Marina Arbetman-Rabinowitz PhD
Sentia Group & Claremont Graduate University

Proposal for the Workshop to Support the Human, Social and Cultural Behavior (HSCB) Modeling Program- July 2008

Time to Measure Up: Dirty Little Secrets in Data Collection[1]
Or on how Econometrics is Easier without data[2]

Summary:
In political economy there are an abundance of theoretical postulates and models, yet very few indicators and measures are developed to test the validity and reliability of the arguments beyond contrasting the results to reality. This is particularly evident in the lack of value placed on data collections, with little to no emphasis attached to maintaining time series data bases that relate to variables beyond the economic realm.
Modern data-collection techniques allow us to **systematically** collect information about our objects of study (people, objects, phenomena) and about the settings in which they occur. The constraints in availability of data are one reason why data is collected haphazardly; and as a consequence makes it difficult to answer our research questions in a conclusive manner. This paper will chart the steps to successful data collections through specification before estimation, operationalization, data collection techniques, data collection and measurement repeatability, reproducibility, accuracy, and stability.

[1] The author want to thanks Jacek Kugler, Kristin Johnson, Mark Abdollahian and Dean Hartley for very useful comments.
[2] Verbeek, Marno. 2000. A Guide to Modern Econometrics. Chichester: John Wiley & Sons. p.1.

Marina Arbetman-Rabinowitz PhD
Sentia Group & Claremont Graduate University

Time to Measure Up: Dirty Little Secrets in Data Collection
Or on how Econometrics is Easier without data[3]

Contents

1. **Introduction**
2. **Theory. Level of Analysis**
3. **Unit of Analysis**
4. **Specification before estimation**
5. **Operationalization. (Semantics)**
6. **Data Collection (Syntax)**
7. **Data Management Strategies (Praxis)**
   a. **A note on expert based data**
8. **Reliability and Validity**
9. **Ready for estimation?**
10. **Conclusion**

Appendix A: Data Archives, Updates and Dissemination
Appendix B: Sources and Data Sets Descriptions

[3] Verbeek, Marno. 2000. A Guide to Modern Econometrics. Chichester: John Wiley & Sons. p.1.

1. Introduction

Data collection is a forensic discovery.

The following paper focuses on data collection techniques to maximize the quality of the input that eventually populates models. By systematizing the data collection step, the modeler can rule out or minimize the "blame" placed on data and variable specification when the output is not what was expected.

Social scientists are very aware of the need to develop and refine theories to grasp the current and future behavior of the political, economic and social world. Reality is too vast and chaotic to comprehend. Therefore we have to simplify our perceptions of reality in trying to understand, forecast and hopefully influence human behavior and policy outcomes. The process of developing theories is the most stimulating part of research; the most rewarding is if our theories stand the scrutiny of testing.

Research[4] can be divided into non-empirical and empirical research. Both types of research require an agreement on what the assumptions are and a deductive approach to providing answers. Within non-empirical research, the focus of normative research is on what should be. Frequently it is not possible to test beliefs and assertions. Here formal theory is helpful because it posits assumptions about human behavior and by logical deduction arrives at unequivocal conclusions. At this point all areas of research interact with each other; for example, formal theory can present the deductions arrived at in a mathematical format and then, if empirical evidence is available or if it is fitting given the objective of the research, such deductions can be tested empirically.

This paper does not concentrate on modeling, nor does it center on econometric methods. And, although it is impossible to talk about data in a vacuum, this paper will focus on the empirical part of theory-oriented and policy-oriented research. Separating research into these two categories is a heuristic exercise, seldom following a set of specific rules as often these categories overlap.

In any case, policy oriented research has a similar structure to engineering research, more oriented towards solving problems with an imbedded normative stand. Theory oriented research, on the other hand, posits questions with the main goal of expanding our knowledge of the political and socio-economic system. The methodologies to advance those substantive approaches vary from descriptive to logical, mathematical and empirical and the subjects can be studied with the same scientific rigor from a single, in depth case study to a large

[4] Shively, Phillips (1990). The Craft of Political Research Prentice Hall: New Jersey. Chapter 1.

cross subject sample. Never dismiss anecdotal evidence which can be very informative to build an inductive theory but remember that it may be an empirical aberration. This type of work does not utilize the systematic data collections described in this paper.

In general, applied empirical research has often been perceived as less elegant than theoretical and mathematical modeling. So why move to a research area that is murky, where results are less certain, based on probabilities and dependent upon the quality of several other steps, from specification to measurement? Why not stop at the theoretical, logical or mathematical modeling? It is an issue of practicality. Models will be useful in the real world only if we can use them to increase our understanding of reality. Statistical models are developed to articulate the theories and test them against reality. Therefore, regardless of the substantive approach, all research requires validation, with empirical testing providing a line of attack to present evidence. Social science is not a lab where simulated conditions can be replicated. Therefore, theory specification, operationalization, data collection, and finally estimation are what allow the researcher to make statements about the world and predictions about the future.

Unfortunately moving from the theory specification phase to the empirical one is not an easy task. The client, more often than not, thinks that the data sets needed exist somewhere "out there", just waiting to be pulled out and used. It is a myth! The common perception is that the only difficulty is to build a theory or combine existing ones, move forward in the modeling and specification. Too little time is devoted to the phase of operationalization and data collection. The reality is that analysts and researchers cannot do analysis without data but deadlines arrive quickly and the time allocated to obtaining data gets shortened. What is important gives way to what is urgent. Data sets are not available and the variables are not defined the way the model is specified, time series are shorter than expected, the methodologies or the samples for the specific data set are questionable, the evidence may or may not reflect directly the concept as originally defined and so on. The reality of having to populate a theoretical model by a certain date forces the project manager to cut corners.

With these problems in mind, the next few sections will address a plan of attack from the initial phases of theory building that will minimize problems during the testing phase. The next three sections will address the issue of level of analysis in theory building, the importance of clear specification before the researcher even thinks about estimation and how to think about operationalization of variables. The following sections will address specific practical data queries: the data collection itself, the problems of validity and reliability in data and determining when the data is ready for the modeler to start estimating. The final two

appendices will present a few pointers for data archives, updates and dissemination of information and a brief description of some available data sets for the social sciences

2. Theory. Level of Analysis

- *Neither – Nor….. we need them all!*

Effective theories require simplicity, importance and predictive accuracy[5] . Of the three elements predictive accuracy will only be tested ex post facto but there are many steps that we need to follow to maximize the probability of success of theories. Importance is addressed from the start in policy or engineering oriented research since the primary goal to start the research is to understand a specific practical problem and in the case of policy, design a strategy. Theory oriented research, on the other hand, is more scholarly based, often reviewing existing theories and is often one of the resources for applied models. Simplicity should allow us to identify the main predictors and build up more complexity as it is deemed necessary. Proper conceptualization of the problem at hand will reduce the time invested later on in the process.

Nicholas Sambanis[6] posits that the gap between micro behavior and macro implications is already evident at the conceptualization stage as the causal paths go both ways. The gap increases in terms of divergence at the testing stage when either micro level data or cross national macro data is used. Simplifying reality of course and separating macro from micro theories to make the testing feasible has trade-offs. Empiricists often characterize this trade-off as: "Part 1 of Errors and Omissions", expecting many more to come as the research process advances. Although there are differences in modeling, it is not necessary to go only one way, either top down to bottom up or, specific to general. Both approaches have pluses and minuses.

