Missing data imputation in large combined cross-sectional and longitudinal data: multilevel multiple imputation and time series imputation.

David Wutchiett
Université de Montréal
david.wutchiett@umontreal.ca
12th International Multilevel Conference
Utrecht, the Netherlands
April 9th, 2019

Topic of research

- The use of multilevel modeling and time series modeling within frameworks of missing value imputation and multiple imputation
- To address questions of interest regarding description and relationships between variables in international comparative research
- With large longitudinal and cross-sectional data.

Points of interest in this research

- Use of multilevel models in reference to an interesting combination of data types, structures and context
  - Large cross-sectional public opinion surveys
    - 536,000 individuals responding to surveys in 27 Eastern European, West Asian and Central Asian countries over 24 years (1993 to 2016).
  - Longitudinal cross-sectional national socioeconomic and political measures
- Use of several methods, models and software of interest
  - Multilevel multiple imputation (R: mice, pan)
  - Multilevel modeling (R: lme4)
  - Time series imputation (R: imputeTS)
- Investigation of important substantive question:
  - How are national and individual level characteristics related to trust in congress or parliament

Missing Data

- Missing data are common in surveys
  - Respondents refusal, don't know or are unsure
  - surveys ask different questions
- Comparative and international studies often use time-series cross-section data
- Missing data are common in longitudinal national socioeconomic characteristic measures
  - Coverage across countries varies by data source/study
  - Developing countries
  - Non-report
- The data have longitudinal and hierarchical properties
  - Relevance of country
  - Relevance of year and longitudinal trends

Several approaches to missing data & imputation

- Complete case analysis
  - Biased estimates if missing values are not missing completely at random (MCAR)
- Multiple imputation
  - Uses Gibbs sampler, iteratively estimating imputation model parameters then simulating missing values from posterior predictive distributions based on variables' conditional distributions' characteristics.
  - Flexible choice of multivariate imputation models for estimation of missing values
  - Statistical analysis of interest is carried out with each imputed data set with estimates pooled to produce a single set of estimate values
  - 'mice' package (Van Buuren and Groothuis-Oudshoorn, 2011)
- Missing at Random (MAR)
- Assumption of Ignorability
  - Missing values and probability of missingness are related to demographics and country-level socioeconomic and political characteristics at the individual and country levels

Challenges in application

- Longitudinal structure of observations:
  - A model-based imputation approach that does not sufficiently take into account longitudinal characteristics tends to produce longitudinal point estimates that differ significantly from adjacent longitudinal observations while respecting the variables' conditional distributional characteristics.
- Large scale data and many relevant variables to include in imputation models
  - Imputation processes can be prohibitively slow to the point where it is not possible to carry out the process with fully specified models.

Imputation modeling longitudinal component

- Multilevel multiple imputation
  - Multilevel data and modeling processes incorporating random intercepts and slopes for time variables (Year and Year^2) by country (cluster variable)

$$y_c = X_c\beta + Z_c u_c + \epsilon_c$$

- Time series imputation
  - Applied to single clusters observed across longitudinal points
  - Single point estimates for missing values
  - Kalman Smoothing to fit basic structural model: seasonal ARIMA(0,2,2)
  - 'imputeTS' package (Moritz and Bartz-Beielstein. 2017)

Specification of the imputation models

- Regression or multilevel models:
- For variables concerning the individual (support for political institutions, demographic characteristics),
  - Country-level socioeconomic and political characteristics and additional individual responses concerning demographics and trust may be relevant.
- For national socioeconomic characteristics
  - Longitudinal trends within the same variable within the same country and others may inform likely observed values.
  - Additional characteristics at the country level as well as survey respondents' demographic characteristics and trust in political institutions may be related to national socioeconomic and political characteristics.

