Supplementary material for the paper Variable selection for hidden Markov models with continuous variables and missing data

Fulvia Pennoni[1], Francesco Bartolucci[2], Silvia Pandolfi[2],

[1]Department of Statistics and Quantitative Methods,
University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8,
20128 Milan, Italy
Email: fulvia.pennoni@unimib.it

[2]Department of Economics, University of Perugia,
Via Alessandro Pascoli 20, 06123 Perugia, Italy
Emails: francesco.bartolucci@unipg.it, silvia.pandolfi@unipg.it

Illustrative application

Here, we provide additional details related to the longitudinal data employed for the applicative example of choosing the best indicators to classify countries according to their development.

In particular, Tables from 1 to 6 describe the 25 macroeconomic World Development Indicators reporting their definition provided by the World Bank[1]. Table 7 lists the names of all countries.

[1]For more details on the World Development Indicators see the webpage: `https://www.worldbank.org/en/home`

Figures from 1 to 4 depict the values of the indicators for each country provided from years 2000 to 2017. We notice intermittent missing responses (in black) for many countries, and that some countries do not provide data during the last periods; see, for example, school enrollment, primary, (Sch1), secondary (Sch2), and tertiary (Sch3) in Figure 1, as well as research and development expenditure (Rese) in Figure 2. The main information we can glean from these figures is as follows:

- the variables with larger intermittent missing values are: Saf, Lit, and Gini;
- the variables with missing values for all the countries in one or more years are: Int, Ren, Saf, Ele, and Comb;
- the variables remarkably heterogeneous between countries across years are the following: Pop, GDP, Int, Ren, Hyd, and Fert;
- some countries do not record data in any year for Edu;
- the variables with remarkably increasing values for many countries across years are Ele and GDP.

Summary statistics of some indicators are reported in Tables 8 and 9, referred to years 2000 and 2017, respectively. Comparing the observed values between 2000 and 2017, we notice that almost all the averages across countries are higher in 2017. The highest relative standard deviation is observed for Int in 2000 and for GDP and Hea in 2017. The largest number of missing values is observed for Saf and Lit in 2000, whereas as previously noticed, Hea and Saf are missing for each country in 2017. Due to the observed skewness, as expressed in the paper, before implementing the variable selection procedure, we applied a logit transformation if the variables are expressed in a percentage scale, a Box-Cox transformation (Box and Cox, 1964) to the other variables, and we scaled all the variables.

Table 10 shows the observed values for each indicator for eight selected countries (Afghanistan, Bhutan, Cambodia, Ethiopia, Nepal, Papua New Guinea, Solomon Islands, Timor-Leste) over the most recent time period and the overall median values across all countries. We observe that Nepal is characterized by particularly high values of Gsav, and Sch1, Timor-Leste by high Imp and Trade values, and Bhutan by high Int.

We also provide in Tables from 11 to 14 the estimated standard errors obtained with the non-parametric bootstrap (Davison and Hinkley, 1997), based on 300 samples, for the transition probabilities reported in the main article in Tables from 5 to 8.

Table 1: *Description of the 25 socioeconomic indicators*

Label	Description
Life	*Life expectancy at birth*: Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
Pop	*Population ages 0-14*: Population between the ages 0 to 14 as a percentage of the total population. The population is based on the de facto definition of population.
Infa	*Infant mortality rate*: Infant mortality rate is the number of infants dying before reaching one year of age, per 1,000 live births in a given year.
Sch1	*School enrollment, primary*: Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education shown. Primary education provides children with basic reading, writing, and mathematics skills along with an elementary understanding of such subjects as history, geography, natural science, social science, art, and music.
Sch2	*School enrollment, secondary*: Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education shown. Secondary education completes the provision of basic education that began at the primary level, and aims at laying the foundations for lifelong learning and human development, by offering more subject- or skill-oriented instruction using more specialized teachers.

Table 2: *Description of the 25 socioeconomic indicators (cont.)*

Label	Description
Sch3	*School enrollment, tertiary*: Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education shown. Tertiary education, whether or not to an advanced research qualification, normally requires, as a minimum condition of admission, the successful completion of education at the secondary level.
Edu	*Government expenditure on education*: General government expenditure on education (current, capital, and transfers) is expressed as a percentage of GDP. It includes expenditure funded by transfers from international sources to government. General government usually refers to local, regional and central governments.
Gedu	*Gross national expenditure*: Gross national expenditure (formerly domestic absorption) is the sum of household final consumption expenditure (formerly private consumption), general government final consumption expenditure (formerly general government consumption), and gross capital formation (formerly gross domestic investment).
Rese	*Research and development expenditure*: Gross domestic expenditures on research and development (R&D), expressed as a percent of GDP. They include both capital and current expenditures in the four main sectors: Business enterprise, Government, Higher education and Private non-profit. R&D covers basic research, applied research, and experimental development.

