Subnational V-Dem in Colombia

Merging V-Dem with external data

Patrick McQuestion[*], Michael Coppedge[†], Matthew Sisk[‡]

First version: February 4, 2024. This version: 29 March 2024.

Mapping

First, we decide on the level of analysis for our study. In this case, we observe that Colombia has fairly complete municipal-level data dating back to at least the 1950s, at least for a relevant amount of time-varying variables (e.g., population density, economic development, voting behavior, etc.).

Second, we collect and format data for mapping purposes. This is a multi-step process that involves grouping responses in some cases (e.g., responses 6-9 that ask to identify 4 cardinal directions). Grouping is necessary for some variables, but not all. Once grouped, the response categories themselves must be "translated" into locally relevant information, which requires a degree of subjective decision-making. For example, how many votes or proportion of votes are necessary to measure "strong" support for the ruling party (question 15)? Or, which municipalities are considered part of the "North" (question 6) versus the "West" (question 8)? We privilege the use of granular data that will allow researchers flexible interpretations of these questions, as well as subsequent validation from country experts.

Third, we export the response grouping dataframes to `.csv` format that can be added as data in GIS software. In ArcMap, this can be done by joining data to the attribute table for the base layer (in this case, the shapefile provided by DANE). We then create and format layers based on the response groupings; for each response group, we visualize the locations where free and fair elections are held (and not held), and where civil liberties are stronger (or weaker).

[*] Department of Political Science & Kroc Institute, University of Notre Dame. Email: pmcquest@nd.edu

[†] Department of Political Science, University of Notre Dame. Email: mcoppedg@nd.edu

[‡] Department of Applied and Computational Mathematics and Statistics, University of Notre Dame. Email: msisk1@nd.edu

Data Wrangling

We undertake a series of steps to wrangle data from V-Dem and external sources. In order to facilitate merging tables to the shapefile maps of Colombia in ArcMap, we use a key ID (either the official DANE code for each municipality, or the ArcMap FID number). In the case we prefer to use the `sf` package, we can proceed to merge along the DANE code.

Note: This code below concentrates on cross-sectional data, but it will expand to time-series cross-sectional (~1900-2022).

Exercise: 2018 cross-section

This exercise is designed test the viability of mapping cross-sectional country-level data before expanding the dataset into panel format. Therefore, internal data from V-Dem is subset to 2018, and last five subnational responses are removed because they are not relevant for Colombia.

Table

External data from Colombia is drawn from a variety of sources. We merge this data at the municipal level using DANE codes. These codes are also used in .dbf tables, such as Colombia's 2018 Census data from DANE. The Census data Shapefile is useful for initial mapping visualization, containing not only the administrative boundaries but also contextual information. The dictionary is in the repository.

Note: To facilitate integration of the datasets across software, we export the Census 2018 attribute table from ArcMaps to .csv format, then load this data into R, preserving only relevant variables for merging purposes. While we use DANE codes to merge external data using R, we use the 2018 Census variable "FID" to merge .csv files into ArcMap. This is because ArcMaps automatically translates some string variables (such as the DANE codes) into a numeric ones. For this reason, we integrate the FID variable for the individual .csv files we write.

0-1: Rurality

The survey asks about elections and civil liberties in rural or urban locations. Rurality or urbanization can be measured in many ways. For example, the World Bank's Rural Access Index (RAI) or European

Commission's Global Human Settlement Layer (GHSL) are exemplary. Similarly, Waldorf and Kim (2015) create an index that considers population size, density, remoteness, and built-up areas. DANE (2015) uses a classification based on OECD criteria that defines rural territories as those with 150 or fewer persons per kilometer squared, finding that over 75% of Colombian municipalities do not meet the threshold of 100 persons per kmˆ2. From another perspective, DANE estimates that close to 4/5 of the population lives in urban centers. In this exercise, we take a demographic approach that considers the proportion of the municipal rural population over the total municipal population. This basic "Rurality Index" measure created by CEDE uses Census data (DANE).

The survey asks two questions, however, the systematized concept can take many forms. In this case, our V-Dem scores reflect the proportion of respondents who selected rural, urban, or both categories to describe election fairness and civil liberties. We can interact these proportions per question with the rurality index from CEDE (also a proportional score) to provide a more granular view of democracy at the municipal level based on our particular systematization. Another option is to create a dichotomous threshold between urban and rural variables and assign V-Dem scores accordingly. We can take the mid-point as a threshold for rural and urban municipalities: if more than half of the population is considered rural inhabitants, then the municipality is considered rural. This can be adjusted later if necessary. - As a first cut a good way to start

2-3: Economic development

DANE provides data for Gross Domestic Product (GDP), known as Producto Bruto Interno (PBI) per capita. They offer a historical "retropolation" of department-level PBI dating back to 1980, and a municipal-level measure of "value added" dating back to 2011 (methodological documentation is included in the repository). Values are in Colombian pesos (COP) and scaled to $1,000,000,000.

