Chapter five

Collecting, Cleaning, and Matching Electoral Data on Candidates and Parliamentarians in Poland, 1985–2011

Zbigniew Sawiński and Joshua Kjerulf Dubrow

This chapter presents the process of collecting, cleaning, and matching the East European Parliamentarian and Candidate data (EAST PaC) for Poland. These data span every national election for the Sejm (1985–2011) and Senat (1989–2011). EAST PaC is based on a former dataset, POLCAN (Polish Candidate data) 1989–2007, that was continually expanded, and was described and used for substantive research questions (e.g. Shabad and Słomczyński 2004; Zielinski et al 2005, Dubrow 2011; Kunovich 2012). EAST PaC Poland has expanded the time span of these data (back to 1985 and forward to 2011) and re-cleaned them. This chapter describes this process in detail.

COLLECTING DATA

The data were collected from official sources and supplemented with publicly available information. Official data include the election report (Wyniki Wyborów do Sejmu), the official webpage of the Polish parliament (Strona Internetowa Sejmu i Senatu), and the annual statistical almanac (Rocznik Statystyczny). When data are incomplete, they are supplemented with publicly available information from personal and media websites. Earlier versions of the data were collected from non-electronic sources. The Polish National Election Commission (PKW) has, since the 2000s, placed all available information online. Demographic variables are limited to gender, age, and occupation.

Electoral and party information includes votes received, whether they won the election and whether to Sejm or Senat, voivodship (administrative district of balloting), list position, and party affiliation. EAST PaC only records this information at the point of election. Countries in Eastern Europe have high levels of party switching (Shabad and Słomczyński 2004) but this is not recorded in EAST PaC between elections.

All characteristics of the candidates are self-reported to the PKW via candidate forms. Thus, the data are subject to errors in reporting. We investigated the case of occupational categories to illustrate the sources of these errors, and thus how best to interpret the variables created out of them.

Occupational Categories

Occupation is self-reported, and reported to the PKW. In most cases the descriptions of occupations are rather general, like "a farmer", "a lawyer" or "in science". Akin to the Ukrainian data (see Pohorila, this volume), occupation is more of a presentation of self than an objective indicator. To understand this, we describe the process of reporting and coding occupation for candidates.

According to the Polish electoral law candidates have to report their occupation [zawód] to the electoral commission[1]. In Polish

[1] As stipulated in the electoral law from 1991 ("Ordynacja wyborcza do Sejmu Rzeczpospolitej Polskiej 1991", Art 68. 4.; attached), 2001 ("Ordynacja wyborcza do Sejmu Rzeczypospolitej Polskiej i do Senatu Rzeczypospolitej Polskiej 2001" Art. 144. 1.) and 2011 (Kodeks Wyborczy Art. 212. § 1.; attached): "Zgłoszenie listy okręgowej powinno zawierać nazwisko, imię (imiona), zawód i miejsce zamieszkania każdego z kandydatów." The English translation is, "The submission of the district list should include surname, name (names), occupation and place of residence of each candidate". The text in the law from 1991 is slightly different, but the meaning is the same. See also the document issued by PKW in 2005 ("Wyjaśnienie Państwowej Komisji Wyborczej z dnia 8 sierpnia 2005 r. w sprawie zgłaszania okręgowych list kandydatów na posłów") according to which: [W opinii Państwowej Komisji Wyborczej, dla zapewnienia wyborcom właściwej informacji o kandydatach, wskazane jest podawanie zawodu wykonywanego (zajęcia) kandydatów, zwłaszcza w przypadku, gdy kandydat nie ma wykształcenia specjalistycznego lub wykonuje zawód niezwiązany z kierunkiem swojego wykształcenia. W przypadku studentów stosowne jest użycie określenia „student", wskazujące zajęcie kandydata]. Translation: "In the opinion of PKW, in order to provide the

"zawód" can refer both to the job that a person is qualified to do and their practiced occupation. "Occupation" can refer to an individual's current job or, really, any job that the candidate wishes to report. The instructions that the PKW issues before every election are based on the electoral law currently in force. There is no disambiguation in these laws as to the way in which the candidates should treat the term "zawód". Oddly, in the case of the presidential election, the law is very precise: the candidates must report their education and practiced occupation.

