SAGE Research Methods Cases
Politics & International Relations

Case Title

Using Data Analysis and Instrumental Variables to Study the Drivers of Corruption

Author Name

Elena Nikolova

Author Affiliation & Country of Affiliation

University College London, School of Slavonic and East European Studies, UK

Lead Author Email Address

Email: e.nikolova@ucl.ac.uk

Discipline: D8

Sub-discipline within Politics & International Relations

Comparative Politics [SD-POLIR-3]

Academic Level of intended readership

Advanced Undergraduate

Contributor Biographies

Elena Nikolova is an Assistant Professor in Economics at the School of Slavonic and East European Studies at University College London. She is also affiliated with the Central European Labour Studies Institute in Slovakia (as a research fellow) and with the Leibniz-Institute for East and Southeast European Studies (IOS) Regensburg in Germany (as an associated researcher). She is a Fellow of the Global Labor Organization. She received her PhD from Princeton University in 2011, and her undergraduate degree from Gettysburg

College. She is interested in political economy, economic development, comparative politics and democratization, and the politics and economics of transition.

Published Articles

Nikolova, E., & Marinov, N. (2017). Do public fund windfalls increase corruption? Evidence from a natural disaster. *Comparative Political Studies*, *50*(11), 1455-1488.

Abstract

This case discusses the use of data analysis and instrumental variables (IV) to study the drivers of corruption. Drawing on a quantitative analysis of the link between municipal flood assistance and local corruption in Bulgaria between 2004 and 2005, it explores the challenges the authors faced when collecting and coding the data used in their work, and how they managed to overcome or minimize them. It also discusses the advantages of corruption measured with objective data as opposed to survey-based perceptions. It highlights the issues of reverse causality and omitted variable bias which arise in a simple regression of corruption on flood assistance. It then explains how the instrumental variables technique can overcome these issues and applies it to the context under study. Finally, it takes a broad view on the practical lessons learned while conducting this research project and concludes that while the research project was full of challenges and frustrations, it was also very useful and rewarding.

Learning Outcomes

By the end of this case, students should be able to . . .

- Differentiate between objective and subjective corruption data
- Understand how the instrumental variable (IV) approach works and what problems it is designed to solve

- Define and give some practical examples of causality, reverse causality and omitted variables
- Evaluate the pros and cons of different data measures, such as corruption or rainfall, and how the researcher can minimize concerns about data reliability

---

Case Study

Project Overview and Context

The project described in this case study focuses on an exploration of the link between unexpected financial windfalls and corruption in local government in Bulgaria. In 2004 and 2005, many Bulgarian municipalities were hit by torrential rains which led to flooding. To deal with the destruction from the floods, the central government awarded funds to 257 (out of 264) affected municipalities, with the average transfer amounting to around 15.6% of municipal income. Reports in the media showed that in many cases the funds were not used for reconstruction but rather ended up in the pockets of corrupt local politicians. Following extensive pressure from opposition parties and the media, in 2006 the incumbent government ordered an unprecedented audit of the municipalities which had received assistance. This is something that rarely happens in Bulgaria, which has long-standing issues with corruption (European Commission, 2017). The research found that municipalities which received more flood-related funds experienced more corruption: a 10% increase in the per capita amount of disbursed funds leads to a 9.8% increase in corruption. More suggestively, corrupt mayors anticipated punishment by voters and dropped out of the next election race.

Our interest in studying this topic was driven primarily by the availability of new and unique corruption data. Since corruption is an illegal activity, objective corruption data are difficult to come by and researchers usually have to rely on surveys which ask respondents to report

their perceptions of corruption. However, corruption perceptions may or may not correspond to actual corruption (Olken, 2009). We exploited information from detailed reports issued by an independent national watchdog on how the flood money was used in each municipality. These publicly available reports chronicle a variety of infringements, including (a) contracts not awarded to the lowest bidder or no bidding, (b) money channeled for the repair of buildings experiencing no damage, and (c) money given for no work. We use this information to create an index of corruption building on the various spending infringements recorded by the auditing agency in each municipality.

As Bulgarians, we were intrigued by the new and unique corruption data and the opportunity to study the drivers of corruption in this setting. Much of the corruption literature focuses on large countries such as India or Brazil (Bertrand, Djankov, Hanna, & Mullainathan, 2007; Ferraz & Finan, 2011) but the lessons from these studies may not travel as easily to a small Eastern European country. We were also hoping that our research would expose corruption and prompt the authorities to punish corrupt local politicians. While this (perhaps understandably) never happened, our work did receive lots of coverage in the Bulgarian and international media and inspired an important debate on how to deal with corruption.