Bottom up research tends to be biased but rich in content as shown in case studies. On the other hand, top down research may be too complex and a general model requires too many variables. It is just not realistic to think that we can estimate initially a general model by incorporating all conceivable variables and functional forms. One way to compromise is to begin with simple models and expand when they fail or incorporate out of sample forecasts when it is a clear case of misspecification. One of the real challenges once the modeling approach is decided lies on the data side. One must keep in mind that case studies

---

[5] Shively *ibid*. Chapter 2

[6] Sambanis, Nicholas. 2004. "Using Case Studies to Expand Economic Models of Civil War" Perspectives in Politics. Vol. 2. No 2. pp 259-279.

complement but do not empirically test theories, one can filter empirical measures extracted from case studies, highlighting explanatory variables with viable parameters.

The selection of the level of analysis is decided by methodological and conceptual reasons. While the system level approach highlights the importance of system, it discounts the importance of each actor.[7] Focusing on the actor tends to highlight ethnocentrism, as well as the individual's goals and motivations. It is a matter of acknowledging these issues. In the end, case studies and generalized theories should cross fertilize, making it possible to identify underlying causal mechanisms that may be more difficult to see at a macro level. Although case studies by their own nature are biased in terms of their selection and are likely to omit variables, they can test the validity of some assumptions. On the other hand a more generalized theory will distill the contradictions of specific cases.

The discussion of the advantages and disadvantages of deductive versus inductive theories resembles the discussion between barefoot empiricism and theoretically driven approaches. Often the best practical strategy is to come at the problem from both angles. In the same way that looking at one case study can enlighten us about the chronological sequence of events or can trace the sequence of independent variables, barefoot empiricism can lead us to identify non-linearity issues. Examination of data, looking at a scatter plot or at the residuals may highlight the need to add to include polynomial variables values of X that were missing from the specification[8] and that a theoretically abstract approach may not have indicated. An illustrative example can be found in the ongoing debate in explaining individual choices and whether or not individuals always choose to maximize utility. Lab research conducted by Daniel Kahneman and Amos Tversky demonstrated that individuals violated axioms of expected utility theory systematically when faced with choices under different conditions of risk (Kahneman et al 1982).[9] Instead of demonstrating consistency in decision making, individuals made different decisions based on whether or not an issue was important, and if that issue was framed in terms of gains or losses. This systematic violation, based on the frame of gains or losses, resulted in the formulation of prospect theory and fueled in large part the field of behavioral economics. These competing theoretical arguments have driven empirical applications and testing of decision making across issues from nuclear deterrence to governance, to US Japanese trade interactions.[10] In the end, theory and

---

[7] Singer, David. 1961. "The Level of Analysis Problem". *World Politics*. Vol. 14. No 1. pp. 77-92. The Johns Hopkins University Press.
[8] A word of caution, when including power of X variables is that it often leads to collinearity.
[9] Kahneman & Tversky. 1984. Choices, Values & Frames. *American Psychology*. 39:341-50.
[10] Mercer, Jonathan. 2005. Prospect Theory & Political Science. *Annual Review of Political Science*. 8:1-21.

empirics go together. Sometimes theory drives empirics and other times, "data talks to us" and, empirics drives theory construction.

3. Unit of Analysis

On how to break the rules and get away with it.

Somewhere between the process of modeling and specification, the researcher focuses on the unit of analysis. The unit of analysis refers to the unit of assessment in the analysis as well at the unit of data collection. Units of analysis commonly chosen in research are: individuals, households, groups (ethnic, age, gender), communities, geographic units (countries) but also the unit of analysis can refer to a characteristic, object or an attribute (newspapers) or a social interaction (dyadic relation).

Traditionally, researchers have been very aware of the problems derived from the ecological fallacy and of the exception fallacy. Ecological fallacy refers to reaching the wrong conclusion about individuals based on a macro level of analysis. Using macro level data to test hypothesis based on micro level behavior is never the answer. In the conflict literature, often cross national statistical analysis is performed when the objective is to explain the micro-macro implications. For example, civil war is often studied at the cross national level, despite domestic motivations. This results in comparisons across a set of countries experiencing varying levels of internal conflict, but does little to examine what is happening within the countries.

The Exception fallacy is the reverse, reaching a conclusion at the macro level based on an individual case that might have been an outlier. The conclusions derived from the analysis at the macro level may not apply at the individual level and vice versa. In principle, the analysis should be done at the level from which generalization can be made and inferences can be valid and reliable. In the real world, focusing on a particular level of analysis may lead to important gaps in understanding. One remedy is to perform assessment at different levels **to** allow any inter-linkages between them to be explored.

In many studies or for different analyses in the same study, the research questions can and should be posed at different levels of analysis using different methodologies, approaches and data sets. This type of hierarchical modeling[11] opens the doors to defining necessary and sufficient conditions. An example can help illustrate this point. In the international system, power transition is a system level theory that posits what distribution of power outlines the necessary

[11] www.socialresearchmethods.net/kb/**unit**anal.php

conditions for conflict. This same theory has been applied to a lower level of analysis, to the provincial level to find out the probability of conflict within a country among different regions. Even though analyzing those two levels concurrently presents a picture of the international threats as well as the domestic threats, neither approach points at the sufficient conditions for conflict. All parties involved need to have positive expectation about their gains from a conflict, ceteris paribus. The unit of analysis to fully answer this query is the individual or group involved in making the final decision. Agent based models at the micro level are the right tool. Therefore by using different levels and units of analysis, the analyst can uncover the necessary and sufficient conditions for conflict and then trace these to the leaders that choose to initiate conflict or preserve peace.

4. Specification before estimation

- *This is not the time to reinvent the wheel!*

Inductive and deductive theories, theory driven and data driven models, all of them need to be tested to ascertain their validity and the explanatory and predictive power of the arguments. Elegant theories and specifications need to meet the criteria of simplicity, importance and predictive accuracy. The process of specification is crucial to the success of the modeling effort. It requires time; cutting corners in this area will require investing a lot more time later, requiring re-specification after the estimation. It is practically impossible to avoid some rethinking of the logical nodes or the nested loops in the original specification but the goal is to minimize these adjustments. This specification is the first step to constructing a data base. From this standpoint there is always a trade-off between complexity and time and simplicity and effectiveness.

The most efficient and effective way of maximizing effort is to avoid reinventing the wheel! Modelers and researchers always want to start from zero but social sciences have accumulates 50 years of literature advanced by experts in the field who have thought through the problems. Social scientists have defined the concepts, worked through data and indicators and know the advantages and limitations of different approaches. Knowing the assumptions and specifications used in previous works, their problems and choices others have made can teach us many lessons. In this era of team work, physicists, econometricians and mathematicians are a major asset to the modeling effort but in a vacuum they tend to ignore that input. Moreover, the surge of studies and the increased attention lent to the estimation technique at the expense of the specification and appropriate data leads to incorrect and problematic inferences.

