POLEVPOP Bio Data Pseudonymization Protocol

Content

The bio data

The data to collect

Each country team has to collect the information listed in Table 1. Please note that some information must only be collected for participating politicians, others must be collected for all.

There are two important differences with regard to the document about bio data collection that has been distributed earlier.

1. Initially, we said that 'Top_politician' and 'Local_mandate' *only* needed to be collected for participating politicians. However, given the potential for this information to identify politicians, it was decided to collect this for all and not just for those participating.
2. Residence has been dropped from the list. The reasons for this is that residence too often reveals the identity of a politician, it cannot be easily recoded into larger categories, and the large number of unique values massively slows down the pseudonymization process. Moreover, we realized the variable was ultimately never used in previous projects. For the latter reason, com_chair has also been dropped.

Table 1: Variables to collect for all politicians

	Variable name	Definition
Gather for all politicians	Parliament	In which federal or regional parliament is the person active? [string]
	National	Is the person active at the national level or at a regional level? [0: regional; 1: national]
	Constituency	Name of the constituency/district in which the politician was elected [string] *>> Code as missing (-99) if the politician did not take part in the previous elections*
	Dm	Number of seats that could be won in the district in which the politician ran for the most recent general election (district magnitude) [numerical variable]
	Total_seat	The total number of seats in the parliament where the politician seats [Numerical variable]
	Party	Party: which political party is the politician a member of? [string]
	Party_seat	Number of seats the politician's party won in the most recent elections at this level of governance in the whole parliament. [Numerical variable] *>> This variable refers to the complete parliament, not to the district and/or individual lists*
	Gov	Is the politician a member of a governing/majority party on the level the politician is active on? [0:no;1:yes]
	Sex	Gender of the politician [0: male; 1: female; 2: other]
	Yob	Year of birth of the politician [year]
	Yofe	The year of first election in the parliament the politician is seated in [year]

	Yofe_total	The year of first election in the same or other relevant parliament (regional/federal/European). Lower levels of government such as municipalities do not count. [year] *>> should only be collected in a country if there are other relevant parliaments apart from the one the politician is seating in (e.g. in Flanders, politicians often switch between the Federal and Regional Flemish parliament, but apart from the topics they work on, their job is almost the same)*
	Top_politician	Is the politician currently speaker or formerly speaker of the house, currently or formerly a parliamentary party group leader, currently or formerly a party leader, or currently or formerly a minister or junior minister? [0: No; 1: yes]
	Local_mandate	Is the politician currently a local mayor, alderperson, or council member? [0: No; 1: yes] These mandates should be held at the municipal level, or the country-equivalent of municipalities.
	Participation	Has the politician participated in the survey? [0: No; 1: yes]
Gather for participating politicians	Com_1(-x)	Parliamentary committees in which the politician is seated [string]

The data to send

Only the bio data collected for all politicians needs to be sent to the Antwerp team for the creation of the master data file. Information on parliamentary committee membership is too specific and will almost certainly result in the identification of politicians. This does not mean it is not important to collect it. At a later stage, it will be used to create measures of issue specialization for different concrete issues dealt with in the survey (but more information on that later). The dataset should also include a unique identifier that will allow for the pseudonymized bio data to be merged with the survey responses. This identifier should be made specifically for the pseudonymization process and should have no relation to other ids.

The variables must be collected and stored precisely in the manner described. Columns containing numeric variables must only contain numeric values, etc. For missing values (because information is unavailable or not relevant), use -99 in all cases, whether it is a numeric or a string variable. Table 3 shows a fictional example of what a dataset should look like. Please be sure to store the data in a .csv format.

This is where the tasks of the various country teams end. Please send the dataset to christophe.lesschaeve@uantwerpen.be. The remainder of this document explains the pseudonymization process. However, it is done entirely by the Antwerp team.

Table 3: Fictional example of the dataset

Parliament	National	Constituency	Dm	Total_seat	Party	Party_seat	Gov	Sex	Yob	Yofe	Yofe_total	Top_politician	Local_mandate	Participation
Parliament A	0	-99	-99	100	Party A	10	1	1	1980	2010	2010	0	1	0
Parliament A	0	Constituency B	10	100	Party A	10	1	0	1985	2005	2005	0	1	1
Parliament A	0	Constituency C	12	100	Party A	10	1	1	1970	2003	2003	0	1	0
Parliament A	0	Constituency D	13	100	party B	5	0	0	1965	2010	1995	1	0	1
Parliament B	1	Constituency E	5	150	party B	8	1	1	1960	2015	2004	1	0	1
Parliament B	1	Constituency F	7	150	party B	8	1	0	1950	1995	1995	1	1	1
Parliament B	1	Constituency G	8	150	party B	8	1	1	1966	1999	1999	1	1	0
Parliament B	1	Constituency H	7	150	Party C	10	0	0	1972	2011	2011	0	1	1
Parliament B	1	Constituency I	3	150	Party C	10	0	1	1976	2020	2020	0	1	1

The pseudonymization algorithm

Pseudonymization will ensure that a single politician is not identifiable with 100% certainty among the total population of politicians. It is not necessary to be anonymous among the sample of participating politicians. For instance, if only one woman from Party A participates, she would be identifiable in the participating sample if one knows her gender and party affiliation. But if there is another woman among the politicians of Party A – who then did not participate in the survey – then the participating politician cannot be identified with 100% certainty, as there is doubt which of the two the responses belong to. Note that it is therefore important never to publish who participated in our research and who did not.