According to the law, the PKW does not have – and has never had – an obligation to issue specific forms of the documents for the candidates. Information that should be given in the documents is specified in the electoral act, but the forms can be created by each party or individual involved in the registration of the candidates. In 2009, the PKW attempted some standardization: they placed on their website examples of the forms for the candidates for the EU Parliament, but there was no obligation for parties or candidates to use the documents. In fact, it is written underneath that these are only examples. In 2011, the PKW started to use registration software in order to facilitate the process of candidates' registration.

We asked the PKW specifically about the extent to which the registration information on the candidates – especially on the occupational characteristics – is verified and if there are penalties for false information. The PKW replied that they only verify this information for a candidate if someone specifically asks them to, i.e. if there is a complaint that the information is inaccurate. They have no formal verification structure other than this.

voters with appropriate information about the candidates, it is advisable to report candidates' practiced profession (activity) [zawód wykonywany], especially in the case when a candidate does not have higher education / vocational training or exercises a profession that has no relation to his/her education. In case of students it would be appropriate use the term "student" as an indication of the candidate's professional activity." The 1989 forms that were produced by the Electoral Committee (as a "response" to the documents that the candidates had to submit) are very similar to those produced in 1991 and 2001.

Coding Occupation

For 1989–2007, the codes of occupations were available in the existing data files. For 2011 the Polish team coded from text descriptions directly into SCO-2009 codes (Domański, Sawiński and Słomczyński 2009). Thus, for the 1985–2007 data, "the old" coding frames were originally used. They were based on 2-digit codes, and a number of codes were slightly different in each election. It was also found that more than one code was assigned to some occupations (See the Appendices in this chapter for the syntax for recoding occupations into SCO-2009 classification).

We designed the coding such that cases can be compared at the level of social classes. This assumes that the class structure is comparable through time; caution must be taken with that assumption (see Słomczyński et al 2007). The comparing of detailed codes must also be done with caution as the numbers of detailed occupations in the original coding frames were different in each election.

An additional, but somewhat minor problem, is that a new "Sales and Service" category has been introduced in the 2011 data. In previous elections, the sales and service occupations were partly coded to manual workers, and partly to non-manual workers of the lowest level. The class, "Sales and Service Workers," did not appear for candidates until 2011, and its members do not appear at all as parliamentarians.

CLEANING AND MATCHING

Cleaning and matching were done iteratively. Valid matching requires data that are free of errors; as such, the initial step was to carefully clean the data in each dataset. But some errors cannot be identified this way, as when the name of a candidate in each data set is recorded in a slightly different way. Matching can reveal some errors that cannot be found during the separate cleaning of datasets. After correcting the errors, the matching starts again, usually by creating some new errors. The process is stopped when the matching process no longer produces errors, and we can assume that the matching is done properly. At this point, we have nothing more to do, because we have exhausted all available means to identify the errors.

Data Cleaning

Data cleaning is a process that removes errors and anomalies from the dataset. Errors and anomalies, which we shorten to "errors," can come from all phases of the data collection life cycle. Errors can come from: (a) The data source: Since EAST PaC is collected from official data sources – national election offices, mainly – and supplemented on occasion from parliamentarian and candidate websites, the data source itself may introduce the error; (b) The data entry: Human error arises from people entering information from the data source into the data set. If automated, there is less chance for human error, but as humans write the programs, human error is always possible; (c) The matching process: EAST PaC is unique in that it follows the same candidates through time. Errors can be introduced in the matching process itself. The matching process that connects candidates from one election to another is frequently done with software. This software is good at reducing duplicate cases, but it does not eliminate them. Thus, error is possible, here; (d) The data cleaning process: If not treated with extreme caution, the data cleaning may introduce new errors.

