Candidate Data in Poland

Zbyszek Sawiński

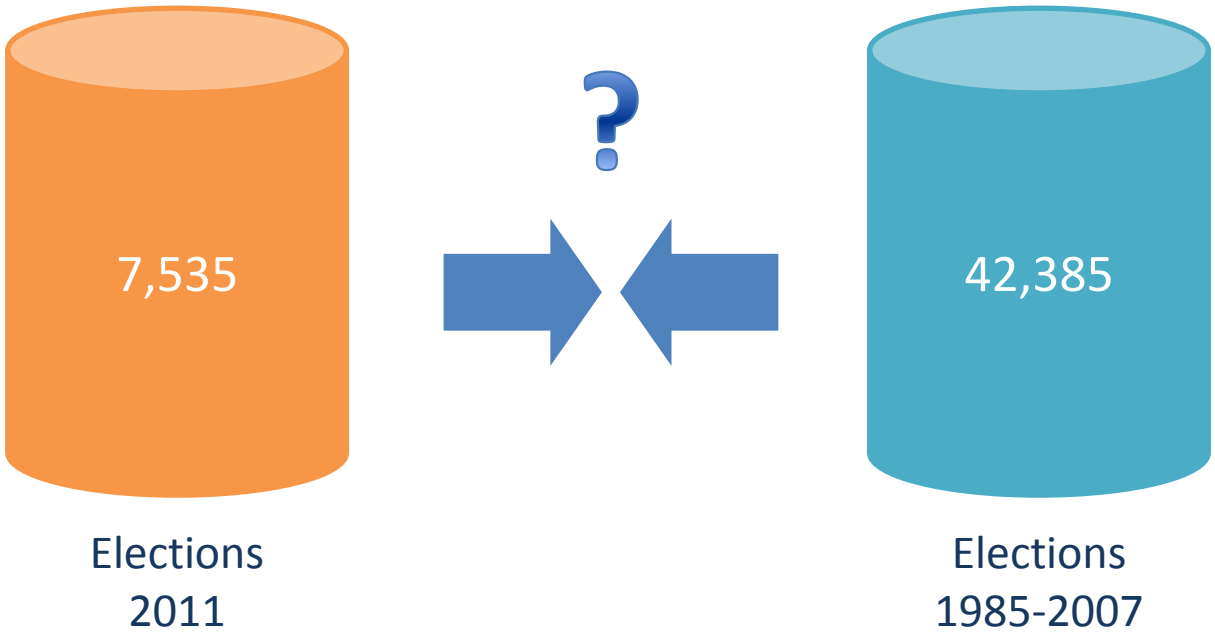
Educational Research Institute
Poland

Agenda

- Merging candidate data: problems and solutions

- The structure of the Polish candidate data file

The starting point (May 2012)