Great effort had been spent in the specification of the model, research and policy questions have been articulated, the model specified, the concepts mapped. Clearly, the fundamentals of applied work require a problem oriented rather than technique oriented approach, the next step is to operationalize these concepts. Build model with data or the other way around, compromise. But the question that remains is: when do I decide what econometric method to use? Once the research design is finished and the system of equations defined or after I know the data that I will use?

Remember that ***"Econometrics is much easier without data"***, but we still need to have a practical answer to the problem posited.

5. Operationalization

Garbage In… Gospel Out!

"The weak and too slowly growing empirical foundation clearly cannot support the proliferating superstructure of pure, or I should say, speculative economic theory" Leontieff (1971)

Leontieff's quote of almost forty years ago is just as valid today. Going from complex reality to modeling is a first leap; going from abstract to empirical models constitutes another big leap. Once the modeling stage is completed, the task is to translate descriptive concepts into measurable ones. Something is always lost in translation; in other words empiricists will say this is "Errors and Omissions Part II", again, expecting many more to come as the research process advances

Operationalization is the neglected child of theory. Once the conceptual side of modeling is finished the mad rush begins. Social scientists have been measuring variables and phenomena for many decades with varying degrees of success. One of the unglamorous projects in social sciences - that would be incredibly useful- but has not been done yet is the development of a library based on past research of concepts and their corresponding variables and measures with an explanation of the problems reported and the success at closing the gap between concept and measurement (metadata). In any case, a less ambitious objective is to make sure that the social science literature is reviewed for each project and an inventory is taken, again before reinventing the wheel, project managers need to build on cumulative knowledge. In many cases, the project will require building a new data set at the risk of measuring variables that are too many degrees removed from the original concept and rendering irrelevant results.

The debate over qualitative vs. quantitative research is obsolete. Each approach has costs and benefits. The nature of the variables and model will determine what is required and the alternative measures. Qualitative variables call for well defined verbal statements to translate into weights, rankings or continua. Some such efforts can be systematic. Content analysis, for example, can provide a useful tool in gauging perceptions and overall attitudes in messages. An interesting example is an examination of the way in which Iran is portrayed in western media sources from the mid 1950's to 1978 by Dorman and Farhang.[12] This study evaluates both, overall positive, negative and neutral framing in discussions of Iran, first as an ally, then as an enemy, and evaluates the degree of similarity between major media reporting and State Department press releases. This type of analysis is a significant undertaking, requiring systematic assessment of all articles in designated newspapers and other publications during the time period under study, and requiring rigorous coding and data collection strategies. Quantitative variables require one less step in translating them into standard units, such as currency, area, people, etc. Although, less quantitative measures offer a greater breadth and capture the complexity of social phenomena, they also open the door for the interpretation of the adjective used. On the other hand, things that can be measured in units may not give us a good idea of magnitude; for example, the number of protests in a country does not illustrate their magnitude or societal friction – which ranges from peaceful democratic engagement to a demonstration of domestic instability.

Precise information tends to be easier to interpret. Established quantitative variables require one less step in translating them into standard units, such as currency, area, people, and so on. For example, summary measures lend little discretion in interpretation. Furthermore, the interpretation of unidimensional concepts is always less problematic than multidimensional concepts, but sometimes it is unavoidable. The same applies to interaction terms. Often they explain social phenomena better than linear unidimensional concepts, but the tricky part is the interpretation. The researcher needs to be aware that regardless of the type of variable used, the results should not mean different things to different people.

When operationalizing variables, the tacit objective is to build evidence for comparability and to answer the original question posited by the project. In this vein, the discussion between macro and micro level data is critical. Using macro level data to test hypothesis based on micro level behavior is never the answer. In the conflict literature, often cross national statistical analysis is performed when the objective is to explain the micro-macro implications. For example, civil

---

[12] Dorman, William A. and Mansour Farhang. (1988). *The US Press and Iran: Foreign Policy and the Journalism of Deference*. University of California Press: Berkeley

war is often studied at the cross national level, despite domestic motivations. This results in comparisons across a set of countries experiencing varying levels of internal conflict, but does little to examine what is happening within the countries.

Rarely are we able to measure our concepts directly, and our measures do not always reflect the richness of our mental constructs. Many of our concepts are complex and do not easily lend themselves to easy interpretation. For example, power is a concept that has different meanings in different settings: national or individual level, and is often multidimensional and can interpreted as everything from military capability to economic prowess and resource mobilization [13].

The semantics of operationalization cannot be taken lightly. What do I want ideally to measure? What are the alternatives? Different meaning of the variables can drive different conclusions. Democracy sometimes increases the risk of conflict between countries, sometimes it is a harbringer of peace and stability according to a number of different analyses. Maybe it is not the structure of the government that the model is after, or perhaps a non linear effect is being represented. The most direct example of this exists in measures of GDP. Economists are lucky to have this overall concept that is used as a big hammer to gage the economic activity. Sometimes GDP represents wealth or power in a model, but often the measure is unsuited to the needs of a particular model. For example, GDP per capita is used as a measure of personal wealth for a country; with the implicit hypothesis that a higher level of GDP per capita is related to a lower level of poverty. This is not the case: without knowing the GINI coefficient or some other measure of distribution of income, the level of poverty is not captured. Other examples in the same category are social class, status, interaction (trade, mail exchanges, phone traffic, alliances, etc.). These concepts ring a different bell for each researcher and reader, and the way the subcomponents are weighted will render different results. I will go back to this issue in Section 7. Theory ends up being adapted to the data at hand and to fit the models, not the other way around. As a consequence, the results of poor data collection and specification seldom fit the theory, when in reality the model tested is not what was put forward. The final product is a poorly done piece of research with no clear causal or explanatory directions.

What if there is no direct measure of the concept the model is trying to capture? For many concepts there are no simple indicators, so we can only use indirect measures. The only answer is to think outside the box. If what we want to

---

[13] Kugler, Jacek & Marina Arbetman. 1989. "Choosing Among Measures of Power: A Review of the Empirical Record", in Michael Ward and Richard Stoll Power in World Politics, Lynne Rienner Publishers: Colorado.

measure is difficult to assess directly, it often can be inferred from other behavior. Ken Organski and Jacek Kugler[14] wrote about the difficult of measuring an elusive concept such as the political capacity of the government. It is clear that knowing how capable a government is, it would be easy to identify governments that would be able to implement their desired policies. Organski equated this problem to knowing that a man is present but being unable to see him. However, even if one cannot see the man if it is possible to see his shadow then it is possible to make an inference. Hence his solution to the query was to follow the logic that every government wants resources to implement their policies but even under similar economic circumstances, some governments are able to extract more resources than others. The answer to this measurement dilemma lies in locating the shadow.

Economics, likewise, can explain part of fiscal extraction but after economics is accounted for, governments that are politically capable are the ones that can extract more. This ingenious way of measuring Relative Political Capacity[15] should be emulated in other areas of research as well. For example, measuring the size of the informal economy, text analysis, where what is left unmeasured or unsaid is as important as the recorded side and, inferences can be made by looking at residuals and correlations.