The pseudonymization process inspects every combination of values on the total set of variables (apart from Participation, obviously) to see if any combination results in the identification of a single politician. The overview below gives an overview of the main principles underpinning the pseudonymization algorithm.

Recoding variables

Continuous variables are, by virtue of the many unique values they contain, very likely to result in the identification of a politician. For that reason, their values are recoded into larger categories. The new categories are country-relative, taking into account contextual differences. The number of categories is determined by whether the variable is a control rather than a variable of key theoretical interest. For the purpose of pseudonymizing the data, large and equally sized groups tends to minimize the information loss.

Age categories

Age is computed by subtracting Yob from 2022. After, age is recoded into two categories based in the median age. This creates two groups of roughly equal size, where in the first Age = low, and in the second Age = high.

Year of first election (Yofe) and year of first election total (Yofe_total)

Yofe(_total) is recoded into three categories based on tertiles. This creates three groups of roughly equal size, where in the first Yofe(_total) = low, in the second Yofe(_total) = medium, and in the third Yofe(_total) = high.

District magnitude (Dm)

District magnitude (Dm) is recoded into the following categories:

1 = 1-3 seats
2 = 4-6 seats
3 = 7-10 seats
4 = 11-20 seats
5 = >20 seats

However, when the results of the pseudonymization show that the loss of information is too great, Dm will recoded into two categories based in the median Dm, within each parliament to take into account

intra-country differences. This will create two groups of roughly equal size, where in the first Dm = low, and in the second Dm = high.

Party seats

The variable Party seats is recoded into three categories based on tertiles, within each parliament to take into account intra-country differences. This creates three groups of roughly equal size, where in the first Party seats = low, in the second Party seats = medium, and in the third Party seats = high.

Step 1: Party affiliation

Party affiliation is the most important variable in the dataset, and the aim is to preserve it as much as possible. However, in instances where there is only one politician from a party, party affiliation results in 100% certain identification. In the example below, 'Id' is the unique identifier for each politician, 'Party' indicates party affiliation, and 'n' indicates how many politicians from that party there are in the dataset.

If there are multiple parties with only one politician, it is enough to set party affiliation for those politicians to 'other' (-95), as shown in Scenario 1 below. Blue marks the records in the (fictional) data that are affected by the pseudonymization process. Politicians 6 & 7 are one-person parties, and knowing their party affiliation reveals who they are. To avoid this, their party affiliation is set to missing, concealing their identities.

However, if only one party has only one politician, setting party affiliation to missing will still make the politician identifiable, by process of elimination. In that case, the politician must be excluded from the dataset, as shown in Scenario 2 below. Party 3 is the only one-person party, and setting that affiliation to 'other' would allow for the identification of the politician. Therefore, that record is removed entirely.

Scenario 1

Before pseudonymization

Id	Party	n
1	1	5
2	1	5
3	1	5
4	1	5
5	1	5
6	2	1
7	3	1

➔

After pseudonymization

Id	Party	n
1	1	5
2	1	5
3	1	5
4	1	5
5	1	5
6	-95	2
7	-95	2

Scenario 2

Id	Party	n
1	1	5
2	1	5
3	1	5
4	1	5
5	1	5
6	2	2
7	2	2
8	3	1

➔

Id	Party	n
1	1	5
2	1	5
3	1	5
4	1	5
5	1	5
6	2	2
7	2	2

Step 2: Party affiliation plus one additional variable

In the example below, 'Variable A' is one of the collected bio characteristics. In the scenario 1, column n indicates that politician 4 & 5 are identifiable if one knows their party affiliation and their value on variable A. The solution is simple: set the values for variable A to 'other' for both, rendering them pseudonymous.

In scenario 2, only one politician is identifiable if one knows their party affiliation and their value on variable A. Setting the value for variable A to 'other' for politician 4 is insufficient, as again by process of elimination, the value can be recovered. A solution would be to set the values of variable A to 'other' for all politicians of party 1. Yet this would be throwing away a lot of potentially useful information. A more elegant solution is proposed. In the case of scenario 2, the value for variable A for politician 4 is set to 'other', and a second politician of party 1 is selected at random whose value for variable A is also set to 'other'.

This approach minimizes the information lost, at the expense of a consistent pseudonymization. If the algorithm is executed a second time on the same dataset, the odds of same politician being selected to ensure pseudonymity are slim. However, we believe the benefits outweigh the costs.

