

Data and estimation procedures for GSJ-18-1100

1 Data file generation

1.1 Association-year panel of macro variables and Elo ratings

We first generate a Stata dta-file, which contains a panel at the level of the calendar year-association. This allows to match the end-of-year Elo rating to macro-economic, historical and climatic variables. The resulting dta-file is called: **"ELOcountries.dta"**. We do this using the program contained in the Stata do-file: **"panelmerger.do"**.

We merge in the following files:

- **uniquecountries.dta**
 - Source: hand collected Wikipedia + Hofstede
 - Specifies each country-year combination and identifiers. Also contains the country's former colonizer and Hofstede measures for cultural values.
 - Variables:
 - Team/op_team: name of national football association
 - Countryid: identifier for country
 - Country: country name harmonized
 - Continent: continent of country
 - weird_country: indicator for defunct or entrant country over panel time period
 - present: indicator for country present in data
 - countrycode/countrycodemac: World Bank country code for matching
 - colony: indicator whether country was colonized
 - colonizer: name of main colonizer
 - colonizercode: World Bank code for main colonizer
 - power distance: Hofstede measure for power distance in country or country region
 - individualism Hofstede measure for individualism in country or country region
 - masculinity: Hofstede measure for masculinity in country or country region
 - uncertainty avoidance Hofstede measure for uncertainty avoidance in country or country region
- **elopanel.dta**
 - Source: <http://www.eloratings.net/>
 - Contains various Elo ratings by year (other variables as defined above):
 - Originalcountry: country name original
 - Year: calendar year
 - elostart_avg: average Elo at start of games played in year
 - eloend_avg: average Elo at end of games played in year
 - elostart_max: maximum Elo at start of games played in year
 - eloend_max: maximum Elo at end of games played in year
 - elostart_min: minimum Elo at start of games played in year
 - eloend_min: minimum Elo at end of games played in year
 - elostart_sd: standard deviation of Elo at start of games played in year
 - eloend_sd: standard deviation of Elo at end of games played in year
 - elo_n: number of games with Elo played in year
 - startyearelo: Elo before first game in year
 - endyearelo: Elo after last game in year

We will mainly use the last variable in further analyses.

- **macrovariables.dta**

- Source: World Bank
- Contains macro-economic and population variables. In the do file we select those with the widest coverage and in agreement with previous literature for later use in our analysis.
- Variables with respective World Bank code (others as defined above):
 - BirthRatePer1000: Birth rate, crude (per 1,000 people) = SP.DYN.CBRT.IN
 - GDPCurrentUS: GDP (current US\$) = NY.GDP.MKTP.CD
 - GDPGrowth: GDP growth (annual %) = NY.GDP.MKTP.KD.ZG
 - GDPperCapitaUS: GDP per capita (current US\$) = NY.GDP.PCAP.CD
 - GDPperCapitaGrowth: GDP per capita growth (annual %) = NY.GDP.PCAP.KD.ZG
 - GNI_US: GNI (current US\$) = NY.GNP.MKTP.CD
 - InflationGDPDeflator: Inflation, GDP deflator (annual %) = NY.GDP.DEFL.KD.ZG
 - LandArea: Land area (sq. km) = AG.LND.TOTL.K2
 - MortalityRatePer1000: Mortality rate, infant (per 1,000 live births) = SP.DYN.IMRT.IN
 - MortalityRateUnder5: Mortality rate, under-5 (per 1,000 live births) = SH.DYN.MORT
 - MortalityRateAdultFemale: Mortality rate, adult, female (per 1,000 female adults) = SP.DYN.AMRT.FE
 - MortalityRateAdultMale: Mortality rate, adult, male (per 1,000 male adults) = SP.DYN.AMRT.MA
 - Pop0to14: Population ages 0-14 (% of total) = SP.POP.0014.TO.ZS
 - Pop15to64: Population ages 15-64 (% of total) = SP.POP.1564.TO.ZS
 - PopOver65: Population ages 65 and above (% of total) = SP.POP.65UP.TO.ZS
 - PopDensity: Population density (people per sq. km of land area) = EN.POP.DNST
 - PopGrowth: Population growth (annual %) = SP.POP.GROW
 - Population: Population, total = SP.POP.TOTL
 - RuralPop: Rural population (% of total population) = SP.RUR.TOTL.ZS
 - UrbanPop: Urban population (% of total) = SP.URB.TOTL.IN.ZS
 - LifeExpectancy: Life expectancy at birth, total (years) = SP.DYN.LE00.IN

- **weather.dta**

- Source: World Bank
- Climatic data for country capital
 - Annual_temp: average annual temperature
 - Annual_precip: average annual precipitation

- **uksplit.dta**

- Source: UK statistics
- Contains sub-national characteristics on the UK to split out the four footballing associations within the country: England, Northern Ireland, Scotland and Wales.
- Variables:
 - Year: calendar year
 - Alluk: overall UK population
 - Pop_eng/sco/ni/wal: population in each part of UK
 - Land_eng/sco/ni/wal: land area in each part of UK
 - Id_eng/sco/ni/wal: identifiers for each part of UK

1.2 Game-level data files on match outcomes and coaches

We generate a data file at the level of the individual game observation. We merge the coaches, coach characteristics and country-level variables into one file called “**ELOgamedata.dta**” using the do-file: “**gamelevelmerger.do**”.