Polish characters

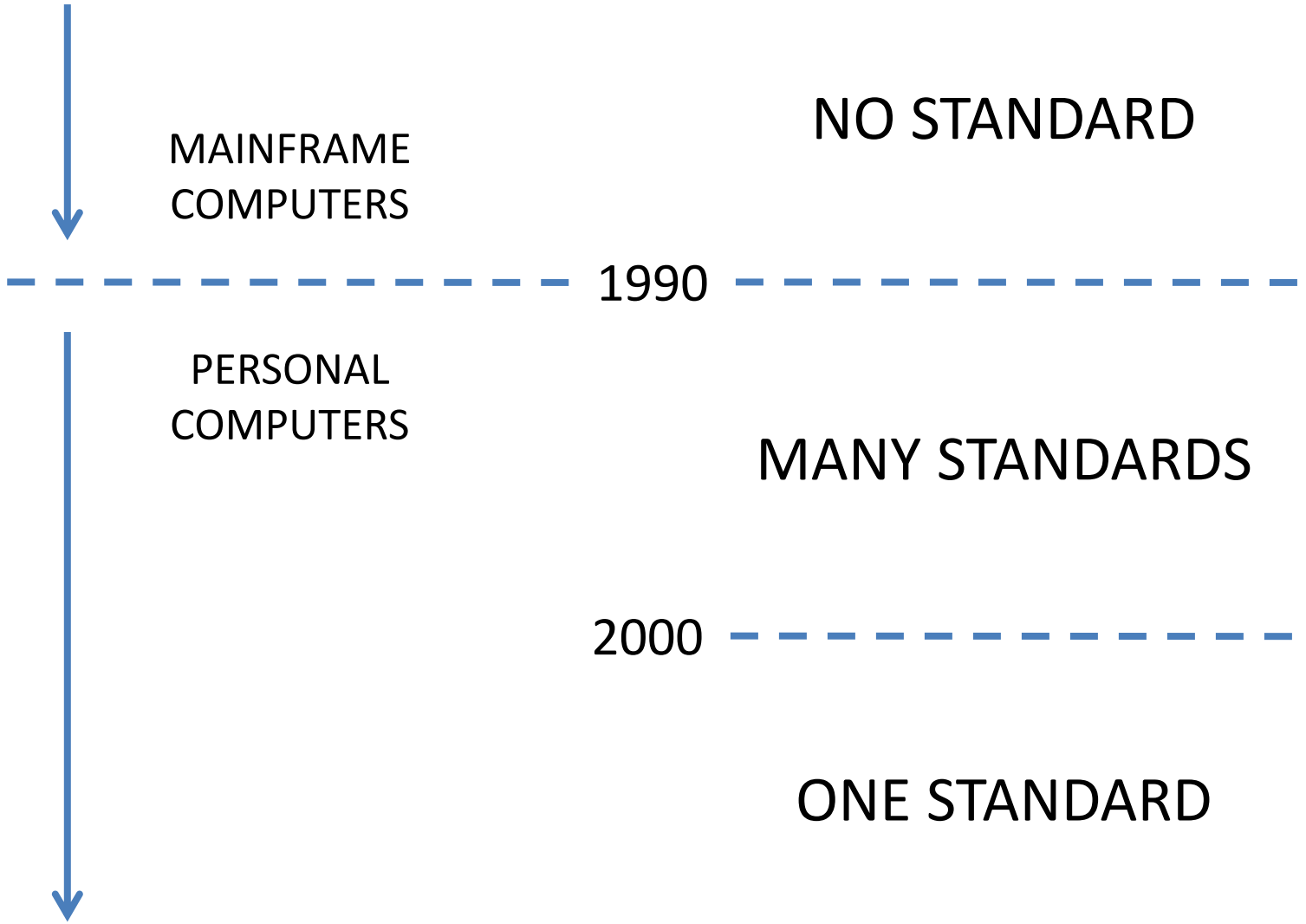
ą ć ę ł ń ó ś ź ż

Ą Ć Ę Ł Ń Ó Ś Ź Ż

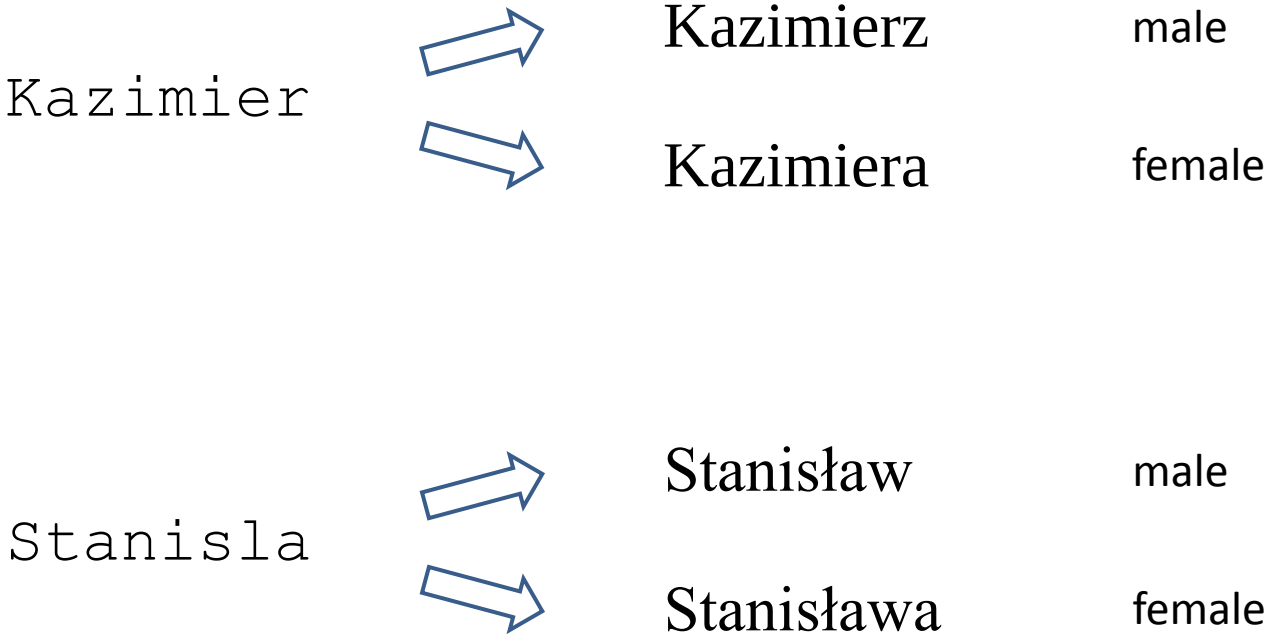
Polish alphabet

a ą b c ć d e ę f g h i j k l ł m n ń o ó p q r s ś t u v w x y z ź ż

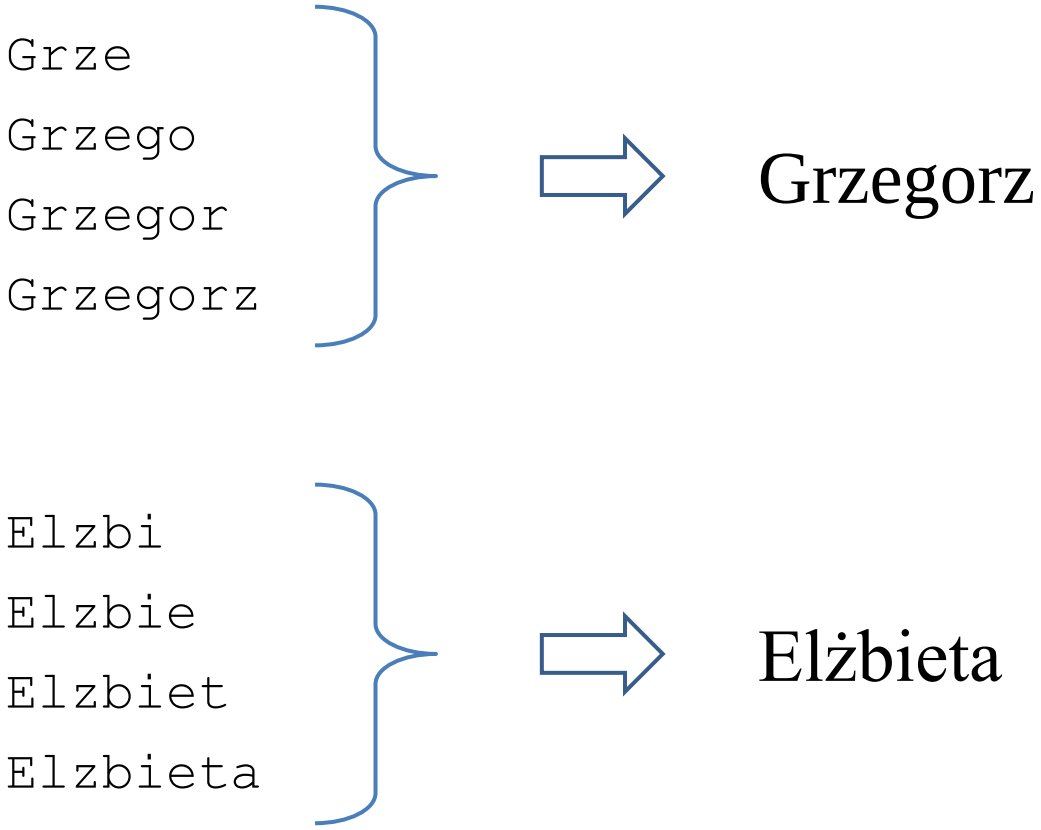
Computer standards for Polish characters



Limited size of data fields containing names

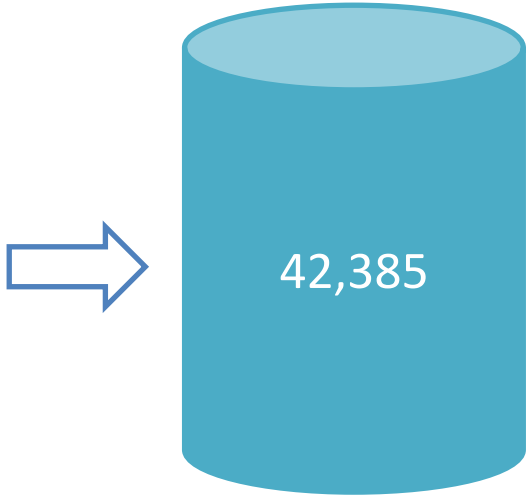


Truncated names



Corrections of names

Family name	13,888	(33%)
First name	9,794	(23%)
Middle name	5,915	(14%)