The Data

- Time series cross-section data
  - 27 countries in Eastern Europe, West Asia and Central Asia
  - Years 1993 to 2016
- 536,000 respondents to survey questions regarding trust in institutions
  - 12 large-scale surveying programs
    - European Social Survey
    - World Values Survey
    - Consolidation of Democracy in Eastern Europe
  - 410 surveys
  - Trust in democratic institutions
  - Basic demographic characteristics
- National socioeconomic and political descriptors
  - World Bank
  - OECD
  - Varieties of Democracy (V-Dem)
  - Cline Center's Composition of Religious and Ethnic Groups (CREG) project
  - Gini, Poverty, Ethnic diversity, Voter participation

Descriptive Statistics

Table 1: Survey respondent descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Trust in Parliament or Congress	527,593	3.18	1.73	1.00	1.60	4.33	7.00
Trust in Political Parties	429,593	2.80	1.60	1.00	1.00	4.00	7.00
Year	535,605	2,006.06	6.53	1,993	2,001	2,011	2,016
Female	535,389	1.55	0.50	1.00	1.00	2.00	2.00
Age	535,605	46.44	17.46	15	32	60	97
Education	496,872	3.33	1.08	1.00	2.00	4.00	5.00

Table 2: Country-level characteristics descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Year	621	2,004.00	6.64	1,993	1,998	2,010	2,015
Population	621	17,357,400.00	29,460,482.00	1,314,545	3,535,961	15,012,985	148,520,000
Ethnic diversity	567	0.74	0.32	0.10	0.52	1.02	1.39
Education (Avg)	529	8.67	1.15	4.70	8.03	9.46	11.01
Gini coef	595	33.32	5.68	19.45	28.94	36.60	50.45
GDP per person	484	6,772.72	3,918.35	824.58	3,890.19	8,842.33	22,108.65
Poverty gap (<5.50)	320	10.99	14.02	0.00	1.00	16.85	65.10
Poverty rate (<5.50)	320	27.59	28.55	0.00	3.88	51.52	98.10
Political Participation	584	46.62	9.46	0.00	39.80	53.20	70.00
Academic and Cultural Expression	621	1.43	1.23	−1.87	0.61	2.42	3.26

Imputation processes

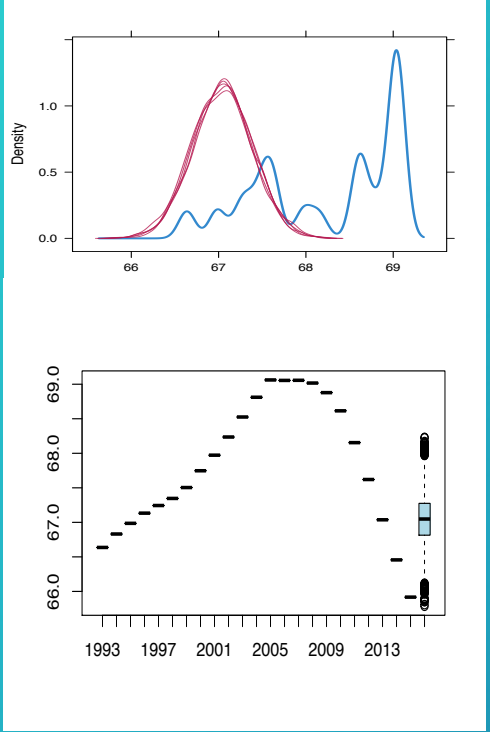
- 1: Multilevel multiple imputation (mice.impute.2l.norm, mice.impute.pmm)
  - Individual-level characteristics (predictive mean matching, all covariates)
  - National-level characteristics (multilevel model with time variable random intercepts and slopes)
- 2: Multilevel multiple imputation with homogenous within group variances (mice.impute.2l.pan)
  - Individual-level characteristics (2l.pan, all covariates (individual & national levels), random intercepts and slopes for countries)
  - National-level characteristics(2l.pan, yearly national characteristics and respondent-level characteristic national means, time variable random intercepts and slopes for countries)
- 3: Time series imputation for national characteristics then multilevel multiple imputation for individuals and remaining national characteristics (na.kalman, mice.impute.2l.pan)
  - Time series imputation for national characteristics (for countries with 3 or more non-missing values)
    - Sequences missing sufficient observations were for 2 countries (Serbia and Poland) for the two poverty variables
  - Multilevel multiple imputation (with homogenous group variances) for individual-level characteristics and national characteristic values for national variable time series with less than 3 observations by cluster
- The approaches above were sequentially less computationally expensive