NOTE: we have anonymized the coach names for privacy reasons.

We merge in the following files:

- **finalgamescoaches.dta**
 - Source for game outcomes: <http://laenderspiel.cmuck.de/>
 - Sources for coaches (hand collected to match games):
 - Wikipedia
 - laenderspiel.de
 - footballdatabase.eu
 - transfermarkt
 - uk.soccerway.com
 - Variables (op_ prefix indicates same variable for opposing team):
 - day: calendar day of game
 - month: calendar month of month
 - year: calendar year of game
 - gameid: identifier of the game
 - firstteam: identifier for team within game
 - firstgame: indicator if team played twice on 1 date, 1=first game, 0=second game
 - team: name of team
 - oldteam: old name of team from previous data collection
 - teamid: identifier for the team
 - home: formal home team indicator
 - homereal: indicator for factual home game based on venue location
 - sc: score obtained by team in game (incl extra time if applicable)
 - godif: goal difference between both teams in game
 - gosum: total number of goals in game
 - penaltywin: indicates game won in penalty shoot-out (if applicable)
 - coach: name of current coach REPLACED BY NUMBER TO KEEP ANONIMITY
 - coachnat: nationality of current coach
 - confederation: confederation in which team plays
 - competition: competition level as defined for elo rating (friendly – other – continental qual – continental final – world qual – world cup final)
 - venue: city in which game is played
 - tournament: name of tournament
 - startday: calendar day of coach appointment
 - startmonth: calendar month of coach appointment
 - startyear: calendar year of coach appointment

Important note: penalty wins are coded as draws unlike what is done at the source webpage. Final score after regular or extra time is taken as the score. Indicator “penaltywin” denotes winner of the shootout.

- **coachcharacteristics.dta**

- Sources: hand collected from Wikipedia, footballdatabase and Transfermarkt. Structured as a panel, where all experience refers to past experience.
- Variables:
 - coach: name of coach REPLACED BY NUMBER TO KEEP ANONIMITY
 - coachnat: nationality of coach as for football purposes, e.g. Scotland!=England
 - yob: year of birth DROPPED TO KEEP ANONIMITY
 - unsearched: coach was not looked up because of very short spell in data
 - birthcountry: country of birth DROPPED TO KEEP ANONIMITY
 - lang1-4: languages spoken DROPPED TO KEEP ANONIMITY
 - playint: played for own national team
 - expint: years played for own national team
 - playpro: played as professional
 - exppro: years played as professional
 - expplaydom: years played in domestic league
 - expplayabroad: year played in leagues abroad
 - play1-9: names of countries played in
 - expplay1-9: years played in country 1-9
 - nat1-21: name of countries in which manager coached national team
 - nat1-21_exp: years of experience in coaching national team in country 1-21
 - nat1-21_start: first year at national team in country 1-21
 - nat1-21_end: last year at national team in country 1-21
 - club1-10: name of countries in which manager coached national team
 - club1-10_exp: years of experience in coaching club team in country 1-21
 - club1-10_start: first year at club team in country 1-21
 - club1-10_end: last year at club team in country 1-21
- **Elo ratings.dta**
 - We calculate Elo scores from the game data collected at <http://laenderspiel.cmuck.de/> in R. This program can be provided on request. The results are stored in this file. We rename the variables for Elo based on the 2000 starting values to elostart and eloend. Other Elo measures are not used in further analysis.
 - Variables:
 - game_id: A key identifying the game ID. This can identify the game and the anchor team for each observation for merging.
 - firstteam: number to identify a team within each game
 - countryid: ID number of team
 - op_countryid: ID number of op_team
 - country_Ro: Elo value for team at the start of the current game using all games in sample (first game initialized at 1325 for all teams)
 - op_Ro: Elo value for op_team at the start of the current game using all games in sample (first game initialized at 1325 for all teams)
 - country_Rn: Elo value for team at the end of the current game using all games in sample (first game initialized at 1325 for all teams)
 - op_Rn: Elo value for op_team at the end of the current game using all games in sample (first game initialized at 1325 for all teams)
 - country_Ro_NF: Elo value for team at the start of the current game using all non-friendly games in sample (first game initialized at 1325 for all teams)
 - op_Ro_NF: Elo value for op_team at the start of the current game using all non-friendly games in sample (first game initialized at 1325 for all teams)

- country_Rn_NF: Elo value for team at the end of the current game using all non-friendly games in sample (first game initialized at 1325 for all teams)
- op_Rn_NF: Elo value for op_team at the end of the current game using all non-friendly games in sample (first game initialized at 1325 for all teams)
- country_Ro_WELOstart: Elo value for team at the start of the current game using only games starting in 2000 and on (first game initialized using most recent pre-2000 World Elo Football value, or World Elo Starting value if country began play after 2000)
- op_Ro_WELOstart: Elo value for op_team at the start of the current game using only games starting in 2000 and on (first game initialized using most recent pre-2000 World Elo Football value, or World Elo Starting value if country began play after 2000)
- country_Rn_WELOstart: Elo value for team at the end of the current game using only games starting in 2000 and on (first game initialized using most recent pre-2000 World Elo Football value, or World Elo Starting value if country began play after 2000)
- op_Rn_WELOstart: Elo value for op_team at the end of the current game using only games starting in 2000 and on (first game initialized using most recent pre-2000 World Elo Football value, or World Elo Starting value if country began play after 2000)

2 Main empirical analyses

2.1 Link ELO to macro variables to infer country technology

Run **linkELOtomacro.do**.

