+0.001	-0.00171429	0.53433333	0.22213358	0.21158951	0.29054
		+0.3	-0.12557727	0.83333333	0.09827059	0.21158951	0.29054

In this table, we present the numerical changes in territorial control that are necessary to compensate a change in the values of monopoly. We can see that the relation is not linear, because it depends on the magnitude of the

135 Note, all tables here are developed by the authors unless otherwise stated.

increment. If it is modest enough not to change the ordering of the dimensions, weights will remain unchanged as well. But if it is big it will change the ordering, and thus the imputation of weights to each box.

Let us observe how this behaves using the values of the example. The fuzzy measure for this year is

	{}	{M}	{C}	{B}	{M, C}	{M, B}	{C, B}	{M, C, B}
2008	0	0.2400	0.2400	0.2400	0.3800	0.3799	0.2400	1

Initial:

$$(1 \times 0.21158951) + 0.38(0.22384787 - 0.21158951) + 0.24(0.53333333 - 0.22384787)$$

Increment: 0.001 (does not change the ordering of the dimensions, thus maintains their weights)

Result:

$$(1 \times 0.21158951) + 0.38(0.22213358 - 0.21158951) + 0.24(0.53433333 - 0.22213358)$$

Increment 0.3 (changes the ordering of monopoly and territorial control, hence changes the weights)

Result:

$$(1 \times 0.09827059) + 0.3799(0.21158951 - 0.09827059) + 0.24(0.83333333 - 0.21158951)$$

Another characteristic of the Choquet integral that deserves to be flagged is that it is not fully compensatory within the [0,1] range on which we are operating. This is a very intuitive and desirable property.

We illustrate this with an example. We want to find the substitution rate for Bahamas between territorial control and monopoly. Bahamas has low values in all dimensions (see below).

Country	Year	Monopoly values	Substitution rates?	Monopoly	Territorial control	Bureaucracy	Choquet
Bahamas	2008	Initial	..	0.18820615	0.13225268	0.23445112	0.16461
		+0.3	-0.14734494	0.58820615	-0.01509227	0.23445112	0.16461

It is not possible to substitute a 0.3 increase in monopoly, because the value of the function falls outside the range. We stress that this expresses a strong intuition. If you are doing very well in two fundamental dimensions, a deterioration in the third one cannot be compensated by marginal improvements in the other two. If you are doing very poorly in all the dimensions, and you fall even further in one of them, you will not be able to compensate with improvements in the other ones.

These two 'boundary non-compensatory behaviours' capture the strong spillovers and interactions between the three dimensions of statehood and thus of fragility.

ANNEX 2:

EXPERIMENTAL DESIGN TO TEST THE IMPACT OF THE VIOLATION OF IRRELEVANT ALTERNATIVES BY DOWNSETS

We describe now the simple experiment through which we tested the magnitude of the violation of IIA by the downset aggregation. The procedure was the following:

1. Calculate the aggregation over the original database (with *n* individuals).

2. Create the matrix of pair-wise comparisons between countries, where the i,j cell of the matrix is 1 if the aggregated value of country is bigger than the aggregated value of country j, 0 otherwise. Delete the rows and columns of the (randomly chosen) *m* elements that are going to be excluded from the set of countries.

3. After these elements are withdrawn, recalculate the aggregation.

4. Calculate the new matrix of pair-wise comparisons.

5. Compare the matrices at steps 2 and 4, and create a new matrix whose i,j value is 1 if the value in both matrices is different, 0 otherwise.

6. Sum the ones and divide by the total of possible pair-wise comparisons in the *n-m* dataset.

We repeat this procedure 100 times for each of the following values of *m*: (1, 2, 3, 4, 5, 10, 20, 30).

ANNEX 3:

FUZZY CLUSTERING

We describe here the procedure utilised to produce fuzzy clusters based on representation of the countries by three number vectors. The algorithm is applied each year and demands from the user the choice of clusters. This choice is critical and must be substantiated, as cluster analysis – both crisp and fuzzy – is quite unstable, and a change in the number of clusters can produce a very different result.

Selection of the number of clusters

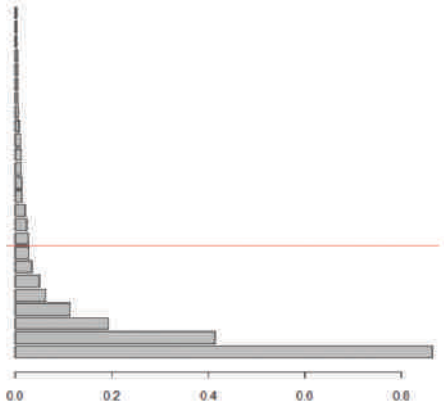
We utilised Ward's criterion (Lebart et al. 1995) to select the number of classes. This criterion takes into account the size of the variation between and within groups. It is known that for a data set

