Challenges of Matching Candidate Data in Hungary

Zsófia Papp, Centre for Social Sciences, Hungarian Academy of Sciences, Hungary

This chapter describes the Hungarian data collection process and discusses the challenges of data matching in Hungary. Although several projects engaged in collecting data on political elites, they rarely choose parliamentary candidates as its basic unit. In the early 2000s, two projects aimed at collecting basic socio-demographic and career information on Hungarian Members of Parliament (MPs) between 1884 and 2006 (Ilonszki, 2005, 2008, 2009)¹. This dataset was later updated with data from the 2010-2014 and 2014-2018 electoral terms (Ilonszki et al., 2016)². As to legislator data, an international research network³ was initiated to harmonize data collection efforts and to create a common codebook that enables comparative analysis. As part of this framework, a new MP dataset was created to make the connection between career paths, parliamentary activities and electoral performance⁴. However, not only legislative elites are of major concern in Hungarian political science. Efforts of data collection focus on the governmental⁵ (Ilonszki, 2011; Ványi, 2015) and local political elites (Várnagy, 2012) as well. The above projects centre around collecting and organizing information that is available to the general public, but was presented either sporadically or in a way that does not support systematic analysis.

The other method for obtaining data on political elites is structured interviews. Legislators and mayors were surveyed in 1992, 1995, 1999⁶, 2007, 2008⁷, 2009⁸ and 2003, 2015⁹ respectively by multiple research groups. Legislative candidates were interviewed systematically in 2010 and 2014 utilizing a web-based survey¹⁰. However, these samples were neither randomly drawn nor representative to the entire population of candidates. A recent major effort to collect candidate information started in 2012¹¹, and covered five national elections (1998-2014). On the upside, the project covered a wide range of variables, thus it offers substantially more information than EAST PaC. On the downside, due to the large number of variables, only candidates of parliamentary parties were taken into account. Although this approach enables the researcher to identify candidate selection patterns, its main problem is that it is difficult to explain candidate performance without having information on all other candidates competing in the district.

The Hungarian EAST PaC data is unique in that it aims to create a longitudinal dataset that contains all electoral data on the basis of candidates. Due to the restricted availability of candidate data in Hungary, this dataset cannot compile a wide range of variables, but designing the dataset in wide form (Weiss, 2006, p. 24) allows researchers to follow the careers and performances of the individual candidates over time from the first democratic elections in 1990 to 2010. The aim of this chapter is to shed light on the major challenges the researcher has to face when she wants to build such datasets in Hungary. The two main problems of data collection are (1) information availability and (2) merging multiple sets that contain electoral data from different elections. These obstacles are common for all data collection efforts for EAST PaC, but each country has its own unique challenges.

Challenges in Collecting Hungarian Data

Data sources and availability

EAST PaC covers the period from 1990 to 2010, and the rules and procedures introduced in this chapter are applicable for this period. However, in 2011 the Hungarian Parliament passed a new legal framework, which changed the procedures of the elections. Given the period covered by EAST PaC, I will discuss rules applicable to the pre-2011 era in the main text, and refer to rules in effect in endnotes.

Electoral data are collected and published by the National Election Office (Nemzeti Választási Iroda - NVI)¹². Parties and independent Single Member District (SMD) candidates have to register to request a recommendation sheet, on which they have to collect 750 signatures¹³ from voters in the given constituency. They have to supply the following information: candidate's full name, candidate's identification number, candidate's date of birth, proof of citizenship, candidate's address and other contact information. Candidates nominated on party lists have to provide the same information as SMD candidates. The National Election Office does not collect any additional information, with the exception of profile photos that the candidates may supply. It is clear that the information demand of

Hungarian candidates is considerably smaller than that of candidates of other Eastern European countries. In Poland, for instance, candidates have to provide information on occupation as well, which makes research on social class possible.