Since EAST PaC data are a continuation of data collected for previous elections, knowing "what should be" is based on these previous data. Of course, there may be errors for the previous elections, and as such when it was possible we tried to refer to the initial sources. Some variables, such as number of votes received, district, and the like, can be checked against the data source from which it came. Proper detection includes a combination of statistical procedures, knowledge of the data sources and its collection, and logic. Detection of numeric errors begins with descriptive statistics (tables) and histograms and scatterplots (graphs). The tables and graphs we produce from the cleaning process may show outliers, illogical or impossible values, and strange patterns. We have been attentive to the possibility that the error we find in a specific case (candidate) is part of a larger pattern. Suspicion of error and confirmation of error have different thresholds. We set a low threshold for suspicion of error: anything that is not immediately clear was flagged as suspicious. This was done in the automation phase of the matching and cleaning process. Confirmation of error has a higher threshold. To confirm an error, we had to be reasonably sure that outliers and strange patterns are truly illogical or impossible.

Confirming error required us to closely examine the individual case in which the error appears and compare it with other, similar cases (those without suspicious errors).

The candidate data was collected over years and, as a result of the Electoral Control project, merged with 2011 data and renamed EAST PaC. The history of these data matters, as technological standards influence the quality of the earliest data and our methods for matching and clean these data. The earliest data, which came from the 1980s and 1990s, had to be transferred from one electronic format to another. The era of personal computers began after 1990, and early mainframe computers, because the equipment came from abroad, did not account for the nine unique Polish letters, or "characters." In Poland, some companies tried to develop standards for reading Polish characters, and many standards were developed. Old databases collected before 2000 used one of the many standards, or they used no standard. The era of one standard began after 2000.

Early computer software put some limitations for string variables, such as the number of characters in strings. When the software arbitrarily sets limits on string length to eight characters, critical information about the name is cut off. For example, traditional gender naming rules in Poland are simple. All given Polish female names have an "a" at the end, and all male names have letters other than "a." There are some exceptions that can be caught, such as names from a foreign origin, but these traditional rules capture the majority. Thus, the case-record "Kazimier" could have been Kazimierz (male) or Kazimiera (female). Candidate gender cannot always be determined if the last letters of the name are not there. From elections 1985 to 2007, there were 13,888 family names corrected, 9,794 first names and 5,915 middle names, and the total number of records was 42,385.

We also needed to correct year of birth. The election commission changed reporting of age such that it was not consistent across elections. In some elections, year of birth data was not published, but age was. We converted age into year of birth (subtraction of age from the year of election). Age matters in electoral politics and candidates may change their ages year to year. A candidate can make themselves one year younger. This causes matching problems, and because there were many such cases, we checked this with software (the software we used was PASCAL based), and others we did manually. If two different cases

were merged together because of this problem, we used internet sources to verify the age. For 1985–2007 data, we found nearly 1400 candidates with year of birth that needed correction, or 3.3 percent of the sample; we also found that 212, or 0.5 percent, have no year of birth.

Even after corrections of this type we encountered duplicate candidates. Data on year of birth were gathered in each election; when data were merged previously, we had as much information about year of birth as there were elections and thus we could compare the records. If the same candidate had two records, but same name and year of birth, then we suspected that it was a duplicate record. To combine two records into one, we needed the right information. We started by keeping each election in a separate data file. It was important to be sure that the data on elections did not overlap. If it were two different persons, then it would be likely that the same person would be in the same election. If they were in different elections, we suspected that these data were not matched properly, and it is truly a duplicate record. The opposite but rarer problem was dividing one candidate into two. Some data about the elections were merged into the wrong candidate. All told, for the dataset 1985 to 2007, there were 1398 candidates moved into one candidate and only 22 candidates divided into two candidates.