This estimates the model of ELO as function of macro variables and the technology parameters for each country. We produce the following output:

- Estimation tables for various 1st stage models:
 - **firststagemodel.xml**: selection of estimates of models above presented in Table 4.
 - **firststageappmodel.xml**: selection of other estimates for appendix table A1
 - OLSmodel.xml: all results for OLS model
 - quantmodel.xml: all results for quantile regression
- **linkELOmacro.log**: log file which contains
 - Country table of various technology measures and Elo
 - Top 20 is presented in Table 6
 - All countries depicted in map in Figure 1
 - Correlation tables for technology measures
- **paneltech.dta**: dataset for model linking technology to migrating coaches
 - Variables refer to different technology estimates:
 - tech3: technology all macro variables included
 - tech1: technology based only on population
 - elo: raw elo scores
 - Variables refer to different time frames:
 - lag1: tech level in year t-1
 - rol5: average tech level in year t-5 through t-1
 - start: average tech level in 5 year window before start of sample period (year 2000)

2.2 Additional analysis: link migration to technology

Run **linktechtomigration.do**.

This estimates the link between source country technology and migration decisions.

Output:

- The data file for this analysis: **migrantcoaches.dta**. This is a panel at the source country – year level. It has macro variables and the number of coaches active/hired in a year in other countries.
- **migrantapp.xml** Model not reported in R1
 - Dependent variables:
 - number of coaches from source country active in other countries in year
 - number of coaches from source country hired in other countries in year
 - Estimation: (zero-inflated) negative binomial
- **migrantextra.xml** Other models with same dependent as baseline model but different measures of technology and timing of technology.

2.3 Link performance to hiring a migrant from high know-how country

Run **linkperformancetomigrant.do**.

This estimates the impact of a high tech manager on the performance in terms of Elo ratings. The input data for this is **ELogamedata.dta** combined with **paneltech.dta**. This merge is done in the first part and some additional variables are created:

- For the technology variables (* = tech1, tech3 or elo):
 - coach*: the technology level of the coach's country of origin estimated at the start of the sample
 - team*: the technology level of the coach's country of origin estimated over the five years leading up to the hire of the coach
 - dif*: difference between coach* and team*
- For the playing career:
 - playabroad: indicator = 1 if played outside home country, 0 otherwise
 - playpro: indicator = 1 if played professional, 0 otherwise
 - playint: indicator = 1 if played for own national team, 0 otherwise
 - playcountry: indicator = 1 if played in country where current coaching jobs is, conditional on being migrant, 0 otherwise
 - playcountry: indicator = number of years played in country where current coaching jobs is, conditional on being migrant, 0 otherwise
- For the coaching career:
 - coachabroad: indicator = 1 if coached outside home country, 0 otherwise
 - coachcountry: indicator = 1 if coached before in country where current coaching jobs is, conditional on being migrant, 0 otherwise
 - coachcountry: indicator = number of years coached in country where current coaching jobs is before current spell, conditional on being migrant, 0 otherwise
- For the technology variables (* = uncertainty masculinity individualism powerdistance):
 - Coach*: level in coach's home country
 - Team*: level in destination country
 - Dist*: Squared difference between team* and coach* divided by variance in *

Then the do-file performs various analyses to evaluate the hypotheses in the paper. We summarize this in the following output tables and figures:

- **perfFEfull.xml**: the baseline results as reported in Table 6 in the manuscript

- **perfFEpop.xml**: the baseline results using the technology measure which only controls for population, not reported
- **perfFEelo.xml**: the baseline results using the Elo as technology measure, reported in Table A2 online appendix
- **crossfull.xml**, **crosspop.xml**, **crosselo.xml**: the same as above, but taken as a cross-section at level of employment spell, shows consistent results, but cannot test H3, not reported.
- **perfFEprev.xml**: makes results for Table A3 in appendix, which groups specifications of the baseline results with controls for the previous manager. We first include the know-how of the previous manager and then do a split-sample analysis.
- **perfFEotherexp.xml**: groups specification of the baseline results with other measures for (international) experience, not reported except in reply
- **betatech3dist.gph** and **betadisttech3.gph**: the interaction coefficient figures, reported in Figure 2 in paper
- **betaelodist.gph** and **betadistelo.gph**, **betatech1dist.gph** and **betadisttech1.gph**: same as above, but for alternative technology measures, not reported.