

Replication files for:
“Democratization, Leader Education and Growth: Firm-level Evidence from
Indonesia”

30 November 2022

When using these data please cite as:

Pelzl, Paul and Steven Poelhekke (2022). Democratization, Leader Education and Growth: Firm-level Evidence from Indonesia. *Journal of Economic Growth*, *forthcoming*.

This document describes how the data used for this project can be obtained and contains relevant details (see DATA SOURCES below, and the folder “Data” in the replication folder), and provides a short guidance on the relevant STATA code (see GUIDANCE ON STATA CODE below, and the folder “dofiles” in the replication folder). Please do not hesitate to reach out to Paul (paul.pelzl@nhh.no) if you have further questions on any of the described content.

DATA SOURCES

Manufacturing (compare Section OA4.1)

- 1) *Plant census data*. Annual manufacturing plant census (IBS) data can be purchased from Indonesia’s national statistical agency (BPS) here: <https://silastik.bps.go.id>. After creating an account and logging in, go to Catalog -> Microdata -> Economic Statistics -> Survey of Large and Medium Manufacturing Industry (note that contrary to what the dataset name suggests, the data source is not a survey but a census).
- 2) *Correspondence tables*. In footnote 22 of the Online Appendix, we discuss how we identify a plant’s four-digit ISIC industry code. This process involves the use of correspondence tables to match industries across two different industry classifications. Specifically, matching industries across ISIC Rev. 3.1 and the Indonesian KBLI2000 classification is possible with the use of the excel file “kbli to isic correspondence” in the replication folder.

In the below table we link the variables used in our analysis to the original variable names in the manufacturing plant census data.

Variable name in STATA code	Name in paper (Tables)	Data source	Source varname(s) (unlogged) and details
ln_empl	ln(# Employees)	IBS	<i>LPRNOU+LNPNOU</i>
ln_rev	ln (Revenue)	IBS	<i>YPRVCU+YISVCU+YRNVCU</i>
ln_tfp	ln(TFP)	IBS	Authors' calculation, see Online Appendix OA2.4 in Pelzl and Poelhekke (2021), DOI: 10.1016/j.jinteco.2021.103457 More detailed information is available from the authors upon request.
ln_inv1	ln(1+Investment)	IBS	<i>FTTLCU</i>
ln_epw	ln (Wage Bill / #Employees)	IBS	$(ZPZVCU + ZNZVCU) / (LPRNOU + LNPNOU)$
indtax_va	Indirect Taxes / Value Added	IBS	<i>ITXVCU / VTLVCU</i>
gifts_va	Gifts, Donations etc. / Value Added	IBS	<i>ICOVCU / VTLVCU</i>
id	NA (this is the plant identifier variable, used e.g. for plant fixed effects)	IBS	<i>PSID</i>
kabID97	kabID97 is the 1997 district identifier variable	IBS, district proliferation data (see next column)	First two digits of the current district identifier: <i>DPROVI</i> . Last two digits: <i>DKABUP</i> . To aggregate to the 1997 district borders and create the unique 1997 district identifier kabID97, we use the files "kabupaten codes 1993 - 2002", "kabupaten codes 2003", and "kabupaten codes 2003 - 2009" in the replication folder.
provID97	provID97 is the 1997 province identifier variable	IBS	<i>DPROVI</i>
isic4	Used for variable indyear, see <i>master.do</i> in replication folder	IBS	First four digits of <i>DISIC5</i> , where ISIC stands for "Indonesia's Standard Industrial Classification" rather than "International Standard Industrial Classification". In Indonesian language, the prior is called KBLI, and the version is KBLI2000. This needs to be translated to the international classification ISIC Rev.3.1 using excel file "kbli to isic correspondence" in the replication folder. For example, the plant census data contains kbli2000 code 2660, which does not exist in ISIC Rev. 3.1.

Timing of mayor appointment (compare Section OA4.2)

Data on the timing of mayor appointment (in the case of Suharto mayors, via direct appointment by the Suharto regime; in the case of democratic mayors, via election) are largely (see below for some exceptions) obtained from the data repository of Monica Martinez-Bravo and Andreas Stegmann, which can be accessed here: <https://www.cemfi.es/~martinez-bravo/mmb/Data.html>.