Comparisons of results

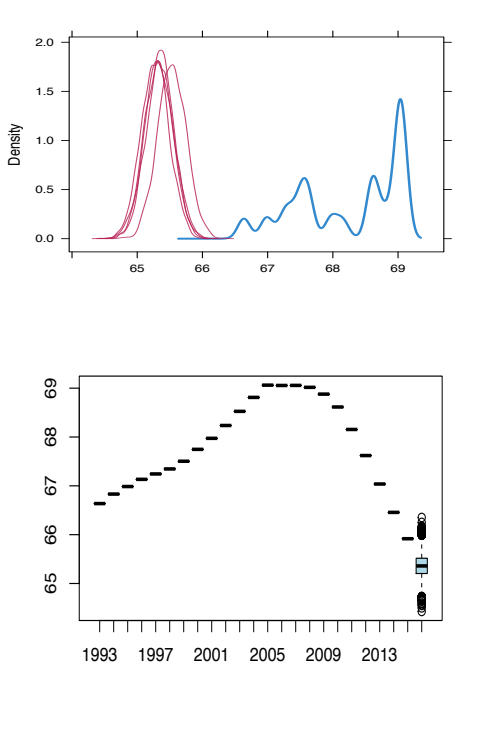
- Descriptive characteristics:
  - Countries with median and greatest proportion of missing values
    - Bulgaria: 2.28 missing values per observation
    - Tajikistan: 5.3 missing values per observation
  - Variables with median, most and least (but greater than 1%) proportion of missing values
    - Population between 15 and 64: 0.105
    - International Poverty Gap: 0.413
    - Trust in Congress/Parliament: 0.015
- Statistical modeling with pooled data:
  - Relationship between trust in congress/parliament and national and individual characteristics
  - Multilevel model with random intercepts by country
  - Four different data to compare:
  - Original: 185,000 complete cases
  - Imputed data: 536,000 complete cases
    - Multilevel 1: time variable random slopes for national characteristics, predictive mean matching with (country intercepts for) individual characteristics
    - Multilevel 2: time variable random slopes, all variables (national, individual, individual national averages)
    - Time series + Multilevel: time series imputation for national characteristics, multilevel for all remaining (all variables)

