Politics of Data in Authoritarian Regimes

Ruth Carlitz and Rachael McLellan

August 16, 2019

Abstract

Data availability has long been a challenge for scholars of authoritarian politics. However, the promotion of open government data – through voluntary initiatives such as the Open Government Partnership, and soft conditionalities tied to foreign aid – has motivated many of the world's more closed regimes to produce and publish fine-grained data on public goods provision, taxation, and more. While this has been a boon to scholars of autocracies, we argue that the politics of data production and dissemination in these countries creates new challenges. Systematically missing or biased data and selective restrictions on data collection may jeopardize research integrity and lead to false inferences. We provide evidence of such risks from Tanzania, comparing data released to the public on local tax revenues with verified internal figures. We find that the latter significantly underestimate opposition performance. This leads to flawed inferences about the determinants of local government performance. We also present an overview of methods for identifying manipulation. In so doing, we demonstrate that caution must be exercised when conducting research using such data. We conclude by proposing ways that scholars can minimize these risks.

1 Introduction

Political science – and the social sciences more generally – has been undergoing a dramatic methodological transformation. The rise of "big data," and computational innovations associated with it ("data science") have been cheered by many prominent scholars. They argue that the "big data revolution" has the potential to improve causal inference by facilitating the design of better experiments and more precise comparisons (Clark and Golder, 2015; Titiunik, 2015; Monroe et al., 2015). Some even hold that big data could help to solve "some of the most important, but previ-

ously intractable, problems that affect human societies"(King, 2014: 166). At the same time, the availability of new and "bigger" data raises new questions and concerns. As we will demonstrate in this piece, scholars studying authoritarian regimes should exercise particular caution.

Until recently, the production and dissemination of fine-grained data on government activities had been the exclusive provenance of developed democracies. However, a growing number of poorer and more closed states have been releasing an increasing amount of information to the public. Such increased openness reflects two parallel trends that serve to reinforce each other. First, the Millennium Development Goals, and their successors the Sustainable Development Goals have oriented the international development community toward the achievement of clearly defined targets, motivating a raft of statistical exercises to improve the tracking of economic and social indicators (Kelley and Simmons, 2019; Jerven and Johnston, 2015; Sandefur and Glassman, 2015). These efforts include a push for national governments to disseminate information under open data protocols. At the same time, the promotion of open government data has pervaded the policy agendas of governments around the world (Davies and Bawa, 2012). For example, the voluntary Open Government Partnership, launched in 2011, counts 79 countries and 20 subnational governments among its members, who together have generated over 3,100 commitments to make their governments more open and accountable.[1] Star performers include not only the usual suspects in Western Europe and North America but also poorer, less democratic countries such as Kenya, Honduras, and Moldova. A more systematic recent analysis finds that democracy no longer provides an advantage with respect to e-government performance (Stier, 2015).

The availability of fine-grained data seems at first to be a boon to scholars of authoritarian politics, making it possible to answer a number of previously intractable questions. Indeed, the ostensible quality and granularity of this data has meant that a growing number of studies on the political economy of development draw conclusions from non-democratic contexts. However, we argue that scholars should treat these newly available data with caution. To begin with, concerns about data quality are pervasive in the developing world (Jerven, 2013; Devarajan, 2013). Sandefur and Glassman (2015) provide compelling evidence that official statistics systematically exaggerate development progress across multiple African countries, reflecting two interlinked principal-agent problems. First, governments misreport to foreign donors, particularly in the context of results-

[1]For more information, see `https://www.opengovpartnership.org/about/about-ogp`

based aid. Second, national governments are themselves frequently misled by frontline service providers tasked with simultaneously providing public services and reporting truthful data on the same. Much of the literature on "Africa's statistical tragedy" (Devarajan, 2013) has focused on such bureaucratic or administrative factors to explain why official statistics may be inaccurate or incomplete. The influence of domestic politics on the production and dissemination of such data is often overlooked.

The nature of electoral competition is particularly important when it comes to the politics of data production and dissemination. On average, democracies tend to be more transparent than authoritarian regimes (Hollyer, Rosendorff and Vreeland, 2011), reflecting that fact that transparency can be dangerous for autocrats. Making information freely available allows citizens to not only update their beliefs about government performance, but also their beliefs about what other citizens believe (Hollyer, Rosendorff and Vreeland, 2015). Journalists, politicians and civil society can leverage open data to criticize the regime and encourage mobilization against it which can foment dissatisfaction with the regime (Reuter and Gandhi, 2011).

However, autocrats do release data. E-government is an increasingly popular form of legitimation – both for external audiences as well as the citizens of authoritarian regimes (Maerz, 2016). An emerging literature suggests that authoritarian regimes may have incentives to allow or even promote certain forms of transparency. Permitting investigative reporting can help central government officials keep lower-level officials in check and reduce local corruption (Lorentzen, 2014; Egorov, Guriev and Sonin, 2009). Berliner (2014) argues that the passage of freedom of information laws allows incumbents (even in autocracies) to ensure that they will not be shut out of access to government information and tools of monitoring if they lose power in the future. Subnational analysis from Mexico confirms that incumbents are particularly likely to pass such reforms when their grasp on power is less secure (Berliner and Erlich, 2015). Transparency can also increase elite cohesion as a strategic response to greater threat from the masses (Hollyer, Rosendorff and Vreeland, 2018). As we will show, the release of statistics can also be used by incumbents to try and discredit their opponents.

This push for ostensible transparency may not produce real transparency. Autocrats can release data that has been manipulated. Authoritarian regimes tend to lack institutions such as a free press, a parliament with meaningful oversight powers, or apolitical bureaucracy, that can hold gov-

ernments to account over fraudulent or missing data. Given the power of information, governments may seek to 'massage' the data they release, altering it in politically expedient ways. Data becomes political because information is political. In this reflection, we conceptualize how scholars should think about data production in authoritarian contexts and propose questions scholars should ask to gauge how and to what extent data is vulnerable to manipulation.

In what follows, we reflect on how data comes to be released in authoritarian regimes to allow scholars to better understand what threats to inference they may face. We provide evidence of these risks from Tanzania, comparing data released to the public with verified internal figures. We put the Tanzanian example in broader context to show how publicly available data on various topics can be manipulated for political reasons – particularly in nondemocratic regimes. In so doing, we demonstrate that caution must be exercised when conducting research using such data. We conclude by proposing ways that scholars can minimize the risks associated with analyzing data from nondemocratic contexts.