An additional technique for dealing with explanatory variables that are difficult to measure or demonstrate estimation difficulties is through the use of an instrumental variable. Sometimes the best direct measure of a variable ends up being highly correlated to error terms in estimation, either due to measurement error in covariates, covariance, or endogeneity. OLS estimation, despite consistent measures, yields biased results due to the correlation. In some instances, researchers can find an instrument, or a variable that does not belong in the equation, that is highly correlated with the explanatory variable, but not correlated with error terms.[16] Certainly, the use of "weak" instruments or instruments that are poor predictors does not resolve endogeneity concerns. Selection and use of instruments should be guided by caution rather than expediency.

Think before you measure!

---

[14] Organski, A.F.K. & Jacek Kugler. 1980. The War Ledger. University of Chicago Press: Chicago

[15] See also Organski et al. 1984. Births, Deaths & Taxes. University of Chicago Press: Chicago. Arbetman, Marina & Jacek Kugler eds. 1997 Political Capacity & Economic Behavior. Westview Press: Colorado.

[16] For an extensive overview of the use and abuse of instrumental variables, see Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference,* Cambridge University Press.

The main lesson of this process is that there is a trade off between data that is available and what the model has specified. The best possible measures may not be available and it may be necessary to collect original data from primary sources. Sometimes, the only variables available may not be the best. For example, in assessing violence within a country, total deaths may be the only available measure that reflects instability, although it may not be representative of the concept we want to measure. In any case the worst case scenario is "Garbage In… Gospel Out". Using the wrong variable will lead to invalid results. Deaths from violence may not be a great indicator of instability but there are not other measures comparable across an entire sample available. Basically there is a trade off between availability of data and how a measure captures a concept. In sum, what do I want to measure and how do I make this leap? Is this the time to start compromising?

You don't want to do this but you have to!

6. Data Collection

- *My Personal Experience: Dirty Little Secrets*

Before starting the data collection it is crucial to know the context of the project, the history of the institutions, as well as the measurement peculiarities, cultural and operating constraints. In this way we can have an idea of the methodology as well as the problems that the researcher is likely to encounter. For example, as an economist, one would think that gathering Gross National Product from international sources, such as the IMF, should not have many problems of comparability since the definitions and methodology are universal. In reality, the fiscal year does not coincide across the board and often series need to be adjusted, particularly when there is inflation. Research and documentation of each of the variables is needed, beginning with formulas used to input data and extending to how the samples were selected and what instructions were given to participants. Stress the definitions of the variables and the accounting conventions followed. Critical to the collection of reliable data is to ensure that specific definitions are consistent, even within a single source such as the IMF, and particularly across sources. This means that researchers should pay close attention to units, timing and small differences in what included or excluded from a particular measure.

One of the first questions to answer is what should be collected if the definitions are similar but not quite the same? For example, we usually talk about GDP but we find series with GNP, GDP, NMP, at market prices, at cost factor, in current prices, in constant prices, in national currency, in international currency, flow or

stock FDI, number of protests or number of individuals present, etc. How do I choose between different measures? There are theoretical and practical reasons to collect more than one series when possible. Not only does it help in case interpolations and extrapolations are needed, but also to conduct reliability and validity tests. How many and what series to choose to put in the project data base is project dependent. Keep in mind that there are a few original data collectors, primarily Statistical Offices and Central Banks and some, very few international organizations. The rest is a rehashing of the previously existing data. So always check the source and get as close to the original source as possible. There is always a trade off: the more series included in the project data set, the more careful the design of the data base should be, as with each additional included similar measure, error introduced by lack of attention to inputs or other human error does increase. Database assembly is painstaking and requires great attention to detail, thinking about the data set requirements and design, and implementing a carefully crafted and systematic approach pays off.

This list includes some useful hints to follow prior to each intervention:

Pre-Data Collection Steps

- Clearly define the goals and objectives of the data collection
- Reach understanding and agreement on operational definitions and methodology for the data collection plan
- Inventory the sources and measures of concepts/variables
- Research definitions, accounting conventions and other descriptors
- Identify sources: original or secondary
- Ensure data collection and measurement repeatability, reproducibility, accuracy, and stability.
- Document! Document!

In sum, before you start the data collection you have found the answers to some practical and theoretical questions: Where does the data come from, the sources and their reputations? If and how the data has been peer reviewed? How were the data collected?

Regardless of quality, the appropriate data needs to be collected, based on the original specification:

A good data collection plan should include:

- A brief description of the project,
- The specific data that is needed,
- The rationale for collecting the data,
- Detailed definition of each variable
- What insight the data might provide and how it will help understand the driving question, and
- What will be done with the data once it has been collected.

A few of these points need to be expanded. Mapping of the conceptual definition of the variable and its rationale for the way in which it was operationalized will help retrace steps back in case the variable needs to be changed, transformed or adjusted due to lack of data. Again the use of GDP is an easy example. National Product is a variable that by now we all think that is piece of information that is produced and updated as we speak. This is not always the case. Equatorial New Guinea, for example[17], did not have the technical capacity to calculate its own GDP until a decade ago. And according to the IMF one of the major issues in data collection is the lack of accuracy and reliability (48% in the National Accounts category, 54% in prices, 47% in Balance of Payments).[18] So here we have one of the major series used in economics that is widely accepted, barely disputed and the experts themselves know that measuring GDP is not a lab experiment under controlled conditions. The examples are frequent. Many countries have barter economies and informal sectors, so monetary transactions go unrecorded and the GDP is underestimated. The way the CPI is calculated also gives room for a large margin of error. A basket of goods for that specific country is surveyed in a sample of cities, and those results extrapolated. What is the degree of confidence that the research community has when it comes to the CPI of China? So, do we stop doing empirical research? Of course, the answer is no. Analysts gather information, aware of the problems, become familiar with detailed definitions of the variables and correlate the variables used with other indicators to validate. In the end, we can attach a certain degree of confidence to the data set in use.

After highlighting the measurement problems of many variables that the research community considers safe, the reader needs to be informed that often they have to build their own datasets, with added technical difficulties, increased budgets and time. In spite of the problems, the researcher should consider themselves lucky if they find the data sets they need. Even if the data set exists, often the sources are not constant, the series, especially polls; are shorter than

---

[17] Klitgaard,…. Tropical Gangsters

[18] International Monetary Fund. 2008. "Assessing the General Data Dissemination System (GDDS) – What has Been Accomplished After 10 Years and Where Do We Go from Here?" Statistics Department. pp.15.

expected, often surveys; or change definitions and methodologies or samples mid way, as is the case of surveys and census. It is feasible to recalibrate. It takes a long time to transform the data set into the needed format but it takes much longer to build a new one. There are immediately trade offs: deadline vs. what is really needed to populate the theoretical model. Again, it may be time to think outside the box and redefine what the model needs to capture or work with what is available.

Until now we have not classified the different types of data or the ways of collecting it. Data-collection techniques allow us to **systematically** collect information about our objects of study (people, objects, phenomena) and about the settings in which they occur. The stress is on systematic because if the data are collected haphazardly, it will be difficult to answer our research questions in a conclusive way. Alternative collection techniques include the following:

Various data collection techniques can be used such as:

- Using available information
- Data mining[19]
- Observing
- Interviewing (face-to-face)
- Administering written questionnaires
- Focus group discussions
- Projective techniques, mapping, scaling
- Key informants- SMEs

Of the many various techniques that can be used for data collection, the next section will present strategies to survey and transform a data set or variables that already exist in some electronic or printed format.