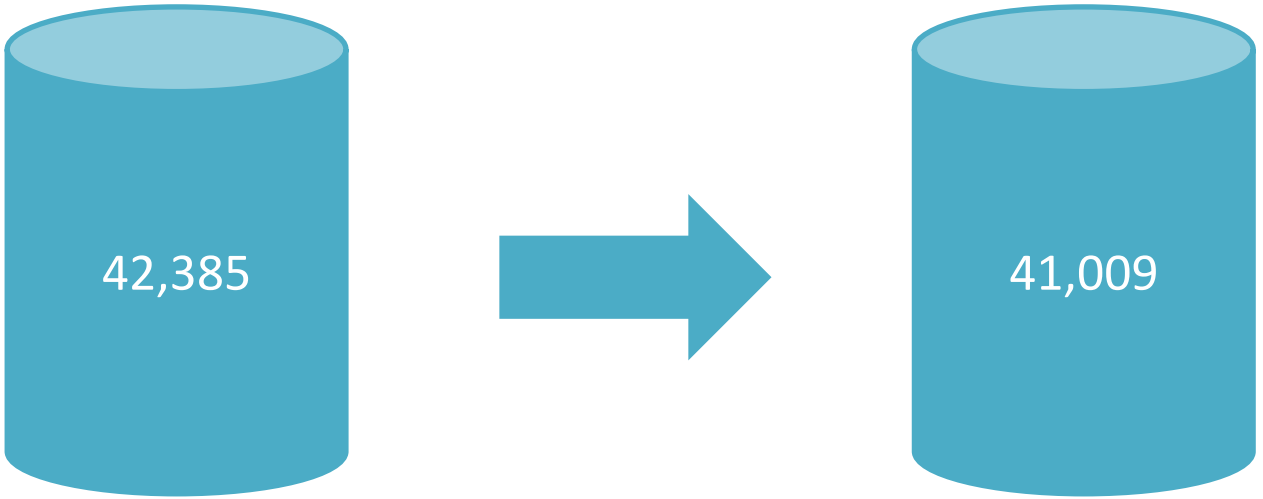
Elections
1985-2007

Duplicated candidates

ID	Last name	First name	Middle	L2007	S2007	L2005	S2005	L2001	S2001	L1997	S1997	L1993	S1993	L1991	S1991
081387	Antonowicz	Bronisław	Walenty							1938					
111179	Antonowicz	Cecylia	Halina	1954											
000108	Antonowicz	Jacek										1951			
050002	Antonowicz	Jan							1956		1956		1956		1956
100087	Antonowicz	Jan					1956								
081797	Antos	Bernard								1969					
021291	Antosiewicz	Antoni	Alfred			1925		1925		1925		1926		1925	
000111	Antosiewicz	Kazimierz										1946			
100088	Antosik	Jarosław				1966									
000112	Antosz	Jan										1935			
100092	Antosz	Zbigniew				1962									
000113	Antoszczyszyn	Leszek										1954			
100089	Antoszczyszyn	Leszek			1954	1954									
082819	Antoszek	Ireneusz	Marian							1970					
111134	Antoszewski	Antoni		1938											
100090	Antoszewski	Jerzy					1930								
000114	Antoszewski	Zbigniew	Edward								1944			1944	
100091	Antoszewski	Zbigniew	Edward				1944								
000115	Antoszkiewicz	Jan												1937	
000116	Anulewicz	Andrzej							1948	1948		1948			
100093	Anulewicz	Andrzej			1948		1948								
020011	Anusz	Andrzej						1965		1965		1965		1965	
083809	Anuszkiewicz	Krzysztof						1962		1962					
000117	Anuszkiewicz	Krzysztof	Zbigniew									1956			

Corrections of data records

Type of correction	#	%
Combining two candidates into one	1398	3.30
Dividing one candidate into two	22	0.05
Moving some elections between candidates	42	0.10



Corrected vs. original data: number of elections for all candidates

# elections	original data	corrected data	over/under estimated	% of corrected
1	34901	32962	1939	6
2	5454	5546	-92	-2
3	1328	1568	-240	-15
4	410	564	-154	-27
5	192	244	-52	-21
6	69	87	-18	-21
7	30	35	-5	-14
8	1	3	-2	-67

Corrected vs. original data: number of times they were elected

# times elected	original data	corrected data	over/under estimated	% of corrected
0	39408	38121	1287	3
1	2090	1973	117	6
2	567	564	3	1
3	198	209	-11	-5
4	76	91	-15	-16
5	28	30	-2	-7
6	13	14	-1	-7
7	5	7	-2	-29

Suspected causes of errors in combining data from elections

Invisible characters

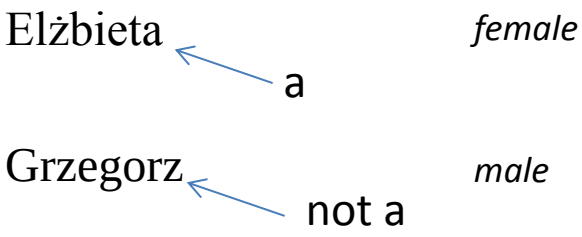
tab-char (ASCII #9)
end-of-line (ASCII #13)
blank (ASCII #32)

} available from the keyboard, but you can't see them on the screen

Dedicated software solutions are needed which enable control invisible characters

Additional verification, recoding and complementing 1985-2007 data

- Gender
  - Combining specific variables for elections into one variable
  - Checking of compliance of the last letter of the first name and of the middle name with the gender



- Year of birth
  - Combining specific variables for elections into one variable
  - Checking whether the candidate completed 21 years in an election year
  - Explaining inconsistencies with external sources of data
  - Complementing missing data wherever possible
    - complemented 1405 candidates with the year of birth (3.3% of all)
    - only 212 candidates still left with the missing year of birth (0.5% of all)