Population ages 15-64, Bulgaria

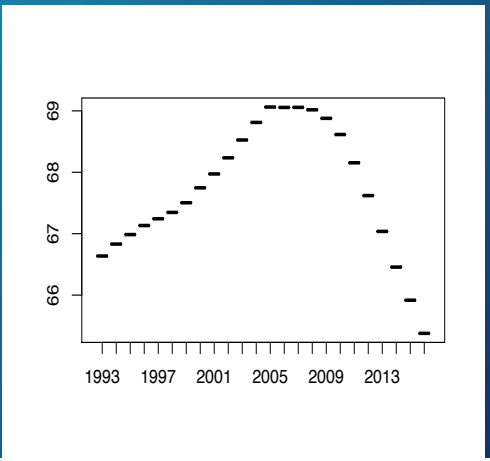
Multilevel 1



Multilevel 2

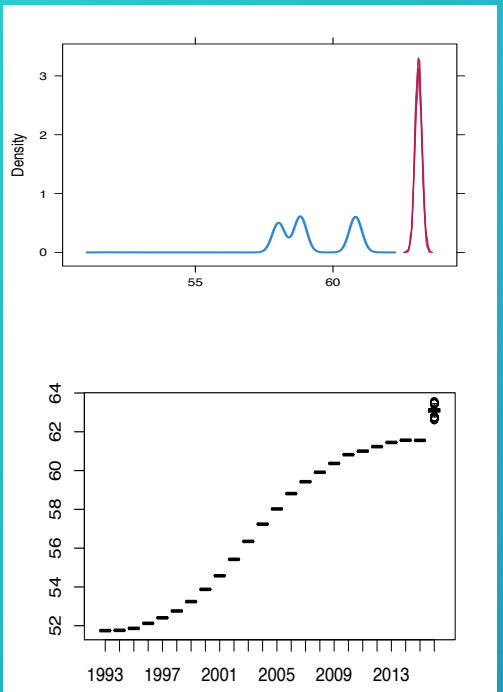


Time series + Multilevel

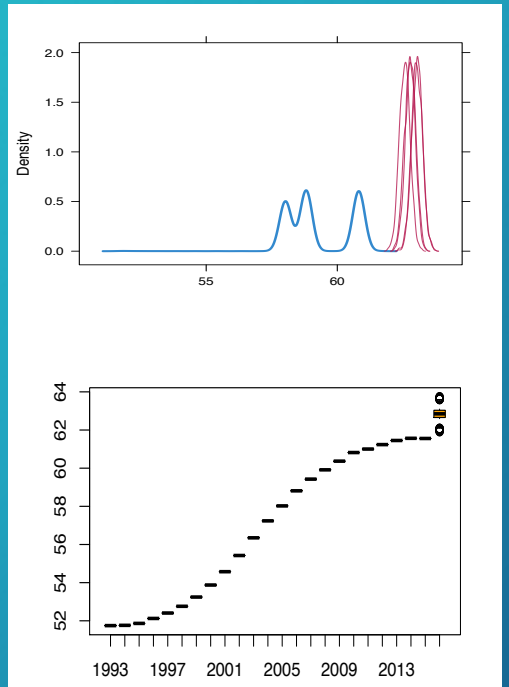


Population ages 15-64, Tajikistan

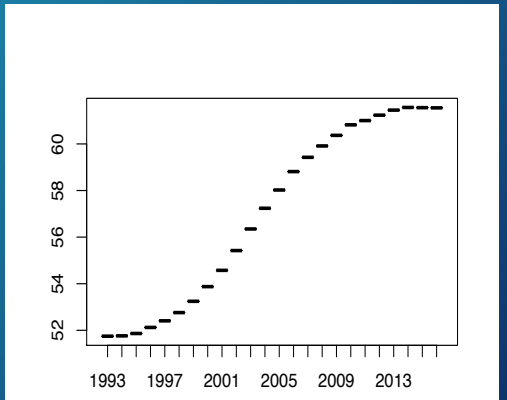
Multilevel 1



Multilevel 2

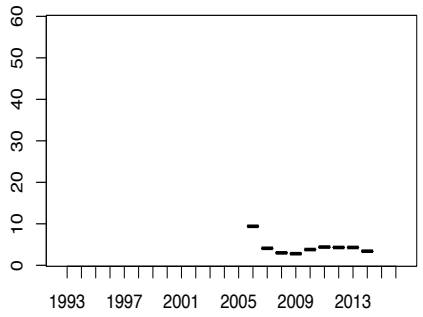


Time series + Multilevel

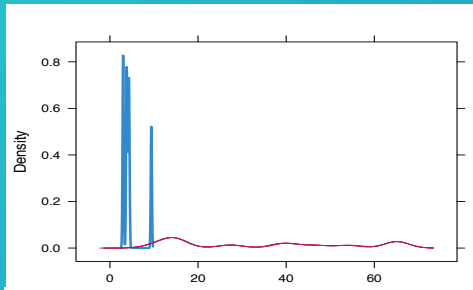


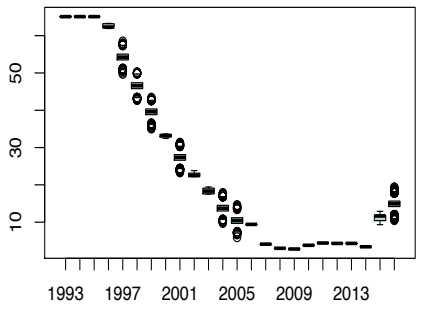
Poverty Gap, Bulgaria

Original

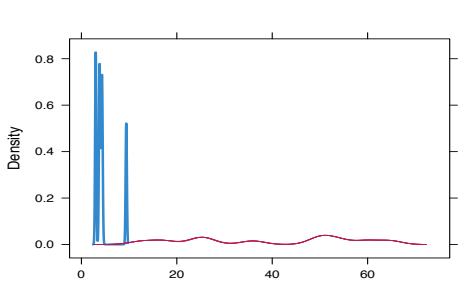


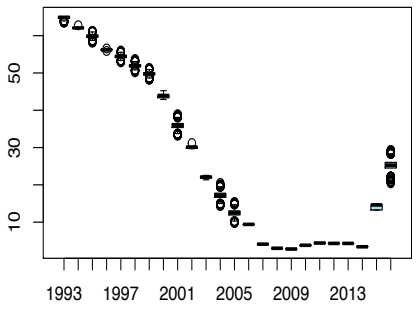