7. Data Management Strategies (Praxis)

"Every number is guilty unless proved innocent"
(Rao, 1997, p.152) quoted in Kennedy

Usually there is a large amount of data that has already been collected by others, although it may not necessarily have been analyzed or published. Locating these

[19] Data mining is very much in vogue now. It is the principle of sorting through large amounts of **data** and picking out relevant information; the extraction of hidden predictive information from large databases.)

sources[20] and retrieving that information is a good starting point. The format of the data set may need to be modified to suit the model for the project. Alternatives formats of the data set are single time series or cross section or panel (cross sections of individuals observed over time). The following is a check list of what to do.

- **First step:** *Define your data matrix*. The two formats most commonly used are cross-sectional, a single point in time across variables; or longitudinal, when the study takes place over time. Usually data in electronic format will be in longitudinal matrices per country with y= item and x= year.). The data will need to be put into panel format for econometric analysis, that is y=country and year and x=item. If there is a change in definitions or methodology, it should be noted in bold or color, a definition of the change and an example will be helpful in the future. The cost of not completing this step is often to have to return to the original sources to be able to explain abrupt changes in the series.

For example, the Government Financial Statistics Yearbook (on line and in printed format) presents their data in the following format:

Longitudinal
Argentina

	1990	1991	……
Primary	10	12	
Secondary	20	25	
Tertiary	30	35	

Needs to be transformed to Panel

Country	Year	Primary	Second	….
Argentina	1990	10	20	
Argentina	1991	12	25	
Argentina	…..			

- **Second Step**: *Data collection from the primary source.* This is the source that has been determined to have the best base series. This is the "**Cut & Paste**" period,

[20] For example, some sources can be analysis of information system data, census data, unpublished reports and publications in archives, newspapers, electronic media and libraries. The use of **key informants** is another important technique to gain access to available information. Key informants could be knowledgeable community leaders

and should be easy, mechanical and fast. There is an issue of precision vs. accuracy. Still attention needs to be paid to the currency, what type of currency, is the data in millions or billions, change of currency.

- **Third Step:** *Quality control of primary sources*. Inspect the data to become familiar with it. The computer revolution and access to internet has made it too easy to push a button and access a data set with the accompanying drawback that researchers are no longer familiar with the data. Inspecting the data involves doing summary statistics, graphs, time lines, residual plots, data cleaning to check and to get a feel for the data. Summary statistics as simple as means, standard errors, maximums and minimums, correlation matrices will help several other tasks such as identifying outliers and other "guilty" points as well as beginning to understand the inferential relationships within the data series. For some researchers it is easier to map the same issues on simple graphs, histograms, scatter plots, or residual data over time. In sum, data cleaning looks for inconsistencies in the data-observations and unrealistic or suspicious values. Some details need to be recorded to avoid future issues. When the econometrics of the model are performed, for example, how is missing data coded, dummies coded zero or one and what other logical constraints are attached to the data configuration. The message is not to ask what technique to use but to summarize and understand the main features of the data set.

- **Fourth Step:** The first batch of corrections, blunders and erroneous transcriptions errors have been solved. A second round of spot checks is needed from the data file to the original sources. Do the data correspond to right country and right year? Another check is to use addition, percentage difference and ratios across columns that reflect, for example, levels of government or percentage of the total economy. Repeat the graphs and summary statistics and ranking of variables per country or identifier. Use other descriptive statistics (In *Stata*: sum, tab, quartiles). If there are further problems, go back to the notes from step one: change of currency, change in reporting units to try to identify errors

- **Fifth Step.** *Data collection from secondary sources*. Cut & Paste? Not so much any more. Here is where the skills and intuition of data analysts are put to the test and where the detective work starts. Add a column next to the one with the variable from the primary source, making sure that the data from primary and secondary sources overlap to allow for interpolation, system between points, and extrapolation, events of data outside the observed data set. The need for overlapping series exists because there is a high possibility that the definitions or the base will be different, therefore the series will need some anchors. After filling in with secondary and tertiary sources the

question still remains, if we should extrapolate or not. For panel data we should since too many observations will be lost to missing values. Do not make the assumption that to fill in the blanks we just need to put predicted values. If we do so, all the texture of time series would be lost. One of the most common techniques in time series data is the *tsfill* command – easy in *Stata*, which does the prediction based on the panel & temporal components. Additional techniques and sophisticated tools, including the program *Amelia II*, now also exist to assist in dealing with missing observations.[21] Amelia II can impute multiple missing data points in cross sections, time series, or in cross sectional-time series data using a bootstrapping based algorithm developed by King et al (2001) that generalizes trends in data over time. Tools like *Amelia* are often considered superior to other statistical imputation techniques (including mean substitution or *listwise* deletion), as it is based on filling data based on trends in the data set as a whole. These tools are a boon to social science research; however researchers should be cautioned to use them wisely and with care. Make sure that the use of tools does not place convenience above consideration of why observations are missing in the first place. Before "filling" the data, close consideration should be given to the context. In general, these are good enough solutions to transform variables that are reported on a bi-annual basis or measures like literacy rate that has proven to show slow and consistent change. It is usually not the best solution for economic growth or violence, which tend to show more abrupt changes. For example, in the Polity data set, a 99 value means missing data or transition year on the -10 to 10 autocracy democracy index. Often, when there are missing values, the country was undergoing a crisis and that is why the reporting ceased. So the advice is: check… do not assume!

The data collection is completed and all data is cleaned; the collection and measurement show repeatability, reproducibility, accuracy, and stability. Before we move to the econometrics of the model, we still need to ensure the validity and reliability of the variables in the data set.

a. A note on Expert based data

A detour is in order. This paper highlighted the importance of incorporating different units and levels of analysis in the modeling to better answer research questions; but the data section has mainly stressed how to deal with published data. The main reason to stress this type of data was a practical one: researchers

[21] Gary King, James Honaker, Anne Joseph, and Kenneth Scheve. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation, *American Political Science Review*, Vol. 95, No. 1 (March, 2001): Pp. 49-69.

assume that because the data is published, it has to be good and ready to go. Conversely, when the modeler uses expert data, it tends to be very tentative. The first issue is to justify why this data is "good enough" for the analysis and decide if the data is reliable and valid.