Total variance = Variance between + Variance within

It is well known that, in the process of data clustering, within variance increases and between variance decreases. So the clustering process seeks to find a 'good' – in some specified sense – level of inter-class variance. To do this, a histogram of level indices is built. Each bar in this histogram represents the loss of between variance when the algorithm goes from a partition of the set in S classes to a partition in S-1 classes

I(s) = Variance within (s) - Variance within (s-1)

The cut-off in the clustering tree is made in the level in which this histogram flags a brusque fall,[136] so that the most homogeneous classes possible are chosen. For our exercise, the histogram of level indices and the number of classes are the following:



Histogram of level indices – Year 2008–9 clusters

As reported above, we used the FCM (fuzzy c-means) algorithm to obtain the classes, choosing the number of classes for each year according to the Ward criterion. The algorithm finds the membership of each individual (country) to each class. After several tests, we found that a good value for the parameter *m* used in FCM was 1.2. Below that value the exercise tends to become a crisp clustering; above, it becomes unstable and distributes the membership uniformly across clusters.[137] Then the countries were divided in the patterns reported below. As was discussed in this section, this partition is only partially hierarchical.

Results:

Cluster	Monopoly mean	Territory mean	Bureaucracy mean
1	0.0956	0.5269	0.3532
2	0.6754	0.3533	0.4648
3	0.5010	0.6237	0.6681
4	0.0511	0.2571	0.3425
5	0.1136	0.5004	0.5354
6	0.3330	0.3179	0.4309
7	0.0458	0.1652	0.5273
8	0.0179	0.1309	0.1840
9	0.4433	0.1753	0.2608
Global mean	**0.1421**	**0.3019**	**0.3837**

136 This criterion is also rather impressionistic in crisp clustering. For comments on this, see Lebart et al. 1995.

137 The sum of the memberships of a case to all the clusters has to add to 1.

BIBLIOGRAPHY:

Aamodt, M. G. (1991). *Applied industrial/organizational psychology*. Belmont, CA/Wadworth Pub. Co.

Al-Awadhi, S., and Garthwaite, P. (2006). Quantifying expert opinion for modelling fauna habitat distributions. *Computational Statistics*, 21(1): 121–140.

Aleinikoff, T. A., and Klusmeyer, D. B. (2002). *Citizenship policies for an age of migration*. Washington, DC: Carnegie Endowent for International Peace.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4): 328–346.

Arrow, K. (1963). *Social choice and individual values*. New York: Wiley.

Ashenfelter, O., and A. C. (2001). Predicting Italian wine quality from weather data and expert's ratings. In M. Pitchery and M. Terraza, *Oenometrie IX 9th annual meeting of the Vineyard Quantification Society*. Montpellier: Cahiers Scie.

Ashenfelter, O., and Storchmann, K. (2003). Are Mosel wine prices determined by experts or fundamentals? In *Oenometrie X 10th annual meeting of the Vineyard Data Quantification Society*. Budapest, Hungary.

Atkinson, A. B., and Bourguignon, F. (2000). *Handbook of income distribution*. Amsterdam: Elseiver.

Baker, P. (2006). The Conflict Assessment System Tool (CAST): an analytical model for early warning and risk assessment of weak and failing state. Washington DC: The Fund for Peace.

Baliamoune-Lutz, M., and McGillivray, M. (2008). State fragility: concept and measurement. United Nations University, Research Paper No. 2008/44.

Ball, P. (1996). *Who did what to whom? Planning and implementing a large scale human*. American Association for the Advancement of Science. http://shr.aaas.org/Ball/contents.html (Accessed: June 2010).

Ball, P. (2001). Making the case: the role of statistics in human rights reporting. *Statistical Journal of the United Nations Economic Commission for Europe*, Vol 18 No 2–3: 163–173.

Bandura, R. (2008). *A survey of composite indices measuring country performance: 2008 update*. New York: Office of Development Studies, United Nations Development Programme.

Bardossy, A. B. (1987). Fuzzy regression for electrical resistivity: hydraulic conductivity relationships. *Proceedings of the North American Fuzzy Information Processing Society Workshop*.

Bardossy, A. B. (1990). Fuzzy regression in hydrology. *Water Resources Research* 26(7): 1497–1508.

Bates, R., Epstein, D., Goldstone, J., Gurr, T., Harff, B., and Kahl, C. (2003). *Political instability task force report: phase IV findings*. Arlington: Science Applications International Corporation.

Beliakov, G., Pradera, A., and Calvo, T. (2007). *Aggregation functions: a guide for practitioners*. New York: Springer.

Bell, W., and Freeman, W. E. (1974). *Ethnicity and nation-building: comparative, international, and historical perspectives*. Beverley Hills: Sage Publications.

Bertelsmann. (2008). *Bertelsmann Transformation Index 2008: political management in international comparison*. http://www.bertelsmann-transformation-index.de/fileadmin/pdf/Anlagen_BTI_2008/BTI_2008_Brochure_EN.pdf (accessed 1 April 2009).

Bertelsmann. (2010). Transformation Index 2010: Political Management in International Comparison. At: http://www.bertelsmann-transformation-index.de/en/bti/ (accessed 1 June 2010).

