AUXIALLARY INFORMATION ON MERGING THE INDIVIDUAL-LEVEL DATASETS ON MALFEASANCE, PREFERENCE VOTES AND MEMBERS OF THE CHAMBER OF DEPUTIES, REPUBLIC OF ITALY, LEGISLATURES I-XI

Assembled by Miriam A. Golden
Department of Political Science
University of California at Los Angeles
under NSF SES-0074860
January 14, 2004

Construction of the Combined Dataset

This appendix describes the procedures used during the creation and cleaning of the combined individual level dataset containing information on members of the Italian Chamber of Deputies from the I to the XI Legislatures, as well as information on preference votes received by all deputies and DC and PSI candidates to those legislatures, and information on all judicial requests to remove parliamentary immunity (RAP) for the same period. The work was performed over the period from fall 2000 to spring 2002 (and following), under the direction of Miriam A. Golden, Department of Political Science, UCLA (**golden@ucla.edu**), with funding from the National Science Foundation.

The main goal in merging the various datasets was to assign each individual person a unique identifier (tid), regardless of how many times s/he appeared in the dataset because of multiple charges of malfeasance or because of having stood as a candidate multiple times, so that individual persons could be tracked over time (for e.g. survival analysis). Numerous discrepancies in the dataset rendered this difficult.

A. SOURCES

Individual level data are drawn from three sources:

1) Requests remove parliamentary immunity from deputies (*richieste di autorizzazione a procedere*, abbreviated in Italian as RAP) were provided in hardcopy by the Gruppo Democratici di Sinistra-L'Ulivo of the Italian Senate. The data were scanned in and corrected manually by Judit Bartha during AY 2000-01.

2) Preference vote data for all DC and PSI candidates to the Chamber of Deputies were input by hand by various research assistants, including David Yamanishi and Jorge Bravo, during AY 2000-01. The data were taken from *La Navicella*, a publication that reports information on individuals elected to the Italian Chamber of Deputies and the Senate for each legislature. Shuhei Kurizaki also subsequently input preference vote figures for deputies accused of malfeasance affiliated with other parties, also using information available in *La Navicella.*

The data on RAP had some fields all in capital letters (including in some instances last names of legislators), whereas the preference vote data did not. This problem was handled from within Excel, using the replacement function, which was used to capitalize all instances of first name, last name, and partyid in order to standardize the information across the two types of files.

3) Dataset of all deputies in the I to XI Legislatures, obtained in SPSS from Francesco Zucchini; original source Centro interdipartimentale di ricerca sul cambiamento politico (CIRCAP), University of Siena (referred to as the Verzichelli-Cotta dataset). For more information, see http://www.unisi.it/ricerca/dip/gips/circap/circap.html.

B. INITIAL CREATION OF A COMPLETE DATASET

In fall 2001, the malfeasance/preference votes file was merged with the Verzichelli-Cotta dataset by Shuhei Kurizaki, with assistance from Elizabeth Stein, using the sas code in the file individual.sas (available upon request). The basic function of the merge was to merge individuals who appeared identical on the basis of four characteristics: last name, party id, electoral district, and legislature. This in principle resulted in a merged file in which individual records contained information both on the number of preference votes a candidate had received and on the charges levied against him, were he named in a RAP once elected.

C. INITIAL ASSIGNMENT OF TIDS TO THE COMPLETE DATASET

The initial merged file contains an ID for each record. These IDs had to be corrected in order to assign a unique ID (tid) to each individual person in the file, so that analysts could track the individual's political career over multiple legislatures. We defined a " person" as an individual candidate and/or deputy, whose characteristics (legislature, party affiliation, etc.) might change.

In making decisions about individuals, errors in the underlying data that were identified were corrected. For instance, coders who looked up an individual in La Navicella often input the first name on the basis of that; spelling errors were corrected as noted; last names were standardized over time, etc.

Difficulties in assigning unique IDs occurred because of missing first names, changes in recorded first names, additions of middle names, changes in electoral district, changes in party affiliation, changes in party id, party mergers, etc.

These problems were analyzed by Shuhei Kurizaki and Liz Stein, who manually evaluated the records of individuals. In cases where individuals in the file were elected (and hence appeared in *La Navicella*), information regarding possibly identical individuals was sufficient to allow the coder to assign a unique ID as warranted. In these cases, the coder simply manually adjusted the ID so that all records for the same individual showed the same ID. (Subsequent numbers were automatically adjusted from within Excel.)