Correcting and complementing 2011 data

- Names
  - Not many improvements: 48 for family name, 0 for 1st name, and 9 for the middle name
- Gender
  - Checking of compliance of the last letter of the first name and of the middle name (almost no errors)
- Year of birth
  - Missing in electoral data!
  - Complemented in 6776 cases (89.9%)using external data and the improved 1985-2007 data
    - 759 candidates still left with the missing year of birth (10.1% of all)
  - Checking whether the candidate was 21 or older in 2011

The two-step strategy for merging data

Step 1. Dividing candidates into groups using an initial key

Initial key: Family name + 1st name

Step 2. Merging candidates inside groups using a complete key

Complete key: Family name + 1st name + Year of birth + Middle name

Correspondence between groups

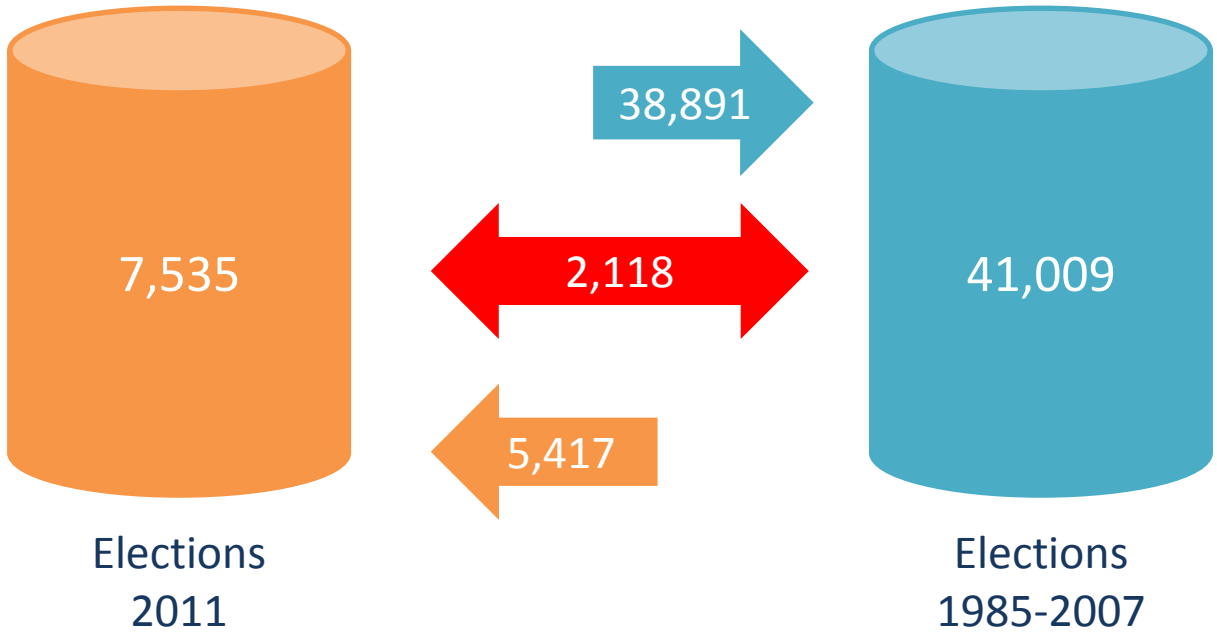
2011 candidates						1985-2007 candidates				
					N	082111	Kowalski	Bogdan	Zbigniew	1956
						004666	Kowalski	Bogdan		1959
312636	Kowalski	Bogusław		1964	D	020507	Kowalski	Bogusław		1964
312848	Kowalski	Bogusław		1964						
316369	Kowalski	Edward		1951	M	110822	Kowalski	Edward		1951
316579	Kowalski	Jacek	Sławomir	1987	N					
315489	Kowalski	Jan	Marcin		C	040416	Kowalski	Jan	Antoni	1939
						012090	Kowalski	Jan		1941
						110512	Kowalski	Jan		1946
						012091	Kowalski	Jan	Bogusław	1953
						081191	Kowalski	Jan	Kazimierz	1953

N – not fits (nothing to do); **M** – merged; **C** – check required; **D** – decision required

The spreadsheet data base of candidates in groups

węzeł	dane	ID	nazwisko	imię-1	imię-2	rok ur.	Ile razy	2011 kand	2011 wyb	2011:okreg	2011:partia	(n-1) kand	(n-1) wyb	(n-1) okręg	(n-1) pa
	d8507	082111	Kowalski	Bogdan	Zbigniew	1956	1					L1997		Krakowskie	UPRz
	d8507	004666	Kowalski	Bogdan		1959	1					L1991		Wrocław	
X7	d8507	020507	Kowalski	Bogusław		1964	6					L2007	L	Siedlce	PiS
Y7W7	d2011	312636	Kowalski	Bogusław		1964	1	L2011		Płock	Ruch Palikota				
Y7W7	d2011	312848	Kowalski	Bogusław		1964	1	L2011		Siedlce	PiS				
	d8507	081803	Kowalski	Cezary		1965	1					L1997		Kieleckie	UPRz
	d8507	081855	Kowalski	Czesław	Stanisław	1945	1					L1997		Konińskie	UP
	d2011	313127	Kowalski	Daniel	Marek	1984	1	L2011		Warszawa I	SLD				
	d8507	055577	Kowalski	Edmund		1947	1					L1989			
X7	d8507	110822	Kowalski	Edward		1951	1					L2007		Konin	LiD
Y7	d2011	316369	Kowalski	Edward		1951	1	L2011		Konin	Ruch Palikota				
	d8507	004667	Kowalski	Eugeniusz		1924	1					L1993		Sosnowiec	PUP
	d2011	316579	Kowalski	Jacek	Sławomir	1987	1	L2011		Piła	PO				
	d8507	040416	Kowalski	Jan	Antoni	1939	1					S1991		Olsztyńskie	
	d8507	012090	Kowalski	Jan		1941	1					L1991		Rzeszów	
	d8507	110512	Kowalski	Jan		1946	1					L2007		Sieradz	LiD
	d8507	012091	Kowalski	Jan	Bogusław	1953	1					L1993		Jeleniogórskie	KLD
	d8507	081191	Kowalski	Jan	Kazimierz	1953	2					L2001		Legnica	PO
Y0W0	d2011	315489	Kowalski	Jan	Marcin		1	L2011		Sosnowiec	Ruch Palikota				
	d8507	021361	Kowalski	Janusz		1933	2					L1993		Warszawa	PWN-PS

The final point (October 2013)