Research Design

Our research question was whether municipalities receiving more flood-related assistance also engaged in more corruption. Our research made use of quantitative methods, and in particular regression analysis. Our dependent variable was calculated from the dataset detailing the various municipality-level spending infringements identified by the Bulgarian National Audit Agency (BNAA). Our corruption index sums all the recorded infractions for each municipality. Interestingly, exactly the same audit of flood-stricken municipalities was

also performed by a second agency, the Public Financial Inspection Agency (PFIA). However, unlike BNAA, PFIA is *not* politically independent (as it is part of the Ministry of Finance), which likely explains why PFIA recorded spending violations in less than a third of the municipalities in which BNAA detected corruption. Perhaps the strongest proof of PFIA's potential bias was the fact that PFIA was nearly five times more likely to underreport corruption in those municipalities in which the local politicians were affiliated with the party that appointed the head of PFIA (which, at the time, was the Movement for Rights and Freedoms, one of the parties in the governing coalition).

Our main independent variable was the amount of flood assistance received by each municipality (per capita), which was also publicly available in the BNAA reports. Although more flood aid should have been given to municipalities which were hit harder by the floods, this may not have been the case. For instance, charismatic mayors may have been more likely to extract flood aid *and* to commit and get away with more spending violations. Therefore, finding that municipalities which received more funds also engaged in more corruption may simply reflect the omission of mayors' personalities (which we do not observe) from the data (the so-called problem of 'omitted variables'). Similarly, inherently more corrupt politicians may be able to extract more flood-related assistance from central government, which leads to reverse causality. Once again, reverse causality means that we cannot interpret the results from a simple regression of corruption on flood aid as causal. The next two sections elaborate more on the approach we used to deal with this issue: the statistical technique of instrumental variables (IV).

Of course, flood assistance may be correlated with other municipality characteristics. Our regressions included a wide range of variables (covering years prior to 2004 to minimize the

issue of reverse causality) obtained from sources such as the 2001 Bulgarian Census and the Bulgarian Electoral Commission. We control for municipal economic conditions, which are log municipal income per capita, unemployment, and the net income the municipality received from privatization. We also account for the strength of media and civil society by including controls for local newspaper circulation per capita, the share of population with university degree, the share of urban population, and voter turnout in the 2003 local elections. Finally, we also include dummy variables for whether the municipal mayor and council belong to the ruling coalition, and whether the mayor and council belong to the party holding the disaster fund portfolio (the Movement for Rights and Freedoms (MRF), a party whose electorate comprises mainly Bulgarian Turks, controlled the ministries allocating the flood aid). In addition, in the robustness checks, we also control for the per capita amount of additional ad hoc funding received by each municipal government from the central government in 2003, as this may indicate prior incidence of politically motivated intragovernmental transfers.

Research Practicalities

In order to deal with the fact that flood assistance was likely *not* allocated based on the degree of flood damage, we implemented an instrumental variable (IV) approach. This is a statistical technique that aims to isolate random variation in the independent variable of interest (in our case, flood assistance). The particular technique we used is called two-stage least squares (2SLS), which, as the name suggests, proceeds in two stages. In the first stage, the independent variable is regressed on the instrument, along with all other controls. This *predicted* value of the independent variable instead of the *actual* independent variable is then included in the second stage. For the IV to work (or 'be valid'), three conditions must be satisfied. First, the instrument must be strongly correlated with the independent variable of

interest in the first-stage regression. Second, the instrument should not be correlated with the error term in the second-stage regression. And third, the instrument should not affect the dependent variable directly. We discuss how these conditions applied to our choice of IV below.

Coming up with a good instrument is sometimes nearly impossible, and designing a plausible instrument is more a matter of serendipity than of skill. In our case, we had luck on our side. While we knew that flood assistance may have been allocated for political reasons, at least some of it must have also been allocated to municipalities which were also hit hardest by the floods. In turn, how hard a municipality was hit by the floods was driven by (1) rainfall, which we knew was random; and (2) its geographic characteristics, such as closeness to river or elevation, which may not have been random. For instance, municipalities close to rivers may engage in more trade and be richer, which may make it easier for them to deal with the disaster even in the absence of flood-related transfers. Therefore, a measure of rainfall, conditional on geographic characteristics, would capture random variation in flood assistance.