Multilevel 1



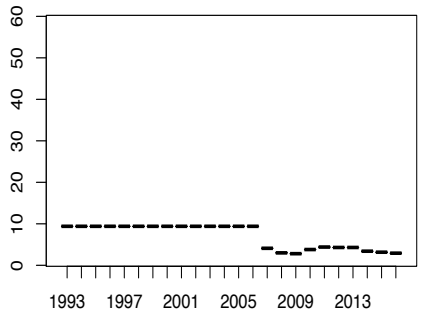


Multilevel 2



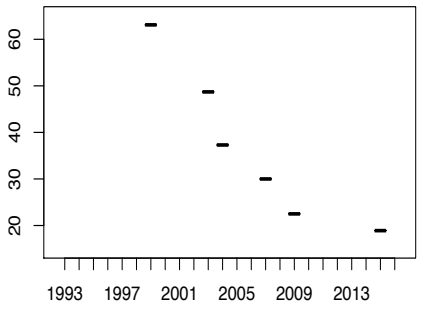


Time series + Multilevel

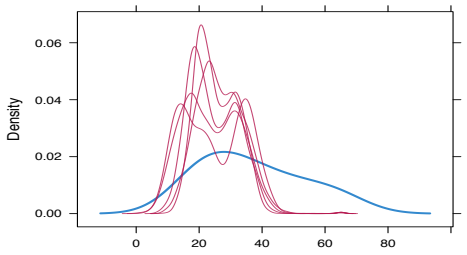


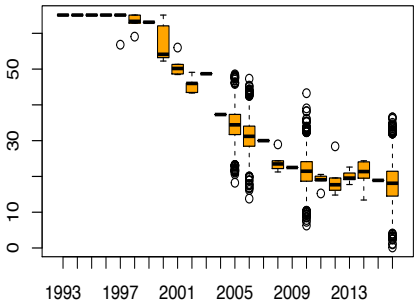
Poverty Gap, Tajikistan

Original

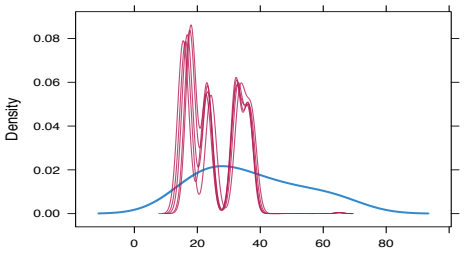


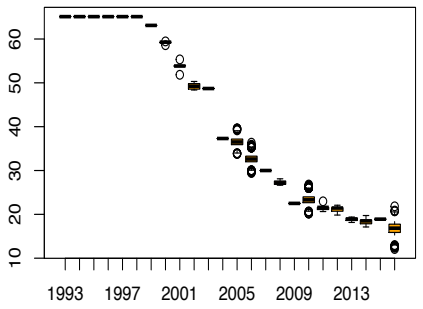
Multilevel 1



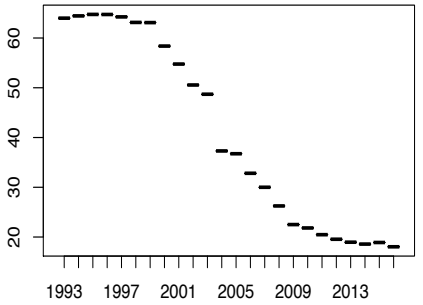


Multilevel 2



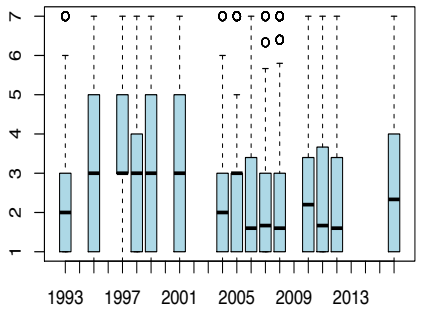


Time series + Multilevel

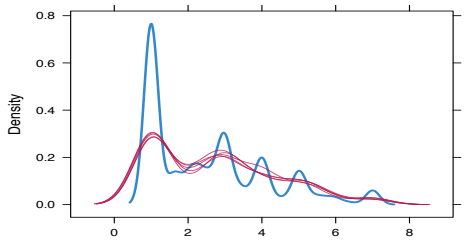


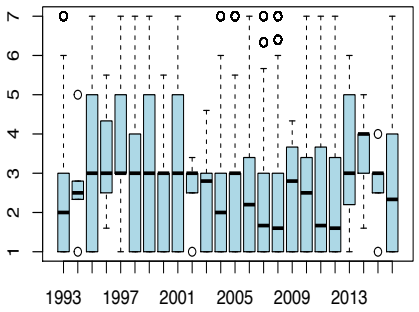