2 Understanding Risks

Administrative data is of value to political scientists because it allows us to test our theories about governments and the decisions they make. Given this data is available, how should scholars approach it? We contend that scholars wishing to take advantage of newly available data in non-democratic regimes should reflect on the politics of data production and dissemination in their countries of study. Even researchers interested in questions that seem far removed from the landscape of political competition and control should take heed since that very landscape may determine the quality of the public data as well as the risks associated with collecting additional information. The salience of the issue, the timing of the data release in the electoral cycle, and the competitiveness of a particular area can all affect the availability and reliability of the data the government chooses to release. These same factors may also increase the risks scholars and their local data collectors face when doing their own data collection.

As a general rule, researchers should understand when data is at most risk of being "cooked" and the types of flawed inference that falsified data may create. Wallace (2016) argues that data is most likely to be manipulated if it is politically sensitive and is released at politically sensitive

times. We concur and argue that scholars should be mindful of the whole process leading up to the data's release. Data has to be commissioned, produced and released. Data may be more or less reliable depending on political incentives at each stage from initialization to dissemination. We urge scholars to think about the process of data production. In Figure 1, we propose a set of questions scholars should ask themselves when trying to understand the risks of manipulation data may face. The proceeding discussion expands upon these questions.

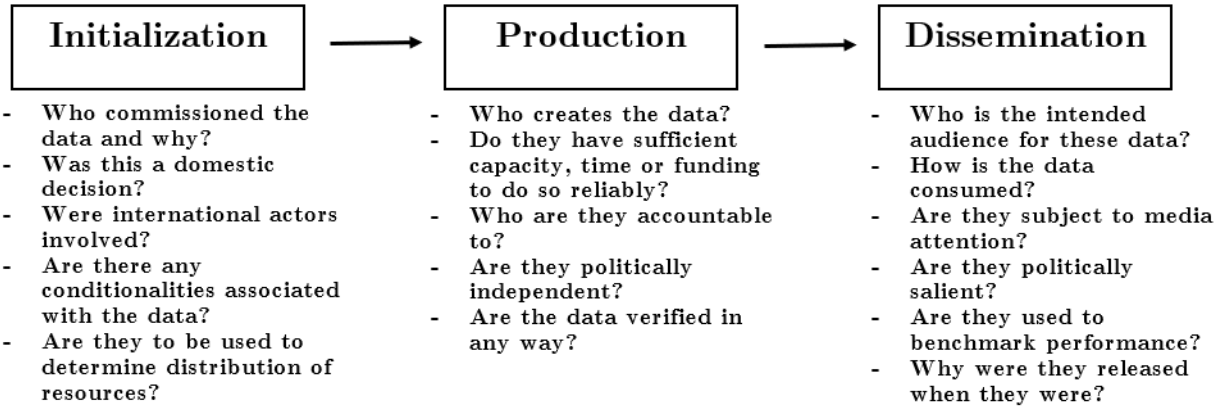


Figure 1: Political considerations at each stage of data production

First, researchers should consider who commissions the data and their motivation for doing so (the initialization stage). This affects the audience(s) for which the data is likely to be important, which in turn affects who stands to benefit or lose from the information that is released. For example, census data is collected to inform allocation of budgets and definition of electoral districts. Local performance data may be collected to benchmark central transfers. Elections in autocracies are partly held to gather information about the distribution of support which may then determine the distribution of resources. Data may also be collected to convey information to international actors. For example, GDP figures matter for international reputation and investment. Data on primary enrolment, HIV/AIDS, hunger etc. are relevant to tracking progress toward the Sustainable Development Goals, determining foreign aid envelopes, and informing external perceptions of domestic performance in developing countries, which can affect foreign direct investment. Data which are likely to influence the allocation of resources to a country or within it are particularly likely to be manipulated.

Second, the production stage – who generates the data, their capacity and incentives – is likely

to influence data quality and the ease with which it can be manipulated. Highly ambitious data collection projects which have little oversight and low budgets are more likely to be unreliable. Data produced by bureaucrats subject to political pressure are more likely to be manipulated. On the other hand, data that is endorsed by civil society organizations or international organizations with in-country presence are less likely to be manipulated because of additional oversight.

Third, dissemination – the audience for the data and its salience – is the last and perhaps most important stage of the process. Data which is primarily for internal use and released on open data platforms as a matter of course is significantly less likely to be manipulated than data which may be covered in the national press. Political salience is also an important consideration. Data may gain salience for a variety of reasons. One example is subnational performance, as discussed in further detail below. Performance may be measured using data which is idiosyncratic to a given country. For example, local tax and school exam results are important metrics of performance in Tanzania. In Rwanda, maternal mortality is regularly reported to measure local improvements (Worley, 2015). In Mexico, murder rates are fundamentally important for the credibility of state governments. Indeed, they are thought to be manipulated in some cases (Telesur, 2017).

We contend that reflecting on data initialization, production and dissemination can help scholars identify potential risks involved in using data from a given context. It is then important for scholars to think about how manipulation can condition the conclusions drawn from empirical analysis. What threats to influence do researchers face? Perhaps the most pernicious are flawed inferences about the relationship between regime support and a variety of "good" outcomes, i.e. those that are in the interest of the regime to artificially inflate. Not only does this risk parroting pro-government propaganda in data form, it can also lead to flawed inferences about the reasons why people support the regime, and the ways in which the regime rewards supporters. Furthermore, public data may underestimate what the incumbent wants to hide. Manipulated data may also underestimate the extent of discrimination in access to public resources. This makes it harder for anti-regime actors to claim credit or point to government abuses of state power. As a result, scholars may fail to appreciate the role of these actors in non-democratic systems.

However, the consequences of manipulation are not limited to inferences about regime strategy and electoral politics. Scholars from disciplines other than political science frequently use data from non-democratic developing countries to explore questions related to education, public health

and the provision of other services. However, a government may seek to manipulate data on school performance, maternal health, infant mortality and so on for reasons related to politics. As a result, bias may affect even "apolitical" studies – leading, e.g, to flawed inferences about the determinants of local development and welfare. Hence, even scholars who are not immediately concerned with electoral politics in their countries of study must be cognizant of them as they affect the politics of data.