As it is stated above, there is limited publicly available official information on Hungarian candidates. An additional problem of data collection is that the National Election Office destroys election related documents 90 days¹⁴ after the election; this makes retrospective investigation extremely problematic. However, even if the documents would be available years after the election, the NVI handles the above information as private data. From the viewpoint of data collection, this privacy policy¹⁵ is especially problematic in the case of year of birth that is often used to match data from different elections. On the official election webpage, the Office publishes aggregate data on the candidates' average age and gender¹⁶. Profile photos and short curriculum vitae (CV) - if supplied by the candidate - are published on the website. Candidate photos turn out to be quite useful in data matching, while CVs provide accurate information for key variables. Unfortunately, only a small minority of candidates submits these additional attachments with their registration. Additionally, data from the early elections (1990 and 1994) is not available on the NVI website. NVI provides election results in a digital format for these elections, but there is no access to candidate CVs.

As opposed to candidate data, information on legislators is well documented. Both profile photos and detailed CVs are provided in digital format and in print by the NVI and the Office of the National Assembly (Országgyűlés Hivatala). This is excellent news from the viewpoint of legislative studies, as we can obtain a wide range of variables about the MPs' socio-economic and career backgrounds. However, this data provides only for a small percentage of all candidates who ran for office.

Due to the problems of data availability, relative to other participating countries, only a restricted amount of variables could be incorporated into the Hungarian part of EAST PaC. With regards to *background variables*, name and gender are available for each candidate. Year of birth was collected from candidate and MP CVs (if provided), web sources and printed almanacs. In the case of candidates placed lower on partly lists and those of small parties, this information was often not available at all.

Before turning to *core variables*, a few words on the Hungarian electoral system are in order. During the period under investigation, Hungary has a three-tier electoral system. On the first tier, candidates are elected in 176 single member districts (SMDs) by the rules of absolute majority in the first round, and should the first round be invalid or unsuccessful, simple majority in the second. A maximum of 152 and a minimum of 58 legislators receive regional and national list mandates from the second and the third tiers, where closed party lists are nominated. Voters cast two votes: one for a candidate in their SMD and one for the party list in their region. The sum of all non-utilized votes from the first valid rounds on both the SMD and the regional list tier forms the basis of mandate allocation on the national level.¹⁷

Based on this, the following core variables were collected using official electoral data: constituency and county of nomination (for SMD and second tier candidates), number and percentage of votes (for SMD candidates), election success and nominating party for all candidates. For the 2010 elections, *additional variables* were coded such as party list information, also using official data.

Merging Election Data

The restricted number of variables is important because it causes substantial problems at the stage of merging. Merging is the process of fitting together candidate data from various elections in wide form. The primary goal of merging is to produce a dataset that contains each candidate only once, and that defines each variable at different time points, so one can clearly see the candidates' electoral path. In other words, merging identifies if a candidate at election T is the same as a candidate in election (T-1). Obviously, the larger the number of variables describing the candidates' background, the easier it is to differentiate between candidates, thus the easier matching will get. The limited number of candidate-specific variables in the Hungarian dataset makes the matching procedure especially difficult. To increase reliability one has to establish two different set of rules: (1) for *automated data matching*, and (2) for *matching problematic cases*.

Automated Data Matching

Upon the start of the project, we were advised to conduct the automated part of matching based on three sets of information: candidate name, gender and year of birth. As male and female names are very distinctive, gender did not prove useful in the Hungarian case. Furthermore, as the availability of the year of birth variable was limited, it could not be used effectively as a matching variable. This left us with matching candidates solely on the basis of candidate names. Thus, the first task was to create a name variable that is suitable for matching. This is more problematic as it sounds, especially as there are more than 15 thousand observations in the dataset.

The first problem arises in every language that uses macrons or special characters: different software handle special characters in different, often not compatible ways. This was especially true in the case of the Hungarian data. While the dataset that contained candidates from 1990 to 2006 came partly without macrons, the 2010 data displayed Hungarian characters correctly. In the case of first names, this problem is relatively easy to solve, because these only have one version. As in the case of family names, different versions exist, the solution was not this straightforward. For example, the family names Hegedűs and Hegedüs look the same without the macrons, but they obviously mark two different names. The same applies to Győrfi - Györfi, Hajdú - Hajdu, Szöllősi - Szöllősi or Szűcs - Szücs, just to name a few. Only unambiguous cases could be corrected while scanning through the list of names. The rest had to be corrected in the stage of matching problematic cases.