Bethany, L., and Gleditsch, N. P. (2005). Monitoring trends in global combat: a new dataset of battle deaths. *European Journal of Population*, 2(3): 145–166.

Bezdek, J., Keller, J., and Krisnapuram, R. (1999). *Fuzzy models and algorithms for pattern recognition and image processing*. Boston: Springer.

Blank, R. M., and Blinder, A. S. (1985). Macroeconomics, income distribution, and poverty. NBER Working Paper Series w1567.

Blanton, R., and Fargher, L. (2008). *Collective action in the formation of premodern states*. New York: Springer.

Bond, D., Bond, J., Jenkins, J. C., Oh, C., and Taylor, C. L. (2003). Integrated Data for Events Analysis (IDEA): an event typology for automated events data development. *Journal of Peace Research*, 40: 733–745.

Bossert, W., and Peters, H. (2000). Multi-attribute decision-making in individual and social choice. *Mathematical Social Sciences*, 40(3): 327–333.

Boumans, M. (ed.) (2007). *Measurement in economics: a handbook*. Amsterdam: Elsevier & Academic Press.

Bouyssou, D., Marchant, T., Perny, P., Pirlot, M., Tsoukias, A. and Vincke, P. (2000). *Evaluation and decision models: a critial perspective*. Boston: Kluver Academic Publisher.

Bouyssou, D., and Perny, P. (1992). Ranking methods for valued preference relations: a characterization of a method based on entering and leaving flows. *European Journal of Operational Research*, 61: 186–194.

Bouyssou, D., and Vansnick, J. (1986). Noncompensatory and generalized noncompensatory preference structures. *Theory and Decision*, 21: 251–266.

Bowles, S. (2004). *Microeconomics: behavior, institutions, and evolution*. Princeton: Princeton University Press.

Bowles, S., Gintis, H., and Gustafsson, B. (1993). *Markets and democracy: participation, accountability and efficiency*. Cambridge: Cambridge Unviersity Press.

Brautigam, D., and Fjeldstad, O.-H. (2008). *Taxation and state-building in developing countries: capacity and consent*. Cambridge: Cambridge University Press.

Bustince, H., Herrera, F., and Monte, J. (2007). *Fuzzy sets and their extensions: representation, aggregation and models.* New York: Springer.

Cammack, D., Christiansen, K., Dihan, M. and Menocal, A., (2006). *Donors and Fragile States agenda: a survey of current thinking and practice.* Report submitted to the Japan international cooperation agenda. London, ODI-JICA.

Campbell, D. E., and Kelly, J. S. (2000). Weak independence and veto power. *Economics Letters*, 66(2): 183–189.

Carbone, E., and Hey, J. (1995). A comparison of the estimates of expected utility and non-expected utility preference functionals. *Geneva Papers on Risk and Insurance Theory*, 20: 111–133.

Carment, D., el-Achkar, S., Prest, S., and Yiagadeesen, S. (2006). The 2006 country indicators for foreign policy: opportunities and challenges for canada. *Canadian Foreign Policy*, 13: 1–35.

Carmines, E., and Zeller, R. A. (1979). *Reliability and validity assessment*. London: Sage.

Chen, S.-J. J., and Hwang, C. (1992). *Fuzzy multiple attribute decision making: methods and applications*. New York: Springer.

Collier, D., and Levitsky, S. (1997). Democracy with adjectives: conceptual innovation in comparative research. *World Politics*, Vol. 49, No. 3: 430–451.

Collier, D., Sekhon, J., and Stark, P. (2010). Inference and shoe leather. In F. David, *Statistical models and causal inference: a dialogue with the social sciences.* Cambridge: Cambridge University Press.

Cooke, R. M. (1991). *Experts in uncertainty : opinion and subjective probability in science.* New York: Oxford University Press.

Cramer, C. (2006). *Civil war is not a stupid thing: accounting for violence in developing countries*. London: Hurst & Company.

Craven, J. (1992). *Social choice: a framework for collective decisions and individual judgements.* Cambridge: Cambridge University Press.

Dahl, R. A. (1990). *A preface to economic democracy*. Berkeley: University of California Press.

Dan, F., and Machover, M. (2000). The measurement of voting power: theory and practice, problems and paradoxes. *Springer Netherlands*, 102(3–4): 373–375.

David, F. (2010). *Statistical models and causal inference: a dialogue with the social sciences.* New York: Cambridge University Press.

de Vaus, D. A. (2002). *Surveys in social research.* London: Routledge.

Dembo, A., and Zeitouni, O. (1998). *Large deviations techniques and applications.* New York: Springer-Verlag.

Upton, G. J. G., and Cook I. (2008). *Dictionary of statistics*. Oxford: Oxford University Press

Di John, J. (2008). Conceptualising the causes and consequences of failed states: a critial review of the literature. Crisis States Working Papers Series 2: 1–51.

Di John, J., and Putzel, J. (2009). Political settlements: issues paper. *Crisis State Research Centre*, 40(2).

