Harassing Deficiencies in Data:
What the Soviet and Russian experience can teach us about extracting information from authoritarian regimes

Yoshiko M. Herrera*

April 15, 2018

For presentation at the "Regime Evolution, Institutional Change, and Social Transformation in Russia: Lessons for Political Science"
Yale University, April 27-28, 2018

Abstract: This paper considers the challenges to getting political and economic information out of authoritarian regimes, with a focus on lessons from the experience of extensive data extraction during the Soviet period. The paper outlines some of the ways that past experience with Soviet data extraction can be useful to researchers interested in using modern content analysis and data extraction techniques in authoritarian or closed settings.

*University of Wisconsin-Madison. Email: yherrera@wisc.edu
1 Introduction

This paper considers the challenges to getting political and economic information out of authoritarian regimes, with a focus on lessons from the experience of extensive data extraction during the Soviet period. The USSR was a closed society of enormous importance to the West and hence despite significant challenges, Western governments devoted enormous resources into figuring out what was happening in Soviet Union especially regarding the economy and military capacity.\footnote{For an overview of economic Sovietology, see Schroeder 1995.} Although the USSR is no longer, and much has changed in information technology, I argue that not all the lessons of information extraction from the Soviet experience should be relegated to the dustbin of history. In the contemporary era, with Russia, China, as well as other even more closed environments like Iran or North Korea, there is a need to get political and economic data in places where fieldwork is not possible and where governments are actively attempting to control what information is available.

Moreover, the technological and methodological advances in content analysis related to “big data,” and the proclivity of many regimes, even closed ones, to document in digitized text opens up avenues to information extraction in ways that either were not possible or required enormous time commitments in the past. The goal of this paper is to outline some of the ways that past experience with Soviet data extraction can be useful to researchers interested in using modern content analysis and data extraction techniques in authoritarian or closed settings.

2 Key lessons from the Soviet experience

The critical starting point for understanding data in Soviet regime is that the regime did not just make up numbers in an \textit{ad hoc} or random way. This fact separates the USSR from simply a low-data environment, or an environment where individuals unsystematically lie. First, there were legitimate ideological differences in approaches to economic data that were consequential, if not necessarily deceptive in intent. But in addition, the Soviets systematically collected and processed data, while simultaneously...
attempting to highlight successes and hide data that would either reflect badly on the regime or reveal information considered to be of strategic importance. Importantly, they did this through a range of complex, rule-bound practices, which, as it turned out, could be discovered with careful analysis. They also limited access to data, which was easier in the pre-digital age, by placing restrictions on the amount of data shared with foreign governments or scholars. The existence of these norms and practices meant that the task was to figure out what Soviet statisticians were doing and why. If that could be done, one could learn a lot about the USSR.

This was exactly the approach arrived at by Abram Bergson, a prominent Harvard economist and pioneer in the study of Soviet statistics. He wrote, “While there are many harassing deficiencies, it seems clear that the Soviet government does not falsify those statistics which it elects to publish. If it were not for this fact, of course, research on the Soviet economy today would be entirely out of the question” (Bergson 1947, 234-35). Similarly, in explaining the “numerous and puzzling inconsistencies,” in Soviet statistics Vladimir Treml concluded, “as a general case, I believe that these puzzles are the product of archaic and often conflicting methodologies and conventions, faulty and biased price and output indexes and formulae, and poorly designed classifications and not of outright ad hoc manipulation and deliberate falsification of data” (Treml 1989, 96).

Given the space constraints of this paper, I focus on one structural factor and four norms among Soviet statisticians working in national accounts that can help us to interpret and understand their work and the data they produced.2 The structural factor concerns the location of the national statistics office (NSO) in the state bureaucracy and the incentives associated with that location. The four norms concern: Marxist ideology; rejection of probability theory; methods for highlighting the positive; and linkages to international statisticians.

---

2 This analysis is largely adapted from material found in (Herrera 2010). The main focus of that work was not the challenge of information extraction, but rather explaining the puzzle of the Soviet/Russian transition to the international System of National Accounts in the 1990s.
2.1 Structure

In the USSR and other centrally planned economies, NSOs were not supposed to be independent of the government, but instead were the state’s eyes and ears for monitoring five-year economic plans. Indeed, Lenin famously remarked, “Socialism is accounting!” and the Central Statistical Administration of the USSR was located within Gosplan, the central economic planning apparatus, for much of its history.

That structural location had implications for the way statistics were organized: it meant that the NSO would be primarily interested in measurement of plan targets rather than other things. And, because firm-level data was used to gauge plan fulfillment, there were well-known incentives for firms to overestimate production to meet plan targets (Treml 1994). Although NSOs were aware of this type of falsification and tried to stop it, it was nevertheless a built-in facet of economic data in the Soviet Union.

2.2 Norms

There are a lot of norms and practices that govern data collection, analysis, and accessibility in NSOs, but here I focus on four important norms that guided the Soviet experience with statistics.