Table 3: *Description of the 25 socioeconomic indicators (cont.)*

Label	Description
GDP	*GDP per capita*: GDP per capita based on purchasing power parity (PPP). PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2011 international dollars.
Une	*Unemployment*: Unemployment refers to the share of the labor force that is without work but available for and seeking employment. Definitions of labor force and unemployment differ by country.
Gsav	*Gross savings*: Gross savings are calculated as gross national income less total consumption, plus net transfers.
Ele	*Access to electricity*: Access to electricity is the percentage of population with access to electricity. Electrification data are collected from industry, national surveys and international sources.
Int	*Individuals using the Internet*: Internet users are individuals who have used the Internet (from any location) in the last 3 months. The Internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV etc..

Table 4: *Description of the 25 socioeconomic indicators (cont.)*

Label	Description
Ren	*Renewable electricity output*: Renewable electricity is the share of electricity generated by renewable power plants in total electricity generated by all types of plants.
Gini	*GINI index*: The Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. Thus a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.
Trade	*Trade*: Trade is the sum of exports and imports of goods and services measured as a share of gross domestic product.

Table 5: *Description of the 25 socioeconomic indicators (cont.)*

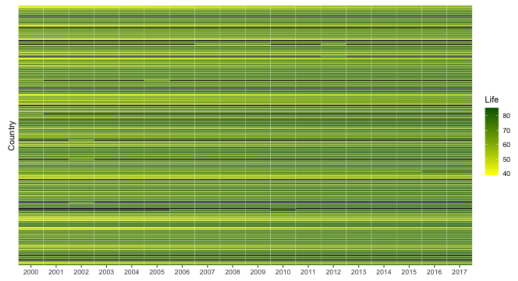
Label	Description
Saf	*Coverage of social safety net programs in poorest quintile*: Coverage of social safety net programs shows the percentage of population participating in cash transfers and last resort programs, noncontributory social pensions, other cash transfers programs (such as child, family and orphan allowances), conditional cash transfers, in-kind food transfers (such as food stamps and vouchers, food rations), school feeding, other social assistance programs (housing allowances, scholarships, fee waivers, health subsidies, and other social assistance) and public works programs.
Lit	*Literacy rate*: Adult literacy rate is the percentage of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life.
Hea	*Current health expenditure*: Level of current health expenditure expressed as a percentage of GDP. Estimates of current health expenditures include healthcare goods and services consumed during each year. This indicator does not include capital health expenditures such as buildings, machinery, IT and stocks of vaccines for emergencies or outbreaks.
Hyd	*Electricity production from hydroelectric sources*: Sources of electricity refer to the inputs used to generate electricity. Hydropower refers to electricity produced by hydroelectric power plants.

Table 6: *Description of the 25 socioeconomic indicators (cont.)*

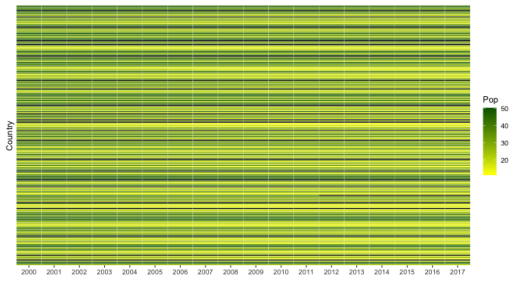
Label	Description
Imp	*Imports of goods and services*: Imports of goods and services represent the value of all goods and other market services received from the rest of the world. They include the value of merchandise, freight, insurance, transport, travel, royalties, license fees, and other services, such as communication, construction, financial, information, business, personal, and government services. They exclude compensation of employees and investment income (formerly called factor services) and transfer payments.
Comb	*Combustible renewables and waste*: Combustible renewables and waste comprise solid biomass, liquid biomass, biogas, industrial waste, and municipal waste, measured as a percentage of total energy use.
Lab	*Labor force participation rate*: Labor force participation rate is the proportion of the population ages 15 and older that is economically active: all people who supply labor for the production of goods and services during a specified period.
Fert	*Fertility rate*: Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.