Di Nola, A., and Gerla, G. (2001). *Lectures on soft computing and fuzzy logic*. New York: Springer.

Diewald, M., Goedicke, A., and Mayer, K. U. (2006). *After the fall of the wall: life courses in the transformation of East Germany.* Stanford, Stanford University Press.

Dubois, D., and Prade, H. (1989). Fuzzy sets, probability and measurement. *European Journal of Operational Research*, 40(2): 135–154.

Dubois, D., Fargier, H., and Prade, H. (1997). Decision-making under ordinal preferences and uncertainty. In D. Geiger and P. P. Shenoy (eds), *Proceedings of the 13th conference on uncertainty in artificial intelligence*,pp. 157–164. Los Altos: Morgan Kaufmann.

Dubois, D., Fortemps, P., Pirlot, M., and Prade, H. (2001). Leximin optimality and fuzzy set-theoretic operations. *European Journal of Operational Research*, 130(28–29): 20–28.

Düntsch, I., and Gediga, G. (1999). Rough set data analysis. *Encyclopedia of Computer Science*, 43: 281–301.

Ehrgott, M., and Gandibleux, X. (2002). *Multiple criteria optimization. state of the art annotated bibliographic surveys.* Boston: Kluwer.

Fabra, J., and Ziaja, S. (2009). *Users' guide on measuring fragility.* Bonn: German Development Institute (DIE), United Nation Development Programme (UNDP).

Finnemore, M. (2003). *The purpose of intervention: changing beliefs about the use of force.* New York: Cornell University Press.

Freedman, D. (2005). *Statistical models: theory and practice.* New York: Cambridge University Press.

Freedman, D. (2010). *Statistical models and causal inference: a dialogue with the social sciences.* New York: Cambridge University Press.

Gharpuray, M. T. (1986). 'Fuzzy linear regression analysis of cellulose hydrolysis'. *Chemical Engineering Communications,* 41: 299–314.

Giddens, A. (1984). *The constitution of society.* Berkeley: University of California Press.

Goldstone, J. A., Bates, R. H., Gurr, T. R., Lustik, M., Marshall, M. G., Ulfelder, J. C., and Woodward, M. (2005). A global forecasting model of political instability. Washington DC: American Political Science Association:

Goldstone, J. A., Gurr, T. R., Harff, B., Levy M., Marshall, M. G., Bates, R. H., Epstein, D. L., Kahl, C.H., Surko, P. T., Ulfelder, J.C., and Unger A. N. (2000). *State Failure Task Force report: phase III findings*. Political Instability Task Force, McLean, VA: Science Applications International Corporation (SAIC).

Grabisch, M. (2000). A graphical interpretation of the Choquet integral. *IEEE Transactions on Fuzzy Systems*, 8(5): 627-631.

Graham, M. (2002). *Moderation of teacher judgments in student assessment*. Discussion paper on assessment and reporting.

Green, D., and Shapiro, I. (1994). *Pathologies of rational choice theory: a critique of applications in political science.* New Haven: Yale University Press.

Gutiérrez, F. (2009). The quandaries of coding and ranking: evaluating poor state performance indexes. *Crisis States Working Papers Series 18*(2): 1–30.

Gutiérrez, F., and Argoty, C. (2010). Order preserving functions from lattices to R: how many are there? *In evaluation*.

Gutiérrez, F., and Buitrago, D. (2010). Multivariate fuzzy regression. *In evaluation.*

Gutiérrez, F., Buitrago, D., González, A., and Lozano, C. (2010). Fragility, democracy, and development: a slightly different story. *In evaluation*.

Gutierrez, F., and González, A. (2009). Force and ambiguity: evaluating sources for cross-national research – the case of military interventions. *Crisis States Working Papers Series* 50(2): 1–32.

Gutierrez, F., González, A., Buitrago, D., Lozano, C. Users´ Guide State Fragility: The Monopoly – Administration-Territory (Mat) Database. 2010. Unpublished Manuscript

Habermas, J. (1975). *Legitimation Crisis*. Boston: Beacon Press.

Haggard, S., and Kaufman, R. R. (1995). *The political economy of democratic transitions*. Princeton: Princeton University Press.

Hald, A. (2003). *A history of probability and statistics and their applications before 1750*. New York: Wiley-IEEE.

Hann, J., and Kamber, M. (2000). *Data mining: concepts and techniques.* San Francisco: Morgan Kaufman Publishers.

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education,* 20(3): 245 –270.

Hay, P., and Macdonald, D. (2008). (Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject. *Assessment in Education: Principles, Policy & Practice*, 15(2): 153–168.

Hewitt, J. J., Wilkenfeld, J., and Gurr, T. R. (2010). *Peace and conflict 2010: executive summary.* College Park, MD: University of Maryland.

Hiroshi, S. (2009). *Rough sets, fuzzy sets, data mining and granular computing*. New York: Springer.

Hirschman, A. (1981). *Essays in trespassing: economics to politics and beyond.* Cambridge: Cambridge University Press.