- 1) We use the file *appointment_dates.dta* in *Appointment_Dates.zip* to obtain information on the **election year of the first democratic mayors**. The relevant variable in the dataset is *apdate_predir*, which we first translate to a new variable *appyear_predir* and then rename it to *elecyear_firstdem* in the dofiles, to make the variable name self-explanatory.
- 2) To obtain information on the **appointment year of the last Suharto mayors**, we first use the variable *apdate_predec* in *appointment_dates.dta*, which is the appointment date of the “last New Order bupati (pre-1999)”, according to *appointment_dates.pdf*. This implies that for the year 1998, only mayors that were appointed before May 21 1998 were appointed by the Suharto regime, and thus belong to the group of “last Suharto mayors”.

As described in *appointment_dates.pdf*, *appointment_dates.dta* only includes data on *apdate_predec* for a limited number of districts. While additional data are included in *appointment_dates_cornell.dta*, this dataset is not publicly available, but only accessible via Cornell University pages. Since we were unable to follow this route, we filled the missing data for all districts in our sample using the district-specific Wikipedia page. The data can be found in the STATA file “district-level variables and selected mayor-level variables.dta”, under the name *appyear_lastsoe*. Rather than only reporting the additional data we collected ourselves, to ease data usage we include the data points for all districts in our sample, thus also data points sourced from the data repository of Monica Martinez-Bravo and Andreas Stegmann. Therefore, when using these data, please cite both

Martinez-Bravo, Monica and Andreas Stegmann (2018). Indonesia data repository.
<http://www.cemfi.es/martinez-bravo/mmb/data.html>.

and

Pelzl, Paul and Steven Poelhekke (2022). Democratization, Leader Education and Growth: Firm-level Evidence from Indonesia. *Journal of Economic Growth*, forthcoming.

- 3) To obtain information on the **election year of the second democratic mayors**, we use the variable *apdate_posdir* in *appointment_dates.dta*, and call the variable *elecyear_seconddem* in the dofiles.

Mayor-level variables (compare Section OA4.3)

Data on mayor characteristics are largely obtained from the data repository of Monica Martinez-Bravo and Andreas Stegmann, which can be accessed here: <https://www.cemfi.es/~martinez-bravo/mmb/Data.html>. Since the repository contains multiple relevant datasets (.dta files), below we specify which dataset is used for a given variable.

- 1) We use the file *appointment_dates.dta* of the repository to obtain information on the **first democratic mayor's education level, gender, and the direction of the college degree** (the latter is inferred indirectly, as described on page 45 of the Online Appendix). In the dofiles, the variables are called *college_firstdem*; *female_firstdem*; and [*econbusfin_firstdem*, *politics_firstdem*, *law_firstdem*], respectively.
- 2) The repository does not explicitly report the **first democratic mayor's age at the time of election** (nor does it explicitly report his/her year of birth), whether or not he/she was born in the district that he/she leads (**nativeness**), whether the mayor **worked in the private sector pre-election**, or the democratic mayor's political **party membership**. However, we infer information on these variables in three ways.

First, *backgrounds_core.dta* and *backgrounds_ext.dta* in *Backgrounds_Mayors.zip* contain information on the backgrounds of Indonesian mayors that held office in the year 2009. Some of these mayors were in fact also the first democratic mayors (they were re-elected and still in power in 2009), such that for these districts we can use the two datasets to learn the first democratic mayor's age at election (use the variable *date_birth* in combination with the variable *appdate_predir* from *appointment_dates.dta* to infer age), nativeness (compare the variable *kab_birth* with the district the mayor leads), private sector status (for this variable, only *backgrounds_core.dta* contains relevant information; use the variables *occup_bef1999* and *occup_bef1999_oth*) and political party membership (use the variable *pol_party*).

Second, *backgrounds_soeharto.dta* contains information on the year of birth and nativeness of the last Suharto mayors. For those of them that were elected as first democratic mayors, we can use these data as well.

Third, for all remaining mayors (districts), we attempted to obtain information on the first democratic mayors' birth year, nativeness, occupation, and political party membership via online search, and were successful for most districts (see the Online Data Appendix). We provide the data on these four mayor-level variables in "district-level variables and selected mayor-level variables.dta". The variables are called *age_elec_firstdem*, *native_firstdem*, *private_firstdem*, and *golkar_firstdem*, respectively, where the latter equals one if the mayor is member of the Golkar party. Rather than only reporting the additional data we collected ourselves, to ease data usage we include the data points for all districts in our sample, thus also data points sourced from the data repository of Monica Martinez-Bravo and Andreas Stegmann. Therefore, when using these data, please cite both

Martinez-Bravo, Monica and Andreas Stegmann (2018). Indonesia data repository.
<http://www.cemfi.es/martinez-bravo/mmb/data.html>.

and

Pelzl, Paul and Steven Poelhekke (2022). Democratization, Leader Education and Growth: Firm-level Evidence from Indonesia. *Journal of Economic Growth*, *forthcoming*.