The question, then, was how to obtain municipal-level data on rainfall over the period 2004-2005. We contacted Dr. Ivan Penkov, head of the Climatology, Hydrology and Geomorphology Department at Sofia University, who has conducted extensive research on torrential rains and the water sector in Bulgaria. He pointed us to the Bulgarian Institute for Meteorology and Hydrology, which collected monthly (but not daily, which we would have preferred) precipitation data in 101 weather stations around Bulgaria for both 2004 and 2005. However, according to its regulations, the Institute could only provide the data in exchange for a very large sum of money which was impossible for either of us to pay. The situation seemed hopeless.

We went on a quest for other sources of rainfall data. We explored precipitation data from NASA only to discard it because it was too coarse and would not be able to capture meaningful variation across small Bulgarian municipalities. Dr. Penkov suggested looking at a historical database recording floods across Bulgarian regions assembled by the Bulgarian Department of Water Resources in order to comply with EU regulations prior to Bulgaria's accession in 2007. We spent a lot of time and resources going through documents and coding the data only to find out that the flood record for 2004-2005 was quite patchy.

We then decided to seek contacts which could help us get the Bulgarian rainfall data free of charge. A friend's aunt who worked at the Bulgarian Academy of Sciences was unfortunately unsuccessful. We spoke to two other colleagues (Erik Berglof and Stefka Slavova whom we thank in the published article based on this research) who in fact knew several people in key political positions in Bulgaria at that time, and both of them offered to help. Although I must say that at some point we had given up hope, one day I received a call from Erik Berglof that the Bulgarian authorities had decided to provide the data for our research project free of charge. Since we had obtained and coded the rest of the variables (corruption, flood assistance and the other control variables described above), this meant that we could finally proceed with putting our research method in action.

Method in Action

Although we were delighted that we had finally managed to obtain the Bulgarian rainfall data which we needed, it (as we knew from the very start) only covered 101 municipalities. To deal with this, we had to adopt an interpolation procedure using a radius of 45 km and weights, which are the inverse of the municipality's distance to a station. We also had to

account for historical rainfall patterns across municipalities as it may always rain more in some municipalities as compared to others. Fortunately, Koleva & Peneva (1990) collected precipitation data for the period 1931-1985, which we interpolated using the same procedure. As this book was only available in the Sofia University library, Nikolay Marinov took a trip to Bulgaria to scan the relevant data. For each month, we were thus able to calculate a monthly rainfall percentage change relative to the monthly historical average. For example, for January 2004 we calculated the following quantity: ($Rainfall_{January\ 2004}$ - $Rainfall_{January\ 1931-1985}$)/ $Rainfall_{January\ 1931-1985}$. However, since the corruption index did not vary over time, in the regressions we had to use a single rainfall quantity. Therefore, for each municipality, we took the average rainfall value for all months in 2004 and 2005 for which the change relative to the historical value was at least 30% (we also experimented with alternative percentage cut-offs and obtained similar results).

Ideally, we would have preferred to have daily, rather than monthly, rainfall data, as intense rainfall usually happens over a period of one or several days. Unfortunately, such detailed data were unavailable, and it is likely that our month-based measure understates the intensiveness of the floods. It would have also been much better to have rainfall data for each municipality (and maybe even several observations per municipality), but such data did not exist.

To capture ground flood risk which would also affect flood damage (and thus presumably at least some of the flood assistance), we included three proxies. The first one is the number of settlements that are located within 1 km of a water body (dam, lake, or river), because households located close to water are more likely to experience flooding when there is extreme rainfall. The second one was average municipal elevation and slope, as flooding may

be more intense in municipalities located at a higher altitude and with a sloping terrain. The third one was latitude and longitude. We would have also liked to have a measure of historical flood management and readiness for each municipality (for instance, information on the extent of levee cleaning). We refer to qualitative evidence that the management of water resources deteriorated throughout the country since the early 1980s, with riverbed cleaning and the upkeep of levees and other protective equipment neglected due to lack of funds.

We also had several challenges when calculating our corruption index. BNAA groups infringements into four broad categories: (a) public procurement (e.g., no public procurement procedure was used by the municipality to select firms), (b) use of funds (for instance, there was payment for activities not listed in the contract), (c) reporting (for instance, no reports on fund use were sent to the Ministry of Finance), and (d) accounting and control (for instance, inaccurate accounting recording of the contracts).