Three aspects of the process of gathering expert data need to be highlighted. The first one is the selection of experts. Different experts will offer different type of information: historical, contextual or descriptive, but in the end agent based models need a complete quantitative data set. Individuals can also suggest details that may lead to revising the continuum and the stakeholders considered influential on the issue at hand. Furthermore, having several experts on an issue when the analysis is done for the first time gives the researcher the possibility of evaluating and weeding out individuals that do not have the profile of good experts, e.g. individuals who cannot make the leap from describing a situation and translating it into a more structured framework, or individuals who cannot separate what it is, from what they wish it would be or individuals who feel that it is in their advantage to offer the "party line". Analysts learn to detect what type of expert is sitting in front of them and to extract any valuable information, even qualitative data which reflects either rankings or relative position and may clarify other questions. The second issue is interviewing skills. In addition, through practice, analysts will probably be increasingly aware of whether experts are providing internally consistent data and, moreover, will become increasingly adept at triangulating on one dataset composed from a number of different expert interviews and more targeted conversations. While some experts may indeed find the technique of assigning numerical values for position, influence, and salience quite easy, many will not, leaving the analyst responsible for translating insights from a conversation or interview into the hard data necessary for analysis. This is a skill learned with practice

The third most relevant issue is the data merge. The expert's data set may show reliability because the data may have internal consistency but that does not assure us that it is valid. So in the case of one, or many experts the issue of validity is the first one to attend to. To do so the expert must be tested and or evaluated by the analyst during an interview. This is a skill that needs to be developed. Being briefed on the situation and having background information on the situation gives extra tools to the analyst who should always be discounting experts that are trying to be strategic. Determining the number of experts needed to populate an issue is not that difficult. How many are enough and how many are too many? The best scenario to populate an issue is to be able to rely on one expert who has been proven to be knowledgeable on the issue and on the technique. A good expert can assess the relative position of the stakeholders, their influence and importance. The problem of selection bias of analysts is different from partial information. There are experts who are very good at

assessing a certain group of stakeholders' influence, position, etc. but not at assessing the whole landscape, and this individual may offer a deep and granular understanding of a good piece of the puzzle. However, it is appropriate to double check the reliability of the expert periodically by interviewing other experts and comparing their inputs.

In general, the analysis' success will place a lot of weight on the input from this one expert. Therefore, the quality of the objective overall knowledge must be a proven fact; the data must be measuring what it is supposed to measure (external and ecological validity). During the data collection interview, analysts must perform checks for internal validity (see internal consistency)[22]. For more than one expert, external validity as well as intra-class correlation needs to be performed for each expert. The scaling, that is the relative distance among different groups along the uni-dimensional issue is as important as the correlation. Although it is customary to remove the upper and lower values when the dispersion is above an accepted level, one must remember that the data may not conform to a normal distribution. It is important to avoid the exclusivity fallacy[23], of a data point considered an outlier which is not so. In the absence of a strong criteria to determine which data points should be excluded, we can rely on the central limit theorem, which states that regardless of the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution, and does it very quickly as N increases. The convergence is approached with a smaller N if we minimize the dispersion by extracting the max and the min numbers from the sample, and then calculating the arithmetic average or the mode, again according to the expert's judgment. Also to keep in mind is that the analyst may need to "slice" ("chunking") the data sets when experts can offer a high quality partial assessment and then interpolate that group of stakeholders in the main data set. In the data merge, the input from those experts may be given a higher relative weight.

As in the published data sets, the researcher doing the data collection needs to have the right skills and training and the final issue is to capitalize on the data collected. A data base architecture should be designed to construct time series per issue, country and expert. This data can be used later to complement models at other levels of analysis.

---

[22] What is gained in expedience, gets lost in having just one excellent expert is the richness of the contextual political economy environment and the nuances of the historical process that led to the present juncture

[23] Delphi validation may deprive the analyst from having flexibility at validity check points but it adds consistency and expediency to the data merge.

8. Reliability and Validity

Errors & Omissions…Recap

Issues of reliability and validity are another way of presenting the practical problems analyzed above. In short, they are problems derived from the correspondence gap between the measure and the concepts. Concepts, as said before, are constructs of the researchers' minds. One of the questions becomes: how can we test if what we measure is what was in the researcher's mind? A measure is valid if it measures what it was conceptually set up to represent. A measure is reliable if the measurement is repeated many times and the results coincide. All valid measures are reliable but not all reliable measures are valid.

Reliability

A great source of lack of reliability in the social sciences is clerical error, another is change in definitions. In surveys unreliable measures are often linked to the wording of the question that is open to interpretations or vague or to the interviewers' mistakes. Reliability problems are often linked to random errors. Careful work and a small pilot study can help clear questions and maximize reliability. Reliability tends to assume that conceptually the variable is correct, and most problems are due to "technical difficulties".

Remember, asssuming that by measuring a large number of times the average will be close to the true value is an invalid solution.

Tests for reliability:

1. Re-test and correlate either by a different researcher or the same researcher at different times.
2. Split and correlate, this strategy is useful for composite measures. Randomly group sub-items, dividing the items that are part of the measure into two groups and correlate (ex: social status: education, income, consumption, house price etc). The summary measures should tend to be the same. If a sub-sample of the measures is very proprietary of one group, or time period, then it will be a source of unreliability for the measure.

Validity

Validity shows the equivalence between a measure and its concept. The errors in validity tend to be systematic, non random errors. It doesn't matter how many times we measure the variable, the same bias persists. Most of problems with validity are derived from translating the concept into a measure; however, it can also be due to the selection bias of the sample. An example is extrapolation from an invalid sample to the population, e.g., considering public opinion by extrapolating the results from a self selected sample of individuals who had sent

letters to the editor or customer service office. A valid measure needs to be free of random and non random errors. A reliable measure needs to be free of random errors. When looking at results, invalid measures will produce no correspondence between what we are measuring and the original concept, therefore the inferences will be erroneous.

Tests or checks for validity:

1. There is no easy way. This is where the "know-how" of the business is evident, having an eye or a feel for the data becomes a must. The simplest way is to think things through during the construction process. This is easier said than done….
2. Think ahead for problems that can occur in the relationship between concept and measure. Often problems occur when the results of the model don't coincide with the hypothesis.

Are we there yet? Unfortunately a good data set is always a work in progress. It is not unusual to keep finding some mistakes, especially as the researcher starts to run your models. When a sign looks suspicious, it necessary to rethink the theory and the operationalization, given that there are no remaining problems of reliability. Otherwise, the first step is to look at the data again. If this loop is to start again, the initial place should be the data specifications, mainly definitions. Econometric techniques have become more sophisticated and glamorous but a good modeling technique needs to focus first on getting better data, looking carefully at what is there, becoming more familiar with the input. There is no econometric technique that will make up for a lousy data set, nor is there an empirical model that will succeed without a reliable and valid data set. And, no, even if we are almost there, we are not there yet. Before the estimation stage starts, in almost all cases, the researcher needs to have a plan for updates of the data set and dissemination. Otherwise the data set becomes obsolete immediately.

9. Ready for estimation? (Praxis)

4th commandment: "Thou shalt inspect the data"
Corollary: Thou shalt place data cleanliness ahead of econometric godliness
The keeper of the truth[24]

We are not there yet. Modelers and econometricians will focus on the specifications of the model and the type of data to make a decision about the appropriate estimation technique. Qualitative probit/logit, limited dependent tobit estimation, time series, unit roots cointegration., count data poissone…use

[24] Kennedy, Peter .2006. A Guide to Econometrics. MIT Press: Boston P. 396

also structural break testing and estimations, dummies. Books and books, article after article have been written about estimation techniques, although Kennedy[25] quoting from Intriligator, Bodkin & Hsiao (1996 p. xiv) restates that "practising econometricians typically spend 20% or less of their time and effort on econometric techniques. Where is the 80% of the time?

10. Conclusions

"It is time to finish. The considered answer to the central question…is already evident. No need for a lengthy summary, only some principal observations are in order." (Organski, 1990. p. 202.)