The metropolition data extends farther back in time, making it more amenable to panel data analysis. Also, there may also be reduced measurement error given the difficulty of isolating economic productivity to municipalities.

Most departments are covered by the retropolation measure, but newer departments (many of which were inducted in 1991) are grouped into one unit, and therefore cannot be mapped based on DANE's Departmental Code system without data intervention. These are: Amazonas (91), Arauca (81), Casanare (85), Guainía (94), Guaviare (95), Putumayo (86), San Andrés, Providencia y Santa Catalina (Archipiélago) (88), Vaupés (97) y Vichada (99).

Nearly all municipalities are measured by the Municipal Value Added (VAM) measure, however, data only exists from 2011 onward.

For experimental purposes, we try both measures for 2018.

After creating the data frame for economic development at municipal and departmental levels, we must conceptualize "less" versus "more" developed areas before merging with V-Dem data. A simple option could be to dichotomize economic development based on median values. A more complex method would interact these municipal measures with V-Dem responses, as we did with rurality, but this would involve scaling and other steps beforehand.

Unclear how to set up the interaction option.

4-5: Inside or outside of capital city

The survey asks about elections and civil liberties "Inside / Outside the capital city". In the context of V-Dem this likely refers to the national capital (Bogota), but it could also refer to municipal districts that are also department capitals. CEDE contains 2 variables of interest for systematization of this concept: distance from Bogota and distance from department capital. Even though we can assume respondents were considering the national capital, We will keep both in case we consider the alternative interpretation. Note: It may be possible to calculate this variable in ArcMaps using the field calculator, but this requires advanced techniques.

Although the survey asks two questions, technically these responses can be grouped into one dichotomous variable. We will therefore create one variable for each survey pair of survey questions that assign dichotomous values (0=capital, >0=non-capital). Nonetheless, we can also consider the relevance of proximity following a continuous scale. We therefore calculate this scale for the national capital distance using the range.

6-9: Cardinal directions

In order to map the four cardinal regions in Colombia, we are guided by external data from DANE who recently classified the country into 6 macro-regions. We find a public-access table of municipalities classified into these regions here. There are other divisions, such as the 5 regions used by CEDE studies (Andina, Caribe, Pacifica, Orinoquia, Amazonia), but for now we go with the prior classification.

We decide to group the regions in the following manner (see code below), creating a new variable called "Cardinal". Then, we merge our V-Dem data related to responses #6-9 for each subnational survey question.

10: Civil unrest

11: Illicit activity

12: Sparse population density

- will likely covary with rurality

13: Remoteness

- will likely covary with rurality

14: Indigenous

15-16: National ruling party

The questions ask "Areas where national ruling party or group is strong / weak." The most straightforward measure is a dichotomous variable where president receives a majority of the vote or not. A more variegated measure can leverage the percentage of votes cast for the president-elect vs. runner-up, for example, however this measure may need to account for blank votes (which may signify protest votes).

The measure for ruling party support used here will be: % of vote that supports ruling party. For first wrangle, I will choose 2018 runoff election (second round) between Petro and Duque (Duque won, so his is ruling party). This makes things easier because there are only 2 candidates, the winner and runner-up. Elections without second rounds (prior to 1994, as well as 2002 and 2006) may require additional coding.

The voting data requires more steps for cleaning. In this exercise, we will take all possible vote entries: top 2 candidates, as well as no mark, blank, or null ballots. We see that Duque won by high margin (over 26% on average) in 2018.

After creating the data frame for voting behavior at the municipal level, we must decide on the threshold for "strong" versus "weak" support of the ruling party. We could take one standard deviation or higher as the threshold, but even so: what is the criteria? In the interest of simplicity, we will create a dichotomous measure, with margin of victory (MOV) above 10 percentage points, applied in either direction (positive for winner Duque; negative for loser Petro). This seems like a standard difference, but further research could guide this threshold setting. - ideally, don't dichotomize. Use a continuous scale. - cl stronger where RP is stronger and experts agree thats

Exporting data

Mapping data can be exported to a shapefile then imported into `sf` or ArcMap

Conclusion

Appendix