Although the corruption measure is based on objective data, it is by no means perfect. First, we only know whether or not a municipality committed a particular violation related to the disaster assistance; there is no information on how much money was actually stolen. Second, a potential concern could be that our measure captures both corruption as well as fund mismanagement and misreporting, with the latter being distinct from corruption. We checked this carefully and found that auditors are provided detailed examples (including from real-world situations) on when misreporting constitutes corruption and are instructed to investigate further if the errors are committed purposefully or are purely accidental. Oversights that are likely to be associated with fraud are then recorded against the audit

criteria in the report, whereas purely administrative slips are listed in a separate section (the information in which we did not use in our analysis).

In addition, a potential concern could be that our corruption index weighs each infringement equally. To deal with this, we also calculated the corruption index using principal component analysis and found that this approach yielded very similar results. Principal component analysis is a data-reduction technique which extracts a common component from a set of variables using data dependent weights. For instance, a researcher can conduct principal component analysis on students' test scores for various subjects in order to extract information on their ability. What is more, we were lucky that a household-level survey with questions on corruption *perceptions* (the first round of the Life in Transition Survey which was jointly administered by the European Bank for Reconstruction and Development and the World Bank) was conducted in Bulgaria in fall 2006, shortly after the data from the audits was made publicly available. Although it is reassuring that corruption perceptions were closely aligned with the information obtained from our objective corruption index, this survey covered only a small subset of municipalities (37), while the BNAA audit covered 227 municipalities. This was unfortunate and we could have done more with the survey if it had covered more municipalities.

The biggest challenge we faced had to do with obtaining and assembling the data used in the analysis, particularly when it comes to data on rainfall and corruption. Analyzing the data via regression analysis (using Stata) was in fact the least difficult part of the research. As we hypothesized, our rainfall measure (the instrument) was highly correlated with the per capita amounts of flood assistance received by each municipality (the instrumented independent variable), which rendered the first-stage relationship strong. To recapitulate, the other two

criteria for instrument validity are that the instrument is uncorrelated with the error term in the second-stage regression and that the instrument does not affect the dependent variable (corruption in our case) directly. To make sure that the two conditions were satisfied, we explored scenarios that could potentially violate them. For example, it could be that heavy rains may make monitoring of reconstruction projects harder, which could create higher corruption, even in the absence of increased fund allocation. This would violate the assumption that the instrument (rainfall) affected corruption only via the instrumented variable (municipal flood assistance). However, most of the flooding episodes were isolated and happened in 2004 and 2005, or at least 4 months before the audit started. This means that each municipality had sufficient time to observe reconstruction activities. In addition, each municipality was required to oversee building works and send reports (along with photographic evidence) to the central government. This process was coordinated locally, suggesting that monitoring was relatively straightforward.

Practical Lessons Learned

Our research was based on a serendipitous idea: to examine the link between municipal financial assistance intended to deal with the destruction following the 2004-2005 floods in Bulgaria, and local corruption. Obtaining the corruption data was relatively easy as the BNAA audit reports were publicly available. However, obtaining the rainfall data (which we used to construct the instrument for flood assistance) was extremely challenging. Although we had a very strong research idea, the lack of rainfall data could have derailed our whole project. Therefore, the first lesson that we learned is that researchers have to be persistent and utilize their networks effectively in order to obtain the data that they need.

The second lesson is that every good idea must be executed properly. A convincing IV regression requires not only an instrument, but also the inclusion of multiple observable characteristics which could be correlated with the instrument, the instrumented variable, or the dependent variable. As explained above, our analysis controlled for a variety of municipal-level characteristics, such as ground flood risk, municipal economic conditions, the strength of media and civil society, and political characteristics. Collecting so much data is not easy. For example, we could only obtain newspaper circulation data at the regional level (which is more aggregated that the municipality level) and had to make several inquiries (and pay a small fee) to the Bulgarian Statistical Office.

The third lesson is that data are imperfect. For instance, our corruption measure only recorded whether a spending infringement had taken place or not, without providing information on how much money was actually stolen. It also raised other questions. Did our index capture fund mismanagement and misreporting, along with true corruption? Was it appropriate for us to give equal weighting to each infringement? If we wanted to weigh some infringements more than others, how would we decide which infringements mattered most? And how did our objective corruption measure square with survey-based measures of corruption perceptions? We spent a lot of time trying to answer these (and other) questions relating to our data, running alternative specifications and looking for alternative data sets (such as the Life in Transition Survey) which could be useful for our story. The key was to put ourselves in the shoes of a critical reader and to anticipate the kinds of questions related to the theory or analysis that this reader may ask. Then, we had to think about how these issues can be resolved or at least minimized. An imperfect solution is still better than no solution.