Scenario 1

Before pseudonymization

Id	party	Variable A	n
1	1	1	3
2	1	1	3
3	1	1	3
4	1	2	1
5	1	3	1

➔

After pseudonymization

Id	party	Variable A	n
1	1	1	3
2	1	1	3
3	1	1	3
4	1	-95	2
5	1	-95	2

Scenario 2

Id	party	Variable A	n
1	1	1	3
2	1	1	3
3	1	1	3
4	1	2	1

➔

Id	party	Variable A	n	random number
1	1	1	2	0.7019
2	1	-95	2	0.3718
3	1	1	2	0.9154
4	1	-95	2	

Step 3-14: Party affiliation plus more than one additional variable

Adding an additional level of complexity, the example below adds a second variable B. Combining Party X variable A, Party X variable B, and variable A X variable B does not result in the identification of any politician. However, combining all three does lead to the identification of politician 9 and 10 in scenario 1. To avoid identification, it is again sufficient to set their values to 'other'.

The second scenario, only politician 12 is identifiable, yet setting variable B to 'other' only for that politician would not solve the problem. Again, another politician of party 2 and value 2 on variable A is selected at random and set to 'other' too. This makes politician 12 no longer identifiable.

Scenario 1

Before pseudonymization

Id	party	Var. A	Var. B	n
1	1	2	1	2
2	1	2	1	2
3	1	2	2	2
4	1	2	2	2
5	2	1	1	2
6	2	1	1	2
7	2	1	2	2
8	2	1	2	2
9	2	2	2	1
10	2	2	1	1

➔

After pseudonymization

Id	party	Var. A	Var. B	n
1	1	2	1	2
2	1	2	1	2
3	1	2	2	2
4	1	2	2	2
5	2	1	1	2
6	2	1	1	2
7	2	1	2	2
8	2	1	2	2
9	2	2	-95	2
10	2	2	-95	2

Scenario 2

Id	party	Var. A	Var. B	n
1	1	2	1	2
2	1	2	1	2
3	1	2	2	2
4	1	2	2	2
5	2	1	1	2
6	2	1	1	2
7	2	1	2	2
8	2	1	2	2
9	2	2	2	3
9	2	2	2	3
11	2	2	2	3
12	2	2	1	1

➔

Id	party	Var. A	Var. B	n	random number
1	1	2	1	2	
2	1	2	1	2	
3	1	2	2	2	
4	1	2	2	2	
5	2	1	1	2	
6	2	1	1	2	
7	2	1	2	2	
8	2	1	2	2	
9	2	2	2	2	0.6437
10	2	2	2	2	0.4265
11	2	2	-95	2	0.2147
12	2	2	-95	2	

Pseudonymization and participation

In the example below, it is clear that politician 4 is identifiable with information on his/her party affiliation and variable A. However, politician 4 did not partake in the survey, and the data does not contain any potentially sensitive information about him/her. For the purposes of the pseudonymization process, there is no need to set any value to 'other'. The previous examples thus only apply to instances where the affected politicians participated.

Before pseudonymization

Id	party	variable A	n	Participation
1	1	1	3	1
2	1	1	3	0
3	1	1	3	1
4	1	2	1	0

➔

After pseudonymization

Id	party	variable A	n	Participation
1	1	1	3	1
2	1	1	3	0
3	1	1	3	1
4	1	2	1	0

Order of importance

In the previous examples, identification is avoided by sacrificing information on one of the variables involved. Which variable is sacrificed is determined by a rank-order of importance. This order is given below in Table 4 below. When a combination of values on a combination of variables results in the identification of one politician, information is sacrificed in the lowest ranked variable.

Table 4: Rank order of variable importance

Variable	**Rank order**
Party	1
Top_politician	2
Gov	3
Yofe	4
Yofe_total	5
Sex	6
National	7
Party_seat	8
Total_seat	9
Parliament	10
Yob	11
Dm	12
Local_mandate	13
Constituency	14

Publishing with the politicians dataset

The pooled politicians dataset will only contain the pseudonymized bio variables. However, country teams still have the additional variables and they have the original, more fine-grained versions of variables such as age and district magnitude. This allows individual countries to publish with the original, detailed data (single country studies). Moreover, it is possible for anyone to ask other teams to run analyses in a 'decentralized manner' on the non-pseudonymized dataset. That is, it should not be much effort for a country team to run a fully prepared script on their data and deliver an outroll of the output.

In such instances (where non-pseudonymized biodata are used for paper analyses), it is not allowed to *make public* (for replication purposes) any biodata that was not pseudonymized. Replication datasets must therefore leave out the extra bio data, or include the recoded and pseudonymized variables instead of the originals, with an obvious negative effect on the ability to repeat the analyses of the paper. Researchers can seek exemptions of replication requirements, which many journals offer. Or, a designated person from the journal can be given access to the non-pseudonymized data on the condition that they sign a confidentiality agreement.