The second problem relates to the structure of Hungarian names. As many candidates have prefixes¹⁸ in their names, which they use quite inconsistently, the easiest appears to be splitting the full names into components like prefix, family name, maiden name, first name and something we called middle name and extra name. Middle name refers to an undefined part between the family name and the maiden name (or first name in the case of male candidates), like the "F." in Török *F*. Tibor. Candidates have extra names if they have multiple first names. The first name listed became the first name in the dataset, and all other given names were coded as extra names (e.g. Takács László *Krisztián*). To obtain the above name components the trimming and splitting procedures of Excel were used. The automated splitting of the full name categorized all name components into the correct category based on their order in three-fourths of all cases. In the remaining fourth of all cases, manual corrections were in order. There were many problems because either that the candidates' family names, consist of multiple parts (e.g. *Tóth Mácsai* Árpád, where Mácsai is part of the family name), or that they are married women (Tarjányiné Bozóki Erzsébet, where Bozóki is the maiden name).

Automated matching was eventually carried out by using all the name components but the prefix. Unfortunately, the automated matching does not match candidates from different elections without error. First of all, there are a lot of candidates with the same name. Second, candidates are allowed to change how they appear on the ballot: they can drop the "Dr." prefix, and one or more of their extra names. The problem with this is that they do this inconsistently throughout the different elections (Vincze László appeared on the ballot as Vincze László Mihály in 2010, whereas earlier, he did not use his extra name). Thus, we lose the advantage of the additional information that prefixes and extra names could provide. Inconsistency might also rise from the changes in names over time. This is especially problematic in the case of female candidates who marry and change their names. Table 1 demonstrates that there are cases in which there is no indication of the candidate's birth name, which makes the identification of these candidates extremely difficult.

Versions	General example
Candidate's maiden name	Tóth Andrea
Candidate's husband's name	Szabó János Gábor
Husband's family name + Husband's first name + "né"	Szabó Jánosné
Husband's family name + Husband's first name + Husband's middle name	Szabó János Gáborné
+ "né"	
Husband's family name + "né" + Maiden family name + First name	Szabóné Tóth Andrea

Table 1. Versions for the names of married female candidates (fictional examples)

Husband's family name + Husband's first name + "né" + Maiden family	Szabó Jánosné Tóth
name + First name	Andrea
Husband's family name + First name	Szabó Andrea
Husband's family name + Maiden family name + First name	Szabó-Tóth Andrea
First letter of husband's family name + Maiden family name + First name	Sz. Tóth Andrea

It is rare for authorities to check if the name on the ballot matches the official name of the candidate. Although the following example is not from the general elections, it beautifully demonstrates how candidates may change their names on the ballot. The problem is that once a candidate is registered under a name, the NVI must put that name on the ballot, even if something clearly went wrong. At the 2006¹⁹ and 2010²⁰ local elections²¹ in a village called Tiszabő, the majority of registered candidates appeared using their nicknames. Local authorities referred to an official statement of the OVB²² from 1998 in which it declares that one can register using their stage-names or names that they use in public affairs. As 60 % of Tiszabő wears the name Mága, the identification of the candidates was very problematic for the voters²³. Therefore, it seemed reasonable to somehow differentiate between candidates. As a result, candidate Mága Gyula appeared on the ballot as "Little Gyula"²⁴, a different Mága Gyula as "The Son of Pipe-smoker Gyuszi"²⁵, Mága Zsolt as "Postman"²⁶ referring to his place of work, Surányi Gusztáv as "The Son of Limping Guszti"²⁷ and Turó Zoltán as "Rooster"²⁸.