The fourth, and most important, lesson: to have fun with research. Research is messy, frustrating, and does not go according to plan. In this project, we spent a long time waiting around for rainfall data or trying rainfall measures that ultimately did not work out. Although theoretically our idea that rainfall could be used as instrument for flood assistance made sense, we were not sure whether the data would confirm our theoretical expectations. As it turns out, they did. If they had not, then we would have had to step back and re-evaluate our theoretical priors, something which can seem frustrating at first but is how really interesting ideas are born. For instance, could the link between rainfall and flood assistance be stronger in municipalities where the population was better able to hold political elites accountable for their behavior? Could this accountability mechanism be captured by characteristics such as municipal size or the educational level of the population? Research setbacks are very often invitations to change perspective, to expand one's thinking and to produce high-quality research.

Conclusions

The project discussed in this case focused on exploring the link between flood-related assistance and corruption at the municipal level in Bulgaria between 2004 and 2005. We were able to obtain unique *objective* corruption data from a municipality audit conducted by an independent national watchdog (the Bulgarian National Audit Agency). Our objective measure is arguably more precise compared to perception-based proxies of corruption which are typically used in the literature due to the lack of other data.

Although more disaster assistance should have gone to municipalities which were more affected by the floods, this may not have been the case. To deal with this, we made use of an instrumental variable approach.

While our research design was theoretically sound, putting it into practice was challenging. Obtaining the corruption data was straightforward as the municipality audit reports were publicly available. Collecting municipal economic, demographic and political characteristics was also not difficult. However, obtaining the rainfall data necessary for the construction of our instrument was difficult and took a lot of time and persistence. In the end, we were fortunate that colleagues from our networks were able to help us obtain the rainfall data free of charge. We also had to make many decisions regarding our imperfect data and coding and to run many robustness specifications. Good research takes time and effort, and all of this work made our research results more convincing.

Exercises and Discussion Questions

1. What is interesting about the corruption data described in this case?
2. Why did the authors use an instrumental variables approach?
3. What was the instrument in this particular research project and how did the authors ensure it was valid?
4. What kind of municipal characteristics did the authors include in the regressions and why did they have to do that?
5. Could you think of other research questions which would be suitable for applying the methods discussed in this case?

Further Readings

Aidt, T. S. (2003). Economic analysis of corruption: A survey. *Economic Journal*, *113*, F632-F652.

Djankov, S., Nikolova, E., & Zilinsky, J. (2016). The happiness gap in Eastern Europe. *Journal of Comparative Economics*, *44(1)*, 108-124.

Nikolova, E., & Marinov, N. (2017). Do public fund windfalls increase corruption? Evidence from a natural disaster. *Comparative Political Studies*, *50(11)*, 1455-1488.

Reinikka, R., & Svensson, J. (2005). Fighting corruption to improve schooling: Evidence from a newspaper campaign in Uganda. *Journal of the European Economic Association*, *3*, 259-267.

Rose-Ackerman, S. (1999). *Corruption and government: Causes, consequences, and reform*. Cambridge, UK: Cambridge University Press.

Web Resources

Transparency International, https://www.transparency.org/

Nikolova, E., & N. Marinov. (2015, April 22). How disaster relief can increase corruption [Blog post]. Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2015/04/22/how-disaster-relief-can-increase-corruption/?noredirect=on&utm_term=.12c12ce12680

References

Bertrand, M., Djankov, S., Hanna, R., & Mullainathan, S. (2007). Obtaining a driver's license in India: An experimental approach to studying corruption. *Quarterly Journal of Economics*, *122*, 1639-1676.

European Commission. (2017, November 15). Report from the Commission to the European Parliament and the Council: On Progress in Bulgaria under the Co-operation and Verification

Mechanism. Retrieved from https://ec.europa.eu/info/sites/info/files/comm-2017-750_en_0.pdf

Ferraz, C., & Finan, F. (2011). Electoral accountability and corruption: Evidence from the audits of local governments. *American Economic Review*, *101*, 1274-1311.

Koleva, E., & Peneva, R. (1990). *Klimatichen Spravochnik (Valeji v Bulgaria)* [Climate Guide (Rainfall in Bulgaria)]. Sofia: Bulgarian Academy of Sciences. (In Bulgarian).

Olken, B. A. (2009). Corruption perceptions vs. corruption reality. *Journal of Public Economics*, 93, 950-964.