3 Evidence of Risks: Comparing public and internal data from Tanzania

Tanzania exemplifies the range of ways that autocrats can strategically censor and manipulate economic indicators for their own benefit. Tanzania is a hegemonic party regime in East Africa that has taken an increasingly authoritarian turn since the election of its current President John Magufuli in 2015 (McLellan, 2018). Magufuli has pursued an unprecedented raft of interventionist economic policies alongside an outright attack on political freedoms and opposition parties. To protect his economic record, Magufuli and his government have engaged in a number of tactics to obscure potentially damaging information. For example, the Tanzanian government refused to allow the International Monetary Fund to publish their annual report which criticized the country's policies (Cotterill, 2019). After the Controller and Auditor General pointed to the mysterious disappearance of $640 million of public money, the President and Parliament refuted these figures and refused to publish the report, breaking with long-standing precedent. Opposition politicians who have pushed for answers about the missing millions have been accused of incitement (Collord, 2019; African Arguments, 2019).

Tanzania publishes a raft of administrative data as part of its open government partnership with the World Bank (World Bank, 2015).[2] To demonstrate how the politics of data can affect the credibility of research, we interrogate one example: tax revenues collected by local governments in Tanzania. Specifically, we compare internal local tax collection data from Tanzania for the 2016/17 fiscal year[3] with that released online through the Local Government Revenue Collection Dashboard,

[2]This data has informed a number recent of working and published papers, including the authors' own work.

[3]This data was collected directly from the President's Office for Regional Administration and Local Government by one of the authors. The reliability of this data was corroborated through interviews with bureaucrats and politicians

an open data portal established in 2017 and hosted on the website of the President's Office for Local Government and Regional Administration (TAMISEMI). These data are published by the offices of presidential appointees in TAMISEMI. Those publishing the data are not politically independent and are directly accountable to the President. [CITATION FOR THIS?]

Data on tax takings is widely used in political science to measure concepts like state capacity and state-society relations among others (Tilly, 1990; Lieberman, 2002). In the context of decentralized government, it is also an important indicator of local government performance. Good opposition performance at the local level can win over voters, institutionalizing support for opposition parties that may ultimately unseat the incumbent (Lucardi, 2016). In Tanzania, local governments raise taxes to supplement central transfers and fund key services (roads, education, health, water) to foster development and so win political support. Over time, local revenue collection has emerged as a yardstick by which politicians and the press appraise local government performance (Malanga, 2019). The government began releasing this data as part of budgetary transparency initiatives but the data has become politically salient over time.

Tanzania's opposition parties, who control around 20 percent of the country's local governments including many key cities, use these figures to show they are delivering on their local mandate. The central government uses tax takings to herald the success of key incumbent areas, notably the administrative capital Dodoma (Chidawali, 2018). Hegemonic parties like Tanzania's Chama Cha Mapinduzi (CCM) maintain support by convincing voters they are the only party which can successfully rule (Magaloni, 2006). This image may be threatened by the release of data which shows that opposition parties outperform them at the local level. These data's role in benchmarking the performance of polarized actors generates incentives for the regime to modify these data before releasing it to the public. The prominence of presidential appointees in its production and dissemination makes manipulation feasible.

Our analysis suggests such manipulation may be taking place. The tax data released by the government underestimates local revenue raising in the vast majority of Tanzania's local government authorities (LGAs). If these differences were fairly consistent, the most plausible explanation would be simple accounting problems.[4] However, there is substantial variation in the magnitude of the

in several regions.

[4]The open data included shillings to two decimal places. The smallest unit of Tanzanian shillings currency is 50 shillings. This suggests that the data entered had been back-converted from US dollars (or another currency).

differences between the internal data and that released to the public. These differences range from an underestimate of 40 billion shillings ($16.5 million) in Kinondoni Municipal Council in Dar es Salaam to an overestimate of 1.25 billion shillings ($550,000) in Kibaha District Council. Per capita differences are similarly skewed. Moreover, these differences appear to vary systematically.

Further investigation suggests that the pattern which emerges can be most plausibly explained as a strategic response by the central government to the dynamics of political competition in Tanzania. Figure 2 plots the differences between total revenue reported in internal and public data by LGA. Opposition local governments' tax takings are underestimated to a far greater (and statistically significant) extent than incumbent local governments. In particular, the public data vastly underestimates the tax takings of prominent opposition strongholds like Moshi, Mbeya, Arusha and central Dar es Salaam. Importantly, where the public data does overestimate tax takings, it is in prominent regime strongholds like the administrative capital Dodoma.

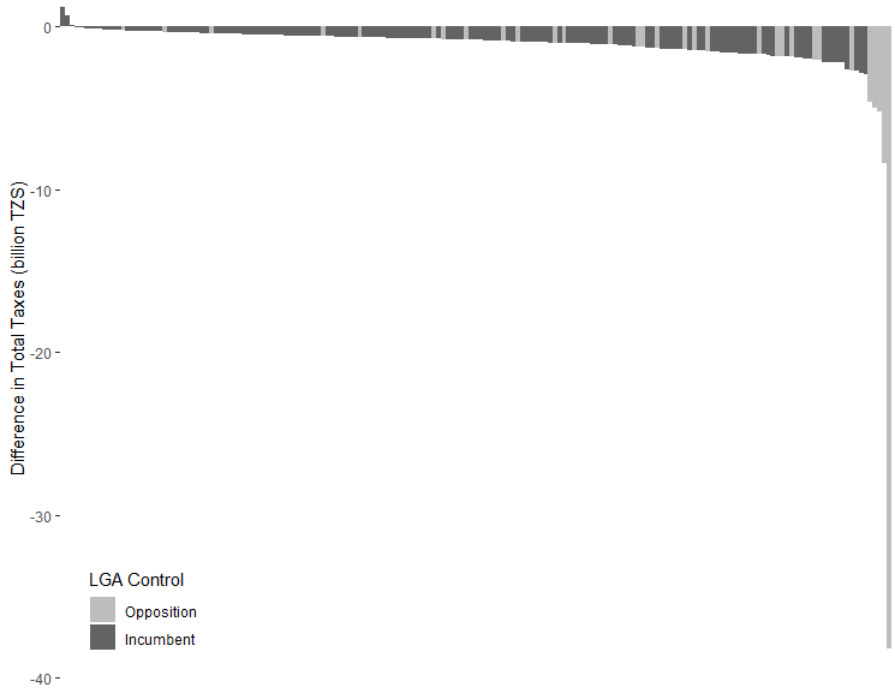


Figure 2: Difference between total revenue reported in internal and public data by LGA