Hojati, M., Bector, C. R., and Smimou, K. (2005). A simple method for computation of fuzzy linear regression. *European Journal of Operational Research,* 116(1): 172–184.

Hokstada, P., Oien, K., and Reinertsen, R. (1998). Recommendations on the use of expert judgment in safety and reliability engineering studies: two offshore case studies. *Reliability Engineering & System Safety*, 61(2): 65–76.

Ingenkamp, K. (1997). *Handbuch der Pädagogischen Diagnostik.* Weinheim: Beltz (Psychologie Verlags Union).

Jackson, R. (1990). *Quasi–states: sovereignty, international relations, and the Third World*. New York: Cambridge University Press.

Jarke, M., and Koch, J. (1984). Query optimization in database systems. *ACM Computing Surveys*, 16: 111–152.

Jawahar, C. V. (2002). *Towards fuzzy calibration: advances in Soft Computing*. London: Springer-Verlag.

Kahraman, C. (2008). *Fuzzy multi-critera decision making: theory and applications with recent developments.* New York: Springer.

Kaufmann, D., Kraay, A., and Mastruzzi, M. (2009). Governance matters VIII aggregate and individual governance indicators 1996–2008. *The World Bank Development Research Group, Macroeconomics and Growth Team.* Policy Research Working Paper 4978.

Keynes, J. (1973 [1921]). *The treatise on probability. Now in The collected writings of J.M. Keynes*, vol. VIII. London: Macmillan for the Royal Economic Society.

Kihl, Y. W. (2006). *North Korea: the politics of regime survival.* Armonk, NY: M.E. Sharpe.

Kim, K. J., Moskowitz, H., and Koksalan, M. (1996). Fuzzy versus statistical linear regression. *European Journal of Operational Research*, 92: 417–434.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing social inquiry: scientific inference in qualitative research.* Princeton: Princeton University Press.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd edn.). New York: Guilford Press.

Kooi, R. P. (1980). The optimization of queries in relational databases. Unpublisjed PhD thesis, Case Western Reserve University.

Kuncheva, L. (1991). Evaluation of computerized medical diagnostic decisions via fuzzy sets. *International Journal of Bio-Medical Computing.* 28: 91–100.

Kuper, A., and Kuper, J. (2003). *The social science encyclopedia*. New York: Routledge.

Kwang Jae Kim, H. M. (1996). Fuzzy versus statistical linear regression. *European Journal of Operational Research* 92: 417–434.

Lakoff, G. (1973). Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2(4): 458–508.

Lebart, L., Morineau, A., and Piron, M. (1995). *Statisitique exploratoire multidimensionnelle*. Paris: Dunot.

Lee, C. (1990). Fuzzy logic in control systems: fuzzy logic controller. *IEEE Transactions on systems, man, and cybernetics,* 20(2): 404–418.

Levi, M. (1988). *Of rule and revenue*. Berkeley: University of California Press.

Little, R., and Rubin, D. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Livat, F., and Vaillant, N. (2006). Expert opinion and brand reputation: an analysis from a French Cuban cigars guidebook. *Applied Economics Letters*, 13(2), 97–100.

Lootsma, F. (1997). *Fuzzy logic for planning and decision making*. London: Kluwer Academic Publishers.

MacLaughlin, J. (2001). *Contested terrains and powerful places: the dynamics of ethnicity nation-building and nationalism in the modern world*. London: Pluto Press.

Mann, M. (1984). The autonomous power of the state: its origins, mechanisms and results. *Archives Européennes de Sociologie*, 25: 185–213.

Manton, K. G., Woodbury, M. A., and Tolley, H. D. (1994). *Statistical applications using fuzzy sets*. New York: Wiley.

Marshall, M. G. (2008). Global report on conflict, governance and state fragility 2008. *Foreign Policy Bulletin; 1–21*.

Martino, J. P. (1983). Technological forecasting for decision making, *Elsevier*: *New York*.

Mas-Colell, A., and Sonnenschein, H. (1972). General possibility theorems for group decisions. *Review of Economic Studies*, 39(2), 185–92.

Mckenzie, D. (2002). An economic analysis of LBRD creditworthiness. World Bank. Policy Research Working Paper 2822.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology,* 88: 355–383.

Ministerio de Defensa. (2010). *Política de Consolidación de la Seguridad Democrática (PCSD)*. Bogotá: Ministerio de Defensa.

Minogue, K. (1994). 'The state'. In A. Kuper and J. Kuper (eds), *The social science encyclopedia*. London and New York: Routledge-Taylor & Francis.

Miyamoto, S., Ichihashi, H., and Honda, K. (2008). *Algorithms for fuzzy clustering*. New York: Springer.

Moore, R. E., and Bierbaum, F. (1979). Methods and applications of interval analysis. Society for Industrial & Applied Math.

Mulaik, S. (2009). *Linear-causal modeling with structural equations*. New York: Chapman & Hall / CRS.