In this paper, I accounted for much of that time in the preparation of the model and appropriate data before the estimation. After testing the econometric model, time is devoted to interpretation of results after estimation. A second iteration of modeling issues may come up again once the researcher looks at the output. There may be unexpected signs in the coefficients, levels of statistical significance of the variables, or the magnitude of the coefficients that make the results consistent with theory anomalies. It is time to explain and rethink where the problems originated and to make a diagnosis: theory, operationalization, data?

Given increasing computational power, we stand at the beginning of a revolution in quantitative computational social science. The theoretical postulates and models in political economy or any other human, social, cultural or behavioral modeling demand the scrutiny of empirical testing, just as any other scientific endeavor. A lot more research needs to be allotted to the development of indicators and measures. And a lot more time has to be allocated to data collection and data base maintenance, in addition to dissemination. This is particularly evident in the lack of value placed on data collections, with little to no emphasis attached to maintaining time series data bases that relate to variables beyond the economic realm, threatening the development of the field through a lack of consistent and systematic analysis.

The nature of human, social and cultural behavioral phenomena is inherently stochastic, and as such, present distinct challenges identified above. However, following the 'dirty little secrets' for data modeling gives us a guide to increase our chances of modeling success. Hierarchical research allows us to do the modeling at different levels of analysis and at different levels of research. Using the right theory and model at each level enables us to differentiate necessary from sufficient conditions for prediction or, for instance, go from a systemic situation at the international level and drill down into the micro motivations for

[25] Kennedy, Peter. 2006. A Guide to Econometrics. MIT Press: Boston pp.405

such behavior. If we cannot, or worse yet, do not, increase the reliability and validity of our data, then the very theoretical foundations of our models are flawed at best. . It is my hope that these glaring data related deficiencies will soon be remedied.

APPENDIX A
Data Archives, Updates and Dissemination

Social scientists use and produce massive amounts of information in different formats from census, opinion polls, surveys and experts, which can be qualitative or quantitative. These data get collected with different periodicity, sometimes on a monthly or annual basis, time series or with no continuity. This approach offers little value beyond that project's specific use. In other cases, data sets that had been collected for many years stop being collected. Often the data gets buried and forgotten, and a massive effort gets lost. Even more common data collection efforts constitute a one time endeavor and as they stop they are not suitable to make inferences beyond that point.

The situation is suboptimal at the moment. One of the many challenges is to create a good data base for assessment, with capabilities to compare pooled time series, capacity building for data updates, and an effective dissemination process for researchers. The data is collected by research institution, government statistical offices, NGOs, and special project teams and then archived without any plan for dissemination. Social scientists, journalists, marketing experts, all of them miss a great opportunity to advance knowledge. Part of the problem is that data has become a public good. Everybody wants to use it and nobody wants to pay for it, or pay to maintain archives or update it. Some issues with dissemination are technical and administrative and sometimes the separation between the metadata and the statistical data. The reasons are many; but, the question is then how do we deal with the issue of quality and dissemination in a world where the Internet has "raised the bar" in terms of quality and reduced the cost? [26] [27]

On the positive side, the IMF put forward an initiative in 1995 to help developing and emerging market countries develop and disseminate their macroeconomic and socioeconomic data (GDDS: General Data Dissemination System). At the moment 81% of the membership participates. The emphasis has been on development and the progress has been slow. IMF is changing the focus to dissemination, which creates its own demand for better statistics and more timely published information. This strategy has also raised the profile and visibility of the statistical agencies.[28]

---

[26] A few commercial undertakings since 1998, like NESSTAR (Network Social Science Tools and Resources) and later on FASTER have tried to focus on this problem but their results are not clear yet
[27] Assini, Pasqualino. NESTAR. A Semantic Web Application for Statistical Data and Metadata. www.nesstar.com/export/sites/default/doc/technical_overview.pdf
[28] IMF. 2008. "Assessing The GDDS- What has been Accomplished after Ten years and Where do we go from Here?" Statistics Department. Working Paper

APPENDIX B
Data Sources

Some potential sources are listed below but keep in mind that the sources keep changing leaving the researcher with incomplete and outdated information. In all, the number of possible data sources is immense. Most data sets encounter grave problems of quality, periodicity and/or timeliness.

Conflict Data: Inter and Intra State
Correlates of War Project: 1816 -2006 State System membership, militarized disputes, formal alliances, national material capabilities. Not updated. Some data sets are historical, for example: Militarized inter and intra state disputes 1816 -2006, Includes national contiguity 1816 -2006. Available at: http://www.correlatesofwar.org/
Conflict Data: Intra State
UCDP/PRIO Armed Conflict v.4- 2007, 1946 – 2006 Conflict-year data within countries where at least one party is the government. Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm
UCDP Armed Conflict Dyadic v.4-2006, 1989 – 2005 Dyadic version of the UCDP/PRIO armed conflict data. Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm
UCDP Conflict Termination dataset v.2.0, 1946 – 2006 Conflict level and conflict year data. Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm
UCDP Peace Agreement Dataset v. 1.0, 1989-2005 Data set that includes peace agreements signed by at least two warring parties. Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm
UCDP Battle-Deaths Dataset v.4.1, 2002-2005 Conflict level and conflict year data on number of deaths in internal conflicts. Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm

UCDP Non-State Conflict Dataset v.1.1, 2002 – 2005 Data set on internal conflicts where neither warring party is the government Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm
UCDP One-Sided Violence Dataset v.1.2 1989 - 2005 Intentional attacks on civilians by the government or third party. Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm
Managing Intrastate Low-intensity Conflict (MILC) v. 1.0 1993-2004 Third party actions in Interstate conflict. Available at: http://www.pcr.uu.se/research/UCDP/data_and_publications/datasets.htm

Politics & Governance Data
Transparency International Corruption Perception Index 1995 – 2007 Transparency across revenue, national accounts, perceived corruption. Available at: http://www.transparency.org/policy_research/surveys_indices/cpi
Bank's Cross National Time Series Data Archive 1815-2006 Demographic, Economic, Development, Stability, Regime Duration data. Available at: http://databanksinternational.com/
Freedom House 1972-2007 Civil, Political and Economic freedom. Available at: http://www.freedomhouse.org/template.cfm?page=15
Human Development Reports 2005-2008 Basic needs and quality of life data (UNDP). Available at: http://hdr.undp.org/en/statistics/
World Values Survey, 1981 – 2004 Surveys conducted in waves, individual perceptions of regime satisfaction, economic satisfaction, well being, general values Available at: http://www.worldvaluessurvey.org/
The Urban Governance Initiative Report Cards World Bank – perception of institutional reliability, corruption, urban & rural development Available at: http://web.worldbank.org/WBSITE/EXTERNAL/WBI/EXTWBIGOVANTCOR/0,,c

ontentMDK :20746471~pagePK:64168445~piPK:64168309~theSitePK:1740530,00.html
World Freedom Atlas 1990-2006 The geo-world atlas is a geo-visualization tool for world statistics Available at: http://freedom.indiemaps.com/