The structure of data records

Candidate unique ID code
Merging profile (2011 elections vs 1985-2007 elections)
Family name, 1st name, middle name
Year of birth and gender
Number of elections in which he/she run (1985-2011)
Elections 2011..1985: participated/elected
History of participation in elections: 2011..1985
Sejm, Senate, Parliament : number of times participated/elected
2011 Elections: detailed data
2007 Elections: detailed data
. . . . .
1989 Elections: detailed data
1985 Elections: detailed data

Auxiliary variables

Auxiliary variables: Results of merging 2011 with 1985-2007 data

merge	Merging profile (2011 elections vs 1985-2007 elections)		
#	%	code	label
38891	83.8	0	Only in 1985-2007 elections
2118	4.6	1	Both in 1985-2007 & 2011 elections
5417	11.7	2	Only in 2011 elections

Auxiliary variables: Number of elections

num_elec	Number of elections in which he/she run (1985-2011)		
#	%	code	label
37428	80.6	1	
5931	12.8	2	
1829	3.9	3	
731	1.6	4	
294	0.6	5	
127	0.3	6	
63	0.1	7	
21	0.0	8	
2	0.0	9	

Auxiliary variables: Elections 2011..1985 participated/elected

elec_2011	[2011 Elections] Participated/elected		
#	%	code	label
38891	83.8	0	Not participated
6575	14.2	1	Elections to Sejm/non-elected
460	1.0	2	Elections to Sejm/elected
400	0.9	3	Elections to Senate/non-elected
100	0.2	4	Elections to Senat/elected

Auxiliary variables: History of participation in elections

303	0,7%	'000000001'
289	0,6%	'000000002'
1000	2,2%	'000000010'
33	0,1%	'000000011'
64	0,1%	'000000012'
172	0,4%	'000000020'
8	0,0%	'000000021'
1	0,0%	'000444400'
7207	15,5%	'001000000'
5	0,0%	'001000010'
2	0,0%	'222222000'
2	0,0%	'222222100'
1	0,0%	'222222110'
4	0,0%	'222222200'
3	0,0%	'222222220'
1	0,0%	'222222232'
1	0,0%	'444443000'

Auxiliary variables:
Number of Sejm+Senate elections (participated vs elected)

36598	78,8%	'10'
830	1,8%	'11'
5160	11,1%	'20'
617	1,3%	'21'
154	0,3%	'22'
1180	2,5%	'30'
347	0,7%	'31'
185	0,4%	'32'
117	0,3%	'33'
322	0,7%	'40'
148	0,3%	'41'
124	0,3%	'42'
98	0,2%	'43'
39	0,1%	'44'
75	0,2%	'50'
55	0,1%	'51'
56	0,1%	'52'
61	0,1%	'53'
38	0,1%	'54'
9	0,0%	'55,
17	0,0%	'60'
18	0,0%	'61'
20	0,0%	'62'
30	0,1%	'63'
19	0,0%	'64'
17	0,0%	'65'
6	0,0%	'66'
3	0,0%	'70'
10	0,0%	'71'
4	0,0%	'72'
8	0,0%	'73'
8	0,0%	'74'
16	0,0%	'75'
7	0,0%	'76'
7	0,0%	'77'
3	0,0%	'83'
5	0,0%	'84'
3	0,0%	'85'
3	0,0%	'86'
4	0,0%	'87'
3	0,0%	'88'
1	0,0%	'97'
1	0,0%	'98'

Veterans of the Polish parliament: Eugeniusz Czykwin

ID	Candidate unique ID code	001556
num_sejm	Number of Sejm elections: participated|elected (1st|2nd digit)	87
num_senate	Number of Senate elections: participated|elected (1st|2nd digit)	10
num_parlam	Number of Sejm+Senate elections: participated|elected (1st|2nd digit)	97
history	HISTORY History of participation in elections: 2011 down to 1985	222213222
name_1st	First name	Eugeniusz
name_mid	Middle name	
name_fam	Family name (Last name)	Czykwin
year_birth	Year of birth	1949
gender	Gender	1 Male
L11_comm	[2011 Sejm] Party electoral list	3 Sojusz Lewicy Demokratycznej
L07_comm	[2007 Sejm] Party electoral list	17 Lewica i Demokraci
L05_comm	[2005 Sejm] Party electoral list	18 Sojusz Lewicy Demokratycznej
L01_comm	[2001 Sejm] Party electoral list	1 Sojusz Lewicy Demokratycznej - Unia Pracy
L97_comm	[1997 Sejm] Party electoral list	15 Mniejszości narodowe
S93_comm	[1993 Senate] Electoral committee	9 KW Prawosławnych
L91_comm	[1991 Sejm] Party electoral list	6 Komitet Wyborczy Prawosławnych

Veterans of the Polish parliament: Waldemar Pawlak

ID	Candidate unique ID code	021581
num_sejm	Number of Sejm elections: participated|elected (1st|2nd digit)	88
num_senate	Number of Senate elections: participated|elected (1st|2nd digit)	0
num_parlam	Number of Sejm+Senate elections: participated|elected (1st|2nd digit)	88
history	HISTORY History of participation in elections: 2011 down to 1985	222222220
name_1st	First name	Waldemar
name_mid	Middle name	
name_fam	Family name (Last name)	Pawlak
year_birth	Year of birth	1959
gender	Gender	1 Male
L11_comm	[2011 Sejm] Party electoral list	5 Polskie Stronnictwo Ludowe
L07_comm	[2007 Sejm] Party electoral list	27 Polskie Stronnictwo Ludowe
L05_comm	[2005 Sejm] Party electoral list	14 Polskie Stronnictwo Ludowe
L01_comm	[2001 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L97_comm	[1997 Sejm] Party electoral list	7 Polskie Stronnictwo Ludowe
L93_comm	[1993 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L91_comm	[1991 Sejm] Party electoral list	2 Polskie Stronnictwo Ludowe - Sojusz Programowy

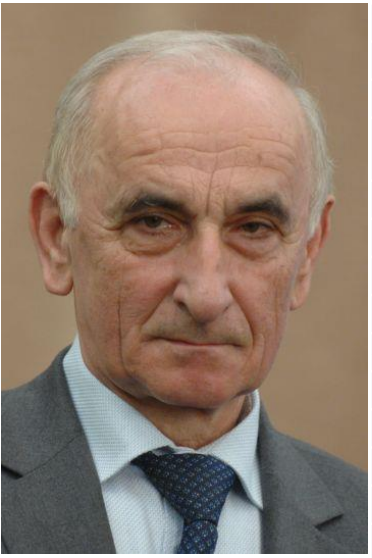
Veterans of the Polish parliament: Franciszek Jerzy Stefaniuk