Munck, G., and Verkuilen, J. (2002). Conceptualizing and measuring democracy evaluating alternative indices. *Comparative Political Studies*, 1(35), 5–34.

Murofushi, T., and Sugeno, M. (2000). Fuzzy measures and fuzzy integrals. In M. Grabisch, T. Murofushi and M. Sugeno, *Fuzzy measures and integrals: theory and applications*, pp. 3–41 Heidelberg: Physica-Verlag.

Nardo, M., Sasaina, M., Saltelli, A., and Tarantola, S. (2005). *Handbook on constructing composite indicators: methodology and user guide.* OECD Statistics Working Paper, 13: 1–55.

Narens, L. (1985). *Abstract measurement theory*. Boston: MIT Press.

Neumann, J. V., and Morgenstem, O. (1944). *Theory of games and economic behavior.* Princeton: Princeton University Press.

Newmann, F. M., King, M. B., and Rigdon, M. (1997). Accountability and school performance: implications from restructuring schools. *Harvard Educational Review*, 67(1): 41–74.

North, D., Wallis, J., and Weingast, B. (2009). *Violence and social orders: a conceptual framework for interpreting recorded human history.* New York: Cambridge University Press.

OECD/DAC. (2007). *Fragile states: policy commitment and principles for good international engagement in fragile states and situations*. OECD/DAC (Organisation for Economic Co-operation and Development / Development Assistance Committee).

Olso, M. (1993). Democracy, dictatorship, and development. *American Political Science Review*, 87(3): 567–575.

Ostrom, E. (1990). *Governing the commons*. New York: Cambridge University Press.

Pearson, F. S., and Baumann, R. A. (1993). *International military intervention, 1946–1988.* Computer file. St. Louis, MO: University of Missouri-St. Louis, Center for International Studies.

Peters, J. F. (2004). *Transactions on rough sets: rough sets and fuzzy sets*. New York: Springer.

Peters, J. F., Skowron, A., and Pawlak, Z. (2007). *Transactions on rough sets VII*. New York: Springer.

Pfeiffer, S. I., and Jarosewich, T. (2007). The gifted rating scales–school form. *Gifted Child Quarterly*, 51(1), 39–50.

Podgursky, M. J., and Springer, M.G. (2007). Teacher performance pay: a review. *Journal of Policy Analysis and Management,* 26(4): 909–949.

Polybius. (1923). *Polybius: The histories, volume III, books 5–8 (Loeb Classical Library No. 138)*. Cambridge, MA: Harvard University Press.

Powell, B. (1982). *Contemporary democracies. participation,stability and violence.* Cambridge, MA: Harvard University Press.

Prebel, J. F. (1984). The selection of Delphi panels for strategic planning purposes. *Strategic Management Journal*, 5: 157–170.

Przeworski, A. (2004). *States and markets: a primer in political economy*. New York: Cambridge University Press.

Przeworski, A., Alvarez, M. E., Cheibub, J. A., and Fernando, L. (2003). *Democracy and development: political institutions and well-being in the world, 1950–1990.* Cambridge Studies in the Theory of Democracy. New York: Cambridge University Press.

Putzel, J. (1997). Policy arena: accounting for the dark side of social capital: Reading in Robert Putman on Democracy. *(Journal of International Development*) 9: 939-49.

Ragin, C. (2000). *Fuzzy-set social science.* Chicago: University of Chicago Press.

Ragin, C. (2008). *Redesigning social inquiry: fuzzy sets and beyond.* Chicago: University of Chicago Press.

Raiffa, H. (1979). *Decision analysis: introductory lectures on choices under.* New York: Addison-Wesley.

Reeves, D., Boyle, W., and Christie, T. (2001). The relationship between teacher assessment and pupil attainments in standard test tasks at key stage 2. *British Educational Research Journal*, 27(2), 141–60.

Regan, P. (2002). *Users manual for Pat Regan's data on interventions in civil conflicts*. Binghamton University.

Reinhardt, B. (1996). Factors affecting coefficient alpha: a mini Monte Carlo study. In B. Thompson (ed.), *Advances in social science methodology.* Greenwich: JAI Press, 3–20.

Rice, S., and Patrick, S. (2008). *Index of state weakness in the developing world.* Washington DC: Brookings Global Economy and Development.

Richardson, G. (1998). The structure of fuzzy preferences: social choice implications. *Social Choice and Welfare*, 15(3): 359–369.

Roman, S. (2008). *Lattices and ordered sets*. New York: Springer.

Rosenau, J. (1969). Intervention as a scientific concept. *Journal of Conflict Resolution*, 13(2), 149–171.

Ross, T. J., Booker, J. M., and Parkinson, W. J. (2002). *Fuzzy logic and probability applications: bridging the gap.* Philadelphia: SIAM.

Rotberg, R. I. (2004). *When states fail: causes and consequences*. Princeton: Princeton University Press.

Saari, D. G. (2000). Mathematical structure of voting paradoxes: II. positional voting. *Social Science Research Network*, 15(1): 55–102.