Matching Problematic Cases

As described above, the five component of the candidate's name was used to match candidates. Due to this strict correspondence criteria -- five variables must coincide in order to reach a valid match -- it is very unlikely that false matches cause substantial problems in the cases of more complicated names (i.e. in cases where most of these variables take a value). In the case of simpler candidate names, the matching process might result in mismatches. For example, Szilassy Gábor Cézár or Tábori Lászlóné Márku Mária are very likely to find their matches, whereas we will most certainly have trouble with the name Szabó József, which consists of two very common parts. The procedure also leaves us with non-matched candidates who are either newcomers (i.e. correctly classified as candidates with no match) or were nominated under slightly (or completely) different names (i.e. incorrectly classified as mismatches). A common example is Deutsch Tamás, who - after getting married - changed his name to Deutsch-Für Tamás. Automated matching will identify him as two different persons. In a more extreme case, Rónaszéki Balázsné became Rónaszékiné Keresztes Mónika. As a consequence, on the stage of matching problematic cases, automated matching must be double-checked.

Table 2 identifies five problems that may result in mismatches at the stage of automated matching. These are the centre of attention at the stage of matching problematic cases. The table shows the results of the automated matching process. In other words, this is what the researcher sees at this stage. In the first case, one of the Kovács Bélás at election T was classified as not being nominated at the election (T-1). At the automated matching stage, the decision of which of the two candidates named Kovács Béla is matched with the Kovács Béla at election (T-1) is quite random: whichever comes first on the list of candidates from election T. Therefore one has to decide which Kovács Béla - if either - from election T has been nominated at the previous election.

Problems	Election (T-1)	Election T	
Problem nr. 1	Kovács Béla	Kovács Béla	
		Kovács Béla	
Problem nr. 2	Szabó József	Szabó József	
	Szabó József		
Problem nr. 3		Szabó József István	
Problem nr. 4	Tóth János	Tóth János	
	Tóth János	Tóth János	
Problem nr. 5	Ágoston Balázs	Ágoston Balázs	

Table 2. Examples for problematic cases after automated matching

The second problem is very similar, but the other way around: the name Szabó József appear twice at election (T-1), and once at election T. The question is whether the one at election T is the same as either those at election (T-1). Note the randomness of matching again: the result of matching these candidates depends on the order in which the two Szabó Józsefs appear at election (T-1). Additionally, there is a candidate called Szabó József István (Problem nr. 3), which is very similar to Szabó József. Remember that candidates can decide whether they want their extra names to appear on the ballot. As a consequence of this decision, Szabó József István at election T might be the same person as one the Szabó Józsefs at election (T-1). Problem nr. 4 demonstrates a combination of Problem nrs. 1 and 2. After the automated matching, both Tóth János's at election T were matched with Tóth János's at election (T-1), but - as the matching is solely based on their names - there is no guarantee that they are matches. They may be matches to different candidates or even be newcomers. Similarly, matched cases with unique names (Problem nr. 5) need to be double-checked to decide whether Ágoston Balázs at election T is the same as the candidate under a similar name at election (T-1), or they just have the same name.

This stage of matching consists of identifying and solving the problematic cases described above. The first step of identification is finding cases that were identified as newcomers (i.e. were not identified as match to any other candidate), but have names that already appear at previous elections (see Problem nr. 1). The second step is to find cases at election T that are classified as matches (see Problem nrs. 2, 4 and 5). The remaining candidates are hypothesized to have unique names, with one exception: similar names must be examined closely (see Problem nr. 3). For the identification of similar names, the family and first names were taken into account. Additionally, the dataset was organized in alphabetical order along the family name, as we tried to find similarities manually. Obviously, the order of these steps can vary, and there might be alternative, perhaps simpler solutions.

To demonstrate the extent of the problems introduced above, let us see a few examples from the Hungarian EAST PaC data. The specific task to solve is merging a dataset that contains candidates from the period 1990-2006 and the 2010 candidate dataset. The biggest problem with the Hungarian data is that there are a lot of candidates with very common Hungarian names. Family names Horváth, Kovács, Tóth, Papp, Farkas, Balogh, Kiss, Szabó, Nagy, Molnár, Németh, Varga, just to name a few, come up often with the most popular first names like László, István, János or József. To name two extreme cases, there are 26 candidates named Horváth László in the 1990-2006 dataset, and 3 were nominated in 2010 under this name. This means that one has to separate all 26 candidates running at the previous elections to be able to tell which of the 2010 Horváth Lászlós have been nominated at previous elections. The same problem creates a similar workload in the case of the name Kovács István, which appears 28 times at previous elections, and twice in 2010. Consequently, solving the above problems is time intensive.