Marxist ideology
Marxism guided soviet statisticians in their work (Guzhvin 1992a). This ideological orientation was evident in publications, which explicitly used Marxist-Leninist terminology. That created a translation issue with Western economics, but was not the primary problem. Rather, the most profound impact of Marxism had to do with the production boundary for national accounts. Gross Domestic Product (GDP) is supposed to measure the total value of goods and services produced in a country in a given year, and is currently based on the System of National Accounts, a highly institutionalized international framework (Herrera 2013). The “production boundary” in GDP specifies what kinds of economic activity are to be included (e.g. monetized transactions, not unpaid household labor, etc.), but the Soviet bloc NSOs, used a different production boundary from the West, which was restricted to “material production,” which was
largely based on Marx’s labor theory of value.\textsuperscript{3} The exclusion of services in Soviet national accounts did not mean that there were no services in the USSR, just that their seemingly equivalent concept to GDP did not include a major part of economic activity, making it incongruent with Western measures. To compare GDP then across countries, the production boundaries had to be aligned, but this one only the start of the challenge.

*Rejection of probability theory*

Until the mid-20th century national accounting relied on counting output from various specified industries. Following important developments in mathematics, statistics and national accounting by John Maynard Keynes, Richard Stone, and others, the UN statistical agency embraced probability theory and the use of sampling for national accounts. The Soviets, on the other hand, from their earliest days clung to the idea of full enumeration of output from all firms (engaged in material production), and they remained deeply skeptical of sampling methodology until the late-1990s (Herrera 2010).

The incredible irony, however, was that an important, perhaps crucial, figure in the development of modern probability theory was Andrei Kolmogorov, a Soviet mathematician. He wrote *Foundations of the Theory of Probability*, which was first published in German in 1933, and remains in print in English. Although the USSR’s contributions to probability theory placed it at the forefront of statistical science, it rejected the practice of survey sampling in favor of full enumeration in its national statistics.

What did this mean for Soviet data? It meant that data would remain dependent on access to data from all firms, because sample surveys and estimation techniques used in the West were rejected. In an authoritarian context with a relatively small number of firms this was possible, but became vastly more difficult after the end of the USSR. And one can see how this rejection of sampling alone would work against the expansion of the production boundary to include services because of the complexity in getting data from all such firms. Hence the norm against the use of probability theory suggests insight into the methods used by Soviet NSOs to gather data from firms, and we also can connect that

\textsuperscript{3} The focus on material production was also congruent with Adam Smith’s conception as well all early models of national accounts (Árvay 1994).
with the location of NSOs in Gosplan to understand more about firm incentives for sharing data with NSOs.

Accentuate the positive
There is no question that Soviet statisticians manipulated methodology to present data in ways that portrayed the Soviet state in the most positive light (Kostinsky and Belkindas 1990; Belkin 1992). But the challenge was to figure out what exactly they were doing. Here I focus on four mechanisms for manipulation: biased price indices, descriptive distortions, silence regarding changes, and lack of access.

Given the general lack of market prices in the USSR, estimating the value of production depended on so called “comparable price indices.” However, official Soviet price lists for both retail and wholesale trade ignored unofficial transactions and neither reflected product improvements nor deterioration. This means the price lists ignored the concept of inflation completely, which was deeply problematic given that with inflation prices may rise without a corresponding rise in output. For decades, the recalculation of these price indices was a central concern of Western economists trying to understand the value of Soviet production over time (Noren 1994; Blades and Harrison 1992).

Gregory Grossman (1960) coined the term “descriptive distortions,” for when indicators are said to mean one thing, but actually mean something else. There are myriad examples, but a few will suffice here to illustrate the problem: Treml (2001) noted that the number of “doctors” might include dentists; “grain weight” might include dirt, water, and other matter not initially separated during the harvest; “meat” might include lard and other inedible animal parts, and so on. A higher-profile example of this obscurantism concerns the category of “defense expenditures,” which turned out to mean only pay and current material expenses, not research, construction, or weapons purchases, which together are estimated to have represented 70 percent of the actual defense budget (Treml 2001, 10). There could have been legitimate reasons for terminological disagreements, but the lack of published methodological descriptions of indicators, suggests these were not unintentional misunderstandings.
Related to the issue of descriptive distortions was consistent silence from Soviet NSOs regarding changes to indicators or methodology over time. One example illustrates the wide reaching effects of some of these changes: in 1985 the Soviet NSO decided to remove alcoholic beverage production from industrial output; this move followed a steep drop in alcohol production due to Gorbachev’s anti-alcohol campaign, and the removal had the effect of making the drop in industrial production appear smaller (Kostinsky and Belkindas 1990, 191). But, because the turnover tax on alcohol was an important source of income for the Soviet state, the forced decrease in supply of alcohol led to lower levels of revenue for the state. To counter this, the NSO then manipulated various formulas (especially the weight of alcohol in total retail sales and past year retail trade figures) to inflate current retail trade figures (Vanous 1987, 2-3). The inflated retail trade figures then affected national income statistics via inflated consumption figures. The change in methodology concerning alcohol and industrial output was not announced; it just appeared in the data, and past data were not updated to reflect the change.