Sartori, G. (1970). Concept misformation in comparative politics. *American Political Science Review*, 64(4), 1033–1053.

Sartori, G. (1984). social science concepts: a systematic analysis. Beverly Hills: Sage.

Scapolo, F., and Miles, I. (2006). Eliciting experts' knowledge: a comparison of two methods. *Technological Forecasting & Social Change*, 73: 679–704.

Schumpeter, J. (1942). *Capitalism, socialism, and democracy.* New York: Harper Colophon.

Schwab, D. P. (2005). *Research methods for organizational studies*. Mahwah, NJ: Routledge.

Sen, A. (1970). Interpersonal aggregation and partial comparability. *Econometrica*, 38(3), 393–409.

Sen, A. (1979). Personal utilities and public judgments: or hat's wrong with welfare economics. *Economic Journal,* (89): 537–588.

Sengupta, A., and Pal, T. K. (2009). *Fuzzy preference ordering of interval numbers in decision problems*. New York: Springer.

Shapiro, G. (1997). The future of coders: human judgments in a world of sophisticated software. In C. W. Roberts, *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts.* Mahwah, NJ: Erlbaum.

Skocpol, T. (2007). *States and social revolutions: a comparative analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

Smelser, N. J., and Baltes, P. B. (2002). *International encyclopedia of the social & behavioral sciences*. Oxford: Elsevier.

Snyder, P., and Lawson, S. (1993). Evaluating results of group quantitative investigations. *Journal of Early Interventions*, 61: 334–349.

Srivastava, U., Munagala, K., Widom, J., and Motwani, R. (2006). Query optimization over web services. *Proceedings of International Conference on Very Large Data Bases*.

Stelios, Z., Solomon, A., and Wishart, N. (1998). Multi-attribute decision making: a simulation comparison of select methods. *European Journal of Operational Research*, 107(3): 507–529.

Sung, H.-E. (2004). State failure, economic failure, and predatory organized crime: a comparative analysis. *International Criminal Justice Review*, 41(2):111–129.

Suppes, P., and Zinnes, J. (1962). *Basic measurement theory.* Stanford: Stanford University Press.

Swedberg, R. (1999). *Max Weber and the idea of economic sociology*. Princeton: Princeton University Press.

Taylor, A. D., and Pacelli, A. M. (2008). *Mathematics and politics: strategy, voting, power and proof.* New York: Springer.

Taylor, R., and Judd, L. (1989). *Delphi method applied to tourism*. In: Witt, S., Moutinho, L. (eds) *Tourism marketing and management handbook,* pp.95-99. Prentice Hall: New York.

Thompson, B. (2003). *Score reliability: contemporary thinking on reliability issues*. London: Sage.

Tilema, H. (1991). *International conflict since 1945. a bibliographical handbook of wars and military intervention.* Boulder: Westview.

Tilly, C. (1978). *From mobilization to revolution.* New York: Random House-McGraw-Hill Publishing Co.

Tilly, C (1989). Collective violence in European perspective. In Gurr, T. D. (ed.), *Violence in America,* pp. 62_100. Berkeley: Sage

Tilly, C. (1990). *Coercion, capital, and European states, AD 990-1990*. Cambridge, MA: Blackwell.

Tilly, C. (1993). *Coercion, capital, and European states AD 990–1992*. Hoboken, NJ: Wiley-Blackwell.

Tilly, C., and Stinchcombe, A. L. (1997). *Roads from past to future*. Lanham, MD: Rowman & Littlefield.

Tolley, R., Lumsdon, L., and Bickerstaff, K. (2001). The future of walking in Europe: a Delphi project to identify expert opinion on future walking scenarios. *Journal of Transport Policy*, 8(4) 307–315.

Tullock, G., and Campbell, C. D. (1970). Computer simulation of a small voting system. *Economics Journal*, 80: 97–104.

Vaillant, N. G., Lesot, P., Bonnard, Q., and Harrant, V. (2008). The use of expert opinion, quality and reputation indicators by consumers: evidence from the French vaulting stallion semen market. *Applied Economics*, 42(6): 739–745.

Vreeland, J. R. (2008). The effect of political regime on civil war: unpacking anocracy. *Journal of Conflict Resolution,* 52(3): 401–425.

Warren, W. (2006). *USAID's approach to fragile states programming in Africa*. Washington, DC: USAID.

Weber, M. (1922). *Economy and society*. México: Fondo de Cultura Económica.

Weber, M. (1968). *Politics as a vocation*. Philadelphia: Fortress Press.

Weber, M., and Parsons, T. (1997). *The theory of social and economic organization*. New York: Simon and Schuster.

Welna, C., and Gallón, G. (2007). *Peace, democracy, and human rights in Colombia*. Notre Dame: University of Notre Dame Press.

World Bank. (2006). *Engaging with fragile states: an IEG review of World Bank support to low-income countries under stress*. Washington, DC: World Bank.

World Bank. (2007). Fragile states and the international aid architecture. Washington, DC: World Bank. http://siteresources.worldbank.org/IDA/Resources/IDA15FragileStates-SectionII.pdf (accessed 1 April 2010).