Trust in Parliament/Congress, Bulgaria

Original

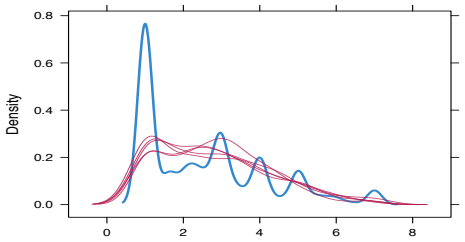


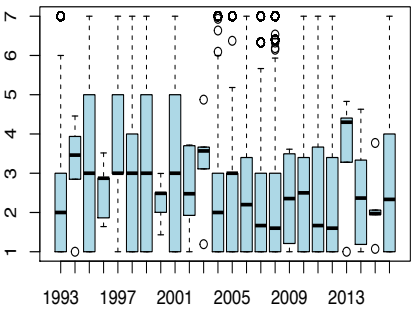
Multilevel 1



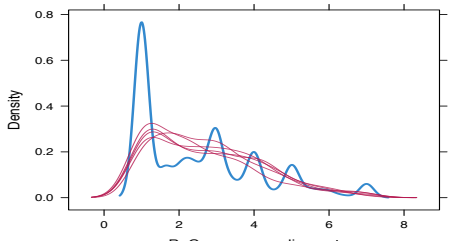


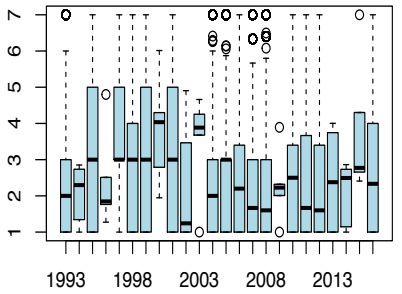
Multilevel 2





Time series + Multilevel

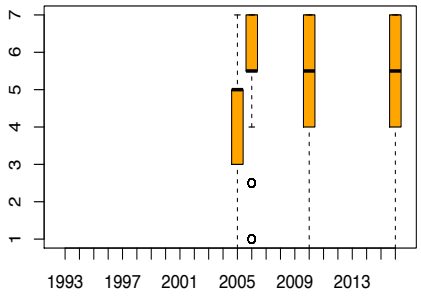




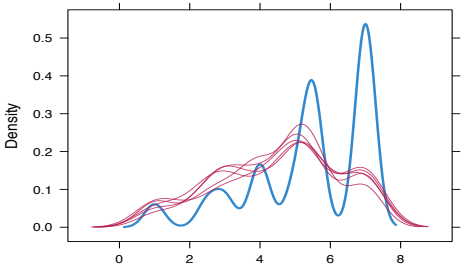
Trust in Parliament/Congress, Tajikistan

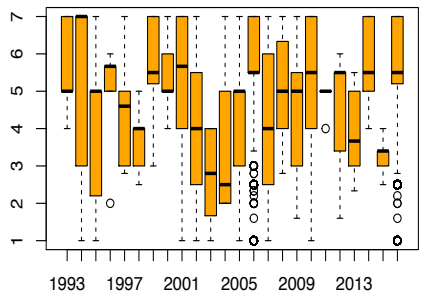
Original

Note*: Tajikistan and Uzbekistan were the only two countries that observed mean yearly trust in parliament greater than 5.