The solution for these issues is problematic as, in most of the cases, double-checking the results of the initial matching must be done manually. An automated solution would be viable if there were additional reliable candidate-specific information available. The year of birth could serve as such auxiliary information, but there are two problems with this. First, as stated earlier, the year of birth data was not available for a certain part of the population. Second, even if it is available, sometimes, it gives no clear solutions. For example, there are two candidates named Király Zoltán born in 1948. Both have been running in 1990 and one was also nominated in 1994 and 1998. The question is what happens if this is the case?

There are several solutions. First is visual information (as in the case of Király Zoltán in the previous example): photos and videos served as extremely important clues in deciding who is who. Second, the details of electoral experience proved to be important as well. Party affiliation and geographical location were the two major types of information to lean on. This, of course, cannot handle party switching and inter-region mobility. As to party switching, during the period between 1990 and 2010 several parties have been formed and disappeared. Very often, if the party disappeared, the candidates re-appeared as candidates of another party at later elections. Inter-region mobility is very often associated with candidates of small parties. The reason for this is that as the nomination of regional party lists is dependent upon the ability to nominate SMD candidates at least a quarter of the constituencies per county (but in at least two), these parties are often not able to nominate a regional list in all counties. Additionally, they have a limited pool of potential candidates to draw from. Therefore, they nominate their candidates wherever they are entitled to nominate a party list.

Thirdly, any other additional information that helped identify the candidates was taken into account. In several cases, newspaper articles helped shed light on candidate identity; in other instances, the website of local governments provided us with information on the candidates' election history and political careers. In the most extreme cases, the candidate's e-mail address gave us the information about the presumed year of birth. Also, publicly available Facebook profiles and personal contacts helped us clear the picture.

Conclusions

In this chapter I discussed the process of data gathering in Hungary. The two most important challenges are related to the limited availability of candidate information and the difficulties of merging the data of different elections. Limited availability is the product of the restricted amount of publicly available systematically collected official information. Additionally, there are a lot of candidates who do not have any trace on the internet. This is particularly true for the candidates of small parties, independent candidates and those who were placed low on party lists. As a consequence of the limited number of variables, the data can only be used to investigate a limited array of questions.

Merging data from different elections also has problems. Naturally, if we know only a few things about a candidate, there are just a few things along which we can differentiate her from other candidates. The other consequence of the short information is that the limit of automated matching comes quite early in the process. Name-based automated matching leaves us with a lot of double-checking to do at the later stages.

Despite all the problems, the data can be considered valid and reliable across elections. The data are a good base to investigate the continuity of political elites, constituency-level election results and female participation in electoral politics. To widen the scale of research, future data collection efforts could focus on the local political background and the occupation of candidates as well as filling in party offices on the different levels of politics.

References

- Ilonszki, G. (2005) Képviselők Magyarországon I. (Hungarian Representatives I.). Budapest: Új Mandátum.
- Ilonszki, G. (2008) Amatőr és hivatásos politikusok Képviselők Magyarországon II. (Amateurs and Professionals Hungarian Representatives II.). Budapest: Új Mandátum.
- Ilonszki, G. (2009) Képviselők és képviselet Magyarországon a 19. és 20. században (Representatives and Representation in Hungary in the 19th and 20th Centuries). Budapest: Akadémiai Kiadó.
- Ilonszki, G. (2011) Szakértők és pártemberek Kormányzati elit Magyarországon 1848-2010 (Experts and Party People The Hungarian Governmental Elite 1848-2010). Budapest: Akadémiai Kiadó.
- Ilonszki, G., Dobos, G., Enyedi, Zs., Hajdú, A., Kovarek, D., Kurtán, S., Nábelek, F., Papp, Zs., Tóth, A., Várnagy, R., Zorigt, B. (2016) Pártok, jelöltek, képviselők (Parties, Candidates, Legislators). Budapest: Szabad Kéz Kiadó.
- Ványi, É. (2015) Jelentés a kormányról Kormányzati karrierút Magyarországon a rendszerváltástól napjainkig (Report on the Government Governmental Career Parths in Hungary after the Transition). Budapest: Századvég Kiadó.
- Várnagy, R. (2012) Polgármesterek a magyar országgyűlésben. Budapest: Budapesti Corvinus Egyetem.
- Weiss, R. E. (2006) Modelling Longitudinal Data. New York: Springer Science & Business Media