World Bank. (2009). Country policy and institutional assessments. http://siteresources.worldbank.org/IDA/Resources/73153-1181752621336/CPIA09CriteriaB.pdf (accessed 15 September 2010).

Wyatt-Smith, C., and Castleton, G. (2005). Examining how teachers judge student writing: an Australian case study. *Journal of Curriculum Studies*, 37(2), 131–154.

Yao, J., Dash, M., and Liu, T. A. (2001). Entropy-based fuzzy clustering and fuzzy modeling. *Fuzzy Sets and Systems,* 113: 381–388.

Yu, D., and Park, W. (2000). Combination and evaluation of expert opinions characterized in terms of fuzzy probabilities. *Annals of Nuclear Energy,* 27: 713–726.

Zadeh, L. (1971). Similarity relations and fuzzy orderings. *Information Sciences,* 3, 177–200.

Zadeh, L. (1975a). Fuzzy logic and approximate reasoning. *Synthese,* 30: 407–428.

Zadeh, L. (1975b). The concept of a linguistic variable and its application to approximate reasoning. *Information Science*, 8(4), 301–357.

Zadeh, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems,* 1: 3–28.

WEBSITES:

100000 Quotes.
www.100000quotes.com
(accessed April 15 2010).

Academic Ranking of World Universities.
www.arwu.org/
(accessed April 15 2010).

ATP World Tour.
www.atpworldtour.com/Rankings/Rankings-FAQ.aspx#pointvalue (accessed June 1 2010).

British Science Association.
www.britishscienceassociation.org
(accessed June 1, 2010).

Carleton. Country Indicators for Foreign Policy.
www.carleton.ca/cifp/
(accessed January 20 2010).

CIA World Factbook
www.cia.gov/library/publications/the-world-factbook/
(accessed January 10 2010).

Dictionary (2010).
http://dictionary.reference.com/browse/database
(accessed March 5 2010).

ELO Rating System.
http://en.wikipedia.org/wiki/Elo_rating_system
(accessed March 1 2010).

Eurostat.
http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes
(accessed February 15 2010).

FIFA world ranking.
www.fifa.com/worldfootball/ranking/lastranking/gender=m/fullranking.html.

Great Quotes.
www.great-quotes.com/
(accessed May 5 2010).

Human Resources, University of Glasgow.
www.gla.ac.uk/services/humanresources/policies/p-z/seniorclinacad/schedule3/
(accessed May 1 2010).

International Road Federation.
www.irfnet.org/
(accessed January 10 2010).

Japan Sumo Association.
www.sumo.or.jp/eng/
(accessed January 15 2010).

Polity IV Project.
www.systemicpeace.org/polity/polity4.htm
(accessed January 20 2010).

Rankopedia.
www.rankopedia.com/
(accessed April 15 2010).

Standard and Poor's.
www.standardandpoors.com/home/en/us
(accessed April 22 2010).

Truth Commission (2003).
www.cverdad.org.pe/ingles/pagina01.php
(accessed 10 June 2010).

University of Minnesota, Human Resources.
www1.umn.edu/ohr/policies/governing/civilrules/rule12/
(accessed June 1 2010).

Universal Postal Union.
www.upu.int/
(accessed January 10 2010).

UNODC Colombia.
www.unodc.org/colombia/es/index.html
(accessed January 25 2010).

UNODC Crime Trends.
www.unodc.org/unodc/en/data-and-analysis/United-Nations-Surveys-on-Crime-Trends-and-the-Operations-of-Criminal-Justice-Systems.html
(accessed January 25 2010).

USAID.
www.usaid.gov/
(accessed January 30 2010).

WHO.
www.who.int/en/
(accessed February 10 2010).

World Bank Indicators.
http://data.worldbank.org/indicator
(accessed January 10 2010).

World Bank, International Development Association.
http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/IDA/0,,contentMDK:22356311~pagePK:51236175~piPK:437394~theSitePK:73154,00.html
(accessed January 30 2010).

World Bank, Tax Revenue.
http://data.worldbank.org/indicator/GC.TAX.TOTL.GD.ZS
(accessed May 1 2010).

CRISIS STATES RESEARCH CENTRE REPORT

www.crisisstates.com

Crisis States Research Centre
LSE
Houghton Street
London WC2A 2AE

e: csp@lse.ac.uk

www.crisisstates.com

THE CRISIS STATES RESEARCH CENTRE

The Crisis States Research Centre (CSRC) is a leading centre of interdisciplinary research into processes of war, state collapse and reconstruction in fragile states. By identifying the ways in which war and conflict affect the future possibilities for state building, by distilling the lessons learnt from past experiences of state reconstruction and by analysing the impact of key international interventions, Centre research seeks to build academic knowledge, contribute to the development of theory, and inform current and future policy making.

The Centre is based within the Department of International Development at the London School of Economics and Political Science and is funded by a grant from the UK Department for International Development.