Variation between internal and public data is most likely because of simple currency fluctuations between the first conversion into US dollars for international consumption and back into Tanzanian shillings whenever the data was then posted online. This would account for the smooth variation between most of the differences. The relevant differences are those in the far tails

Why might the central government manipulate the data in this way? Changes in the data affect LGA rankings and so affect the extent to which opposition and ruling parties can claim credit for good performance. Table 1 shows the top five LGAs in total and per capita collection for the public and internal data. The top five LGAs according on the internal data are all opposition councils (shown in bold italics).[5] Opposition councils also dominate the top five from the public data but there are some notable shifts. First, Dodoma moves into the top five for total revenue collection from 11th place in the internal data.[6] Second, Chadema councils fall out of the top five in per capita collection completely when the public data is used. In contrast, LGAs from other opposition parties remain among the top performers.

Table 1: Top performing local governments in 2016/17

Top 5 LGAs (total collection)	**Top 5 LGAs (per capita collection)**
Internal	*Internal*
1. ***Kinondoni MC***	1. ***Kinondoni MC***
2. ***Arusha CC***	2. ***Mtwara MC***
3. ***Tanga CC***	3. ***Tanga CC***
4. ***Ubungo MC***	4. ***Arusha CC***
5. ***Mbeya CC***	5. ***Moshi MC***
Open data	*Open data*
1. ***Kinondoni MC***	1. ***Mtwara MC***
2. ***Arusha CC***	2. Kibaha DC (CCM)
3. ***Ubungo MC***	3. Geita TC (CCM)
4. Dodoma MC (CCM)	4. ***Tandahimba DC***
5. ***Tanga CC***	5. Bagamoyo DC (CCM)

Opposition councils are shown with bold italics; CCM councils are labeled as such.

These changes are politically useful for the CCM. First, Dodoma's rise to number 4 in 2016/7 then to number 1 in the 2017/18 data has been used as proof that Dodoma can handle the movement of government functions from Dar es Salaam to the small official capital. Second, Chadema (the largest opposition party in Tanzania and the biggest threat to the ruling party (Mtulya, 2015)) cite opposition LGAs' performance in places like Moshi and Arusha as evidence that they are a serious party of government. All these councils' rankings drop significantly in the public data. Moshi goes

[5] We define opposition councils as those with a majority of opposition councilors.

[6] Dodoma is the political capital of Tanzania but has long paled into irrelevance compared to the financial, cultural and administrative capital of Dar es Salaam. Dodoma is a town of around two hundred thousand people with markedly worse infrastructure than Dar es Salaam. President Magufuli has made it a priority to move ministries, embassies and other political offices to Dodoma. Any ministry or embassy moves or other success stories about Dodoma are heralded by pro-government newspapers (Lugongo, 2019).

from 5th in per capita and 7th in total earnings in internal data to 42nd and 38th respectively in the public data. The credibility that opposition parties can build through good performance is dangerous to the CCM. The patterns in the data suggest this performance has been strategically underestimated.

The publicly released figures allow the ruling party to claim they are performing well and undermines opposition parties' ability to do the same. Importantly, the manipulation our analysis suggests has significant effects on the inferences one can draw from the tax data as a whole. First, the differences will bias the results of any study which seeks to understand the effect of opposition control on local government performance. Table 2 plots the results of regression analysis investigating the effect of LGA partisanship on local revenue collection using both the internally validated and publicly available data.[7] The analysis based on internally validated data shows that opposition majority LGAs collect significantly higher total and per capita taxes. Likewise an increase in opposition share increases per capita and total tax collection. However, the significance of these relationships disappear when the public data is used. A researcher using the internal data would conclude that opposition councils indeed collect more taxes while a researcher using the public data would not find any significant effect of opposition control.

Manipulation of this sort can also change the results of studies which ostensibly have nothing to do with party politics. Consider a researcher who is interested in how increased bureaucratic resources affect tax collection. The researcher may try to leverage administrative status (rural, town, city etc) to test this. As part of their study, the researcher may test the effect of administrative status on tax collection.[8] Because a disproportionate number of opposition councils are urban, the results of this study may be biased because opposition performance and hence urban performance is underestimated.[9] Because the manipulation may not be consistent across units or over time, it is difficult to control for it directly or through fixed effects. Therefore, the politics of data even affects the reliability of inferences drawn by researchers who may study questions which are substantively removed from electoral politics in non-democracies.

[7]We regress local tax collection on ruling party control of local government, controlling for LGA characteristics and region fixed effects.

[8]The central government decides which LGAs are administratively upgraded based on population, population density and their own discretion about 'readiness'. When LGAs are upgraded, they gain additional transfers and an increased workforce of bureaucrats, among other resources

[9]Indeed, the two datasets generate significantly different results as shown in Table A1 in the Appendix.

Table 2: Effect of LGA partisanship on revenue raising

	Dependent variable:			
	Internal data		*Public data*	
	log(Local Tax Revenue Per Capita)			
	(1)	(2)	(3)	(4)
CCM majority	−0.281**		−0.288	
	(0.111)		(0.242)	
Share of opp councilors		0.580***		0.511
		(0.195)		(0.419)
Observations	179	173	179	173
R^2	0.711	0.745	0.408	0.468
Adjusted R^2	0.652	0.692	0.288	0.356
Residual Std. Error	0.412	0.392	0.897	0.842
F Statistic	12.118***	13.864***	3.395***	4.163***
Note:				*p<0.1; **p<0.05; ***p<0.01