Table 7: *List of the 217 countries providing data for the World Development Indicators*

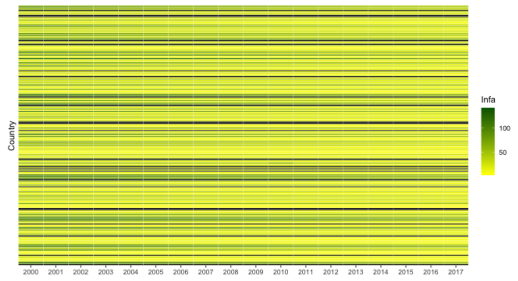
Country name			
Afghanistan	Djibouti	Lao PDR	Rwanda
Albania	Dominica	Latvia	Samoa
Algeria	Dominican Republic	Lebanon	San Marino
American Samoa	Ecuador	Lesotho	Sao Tome and Principe
Andorra	Egypt, Arab Rep.	Liberia	Saudi Arabia
Angola	El Salvador	Libya	Senegal
Antigua and Bar.	Equatorial Gui.	Liechtenstein	Serbia
Argentina	Eritrea	Lithuania	Seychelles
Armenia	Estonia	Luxembourg	Sierra Leone
Aruba	Eswatini	Macao SAR, China	Singapore
Australia	Ethiopia	Madagascar	Sint Maarten (D)
Austria	Faroe Islands	Malawi	Slovak Republic
Azerbaijan	Fiji	Malaysia	Slovenia
Bahamas, The	Finland	Maldives	Solomon Islands
Bahrain	France	Mali	Somalia
Bangladesh	French Polynesia	Malta	South Africa
Barbados	Gabon	Marshall Islands	South Sudan
Belarus	Gambia, The	Mauritania	Spain
Belgium	Georgia	Mauritius	Sri Lanka
Belize	Germany	Mexico	St. Kitts and Nevis
Benin	Ghana	Micronesia, F. S.	St. Lucia
Bermuda	Gibraltar	Moldova	St. Martin (F.)
Bhutan	Greece	Monaco	St. Vincent and the G.
Bolivia	Greenland	Mongolia	Sudan
Bosnia and Herzegovina	Grenada	Montenegro	Suriname
Botswana	Guam	Morocco	Sweden
Brazil	Guatemala	Mozambique	Switzerland
British Virgin I.	Guinea	Myanmar	Syrian Arab R.
Brunei Dar.	Guinea-Bissau	Namibia	Tajikistan
Bulgaria	Guyana	Nauru	Tanzania
Burkina Faso	Haiti	Nepal	Thailand
Burundi	Honduras	Netherlands	Timor-Leste
Cabo Verde	Hong Kong (SAR)	New Caledonia	Togo
Cambodia	Hungary	New Zealand	Tonga
Cameroon	Iceland	Nicaragua	Trinidad and Tobago
Canada	India	Niger	Tunisia
Cayman Islands	Indonesia	Nigeria	Turkey
Central African Rep.	Iran, Islamic Rep.	North Macedonia	Turkmenistan
Chad	Iraq	Northern Mari. I.	Turks and Caicos I.
Channel Islands	Ireland	Norway	Tuvalu
Chile	Isle of Man	Oman	Uganda
China	Israel	Pakistan	Ukraine
Colombia	Italy	Palau	United Arab E.
Comoros	Jamaica	Panama	United Kingdom
Congo, Dem. Rep.	Japan	Papua New Guinea	United States
Congo, Rep.	Jordan	Paraguay	Uruguay
Costa Rica	Kazakhstan	Peru	Uzbekistan
Cote d'Ivoire	Kenya	Philippines	Vanuatu
Croatia	Kiribati	Poland	Venezuela, RB
Cuba	Korea, Dem. P. R.	Portugal	Vietnam
Curacao	Korea, Rep.	Puerto Rico	Virgin Islands (U.S.)
Cyprus	Kosovo	Qatar	West Bank and Gaza
Czech Republic	Kuwait	Romania	Yemen, Rep.
Denmark	Kyrgyz Republic	Russian Federation	Zambia
Zimbabwe			