- 3) We use the file *backgrounds_soeharto.dta* to measure the **education level of the last Suharto mayor**, as described on pages 45-46 of the Online Appendix. The variable is called *college_lastsoe* in the dofiles.
- 4) We use the file *appointment_dates.dta* to infer the **education level of the second democratic mayor**. Specifically, we analyze the title(s) (if any) of the second democratic mayor, using the variable *bup_name_posdir*; see page 46 of the Online Appendix for details.
- 5) The data on **first democratic mayor-level corruption involvement** can be found in “district-level variables and selected mayor-level variables.dta” in the replication folder (variable name = *corr_status*; compare page 47 of the Online Appendix). As the variable label in the STATA file also indicates, the corruption status can take one of the following realizations: 1=research, 2=official investigation, 3=declared official suspect, 4=convicted, 5=acquitted. For details, please see pages 46-48 of the Online Appendix. For the original mayor-specific data source(s), see the excel file “mayor-level corruption outcomes and sources” (note that this file also includes data on the duration of the prison sentence in years, for convicted mayors) and complementary documents, all of which can be found in the folder “Mayor-level corruption dataset (incl sources)” in the replication folder. All source content is in Indonesian language (Bahasa Indonesia); we translated the content to English using Google Translate.

District-level variables (compare Section OA4.4)

- 1) The outcomes of the 1999 legislative elections at the district level (DPRD) are obtained from Kevin Evans’ website <http://pemilu.asia/>. While the data are presented as one pie chart per district displaying the vote share of each party, we bring the data to excel format (not only for the districts in our final sample, but for all districts on the website) and include the resulting excel file “DPRD 1999 election results” in the replication folder. Note that for a few districts, the sum of vote shares across all parties does not equal 100 in the raw data, and the difference to 100 is too large to be explained by rounding issues (see variable *Total* in the excel file). Therefore, for our analysis we compute a specific party’s vote share as the ratio of the reported vote share value and the variable *Total*. In our analysis, we use the data to compute the two variables **Golkar wins 1999-elections** and **1999-election vote share HHI**, and provide them in the STATA file “district-level variables and selected mayor-level variables.dta” in the replication folder, for the 96 districts in our final sample. In that dataset the variables are called *golkar_win* and *HHI_elec*, respectively.
- 2) **Population** data come from the Indonesian Census of 2000. The data can be accessed free of charge at IPUMS International: <https://international.ipums.org/international/index.shtml>. Using the appropriate weights (the variable *PERWT*), we use the micro-data to compute district-level population of 1997-districts. To match the district identifier in IPUMS International (which is called *GEO2_ID2000*) with our 1997 district identifier *kabID97* (see also the above table), we use the label of the *GEO2_ID2000* district codes, indicating the current name of the district. The list of district labels and *GEO2_ID2000* codes can be obtained at the following link: <https://international.ipums.org/international->

[action/variables/GEO2_ID2000#codes_section](#). We include data on the population variable in “district-level variables and selected mayor-level variables.dta”; the variable is called *pop_00*.