Though they represent a small proportion of the total records, there are consequences of incorrectly matched data. Table 1 shows the differences between unclean, original data and the newly cleaned, corrected data. In the uncorrected data, there were 34,901 candidates who participated in only one election. But this is overestimated, since after correcting these data, the number is 32,962. Perhaps six percent of overestimated cases are not much, but in the cases of underestimation, the problem can be large. For example, of the candidates who appear in four elections, the underestimation was 27 percent, or over a quarter of the population. We thought there was only one candidate who participated in all elections, but after corrections, we saw that there were three such candidates.

Table 1. *Differences between Uncorrected and Corrected Data by Number of Elections Participated, 1985–2007*

Number of Elections Participated	Uncorrected data	Corrected data	Number of Over- and Underestimated	Over- and Underestimation in %
1	34,901	32,962	+1,939	+5.9
2	5,454	5,546	-92	-1.7
3	1,328	1,568	-240	-15.3
4	410	564	-154	-27.3
5	192	244	-52	-21.3
6	69	87	-18	-20.7
7	30	35	-5	-14.3
8	1	3	-2	-66.7
Total	42,385	41,009	+1,376	+3.4

Matching Process

When we tried to merge the old data file with candidates from 1985 to 2007, with the new one from 2011, we got zero matches. Why? The answer was invisible characters. These are characters available from the keyboard, but you cannot see them on the screen, e.g. tab-char (ASCII #9), end-of-line (ASCII #13), and blank (ASCII #32). We decided that we cannot use the standard software for this project, and that it was necessary to develop dedicated software that controls for invisible characters. We used some procedures that are not available in standard software, so that we could delete blank spaces at the beginning and the end of data fields, replace double blanks with single blanks, and delete "enter" keystrokes.

We then merged the newly corrected 1985–2007 data with the newly corrected 2011 data. Thankfully, not many improvements were made to the 2011 data: we corrected 48 for family name, none for first name and nine for middle name. Gender was compared with last letter of first name with almost zero errors. However, the electoral commission did not report year of birth in 2011. We asked the PKW for an explanation, but the PKW did not say why age was missing for 2011. We

were forced to complement 89.9 percent of the 2011 cases (or, 6776 cases) using external data, that – in some cases – required us to improve 1985–2007 data, also. All told, 10 percent of the 2011 candidates (or, 759 cases) are missing year of birth.

We then merged the old and new data with a two-step procedure. First, we divided candidates into groups using an initial key, which was family name plus their first name (because not everyone had a middle name). Next, we created a complete key merging the first step with middle name and then year of birth. When the complete key produced a duplicate record, we thought that, most likely, it was the same person. But no computer program given this information can distinguish the two of them, so a human researcher must make the decision. Such suspicious cases sometimes required referral to internet sources.

Table 2 presents the basics of the merged, matched and clean data: there were 46,426 candidates between 1985–2011, and nearly 81 percent of them ran only once.

Table 2. *Electoral Participation in Parliament, 1985–2011*

Number of Elections Participated	Number of Cases	Percent of Total
1	37,428	80.6
2	5,931	12.8
3	1,829	3.9
4	731	1.6
5	294	0.6
6	127	0.3
7	63	0.1
8	21	Below 0.1
9	2	Below 0.1
Total	46,426	100.0

CONCLUSION

In this chapter, we describe how EAST PaC Poland 1985–2011 data was collected, cleaned, and matched. The goal of EAST PaC was to create a high quality dataset of the universe of candidates who ran for national office and in which users can track candidates across elections.

The problems and errors we encountered were of various kinds: technological, bureaucratic, and social and political. Technological problems included merging data created with computer systems designed in different eras and thus had to be retrofitted to be compatible. Bureaucratic problems included the vagueness of what is reported about the candidates to the PKW, and what the PKW provides as data, all of which changed over time. For example, for unknown reasons, the PKW declined to publish the age of the candidates in 2011. This caused problems in matching and required extra effort from the research team. Social and political problems include not only the vagueness of the laws and policies that govern the PKW, but also the decisions of the parties and candidates in what and how they report to the PKW. According to social norms, after marriage, many women change their family names. A changed name causes difficulties in matching candidates across elections. One can suspect, but cannot prove, that for political reasons parties and candidates change what they report to the PKW as to the age of the candidates. The technological, bureaucratic, and social and political issues, and their combination, were the main challenges of this project.