¹ "Circulation of the Parliamentary Elite in Hungary 1884-1998" (Principal Investigator: Gabriella Ilonszki; Grant Number: OTKA T32829)

² "Candidates and Parliamentarians. Individualistic and Partisan Representation" (Gabriella Ilonszki, OTKA K106220)

³ Parliamentary Activities, Career Paths and Accountability (PACTA) initiated by Julien Navarro and Federico Russo.

⁵ "Government ministers in Hungary 1867-2006: inertia and mobility, modernization and Europeanization" (Gabriella Ilonszki, OTKA K62628)

Centre for Elite Research at the Corvinus University of Budapest; DKMKA

⁷ Integrated and United project (INTUNE)

⁸ Participation and Representation project (PARTIREP), <u>http://www.partirep.eu/</u>

⁹ Political Leaders in European Cities (1st and 2nd round).

¹⁰ The data collection was administered by the Hungarian Election Study in 2010 and by the OTKA project "Candidates and Parliamentarians" in 2014.

¹¹ "Candidates and Parliamentarians. Individualistic and Partisan Representation" (Gabriella Ilonszki, OTKA K106220)

¹² http://valasztas.hu/.

¹³ From 2011 on, each candidate has to collect 500 signatures. However, contrary to the previous practice, one voter may sign for multiple candidates.

¹⁴ 28/2013. (XI. 15.) KIM rendelet

¹⁵ Based on correspondence with the chair of the Election Information Service, that is responsible for citizen orientation.

¹⁶ In the case of the 2010 election, the Office did not report on age and gender on the aggregate level.

¹⁷ From 2014 on, 199 representatives are elected in SMDs (106) and national party lists (93) in one round. Voters still have two votes, one cast for a candidate and one for a party list. All non-utilized votes on the SMD tier pools to the national level.

¹⁸ Hungarian prefixes are: Dr., Id. (Sen.), Ifj. (Jr.), Özv. (Widower). With some exceptions, prefixes come before the family name. In the case of married women, the Dr. prefix might be placed after the family name, if the doctorate is attached to the maiden name (e.g. Alakszainé Dr. Oláh Annamária). Furthermore, one candidate might have multiple prefixes in her name (e.g. Dr. Vidorné Dr. Szabó Györgyi, which indicates that Dr. Szabó Györgyi married someone named Vidor who also happen to hold a Dr. title).

¹⁹ http://valasztas.hu/ov06vt/maz/11/taz/053/j42112.html; Webpage accessed: February 17, 2016

²⁰ http://valasztas.hu/dyn/ov10vt/vertaj-static/ov10vt/maz/11/taz/053/j42112.html; Webpage accessed: February 17, 2016

²¹ http://www.origo.hu/itthon/onkormanyzati-valasztas-2010/hirek/20100908-beceneves-jeloltek-valasztasi-riporttiszaborol.html; Article accessed: January 15, 2016

http://www.szon.hu/becenevekkel-indul-a-jel246ltek-t246bbsege-a-valasztason-tiszabon---rtl-hirado/haonnews-charlotteInform-20100905-0745489750; Article accessed: January 15, 2016 ²² National Election Committee (Országos Választási Bizottság - OVB)

²³ Voters could not have known who they vote for, because all candidates appeared on the ballot as independent. Although the solution created quite funny results, it was necessary for voters to be able to decide who is who.

²⁴ Appeared on the ballot as "Mága Gyula Kis Gyula".

²⁵ Appeared on the ballot as "Mága Gyula Pipás Gyuszi Fia".

²⁶ Appeared on the ballot as "Mága Zsolt Postás".

²⁷ Appeared on the ballot as "Surányi Gusztáv Sántaguszti Fia".

²⁸ Appeared on the ballot as "Turó Zoltán Kakas".

[&]quot;Legislator activities and electoral outcome" (Zsófia Papp, NKFIH PD115747)