- 3) **Population density** is computed as the ratio of population and district area in square miles. The latter is computed based on shapefiles provided by the BPS, which we make available in the replication folder (see the folder “shapefiles etc. (for district area)”). We also provide area data directly in the replication folder, in “district-level variables and selected mayor-level variables.dta”, under the name *area_in_sqmiles*.
- 4) GDP data are obtained from the Indonesia Database for Policy and Economic Research (INDO DAPOER): <https://databank.worldbank.org/source/indonesia-database-for-policy-and-economic-research>. Since we measure GDP in 2000 and 2000 is the base year in the dataset, GDP in current IDR and GDP in constant IDR are equivalent, and thus both can be used (for current prices for example, the variable name in INDO DAPOER is *Total GDP including Oil and Gas (in IDR Million), Current Price*). We include GDP in 2000 in “district-level variables and selected mayor-level variables.dta”; the variable is called *GDP_00*. **GDP per capita** is easily computed by dividing GDP in 2000 by population in 2000.
- 5) Data on the **average education level of the working age population** come from the Indonesian Census of 2000, via IPUMS International (see above). Specifically, we use the variable *EDATTAIN*. Note that in the raw data this variable takes one of the following realizations: 0=Not in Universe, 1=Less than primary completed, 2=Primary completed, 3=Secondary completed, 4=University completed. We replace 0=Not in Universe by missing, and subtract *EDATTAIN* by one, such that 0=Less than primary completed, 1=Primary completed, 2=Secondary completed, 3=University completed. We then compute the average realization per district among the working age population (age 15-65; here we use the variable *AGE*), using the appropriate weights (*PERWT*). The variable is called *educ_workpop_00* in “district-level variables and selected mayor-level variables.dta”.
- 6) **Religious fractionalization** is computed as a Herfindahl-Hirschmann Index (HHI) based on the district-specific shares of each religion (muslim, buddhist, hindu, christian, other). The data are from the Census of 2000, via IPUMS International (see above), where the relevant variable is called *RELIGION*. In “district-level variables and selected mayor-level variables.dta”, we call the variable *HHI_relig*.
- 7) District-level **expenditure data** are from Indonesia’s ministry of Finance: http://www.djpk.kemenkeu.go.id/?page_id=321. For the year 2000, the dataset is called “Data per Daerah”. For the other years, the dataset is called “Realisasi”. All relevant variables can be found in “district-level variables and selected mayor-level variables.dta”; type “des” in STATA to see the corresponding variable labels.
- 8) Data on **local institutional quality** and the **quality and availability of physical infrastructure** are from the Regional Autonomy Watchdog KPPOD. The data can be accessed here: (for the general page containing research output of KPPOD, see here: <https://www.kppod.org/penelitian/>)

2002 data: https://www.kppod.org/backend/files/laporan_penelitian/rating2002.pdf (pp179-183; pdf pp 187-191)
2003 data: https://www.kppod.org/backend/files/laporan_penelitian/rating2003.pdf (pp219-224; pdf pp 229-234)
2004 data: https://www.kppod.org/backend/files/laporan_penelitian/rating2004.pdf (pp 102-105; pdf pp 119-122)

The above-linked pdf files can also be found in the folder “KPPOD” in the replication folder. We digitize these data for all districts that are surveyed by KPPOD and provide them in excel and STATA format in the “KPPOD” folder. The digitized dataset does not only include data used in the paper, but also data on additional categories, specifically: *Social Politics, Regional Economy, and Labor & Productivity* (for variable-specific data sources and computation details, see e.g. Appendix 3 (pdf page 111ff) and Appendix 4 (pdf page 116ff) in rating2002.pdf). For the 96 districts in our final sample, we also include the variables used in our analysis in “district-level variables and selected mayor-level variables.dta”.

- 9) In a robustness check (Table OA10, column 7), we drop districts that split over 1990-1997. For 1990-1993, information on district splits comes from Matthew Gudgeon and Samuel Bazzi, and is listed below.

SIJUNJUNG (kab id in 1990: 1303) split into SIJUNJUNG and KOTA SAWAHLUNTO in dec 1990.
LAMPUNG UTARA (1990 id: 1803) split into LAMPUNG UTARA and LAMPUNG BARAT in Aug 1991.

TANGERANG (3219) split into TANGERANG and KOTA TANGERANG in Feb 1993

BADUNG (5103) split into BADUNG and KOTA DENPASAR in Jan 1992.

LOMBOK BARAT (5201) split into LOMBOK BARAT and KOTA MATARAM in Dec 1993.

MINAHASA (7103) split into MINAHASA and KOTA BITUNG in Aug 1990.

DONGGALA (7203) split into DONGGALA and KOTA PALU in Jul 94

KENDARI (7403) split into KENDARI and KOTA KENDARI in Aug 95

MALUKU UTARA (8103) split into MALUKU UTARA and HALMAHERA TENGAH in Dec 1990

JAYAPURA (8203) split into JAYAPURA and KOTA JAYAPURA in Aug 1993.

GUIDANCE ON STATA CODE

The master dofile is “master.do”, which can be found in the folder “dofiles” in the replication folder. It calls the file “datafile_preregs.dta” which is a placeholder for the file that merges the proprietary manufacturing plant-level data with all other data. The dofile calls separate dofiles, one dofile per table, all of which can also be found in the “dofiles” folder.