This process resulted in the creation of a file with fewer than 10,000 individual TIDs out of a total of more than 16,000 records. The output files were named individual.csv and individual.xls.

D. SUBSEQUENT CLEANING OF THE DATASET

Once an initial set of TIDs were assigned, Shuhei identified an apparent error in the underlying data deriving from the merge he had performed earlier. He corrected this by running a small piece of code (founded located at the bottom of individual.sas) on the dataset that he and Liz had worked on; i.e. on individual.csv. The output file was called individual2.csv.

E. ADDITIONAL CLEANING AND RECODING OF TIDS

Subsequently, David Yamanishi (Feb. 2002) again manually evaluated the dataset and removed additional errors and duplications in the TIDs for the 11th legislature. The resulting dataset is called individual3.xls. The corrections he made are documented in david-corrs.xls.

Another partial cleaning was performed by MG in March-April 2002 over all legislatures but only up through the C's (working alphabetically). This involved manually correcting TIDs and typographical errors. The output dataset is called individual4. Relevant notes are in miriam-corrs.txt.

A final cleaning was performed in June 2002 by David Yamanishi and Carmelo Barbera, who looked up all the discrepancies that appeared. For instance, they verified whether two deputies with the same first and last names but elected in two different districts from the same party was the same person or not by examining either *La Navicella* or another source identified below. They corrected typos in first and last names, etc. The output dataset is called individual5.xls. The file individual5-corrs.log contains all the notes about manual changes undertaken by Miriam Golden, David Yamanishi, or Carmelo Barbera.

Carmelo consulted the following sources to resolve discrepancies:

Istituto nazionale dell'informazione. *Repubblica italiana, 1948-1998. 50 anni di Parlamento, governi, istituzioni.* . INI/La Navicella. Rome: Editoriale Italiana, 2000.

1959-1963: una difficile transizione: verso il centro-sinistra. Vol. 18 of *Il Parlamento italiano, 1961-1988*. Milan: Nuova CEI, c1991.

1964-1968: il centro-sinistra, la stagione di Moro e di Nenni. Vol. 19 *of Il Parlamento italiano, 1861-1988*. Milan: Nuova CEI, 1992.

Giuseppe Restifo, ed. *La quarta legislatura del Parlamento della Repubblica: il centro-sinistra.* Vol. 19 *of Storia del Parlamento italiano*. Palermo: S.F. Flaccovio, 1983.

F. MERGER OF DATA ON MALFEASANCE WITH DATA ON ALL DEPUTIES IN LEGISLATURES I-XI

Judit Bartha then performed a merge between individual5.csv and the Verzichelli-Cotta dataset (partlidbis92.dta, a sorted Stata version of the original SPSS file parlitbis92.sav). The merge originally failed in more than 1,000 cases due to misspellings in last and first names. These were identified by Judit Bartha and Carmelo Barbera (using *La Navicella* and the Pasquino database of MPs, Gianfranco Pasquino, ed., *1945-1996. Archivio della politica in Italia*. Rome: Laterza Multimedia, 1996) and then corrected through code written in a stata "do" file (mergeindivparlit.do). The remaining 13 records

(concerning 8 deputies) that failed to merge did so because they were included in the RAP dataset but not the Verzichelli-Cotta dataset. The file questionmerge.xls contains a list of these MPs, indicated by "missing from parlisbis." The resulting dataset is called individual6.dta.

In February 2003, MG manually recoded instances where opinion had been left missing when charges were listed as "vedi allegato" as 1, on the argument that if the charges were too numerous to list, they were probably more serious than mere opinion charges. The resulting dataset is called individual7.dta.

In late 2003 and early 2004, MG made some additional corrections to that dataset of errors identified by users. The corrections are recorded in corrections_2003_12_13.do (generating individual8.dta) and corrections_2004_01_14.do (generating individual9.dta)

Notes recorded by Miriam Golden, Nov. 29, 2001; with assistance from Elizabeth Stein and Shuhei Kurizaki; rev. March 2002; rev. Oct. 2002.; rev. Nov. 2002; rev. Feb. 2003; rev. Jan. 2004.

Please submit questions or corrections to Miriam Golden (golden@ucla.edu).