(a) Life expectancy at birth

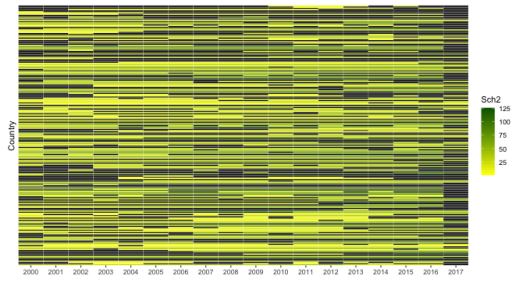


(b) Population ages 0-14

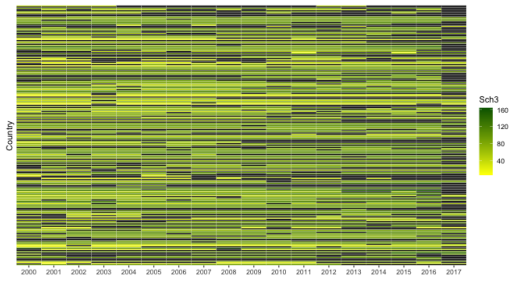


(c) Infant mortality rate

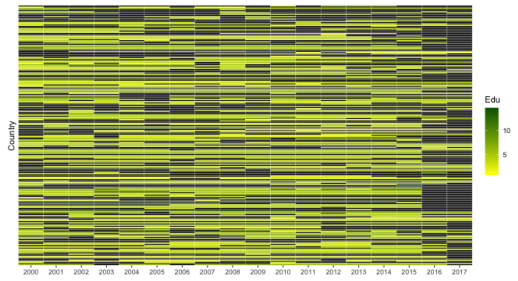
(d) School enrollment, primary



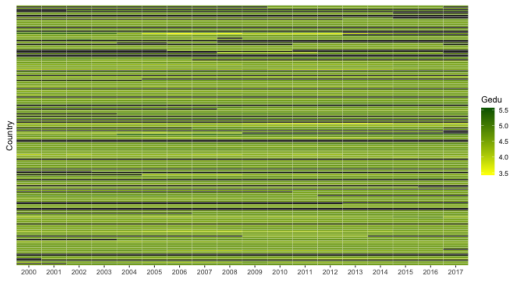
(e) School enrollment, secondary



(f) School enrollment, tertiary



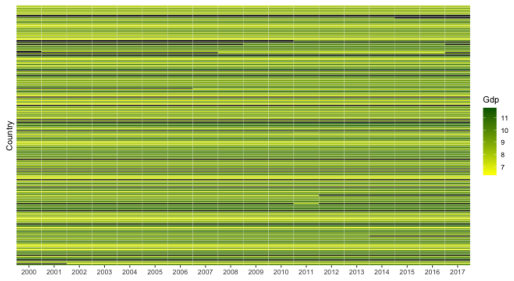
(g) Government expenditure on education



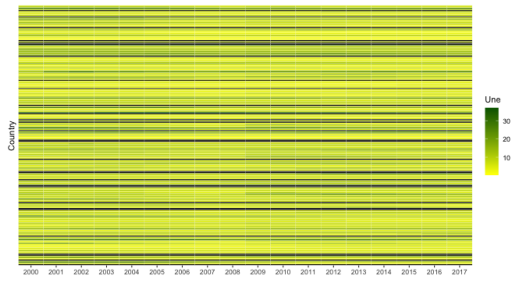
(h) Gross national expenditure

Figure 1: *Observed values of the 217 countries across time occasions from 2000 to 2017, missing values depicted in black*