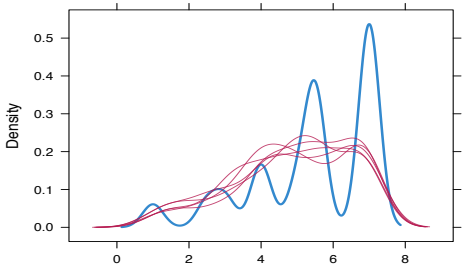


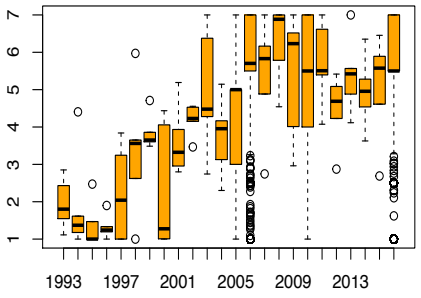
Multilevel 1 (pmm)



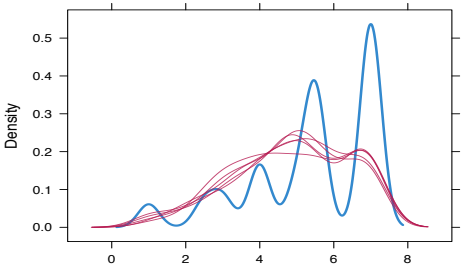


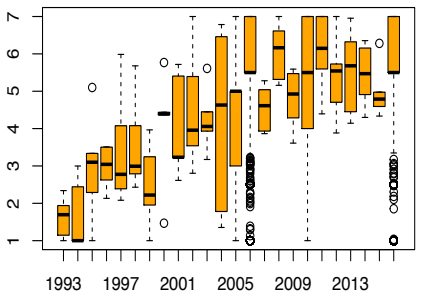
Multilevel 2





Time series + Multilevel





Pooled multilevel modeling: Trust in Congress/Parliament

	Complete Cases		Multilevel Multiple Imputation 1		Multilevel Multiple Imputation 2		Time Series + Multilevel Multiple Imputation	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
(Intercept)	7.105	6.594	-42.063	1.456	-33.619	1.264	-33.238	2.822
Year	-0.008	0.004 **	-0.038	0.002 ***	-0.030	0.003 ***	-0.017	0.003 ***
Year2	0.003	0.000 ***	0.001	0.000 ***	0.002	0.000 ***	0.001	0.000 ***
Individual-level								
Sex	0.037	0.007 ***	0.032	0.005 ***	0.032	0.004 ***	0.031	0.005 ***
Age	0.003	0.000 ***	0.003	0.000 ***	0.003	0.000 ***	0.003	0.000 ***
Education	0.004	0.003	0.007	0.002 **	0.007	0.002 **	0.005	0.002 *
Country-level								
% Ages 15-54	0.096	0.011 ***	0.075	0.003 ***	0.072	0.009 ***	0.051	0.007 ***
% Ages 65+	-0.114	0.013 ***	-0.042	0.005 ***	-0.041	0.014 **	-0.058	0.009 ***
Population (ln)	0.270	0.127 **	2.229	0.089 ***	1.818	0.073 ***	1.912	0.179 ***
Ethnic Diversity	0.782	0.125 ***	-0.199	0.077 **	-0.224	0.065 ***	-0.057	0.160
Education (average)	0.158	0.026 ***	0.017	0.009	0.032	0.021	-0.047	0.005 ***
Gini	0.018	0.002 ***	0.011	0.001 ***	0.013	0.002 ***	0.015	0.001 ***
GDP per person (ln)	-0.016	0.048	0.592	0.032 ***	0.368	0.038 ***	0.364	0.068 ***
Poverty Gap	0.000	0.003	0.001	0.004	-0.002	0.002	0.005	0.011
Poverty Rate	0.003	0.001 **	0.001	0.002	0.001	0.002	-0.002	0.005
Voter Participation	0.016	0.001 ***	0.006	0.000 ***	0.006	0.001 ***	0.010	0.000 ***
Freedom of Academic & Cultural Expression	-0.378	0.018 ***	-0.172	0.010 ***	-0.170	0.010 ***	-0.172	0.014 ***