Economic and Socio Demographic Data at the National Level
Central Bank Websites URL for Central Banks of most countries in the world Available at: http://www.bis.org/cbanks.htm
Earth Trends A compilation of a variety of social. Economic and environmental data http://earthtrends.wri.org/searchable_db/index.php?step=countries&cID[]=6&theme=5&variab le_ID=353&action=select_years
Government Financial Statistics Available at: www.imf.org/external/pubs/ft/gfs/manual/gfs.htm
International Financial Statistics Available at: http://www.imfstatistics.org/imf/logon.aspx
International National Statistics 1950-1993 Available at: "International National Statistics, UK, MacMillan Reference, LTD. 1998 [Europe; Africa, Asia & Oceana; The Americas, 1950-1993"
Penn World Tables Provides purchasing power parity and national income accounts converted to international prices for 188 countries for some or all of the years 1950-2004. Available at: www.imf.org/external/pubs/ft/gfs/manual/gfs.htm
World Development Indicators Available at: www.**world**bank.org/data/wdi2004/
UN Statistical Data bases, Yearbooks & Reports Available at: http://unstats.un.org/unsd/databases.htm
Region or Country Specific
Europa World Yearbook

Available at: www.**europaworld**.com/ "The Europa World Yearbook, London: England, Europa Publications Limited,"
Africa South of the Sahara Available at: http://www-sul.stanford.edu/depts/ssrg/africa/ www.unesco.org/unesdi/index.php/eng/repertoire/tous "Africa South of the Sahara, Europa Publication (London), 1971-2000"
"African Statistical Yearbook "[ALSO African National Statistics, ASN=ASY]" https://unp.un.org/details.aspx?pid=15447
The Far East and Australasia 1970-2000 Reference Library Book "The Far East and Australasia, London, Europa Publications, 1970-2000"
The Middle East and North Africa 1948-59, 1974-2000 Reference Library Book "The Middle East and North Africa, London, Europa Publications, 1974-2000"
Statistical Yearbook for Asia and the Far East1968-1969 Reference Library Book "Statistical Yearbook for Asia and the Far East, Bangkok, Thailand, 1968-1969, United Nations"
"Statistical Yearbook for Asia and the Pacific, United Nations, Bangkok, Thailand, Economic and Social Commission for Asia and the Pacific, 1974-84, 1986-90, 1992-1999" Reference Library Book
"Statistical Yearbook for Latin America, Santiago, Chile, United Nations, Economic Commission for Latin America, 1975-84" Reference Library Book
"Statistical Yearbook for Latin America and The Caribbean, Santiago, Chile, Economic Commission for Latin America and the Caribbean, 1986-1999" Reference Library Book
"The Statistical Abstract of Latin America CEPAL" Reference Library Book
"International Historical Statistics" Mitchell Reference Library Books

Example of notes on a data set before starting upgrades/ updates

Time Span 1960-2003
Sample: 118 countries. Sample includes Poland 86-on, Hungary 89-on, Romania 88 on. These countries use NMP in their National Accounts before those dates so the data is not comparable to GNP/GDP methodology). The sample includes China.

The main source until the 1970s was the World Bank with different methodology from the IMF. In 1975 the World Bank handed the collection to the IMF and although the definition of the accounts remained unchanged, the data was not always consistent. The IMF reports the fiscal data disaggregated by central, state and local revenues, while the WB reported General revenues. The quality of the IMF series was improved. Also the IMF has relied more on country reports, which has made the African data more spotty. We have used IMF as the main source and adjusted back using the rates of change from WB (WDI). The data from OECD, although more complete does not always coincide with IMF, the same criteria have been applied. The alternative sources to fill in blanks has been Europa Yearbook and Central Bank reports.

For missing points, after going through the previous sources, we used regression analysis and adjusted for specific policy changes and other intervening factors that may have impacted the variables.

Conceptually we need to rethink social security. The argument to exclude social security has been that those transfers have been allocated previous to the collection and therefore there is little room for any further manipulation by the government. The issue of deficit as a variable that can be manipulated for short term political gain needs to be taken into account.

DATA DESCRIPTION

FILE : XXJULY14. DTA

<u>VARIABLES:</u>
code
country : name
year : 1960-2004
totrev = total revenue divided by GDP (each of the components of the ratio is measured in current national currency)
nontax = non tax revenue divided by GDP (each of the components of the ratio is measured in current national currency)

socsec = social security revenues divided by GDP (each of the components of the ratio is measured in current national currency)
agri = agriculture GDP divided by Total GDP
exp = exports divided by Total GDP
mining = mining production divided by Total GDP
realgdp = Real GDP per capita in constant dollars from PWT 2000 plus updates
taxratio = totrev – nontax - socsec

SOURCES for each variable

Totrev = Main Source: World Bank 1960 – 1972 and Government Financial Statistics 1970 – 2003. Overlapping structures were based on the IMF (Table 1 Line 1, 11); Statistical Annex/Country Reports, International Monetary Fund, Washington DC, available at http://www.imf.org/external. When data was lacking we used National sources such as Central Bank Reports

nontax = Main Source: World Bank 1960 – 1972 and Government Financial Statistics 1970 – 2003. Overlapping structures were based on the IMF (Table 1 Line 1, 11; Statistical Annex/Country Reports, International Monetary Fund, Washington DC, available at http://www.imf.org/external. When data was lacking we used National sources such as Central Bank Reports

socsec = Main Source: Government Finance Statistics Year book 1970 - 2003, International Monetary Fund, Washington DC, Table 1, line 121. Overlapping structures were based on Statistical Annex/Country Reports, International Monetary Fund, Washington DC, available at http://www.imf.org/external. When data was lacking we used National sources such as Central Bank Reports.

agri = Main Source: International Financial Statistics 1960 – 2003; Overlapping structures were based on Indonesia: Statistical Annex, International Monetary Fund Staff Country Report No. 00/133. Table 2 line 31. October 2000; Statistical Annex, International Monetary Fund Staff Country Report: 03/388. Table 2 Line 1. December 2003; The Europa World Yearbook, London: England, Europa Publication Limited, page 2081-83 (2003 Volume) http://www.imf.org/external. When data was lacking we used National sources such as Central Bank Reports.

exp = Main Source: International Financial Statistics 1960 – 2003; Overlapping structures were based on Indonesia: Statistical Annex, International Monetary Fund Staff Country Report No. 00/133. Table 4 line 27. October 2000; Statistical Annex, International Monetary Fund Staff Country Report: 03/388. Table 2 Line 2. December 2003. The Europa World Yearbook, London: England, Europa

Publication Limited, table on Export, page 2081-83 (2003 Volume) for Indonesia and page 4039-4043 (2003 volume) International Finance data.

mining = Main Source: International Financial Statistics 1960 – 2003; Overlapping structures were based on Indonesia: Statistical Annex, International Monetary Fund Staff Country Report No. 00/133. Table 2 line 31. October 2000; Statistical Annex, International Monetary Fund Staff Country Report: 03/388. Table 2 Line 1. December 2003. The Europa World Yearbok, London: England, Europa Publication Limited, table on Mining, page 2081-83 (2003 Volume) for Indonesia and page 4039-4043 (2003 volume). International Finance data.

realgdp = Main Sources: Penn World Tables 2000 (Summers and Hestons) and World Development Indicators, World Bank 1960 – 2004. When data was lacking we used National sources such as Central Bank Reports.