(a) Research and development expenditure

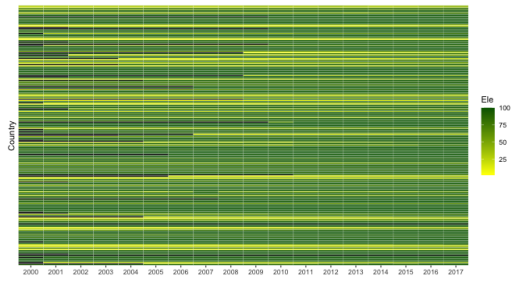


(b) GDP per capita

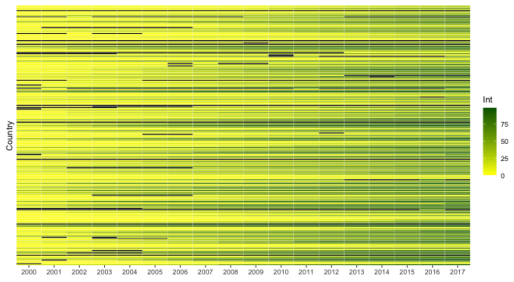


(c) Unemployment

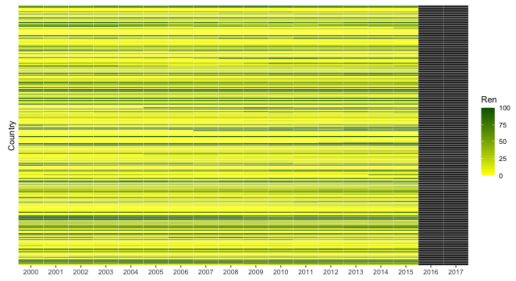
(d) Gross savings



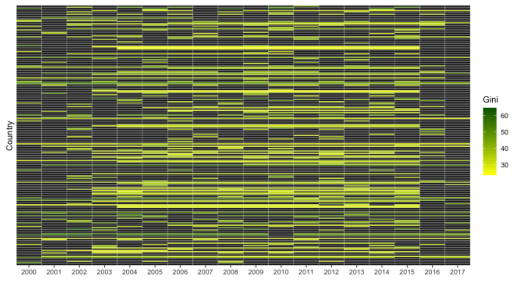
(e) Access to electricity



(f) Individuals using the Internet

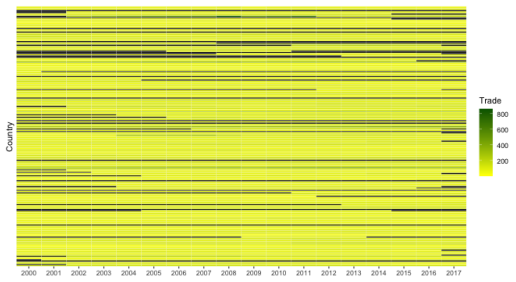


(g) Renewable electricity output

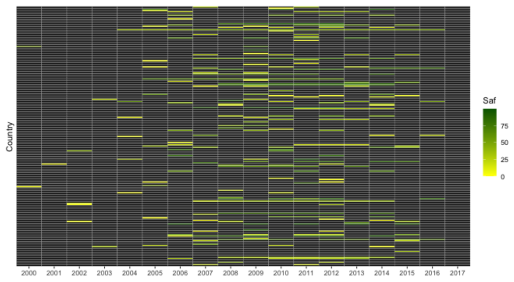


(h) GINI index

Figure 2: *Observed values of the 217 countries across time occasions from 2000 to 2017, missing values depicted in black*

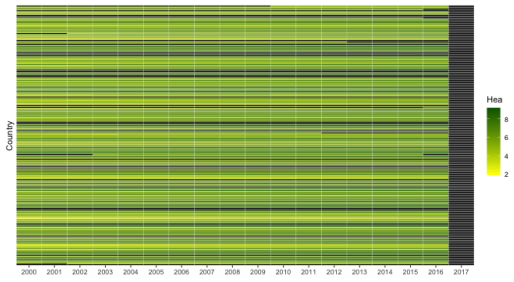


(a) Trade



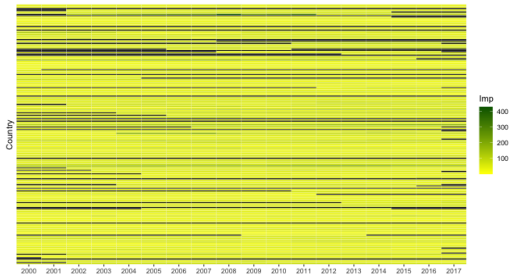
(b) Coverage of social safety net programs in poorest quintile