ID	Candidate unique ID code	021012
num_sejm	Number of Sejm elections: participated|elected (1st|2nd digit)	88
num_senate	Number of Senate elections: participated|elected (1st|2nd digit)	0
num_parlam	Number of Sejm+Senate elections: participated|elected (1st|2nd digit)	88
history	HISTORY History of participation in elections: 2011 down to 1985	222222220
name_1st	First name	Franciszek
name_mid	Middle name	Jerzy
name_fam	Family name (Last name)	Stefaniuk
year_birth	Year of birth	1944
gender	Gender	1 Male
L11_comm	[2011 Sejm] Party electoral list	5 Polskie Stronnictwo Ludowe
L07_comm	[2007 Sejm] Party electoral list	27 Polskie Stronnictwo Ludowe
L05_comm	[2005 Sejm] Party electoral list	14 Polskie Stronnictwo Ludowe
L01_comm	[2001 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L97_comm	[1997 Sejm] Party electoral list	7 Polskie Stronnictwo Ludowe
L93_comm	[1993 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L91_comm	[1991 Sejm] Party electoral list	2 Polskie Stronnictwo Ludowe - Sojusz Programowy

Veterans of the Polish parliament: Józef Zych

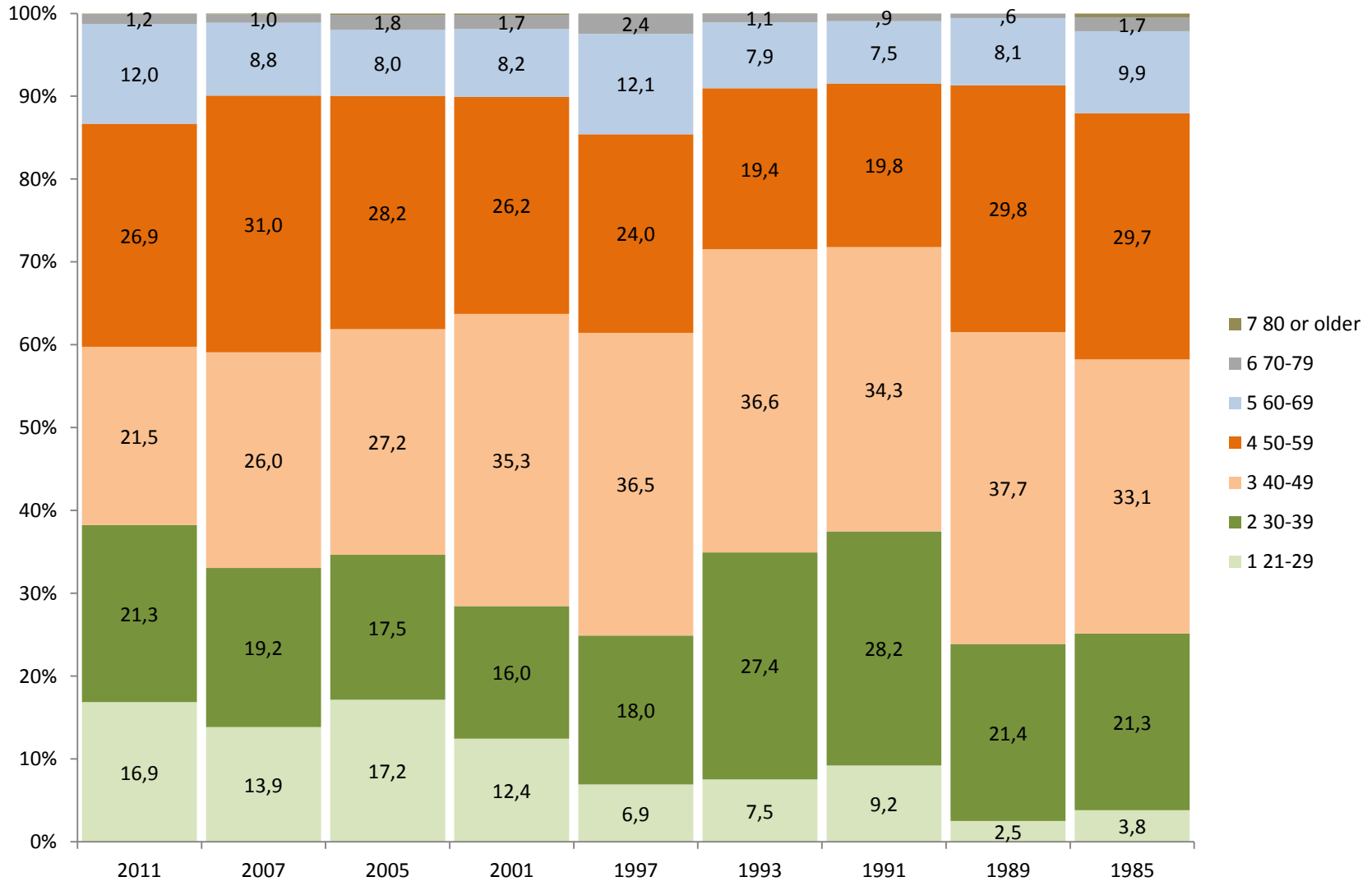
ID	Candidate unique ID code	021542
num_sejm	Number of Sejm elections: participated|elected (1st|2nd digit)	88
num_senate	Number of Senate elections: participated|elected (1st|2nd digit)	0
num_parlam	Number of Sejm+Senate elections: participated|elected (1st|2nd digit)	88
history	HISTORY History of participation in elections: 2011 down to 1985	222222220
name_1st	First name	Józef
name_mid	Middle name	
name_fam	Family name (Last name)	Zych
year_birth	Year of birth	1938
gender	Gender	1 Male
L11_comm	[2011 Sejm] Party electoral list	5 Polskie Stronnictwo Ludowe
L07_comm	[2007 Sejm] Party electoral list	27 Polskie Stronnictwo Ludowe
L05_comm	[2005 Sejm] Party electoral list	14 Polskie Stronnictwo Ludowe
L01_comm	[2001 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L97_comm	[1997 Sejm] Party electoral list	7 Polskie Stronnictwo Ludowe
L93_comm	[1993 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L91_comm	[1991 Sejm] Party electoral list	2 Polskie Stronnictwo Ludowe - Sojusz Programowy