The collecting, cleaning, and matching required multiple technological solutions, some of them dedicated to this task and all of which we developed over time and improved our data. We are confident that these data are the best that they have ever been. We hope that by describing these data and the process, future scholars can build on these data for future elections.

REFERENCES

Domański, Henryk, Zbigniew Sawiński, and Kazimierz M. Słomczyński. 2009. *Sociological Tools Measuring Occupations: New Classification and Scales*. Warsaw: IFiS Publishers.

Dubrow, Joshua Kjerulf. 2011. "The Importance of Party Ideology: Explaining Parliamentarian Support for Political Party Gender Quotas in Eastern Europe." *Party Politics* 17(5): 561–580.

Kunovich, Sheri. 2012. "Unexpected Winners: The Significance of an Open-List System on Women's Representation in Poland." *Politics & Gender* 8(2): 153–177.

Shabad, Goldie and Kazimierz M. Słomczyński, "Inter-Party Mobility among Parliamentary Candidates in Post-Communist East Central Europe." *Party Politics* 10: 151–176.

Słomczyński, Kazimierz M., Krystyna Janicka, Goldie Shabad, and Irina Tomescu-Dubrow. 2007. "Changes in Class Structure in Poland, 1988–2003: Crystallization of the Winners–Losers' Divide." Pp. 25–46 in *Continuity and Change in Social Life: Structural and Psychological Adjustment in Poland*, edited by K. M. Słomczyński and S. T. Marquart-Pyatt. Warsaw: IFiS Publishers.

Zielinski, Jakub, Kazimierz M. Słomczyński and Goldie Shabad. 2005. "Electoral Control in New Democracies: The Perverse Incentives of Fluid Party Systems." *World Politics* 57(3): 365–395.

APPENDIX A: SYNTAX FOR RECODING 1989–2007 OCCUPATIONS INTO SCO-2009

For 1989–2007 data, we calculated 75 different codes. We present them in the format of a table which was applied to recode original codes into the SCO-2009 categories. In the first two columns you find an original 2-digit code (or two codes in some cases) and a corresponding original label. In the last two columns the SCO-2009 categories are presented which, in our opinion, suited the original category in the best way. The rows are ordered according to SCO-2009 categories.

Original codes	Categories in 1989–2007	SCO codes	SCO-2009 categories
70	Parliamentarians	0111	Parliamentarians
31; 71	Politicians	0170	Politicians
07	Directors	0290	Top management, directors
41	Managers	0300	Managers
32	Professionals	1000	Professionals and specialists
03	Artists	1110	Artists
02	Journalists	1113	Journalists, and commentators in TV and other media
01	Scientists	1120	Research scientists and faculty of universities and colleges
13; 49	Teachers	1130	Teachers and school inspectors
09	Political scientists	1141	Sociologists and political scientists
10	Sociologists		
11	Psychologists	1142	Psychologists
06	Economists	1144	Economists, and specialists in banking and finances
45	Consultants		
08	Historians	1149	Specialists in social sciences and humanities
12	Philologists		
05	Law professionals	1150	Law professionals
04	Lawyers	1157	Lawyers, attorneys at law
37	Chemists	1162	Chemists