(c) Literacy rate

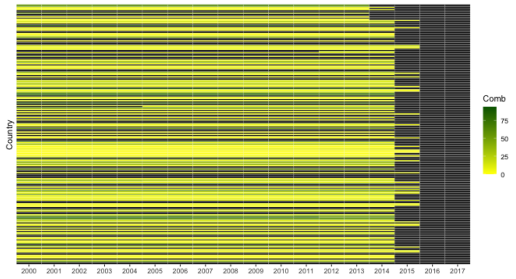


(d) Current health expenditure

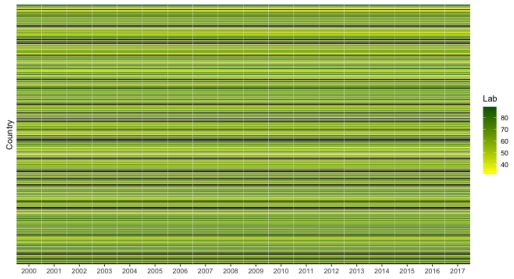
(e) Electricity production



(f) Imports of goods and services

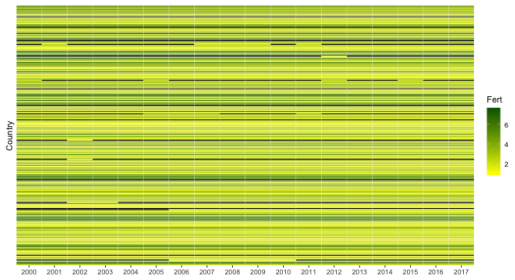


(g) Combustible renewables and waste



(h) Labor force participation rate

Figure 3: *Observed values of the 217 countries across time occasions from 2000 to 2017, missing values depicted in black*



(a) Fertility rate

Figure 4: *Observed values of the 217 countries across time occasions from 2000 to 2017, missing values depicted in black*

Table 8: *Summary statistics of 15 indicators collected in 2000*

Summary	Ele	GDP	Hea	Life	Gsav	Imp	Sch3	Rese
Min.	2.845	572.813	6.596	38.702	-27.661	0.627	0.269	0.045
1st Qu.	55.523	2930.057	103.714	60.382	14.897	27.206	5.589	0.239
Median	97.919	7938.751	307.694	70.315	20.773	38.980	21.861	0.542
3rd Qu.	100.000	20782.434	788.947	74.413	27.221	55.290	41.146	1.174
Max.	100.000	108287.154	4559.888	81.076	50.650	176.014	82.330	3.933
Sd	33.055	19350.563	828.023	9.985	10.777	25.604	20.843	0.869
NA's	41	27	33	16	86	41	103	145

Summary	Trade	Edu	Sch1	Int	Sch2	Saf	Lit
Min.	1.166	1.012	21.872	0.000	6.112	4.616	25.654
1st Qu.	52.424	2.974	96.387	0.380	43.836	15.244	63.117
Median	74.630	4.116	101.636	2.460	77.870	25.872	86.254
3rd Qu.	104.789	5.301	107.273	9.515	93.085	36.501	92.623
Max.	364.365	11.186	134.029	52.000	151.851	47.129	99.767
Sd	49.148	1.934	17.901	13.309	31.695	30.061	19.832
NA's	41	97	55	21	78	215	174

Table 9: *Summary statistics of 15 indicators collected in 2017*

Summary	Ele	GDP	Hea	Life	Gsav	Imp	Sch3	Rese
Min.	9.300	670.777	-	52.214	-48.782	11.571	4.018	0.015
1st Qu.	80.569	4452.428	-	66.894	14.845	31.072	17.954	0.131
Median	100.000	12333.663	-	74.300	22.084	43.270	37.210	0.225
3rd Qu.	100.000	27512.682	-	77.738	28.211	60.162	58.059	0.470
Max	100.000	113262.182	-	84.680	55.640	189.789	113.769	1.530
Sd	24.786	20462.594	-	7.729	11.328	26.901	26.865	0.345
NA's	2	26	217	18	67	39	151	193

Summary	Trade	Edu	Sch1	Int	Sch2	Saf	Lit
Min.	21.507	0.981	49.422	1.309	17.546	-	51.900
1st Qu.	56.658	3.201	98.638	29.198	54.987	-	76.790
Median	78.403	4.349	102.070	58.317	85.838	-	92.143
3rd Qu.	111.520	5.246	107.857	78.914	98.573	-	95.564
Max.	412.869	7.432	139.945	98.871	150.989	-	98.616
Sd	55.118	1.561	12.596	28.381	27.245	-	15.132
NA's	39	162	120	13	135	217	207

Table 10: *Overall median of 15 indicators in 2017 (in italic), and observed values for 8 countries*