4 Additional Evidence of Risks and Established Best Practice

Manipulation of data on tax takings and other macroeconomic indicators is by no means limited to Tanzania. Scholars have documented considerable evidence of similar phenomena in other non-democratic contexts. For instance, Martinez (2019) compares World Bank GDP statistics with nightlight data from democracies and autocracies all over the world. He finds that non-democracies inflate GDP figures by an average of 1-1.3 per cent annually, more so in election years, years of low growth and periods when the country loses foreign aid (Martinez, 2019). These practices have been identified in non-democracies across the income distribution. In China, scholars have questioned the veracity of official statistics from top-level GDP and productivity statistics to subnational growth rates, all the way down to village statistics (Rawski, 2001; Tsai, 2008; Wallace, 2016). The Rwandan government has also been accused of manipulating official statistics to exaggerate the extent of poverty alleviation – a pillar of the Kagame regime's legitimacy (ROAPE, 2019).[10]

Beyond macroeconomic indicators, scholars have also documented manipulation of data on election results and population statistics. The fact that data on turnout and vote counts face risk

[10]The National Institute of Statistics of Rwanda has refuted these claims: `http://www.statistics.gov.rw/press/press-release/nisr-refutes-financial-times-article-rwanda\%E2\%80\%99s-poverty-trends`

of falsification is unsurprising. In any country where leadership is determined to some extent by elections, these numbers are of profound importance. Indeed, falsifying election returns is one of the less costly options available to nondemocratic regimes (Harvey, 2016). It is in just such places that direct observation of elections is most difficult (Deckert, Myagkov and Ordeshook, 2011).

Population statistics also face a risk of manipulation by government officials. First, the mere fact of *who* gets counted can serve to reify certain identities in a way that may run counter to regime interests. In India, caste is not recorded in the census since counting it is seen as a threat to creating an integrated national identity (Gill, 2007). In a number of African countries, questions about ethnicity have been excluded from the national census as a symbolic rejection of the colonial past (Christopher, 2006). The relationship between subnational population counts and the distribution of federal resources is one of the prime reasons why population data becomes politicized. In Nigeria, as Elemo (2018) argues, the regime has used its access to oil revenues (which account for about 10 percent of the country's GDP) as a key feature of clientelistic control. Over a quarter of these revenues are distributed to non-oil producing states, using a formula that relies in part on state population (Elemo, 2018). Public awareness of the political and financial implications of census results for subnational units make Nigeria's census results highly controversial and subject to manipulation. Indeed, Akinyoade, Appiah and Asa (2017)'s survey of all censuses in Nigeria from 1866-2006 finds them all "grossly inadequate."

Scholars have thus developed a range of methods for detecting fraud in official statistics. In general, these methods make certain assumptions about the data generating process and then analyze whether official government statistics deviate from what would be expected in the absence of manipulation. These include various forms of digit analysis (Mebane, 2010; Deckert, Myagkov and Ordeshook, 2011). For instance, Beber and Scacco (2012) demonstrate that fair elections produce returns where last digits occur with equal frequency, whereas the experimental literature suggests that people do not tend to select digits in this manner. Thus an election official working to meet a quota of votes for a particular candidate or party would be unlikely to generate numbers that follow the expected pattern in the absence of fraud. They find substantial evidence to suggest manipulation in Nigeria as well as in Senegal in 2007. Similar approaches have been applied to detect fraud in recent Russian elections (Harvey, 2016; Rundlett and Svolik, 2016; Skovoroda and Lankina, 2017). Scholars have also developed a variety of regression-based approaches that involve

the estimation of statistical models and examine irregularities, anomalies, or outliers (Alvarez and Katz, 2008; Myagkov, Ordeshook and Shakin, 2009; Wand et al., 2001).

Detecting falsification in population statistics follows a similar logic to the methods for identifying election fraud discussed above. Two of the most commonly applied tests used to determine the extent of digit preference are the Whipple and Myers Indices, which provide summary measures of "heaping" on numbers ending in 0 or 5 (Borkotoky and Unisa, 2014). These measures are most commonly applied to detect incorrect information on age but could presumably also detect falsified data on other population statistics. We contend that these techniques developed to detect falsification of population and election data should be applied more broadly to administrative data released by authoritarian governments.

While government statistics are the result of political processes in both democracies and autocracies, the risk is appreciably higher in the latter type of regime given constraints on independent oversight. The preceding discussion clearly illustrates the risks associated with uncritically downloading newly available data from the web portals of less than fully democratic regimes. Incorporating such data into one's analysis can lead to flawed inferences akin to what Ross (2006) demonstrates in his seminal re-examination of prior studies of democracy and poverty reduction. Prior cross-national studies tended to exclude nondemocratic states that performed well. He shows that once these cases are included, democracy has little or no effect on infant and child mortality rates. In what follows we discuss when these risks are likely to be highest in order to guide researchers wishing to take advantage of newly available data from non-democratic regimes.

5 Minimizing risks

Given these threats to inference, what are scholars to do as individuals and as a discipline? This reflection is not intended to discourage scholars from using large-N data from non-democracies. Instead, this piece is intended as a reminder to scholars that that the data they use, regardless of the subject of their research, is inherently political. In particular, data published by non-democratic governments is liable to manipulation. In what remains, we provide recommendations for how scholars can ameliorate and account for these risks. We recommend that researchers do as much as possible to know their case and know their data in order to be able to identify when

inference may be jeopardized by the politics of data.

As we have shown, data is not produced in a vacuum. What appear to be 'great data' may be fundamentally compromised by the politics of data. Without a good grasp on these politics, it is impossible for researchers to understand what data is most likely to be manipulated. Without case knowledge, it is also impossible for researchers to preempt the kinds of manipulation that incumbents might have incentive to make and hence what threats to inference they face when analyzing data. We therefore encourage researchers, even those who are not deeply invested in a given case, to gain as much familiarity with that case as they can. Scholars should talk to and co-author with local scholars and case experts if they plan to use evidence from a given case. They should try and validate subnational data with stakeholders like local bureaucrats, politicians or civil society. They should make use of comparable questions in validated, independent data like the Barometer surveys or Demographic and Health Studies. All of these strategies allow scholars to generate a baseline idea of 'what looks right' against which they can appraise the public data they are using. These kinds of simple 'sniff tests' protect researchers from publishing work based on manipulated data. Indeed, additional case knowledge is likely to enrich the study in other ways.