Veterans of the Polish parliament: Stanisław Żelichowski

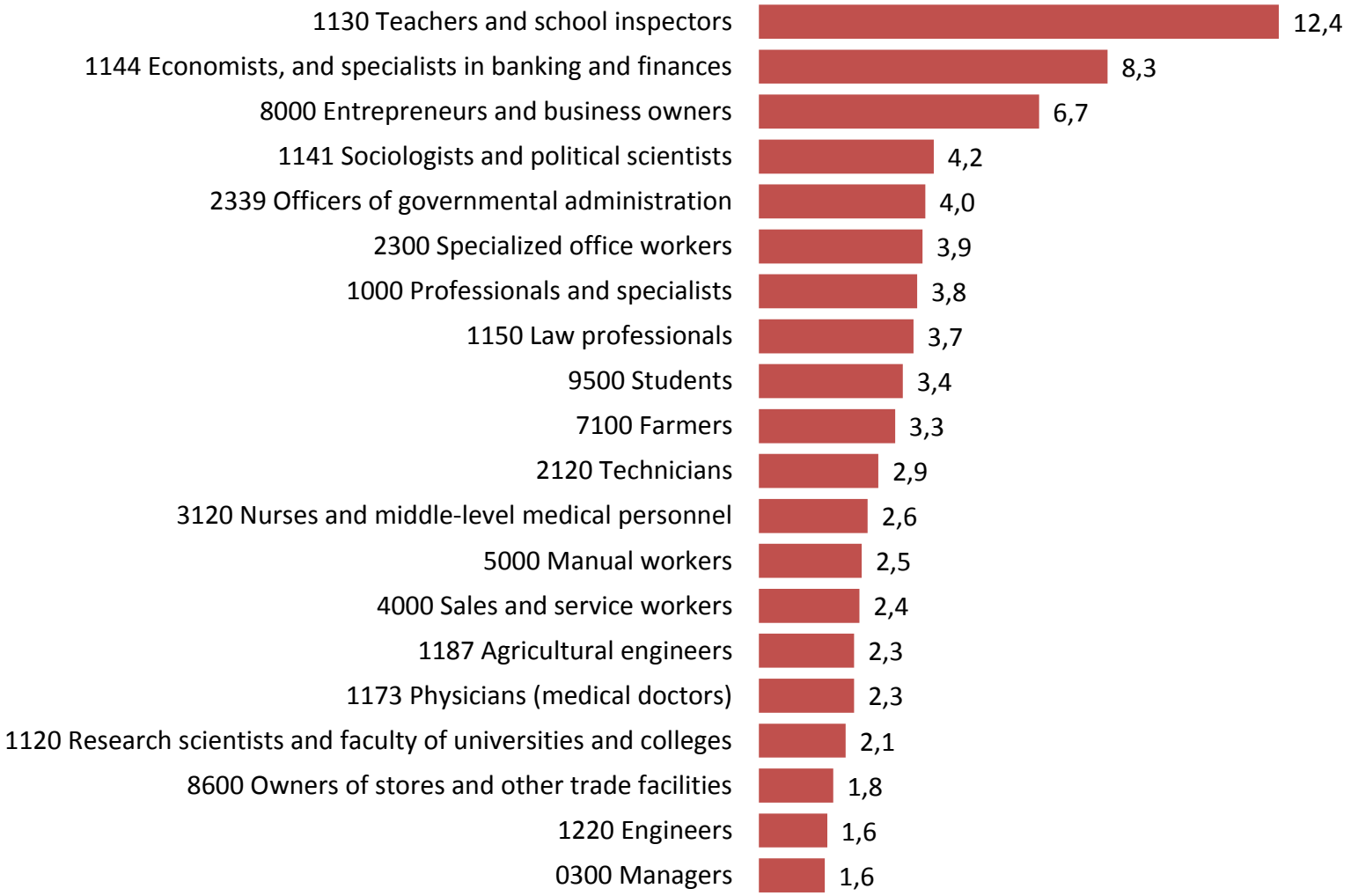


ID	Candidate unique ID code	060426
num_sejm	Number of Sejm elections: participated|elected (1st|2nd digit)	88
num_senate	Number of Senate elections: participated|elected (1st|2nd digit)	10
num_parlam	Number of Sejm+Senate elections: participated|elected (1st|2nd digit)	98
history	HISTORY History of participation in elections: 2011 down to 1985	222222232
name_1st	First name	Stanisław
name_mid	Middle name	
name_fam	Family name (Last name)	Żelichowski
year_birth	Year of birth	1944
gender	Gender	1 Male
L11_comm	[2011 Sejm] Party electoral list	5 Polskie Stronnictwo Ludowe
L07_comm	[2007 Sejm] Party electoral list	27 Polskie Stronnictwo Ludowe
L05_comm	[2005 Sejm] Party electoral list	14 Polskie Stronnictwo Ludowe
L01_comm	[2001 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L97_comm	[1997 Sejm] Party electoral list	7 Polskie Stronnictwo Ludowe
L93_comm	[1993 Sejm] Party electoral list	6 Polskie Stronnictwo Ludowe
L91_comm	[1991 Sejm] Party electoral list	2 Polskie Stronnictwo Ludowe - Sojusz Programowy