In addition to changes in current formulae, changes were made retroactively to published data without explanation. For example, in January 1992 the Russian NSO published a net 2.2 percent decline in industrial production for 1991, but in 1993 they published a net 8 percent decline for 1991 without any commentary on the discrepancy between the 1992 and 1993 publications (Kudrov 1993, 128). This problem continued for SNA data in the 1990s (Tabata 1996) and in many other areas.

Finally, inaccessibility was a prevailing feature of Soviet statistics. The government released a fraction of what was available in Western countries. They limited the scope of data collected, published aggregates rather than detailed data sets or underlying components, limited the number of copies of works published, and then limited access to those publications (Schroeder 1995, 200). It almost goes without saying that some politically sensitive activities, such as those related to the military-industrial complex, were in practice excluded or only partially included as noted above.

---

4 For more on data restrictions, see Heleniak and Motivans 1991; Kudrov 1993.
It is worth noting that all of these issues with Soviet data manipulation were discovered through painstaking collection, comparison, and analysis of published paper documents, largely in an era before powerful personal computers. In the current era, it would be much easier to compare and contrast digitized statistical handbooks to look for anomalies. Similarly thousands of computer simulations could be done to figure out different formulae for calculation of aggregates. Nevertheless, it helps to understand the overall context to be able to know what to look for.

*Professional identities*
So far I have discussed ways in which understanding the structure and norms of the Soviet NSO reveals the source of biases or omissions in Soviet data. But another factor that has had a more positive impact on Soviet data was the professional identities of Soviet statisticians. Even during the Cold War, many Soviet statisticians interacted with their Western counterparts. For example, Mikhail Korolev began his career as an academic statistician in the 1960s at the Moscow Institute of Economics and Statistics. From 1972 until 1989 he worked at the Soviet NSO, eventually becoming director, and following that he headed the NSO for the CIS. Korolev was as closely linked to the hierarchy of Soviet statistics as any person could be, but at the same time, he was also a member of the UN Economic Commission for Europe’s Conference of European Statisticians, as well as a member of the UN Statistical Commission, to which he was elected chairman in 1979.

Moreover, it is through these international linkages that consequential international comparison projects emerged. The first formal UN comparison project began in the 1970s, with the idea that both sides could learn from the experience, and the UN published its methodology for comparing the SNA and the Soviet Material Product System in 1977. Work on exchange rates and concepts such as purchasing power parity also developed out of efforts to compare Soviet and Western national accounts. Indeed many of the questions and problems outlined above could be addressed via discussion with colleagues from different countries. And it was the sustained international interaction and comparison projects that eventually led to the development of a single, hegemonic system for national accounts around the world in the 1990s (Herrera 2010).
3 Implications for the current era

The Soviet Union was at the forefront of state data collection; they kept records on all manner of things. While that was in the pre-digital age, currently we are in an unprecedented time of data collection in which the state is not the only actor (and maybe not even primary actor), but in which states nevertheless participate more than ever. And, state data are not just written records, published in books that require espionage to acquire: they are available online, in a variety of digitized formats, even in closed or authoritarian regimes. For example, Russian census data from 2002 and 2010 is mostly online. Similarly at the international level, the System of National accounts was amended in the 2000s and virtually all meetings, documents, and decisions were posted on the SNA website.

Our challenge now, however is to wade through the sea of data. How do we make sense of all the digitized data we can find about today’s Russia, or that is even available on Russian state websites? Though it may be more limited, we can ask the same about North Korea or Iran. Robotic topic-modeling is one approach where we just see what comes up in terms of frequency, but the discussion above suggests a more targeted approach.

- First, for a state agency, we should ask, where is it located? What are its goals? How does that structural position affect what kind of data it collects, the way it collects the data, and its publications?
- Second what are the organizational norms in a data collection agency?
  - What data are considered relevant to the agency; what do we think would be included or excluded?
  - What methods does the agency use to collect data
  - What are methodological tools (or tricks depending on your perspective) that the agency might employ to present data in a more positive light?
  - What can we learn from comparing versions of agency documents over time?
  - Are there ways to interact with those individuals who collect and analyze data that could provide insight into their practices?
Whereas in the USSR foreigners were not even allowed in NSO buildings, and CIA staff had to pour over published books by hand, trying to figure things out, in the 1990s statisticians from the World Bank literally set up offices in the NSO buildings and would make phone calls to relevant statisticians to ask questions. In many countries the boundaries today, even in authoritarian countries, are much more permeable. Professional organizations have members from a variety of countries who can discuss their experiences and ways of working with data. The regular international SNA conferences, for example, are highly documented events with members from all over the world regularly addressing data challenges. In addition we can use computers to analyze thousands of pages of digitized data instantly, once we know what we are looking for.

By understanding the structure of a country's NSO or other state agencies in a country, as well as and its associated norms and practices we can gain insight into the way in which data is collected and constructed for publication. In this way many of the tools used by Sovietologists could remain relevant and useful in an increasingly digitized, data-rich world.
References