Similarly, scholars should also get to know their data before diving into analysis.[11] We recommend that researchers use their priors about the data generating process relevant to their area of interest in order to look for deviations from it that may suggest manipulation. Exploratory analysis – initial investigations to discover patterns, spot irregularities, and check assumptions with the help of summary statistics and graphical representations – is especially important when dealing with potentially risky data. Scholars should also make use of established techniques developed for census and election data described above to check for evidence of manipulation.[12]

Using this kind of data puts an additional burden on scholars to convince the reader that these concerns have not undermined inference in their study. We call on the discipline to engage with the 'politics of data' as good practice when describing their data. As scholars outline their survey design to convince the reader the sample is representative or demonstrate balance across samples when

[11] This is advisable for data from all contexts. As the anthropologist Crystal Biruk (2018) points out, "all data – even that verified as clean by demographers –are cooked by the processes and practices of production (5)."

[12] We run first and last digit analysis on both tax datasets (internal and public figures). Our data is small at around 180 observations and so is not the most amenable to this kind of data forensic analysis. However, Figures A1 and A2 in the Appendix show that the internal data is more consistent with theoretical expectations about digit distribution than the public data.

leveraging natural experiments, scholars should convince the reader why and to what extent they are convinced of the quality of their data. Furthermore, they should acknowledge what direction any possible bias they suspect may run. By being upfront about the risks as well as the benefits of this new data availability, scholars demonstrate an awareness of these risks and signal an effort to minimize them. This makes it easier for the reader to transparently engage with the work.

Finally, we call on scholars and civil society to track data censorship and manipulation. Organizations like V-Dem and Freedom House should consider incorporating data censorship into their existing work tracking other forms of censorship. Measures of data censorship are critical for scholars to understand the likelihood that the data they are using is compromised and therefore the level of caution they should exercise when using it. Furthermore, as data journalism becomes increasingly prominent, these kinds of measures are even more important for civil society and the press to be able to appraise what conclusions can be drawn from official statistics and what claims should be questioned.

Open data released by non-democratic governments is not necessarily 'bad data', unusable by academics. The proliferation of this data provides great opportunities for scholars of non-democratic politics and social scientists more broadly to advance our knowledge. But we argue that the politics that led to the release of these data cannot be separated from the data itself. These data pose threats to inference which have thus far not been systematically explored. In this reflection, we demonstrate the risks posed by open data and suggest when data are most likely to be manipulated. In line with the broader push towards transparency in the discipline, we call on scholars to more openly engage with the risks that their data has been manipulated. We propose ways that scholars can better understand the politics of data and indeed the data itself in order to minimize the risks of drawing inferences based on manipulated data.

References

African Arguments. 2019. "Tanzania search for missing millions raises questions over $1 billion - African Arguments.".

URL: *https://africanarguments.org/2019/02/13/tanzania-search-missing-millions-reveals-missing-billion/*

Akinyoade, Akinyinka, Eugenia Appiah and Sola Asa. 2017. "Census-taking in Nigeria: The good, the technical, and the politics of numbers." *African Population Studies* 31(1).

Alvarez, R. Michael and Jonathan N. Katz. 2008. The Case of the 2002 General Election. In *Election Fraud: Detecting and Deterring Electoral Manipulation*, ed. R. Michael Alvarez, Thad E. Hall and Susan D. Hyde. Washington, DC: Brookings Institution Press pp. 149–61.

Beber, Bernd and Alexandra Scacco. 2012. "What the numbers say: A digit-based test for election fraud." *Political analysis* 20(2):211–234.

Berliner, Daniel. 2014. "The political origins of transparency." *The Journal of Politics* 76(2):479–491.

Berliner, Daniel and Aaron Erlich. 2015. "Competing for transparency: political competition and institutional reform in Mexican states." *American Political Science Review* 109(1):110–128.

Biruk, Crystal. 2018. *Cooking data: Culture and politics in an African research world.* Duke University Press.

Borkotoky, Kakoli and Sayeed Unisa. 2014. "Indicators to examine quality of large scale survey data: an example through District Level Household and Facility Survey." *PLoS One* 9(3):e90113.

Chidawali, Habel. 2018. "Revealed: What gave Dodoma City a boost in revenue earnings.".

URL: *https://www.thecitizen.co.tz/News/Land-sale–transparency-boosts-revenue-earnings-for-Dodoma-City/1840340-4694296-14j1m3ez/index.html*

Christopher, A. J. 2006. "Questions of identity in the millennium round of Commonwealth censuses." *Population Studies* 60(3):343–352.

URL: *https://doi.org/10.1080/00324720600896163*

Clark, William Roberts and Matt Golder. 2015. "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?" *PS: Political Science and Politics* 48(01):65–70.
URL: *http://dx.doi.org/10.1017/S1049096514001759*

Collord, Michaela. 2019. "Tanzania The politics of being Auditor General.".
URL: *https://presidential-power.com/?p=9503*

Cotterill, Joseph. 2019. "Tanzania president blocks critical IMF report on economy.".
URL: *https://www.ft.com/content/cb51db44-61f8-11e9-a27a-fdd51850994c*

Davies, Tim G and Zainab Ashraf Bawa. 2012. "The promises and perils of Open Government Data (OGD)." *The Journal of Community Informatics* 8(2):1–8.

Deckert, Joseph, Mikhail Myagkov and Peter C Ordeshook. 2011. "Benford's Law and the detection of election fraud." *Political Analysis* 19(3):245–268.

Devarajan, Shantayanan. 2013. "Africa's Statistical Tragedy." *Review of Income and Wealth* 59:S9–S15.
URL: *http://dx.doi.org/10.1111/roiw.12013*

Egorov, Georgy, Sergei Guriev and Konstantin Sonin. 2009. "Why resource-poor dictators allow freer media: A theory and evidence from panel data." *American Political Science Review* 103(4):645–668.

Elemo, Olufunmbi M. 2018. Fiscal Federalism, Subnational Politics, and State Creation in Contemporary Nigeria. In *The Oxford Handbook of Nigerian Politics*. Oxford University Press p. 189.

Gill, Mehar Singh. 2007. "Politics of population census data in India." *Economic and Political Weekly* pp. 241–249.

Harvey, Cole J. 2016. "Changes in the menu of manipulation: Electoral fraud, ballot stuffing, and voter pressure in the 2011 Russian election." *Electoral studies* 41:105–117.

Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2011. "Democracy and Transparency." *The Journal of Politics* 73(4):1191–1205.
URL: *http://dx.doi.org/10.1017/S0022381611000880*

Hollyer, James R, B Peter Rosendorff and James Raymond Vreeland. 2015. "Transparency, protest, and autocratic instability." *American Political Science Review* 109(4):764–784.

Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2018. "Transparency, protest and democratic stability." *British Journal of Political Science* pp. 1–27.

Jerven, Morten. 2013. *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press.

Jerven, Morten and Deborah Johnston. 2015. "Statistical tragedy in Africa? Evaluating the data base for African economic development." *The Journal of Development Studies* 51(2):111–115.

Kelley, Judith G and Beth A Simmons. 2019. "Introduction: The Power of Global Performance Indicators." *U of Penn Law School, Public Law Research Paper* (19-06).

King, Gary. 2014. "Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science." *PS: Political Science and Politics* 47(01):165–172.
URL: *http://dx.doi.org/10.1017/S1049096513001534*

Lieberman, Evan S. 2002. "Taxation data as indicators of state-society relations: possibilities and pitfalls in cross-national research." *Studies in Comparative International Development* 36(4):89–115.

Lorentzen, Peter. 2014. "China's strategic censorship." *American Journal of Political Science* 58(2):402–414.

Lucardi, Adrian. 2016. "Building Support From Below? Subnational Elections, Diffusion Effects, and the Growth of the Opposition in Mexico, 1984-2000." *Comparative Political Studies* 49(14):1855–1895.

Lugongo, Bernard. 2019. "Tanzania: Government Move to Dodoma Now At 86 Per Cent." *Tanzania Daily News (Dar es Salaam)* .
URL: *https://allafrica.com/stories/201902060401.html*

Maerz, Seraphine F. 2016. "The electronic face of authoritarianism: E-government as a tool for

gaining legitimacy in competitive and non-competitive regimes." *Government Information Quarterly* 33(4):727–735.

Magaloni, Beatriz. 2006. *Voting for autocracy: Hegemonic party survival and its demise in Mexico*. Cambridge University Press Cambridge.

Malanga, Alex. 2019. "LGA have collected only 55 per cent of targeted revenue.".
URL: *https://www.thecitizen.co.tz/News/LGA-have-collected-only-55-per-cent-of-targeted-revenue/1840340-5066774-eaftcr/index.html*

Martinez, Luis R. 2019. How Much Should We Trust the Dictator's GDP Growth Estimates? SSRN Scholarly Paper ID 3093296 Social Science Research Network Rochester, NY: .
URL: *https://papers.ssrn.com/abstract=3093296*

McLellan, Rachael. 2018. "Why is once-peaceful Tanzania detaining journalists, arresting schoolgirls and killing opposition leaders? - The Washington Post.".
URL: *https://www.washingtonpost.com/*

Mebane, Walter R. 2010. "Fraud in the 2009 presidential election in Iran?" *Chance* 23(1):6–15.

Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen and Betsy Sinclair. 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS: Political Science and Politics* 48(01):71–74.
URL: *http://dx.doi.org/10.1017/S1049096514001760*

Mtulya, Athuman. 2015. "2015 elections: CCM loses key local authorities to opposition.".
URL: *https://www.thecitizen.co.tz/News/2015-elections–CCM-loses-key-local-authorities-to-opposition/1840340-2938274-rk4mlfz/index.html*

Myagkov, Mikhail, Peter Ordeshook and Dimitri Shakin. 2009. *The forensics of electoral fraud*. New York: Cambridge University Press.

Rawski, Thomas G. 2001. "What is happening to China's GDP statistics?" *China Economic Review* 12(4):347–354.

Reuter, Ora John and Jennifer Gandhi. 2011. "Economic performance and elite defection from hegemonic parties." *British Journal of Political Science* 41(1):83–110.

ROAPE. 2019. "A Straightforward Case of Fake Statistics.".
URL: *http://roape.net/2019/04/18/a-straightforward-case-of-fake-statistics/*

Ross, Michael. 2006. "Is democracy good for the poor?" *American Journal of Political Science* 50(4):860–874.

Rundlett, Ashlea and Milan W Svolik. 2016. "Deliver the vote! micromotives and macrobehavior in electoral fraud." *American Political Science Review* 110(1):180–197.

Sandefur, Justin and Amanda Glassman. 2015. "The political economy of bad data: evidence from African survey and administrative statistics." *The Journal of Development Studies* 51(2):116–132.

Skovoroda, Rodion and Tomila Lankina. 2017. "Fabricating votes for Putin: new tests of fraud and electoral manipulations from Russia." *Post-Soviet Affairs* 33(2):100–123.

Stier, Sebastian. 2015. "Political determinants of e-government performance revisited: Comparing democracies and autocracies." *Government Information Quarterly* 32(3):270–278.

Telesur. 2017. "Are Mexican States Lying About Crime and Homicide Rates?".
URL: *https://www.telesurenglish.net/news/Are-Mexican-States-Lying-About-Crime-and-Homicide-Rates–20170419-0020.html*

Tilly, Charles. 1990. *Coercion, capital, and European states, AD 990*. Cambridge: Basil Blackwell.

Titiunik, Roco. 2015. "Can Big Data Solve the Fundamental Problem of Causal Inference?" *PS: Political Science and Politics* 48(01):75–79.
URL: *http://dx.doi.org/10.1017/S1049096514001772*

Tsai, Lily L. 2008. "Understanding the falsification of village income statistics." *The China Quarterly* 196:805–826.

Wallace, Jeremy L. 2016. "Juking the stats? Authoritarian information problems in China." *British Journal of Political Science* 46(1):11–29.

Wand, Jonathan N, Kenneth W Shotts, Jasjeet S Sekhon, Walter R Mebane, Michael C Herron and Henry E Brady. 2001. "The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 95(4):793–810.