18	Physicians	1173	Physicians medical doctors
69	Pharmacists	1175	Pharmacists
19	Veterinarians	1185	Veterinarians
14	Agricultural engineers	1187	Agricultural engineers
17	Clergy	1190	Clergy
64	Specialists not classified	1200	Specialists in technical fields
16	Other engineers	1220	Engineers
50	Mechanical engineers	1222	Mechanical engineers
38	Electronic engineers	1223	Electrical, electronic, and power industry engineers
39; 65	Electrical engineers		
57	Construction engineers	1224	Architects and construction engineers
15	Mining engineers	1225	Geodesy, geology, and mining engineers
58	Geodesy engineers		
53	Technologists	1230	Engineering specialists, technologists, constructors
22	Technicians	2120	Technicians
52	Mechanical technicians	2122	Mechanical technicians
55	Electromechanical technicians	2123	Electrical, electronic, and power industry technicians
51	Construction technicians	2125	Construction technicians
54	Construction middle level specialists		
36	Administrators	2300	Specialized office workers
43	Office workers		
44	Banking specialists		
47	Financial advisor	2325	Finance inspectors and advisors
46	Computer specialists	2326	Computer operators and DP technicians
42	Self-government activists	2339	Officers of governmental administration
23	Nurses	3120	Nurses and middle-level medical personnel
21	Technicians in animal rearing	3140	Middle-level specialists in agriculture and forestry

59	Agricultural technicians		
73	Real estate managers	3150	Real estate, insurance, and trade agents
48	Bookkeeping clerks	3211	Bookkeeping clerks
67	Economic middle level specialists	3212	Clerks in statistic and economic departments
28	Policemen	3310	Policemen
27	Military officers	3320	Military officers
68	Sportsmen	3420	Athletes, sportsmen, sport officials
30	Manual workers	5000	Manual workers
66	Foresters	5100	Foremen and supervisors of manual work
63	Miners	5210	Miners
56	Electricians in construction	5220	Electricians
62	Mechanics	5240	Mechanics
61	Railway workers	5272	Railway workers
40	Drivers	5274	Car, truck, and bus drivers
60	Motor vehicle operators		
29	Farmers	7100	Farmers
25	Entrepreneurs	8000	Entrepreneurs and business owners
24	Craftsmen	8110	Self-employed craftsmen
26	Owners of stores	8600	Owners of stores and other trade facilities
34	Without occupation	9000	Without occupation
35	Retired	9100	Retired
72	Pensioners	9200	Pensioners
74	Unemployed	9300	Unemployed
75	Running the household	9400	Running the household
20	Students	9500	Students
33	Others	9900	Other not classified

APPENDIX B:
SYNTAX FOR RECODING SCO-2009 OCCUPATIONS INTO SOCIAL CLASS

SCO-2009	Code	Social Class
0000..1199	1	Non-technical intelligentsia
1200..1999	2	Technical intelligentsia
2000..3999	3	Middle and low-level non-manual workers
4000..4999	4	Sales and service workers
5000..6999	5	Manual workers
7000..7999	6	Farmers
8000..8999	7	Business owners, self-employed in sales and service
9000..9900	8	Without occupation, not classified
other	9	Missing data

Appendix C presents the percentage of candidates within each class category for each election since 1989.

Table C1. *Social Classes in EAST PaC Poland, 1989 to 2011*

Candidate Class	1989	1991	1993	1997	2001	2005	2007	2011
Intelligentsia non-technical	60.5	46.5	46.0	48.6	46.6	41.2	43.9	47.0
Technical intelligentsia, engineers	16.9	16.9	15.0	13.7	10.6	9.2	9.6	7.1
Middle and low-level nonmanual workers	8.3	13.7	14.7	12.2	13.5	16.1	16.1	20.4
Sales and service workers	–	–	–	–	–	–	–	2.4
Manual workers	2.1	4.7	3.9	3.5	5.1	8.1	6.8	5.6
Farmers	9.9	6.7	7.9	4.9	7.6	4.1	4.9	3.3
Business owners, self-employed in sales and service	.9	5.8	5.4	6.6	6.5	9.0	8.7	8.6
Not working, without occupation	1.3	5.8	7.0	10.6	10.2	12.2	10.0	5.7