Countries	Ele	GDP	Hea	Life	Gsav	Imp	Sch3	Rese
Overall median	*100*	*12333.663*		*74.300*	*22.084*	*43.270*	*37.210*	*0.224*
Afghanistan	97.701	1758.466	-	64.047	20.791	45.332	-	-
Bhutan	97.700	9246.677	-	70.565	29.396	49.553	-	-
Cambodia	89.070	3653.641	-	69.331	23.193	64.106	13.138	-
Ethiopia	44.301	1724.483	-	65.874	30.820	23.477	-	-
Nepal	95.507	2605.510	-	70.604	45.624	42.887	11.793	-
Papua New Guinea	54.427	3880.650	-	65.705	-	-	-	-
Solomon Islands	62.895	2126.353	-	71.006	-	-	-	-
Timor-Leste	80.381	6740.890	-	69.199	12.300	59.916	-	-

Countries	Trade	Edu	Sch1	Int	Sch2	Saf	Lit
Overall median	*78.403*	*4.349*	*102.069*	*58.317*	*85.838*	-	*92.142*
Afghanistan	51.237	3.927	103.924	11.448	54.813	-	-
Bhutan	78.645	7.050	92.575	48.106	86.096	-	-
Cambodia	124.788	-	107.835	34.000	-	-	-
Ethiopia	31.107	-	-	18.618	-	-	-
Nepal	51.983	5.096	134.121	21.403	71.209	-	-
Papua New Guinea	-	-	-	11.209	-	-	-
Solomon Islands	-	-	114.351	11.924	-	-	-
Timor-Leste	121.005	-	100.583	27.493	79.303	-	-

Table 11: *Standard errors obtained with the non-parametric bootstrap for the estimated transition probabilities from 2000 to 2001 under the HM model with $k = 6$ hidden states*

	1	2	3	4	5	6
$se(\widehat{\pi}_{u\|1})$	0.045	0.005	-	-	-	-
$se(\widehat{\pi}_{u\|2})$	0.003	0.033	0.001	-	-	-
$se(\widehat{\pi}_{u\|3})$	-	-	0.047	0.043	-	0.002
$se(\widehat{\pi}_{u\|4})$	-	-	-	0.008	-	-
$se(\widehat{\pi}_{u\|5})$	-	-	0.003	0.004	0.035	-
$se(\widehat{\pi}_{u\|6})$	-	-	-	-	-	0.001

Table 12: *Standard errors obtained with the non-parametric bootstrap for the estimated transition probabilities from 2005 to 2006 under the HM model with $k = 6$ hidden states*

	1	2	3	4	5	6
$se(\widehat{\pi}_{u\|1})$	0.032	-	-	-	-	-
$se(\widehat{\pi}_{u\|2})$	-	0.005	-	-	-	-
$se(\widehat{\pi}_{u\|3})$	-	-	0.053	0.005	0.005	-
$se(\widehat{\pi}_{u\|4})$	-		-	0.060	0.005	-
$se(\widehat{\pi}_{u\|5})$	-	-	-	-	0.035	-
$se(\widehat{\pi}_{u\|6})$	-	-	-	-	-	0.001

Table 13: *Standard errors obtained with the non-parametric bootstrap for the estimated transition probabilities from 2010 to 2011 under the HM model with $k = 6$ hidden states*

	1	2	3	4	5	6
$se(\widehat{\pi}_{u\|1})$	0.072	-	-	-	-	-
$se(\widehat{\pi}_{u\|2})$	0.037	0.062	0.053	-	-	-
$se(\widehat{\pi}_{u\|3})$	-	-	0.055	0.004	-	0.004
$se(\widehat{\pi}_{u\|4})$	-		-	0.082	-	0.008
$se(\widehat{\pi}_{u\|5})$	-	-	-	-	0.012	-
$se(\widehat{\pi}_{u\|6})$	-	-	-	-	-	0.001

Table 14: *Standard errors obtained with the non-parametric bootstrap for the estimated transition probabilities from 2016 to 2017 under the HM model with $k = 6$ hidden states*

	1	2	3	4	5	6
$se(\widehat{\pi}_{u\|1})$	0.001	-	-	-	-	-
$se(\widehat{\pi}_{u\|2})$	-	0.035		-	-	-
$se(\widehat{\pi}_{u\|3})$	-	-	0.053		-	-
$se(\widehat{\pi}_{u\|4})$	-		-	0.104	-	0.161
$se(\widehat{\pi}_{u\|5})$	-	-	-	0.025	0.025	-
$se(\widehat{\pi}_{u\|6})$	-	-	-	-	-	0.001

References

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–243.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, MA.