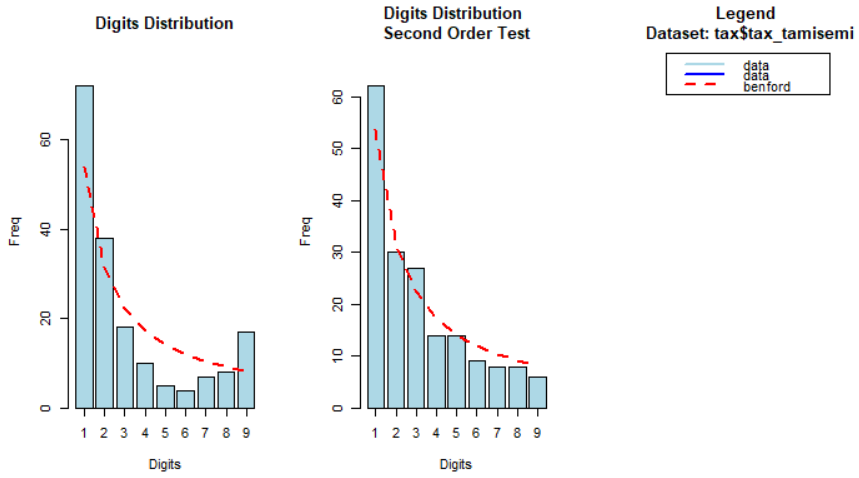
World Bank. 2015. "Tanzania Conference Places Open Data at Center of Development Agenda.". **URL:** *http://www.worldbank.org/en/news/press-release/2015/09/04/tanzania-conference-places-open-data-at-center-of-development-agenda*

Worley, Heidi. 2015. "Rwandas Success In Improving Maternal Health Population Reference Bureau.". **URL:** *https://www.prb.org/rwanda-maternal-health/*

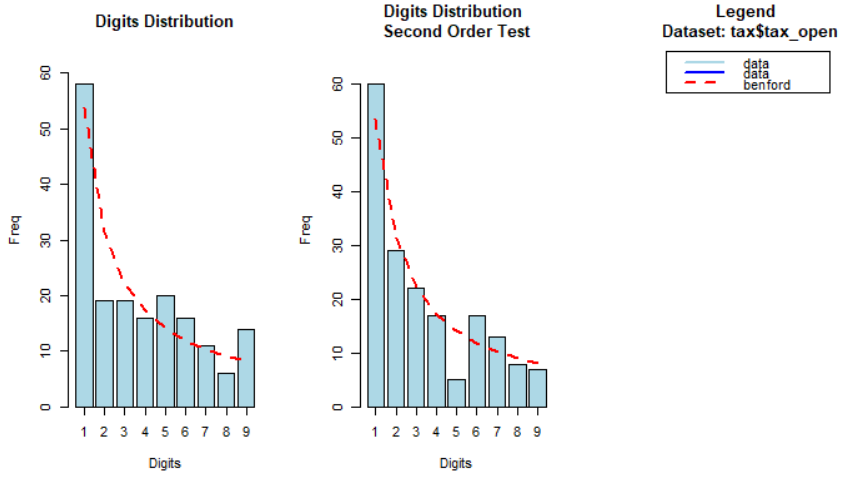
Appendix

Table A1: Effect of administrative type on revenue raising

	Dependent variable:			
	Internal data		*Public data*	
	log(Local Tax Revenue)			
	(1)	(2)	(3)	(4)
Rural council	-1.445^{***}	-1.173^{***}	-1.816^{***}	-1.538^{**}
	(0.277)	(0.285)	(0.567)	(0.597)
Municipal council	-0.792^{***}	-0.617^{**}	-1.179^{*}	-1.000
	(0.303)	(0.301)	(0.622)	(0.632)
Town council	-1.156^{***}	-0.904^{***}	-1.570^{**}	-1.313^{**}
	(0.302)	(0.306)	(0.619)	(0.642)
Region fixed effects	Y	Y	Y	Y
LGA majority controls	N	Y	N	Y
Observations	179	179	179	179
R^2	0.656	0.675	0.402	0.410
Adjusted R^2	0.588	0.609	0.285	0.291
Residual Std. Error	0.440	0.429	0.902	0.899
F Statistic	9.777^{***}	10.241^{***}	3.452^{***}	3.431^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

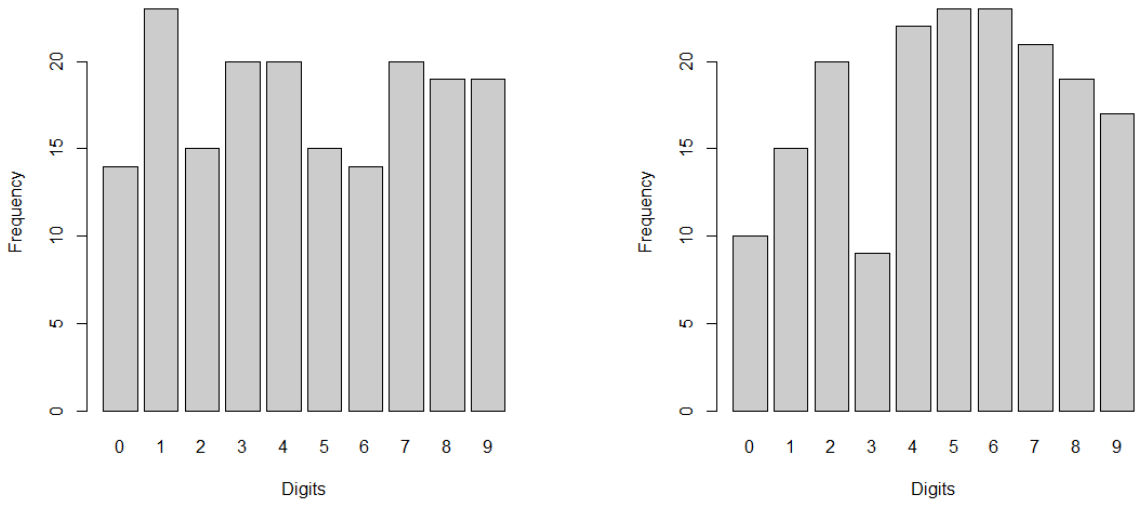


(a) Benford Law analysis of internal data



(b) Benford Law analysis of public data

Figure A1: Digit analysis of first digits of tax data. Benford's law states that the distribution of first digits should conform to the red line shown



(a) Internal data

(b) Public data

Figure A2: Digit analysis of last digits of tax data. If the data reported is the true data, we would expect broadly consistent frequencies across all digits.