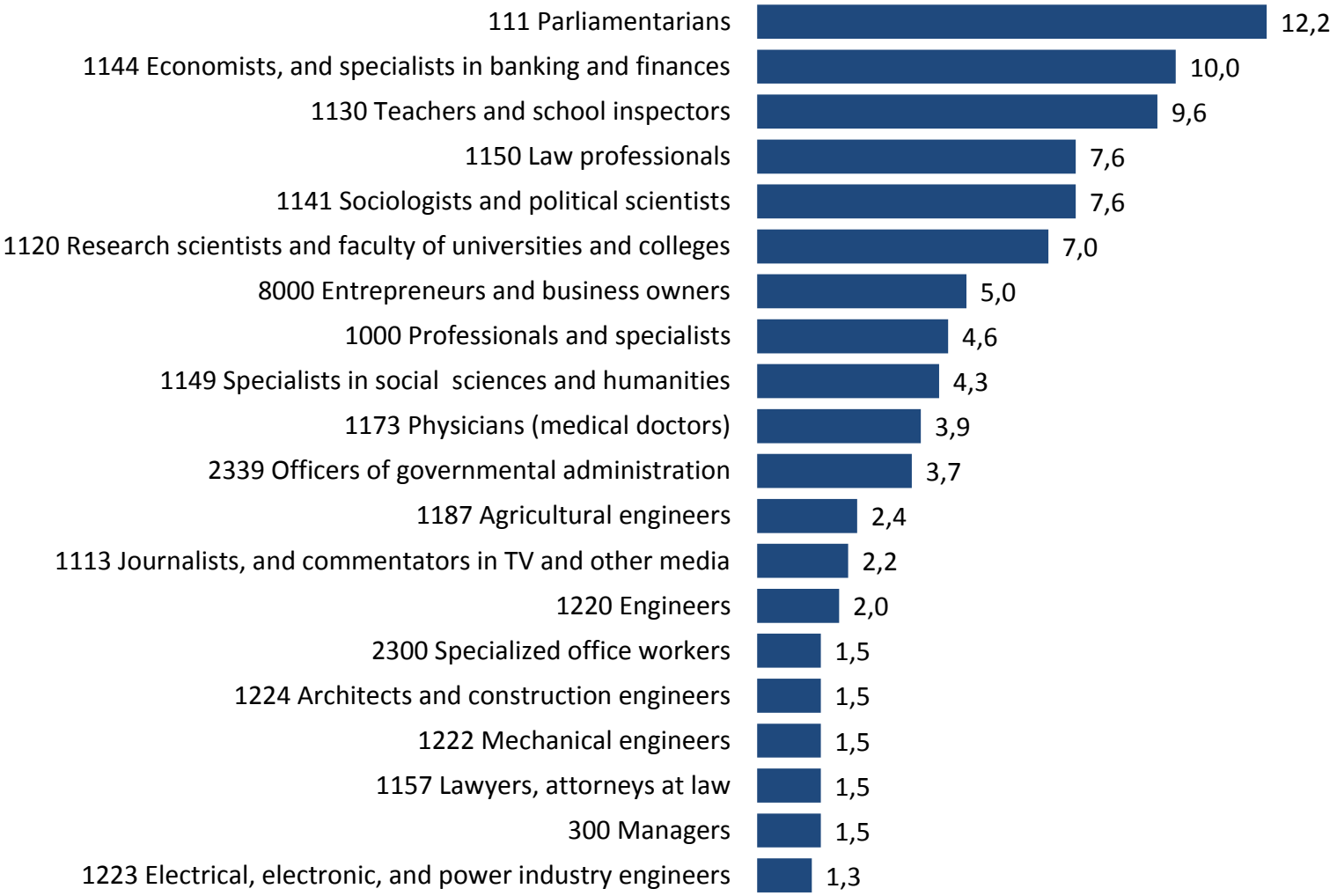
Age of candidates: 2011..1985



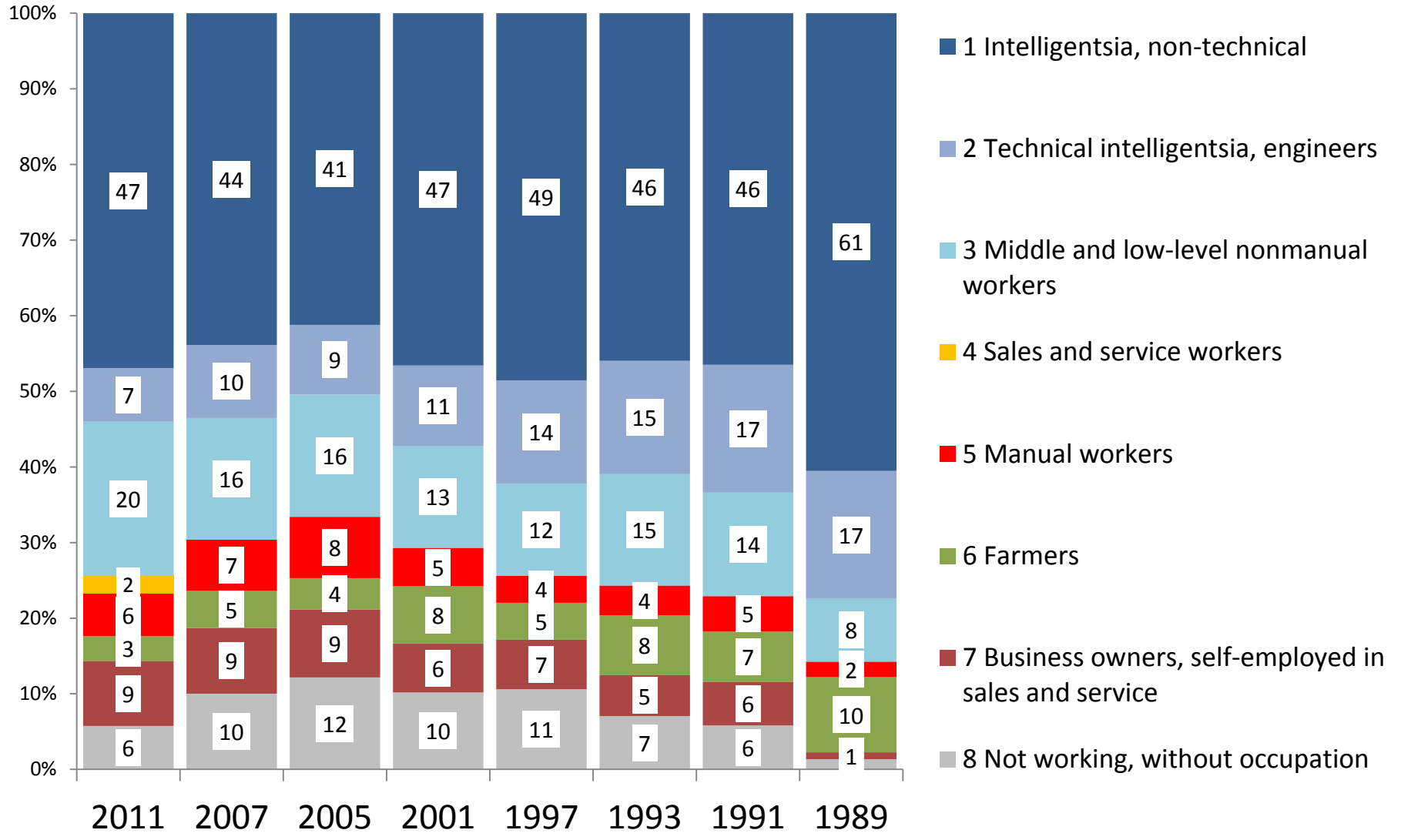
Occupation of all candidates to Sejm in 2011 elections



Occupation of candidates elected to Sejm in 2011



Occupational category of candidates to Sejm: 2011..1989



Conclusions

- Newer standards are more useful than older
- Merging candidates requires dedicated solutions
- Auxiliary variables help to get the audience

Thank you for your attention.

Comments are welcome!