Pooled multilevel modeling: Trust in Congress/Parliament

	Complete Cases		Multilevel Multiple Imputation 1		Multilevel Multiple Imputation 2		Time Series + Multilevel Multiple Imputation	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
(Intercept)	7.105	6.594	-42.063	1.456	-33.619	1.264	-33.238	2.822
Year	-0.008	0.004 **	-0.038	0.002 ***	-0.030	0.003 ***	-0.017	0.003 ***
Year2	0.003	0.000 ***	0.001	0.000 ***	0.002	0.000 ***	0.001	0.000 ***
Individual-level								
Sex	0.037	0.007 ***	0.032	0.005 ***	0.032	0.004 ***	0.031	0.005 ***
Age	0.003	0.000 ***	0.003	0.000 ***	0.003	0.000 ***	0.003	0.000 ***
Education	0.004	0.003	0.007	0.002 **	0.007	0.002 **	0.005	0.002 *
Country-level								
% Ages 15-54	0.096	0.011 ***	0.075	0.003 ***	0.072	0.009 ***	0.051	0.007 ***
% Ages 65+	-0.114	0.013 ***	-0.042	0.005 ***	-0.041	0.014 **	-0.058	0.009 ***
Population (ln)	0.270	0.127 **	2.229	0.089 ***	1.818	0.073 ***	1.912	0.179 ***
Ethnic Diversity	0.782	0.125 ***	-0.199	0.077 **	-0.224	0.065 ***	-0.057	0.160
Education (average)	0.158	0.026 ***	0.017	0.009	0.032	0.021	-0.047	0.005 ***
Gini	0.018	0.002 ***	0.011	0.001 ***	0.013	0.002 ***	0.015	0.001 ***
GDP per person	-0.016	0.048	0.592	0.032 ***	0.368	0.038 ***	0.364	0.068 ***
Poverty Gap	0.000	0.003	0.001	0.004	-0.002	0.002	0.005	0.011
Poverty Rate	0.003	0.001 **	0.001	0.002	0.001	0.002	-0.002	0.005
Voter Participation	0.016	0.001 ***	0.006	0.000 ***	0.006	0.001 ***	0.010	0.000 ***
Freedom of Academic & Cultural Expression	-0.378	0.018 ***	-0.172	0.010 ***	-0.170	0.010 ***	-0.172	0.014 ***

Conclusions

- Multilevel multiple imputation is a promising and effective approach to imputing missing values across individual and country-level data.
- Time variables (within the framework of multilevel modeling) and time series approaches improve estimates
- Multivariate approaches improve estimates, particularly where longitudinal data is lacking
- Computational time continues to be a challenge
- The ‘2l.pan’ multilevel multiple imputation approach in ‘mice’ with more predictor variables and time-variable random slopes performed well

Thank you!

- Questions?

- Email: david.wutchiett@umontreal.ca

References

- Harvey, Andrew C. 1990. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge university press.
- Lindberg, Staffan I., Michael Coppedge, John Gerring, and Jan Teorell. 2014. "V-Dem: A New Way to Measure Democracy." *Journal of Democracy* 25(3):159–69.
- Moritz, Steffen and Thomas Bartz-Beielstein. 2017. "ImputeTS: Time Series Missing Value Imputation in R." *The R Journal* 9(1):207–18.
- Nardulli, Peter F., Cara J. Wong, Ajay Singh, Buddy Peyton, and Joseph Bajjaliegh. 2012. "The Composition of Religious and Ethnic Groups (CREG) Project." *Cline Center for Democracy, University of Illinois, Urbana-Champaign*.
- Rubin, D.B., Multiple Imputation for Nonresponse in Surveys. 1987, New York: John Wiley.
- Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software". 45(3): p. 67.
- The World Bank, World Development Indicators. 2018.