A Ph.D. Student's Approach to Data Availability Challenges

Description

By Adriana Cassis

As is the case in many other areas of research, scholars of immigrant political participation face challenges of data availability and accessibility. This problem affects the type of questions that can be researched and the methods that can be applied. One solution is to collect one's own data, but for doctoral students like myself, a more plausible answer to getting around this problem is to use existing data in innovative ways. In this blog entry, I will describe how I merged existing data sets to solve my data needs.

Initiatives such as the Ethnic and Migrant Minorities (EMM) Survey Registry demonstrate the deliberate efforts of scholars in the field of migration studies to make data available and accessible. The thousands of surveys collected in the survey registry cover a wide range of topics and methodologies. Many are EU-funded projects with multiple survey waves in several countries while others are nationally funded and have a subnational scope. This demonstrates that the problem is not an absence of data on the immigrant population.

The challenges are the availability and scope of the data. When data is available, there are issues regarding the themes and populations covered, which hinders the ability to make comparisons across locations, time periods, and topics. For example, if you are interested in the political behavior and participation of immigrants, you will quickly find that many of the surveys that cover this topic are national in scope. Consequently, to compare across countries you may need to rely on multiple surveys from different timepoints and with varying methodology. Or, if there is a longitudinal multi-country survey, chances are that some key measurements are missing.

Faced with such problems, I decided that my best bet was to take different existing surveys and essentially merge them into a new dataset. That way I could pick and choose the data that best fit my needs. However, this was much easier said than done, since Ifirst had to find surveys that had data I was interested in, make sure that it was actually available in full, and analyze whether the surveys were similar enough to merge.

For my research purposes, I needed data on the transnational political participation of Turkish immigrants that explicitly controlled for immigrant-based characteristics. I decided to use the Immigrant German Election Study I (IMGES I) and the Dutch Ethnic Minority Election Study (DEMES). This decision came after a review of countless codebooks, which were my main source of information and through which I assessed how similar the surveys were, both methodologically and substantively. Both DEMES and IMGES sampled the population I wanted to study, took place around the same time period, and were comparable in terms of my main research topics.

After finding the surveys came the most difficult task: wrangling the data. This requires the researcher to become fluent in two separate datasets in order to merge them into a new dataset. In practice, this means relabeling and recoding variables so that they make sense in both content and form. For example, to construct a variable of self-identification, I had to choose from several questions in the DEMES questionnaire that asked respondents about their connection to the Netherlands, their pride in being Dutch, their sense of belonging to their country of origin, or their view of themselves as citizens. In the IMGES questionnaire, on the other hand, I could only rely on one question about how respondents would describe themselves, which again had several possible answers. After selecting the appropriate IMGES and DEMES self-identity variables and relabeling them, I had to recode them so that they were on the same scale. This process was repeated for all the variables of interest.

Although merging the data sets is a cost-effective solution, it is still a time-consuming process that has inherent limitations. One limitation of this approach is that some topics cannot be researched because they are not measured in both surveys. Another limitation is the appropriateness of merging variables that collect information from differently worded questions. For instance, the question of whether one has a strong sense of belonging to a country is distinct from the question as to whether one would describe oneself as a national of that country. This raises questions about the validity of assuming that they are

measuring the same concept. Lastly, there is the issue of proper weighting after merging datasets, which must be addressed on a case-by-case basis.

While merging data sets is neither a foolproof nor a revolutionary method, it provides a solution to the problem of data availability. Furthermore, it is a practical solution for times when gathering one's own data is not possible, yet the research questions require data that cannot be found in one single dataset. It is indeed this technique that makes the analys of patterns of transnational political participation, with data that has never before been used together, possible.

References

Goerres, Achim, Spies, Dennis C., & Mayer, Sabrina (2023). Immigrant German Election Study (IMGES). *GESIS, Cologne. ZA7495 Data file Version 2.0.0, https://doi.org/10.4232/1.14187*

Morales, Laura, Saji, Ami, Méndez, Mónica, Bergh, Johannes, & Bernat, Anikó. (2020, May 25). The EMM (Ethnic and Migrant Minorities) Survey Registry. Zenodo. http://doi.org/10.5281/zenodo.3841433

M. Lubbers; T. Sipma; N. Spierings (2021): DUTCH ETHNIC MINORITY ELECTION STUDY 2021 (DEMES 2021). V2: DANS Data Station Social Sciences and Humanities.

Date Created
Oktober 28, 2024
Author
politikwissenschaft_h1c5yk