Statistical issues in analyzing harmonized data

By Claire Durand,
Professor, department of sociology,
University of Montreal,

Presented at the
Workshop “Building Multi-Source Databases for Comparative Analyses” of the Survey Data Recycling Project,
Warsaw, Poland,
December 16-20, 2019

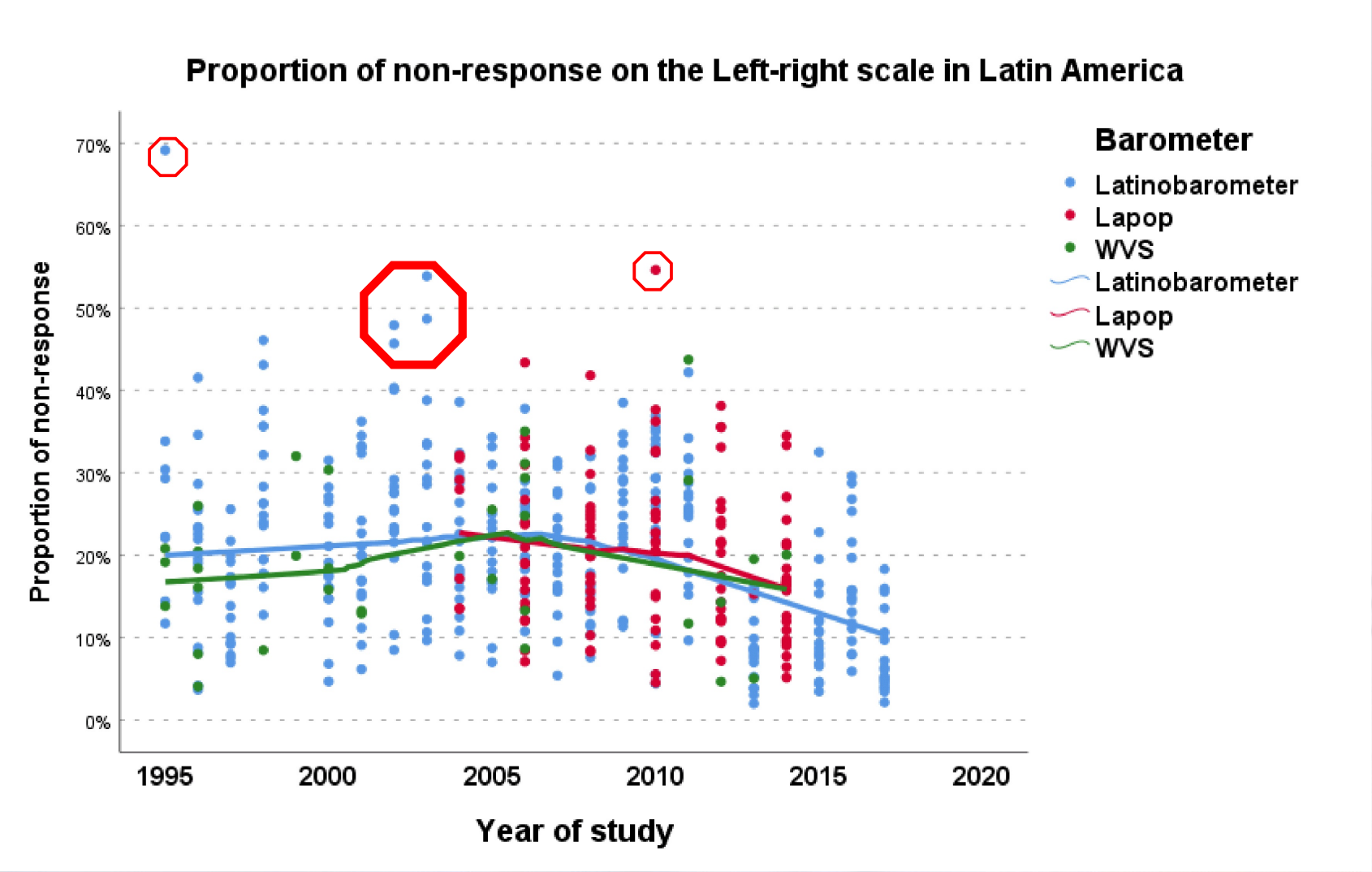
Statistical issues: Four main challenges

- "Missing data" may be found at the survey level as well as at the individual respondent level.
- Cross-sectional data is available at many time points. How should we take it into account?
- Disentangling variation due to country-level effects and to the presence and methodological features of survey-projects.
- Should we weight, why, at what level?

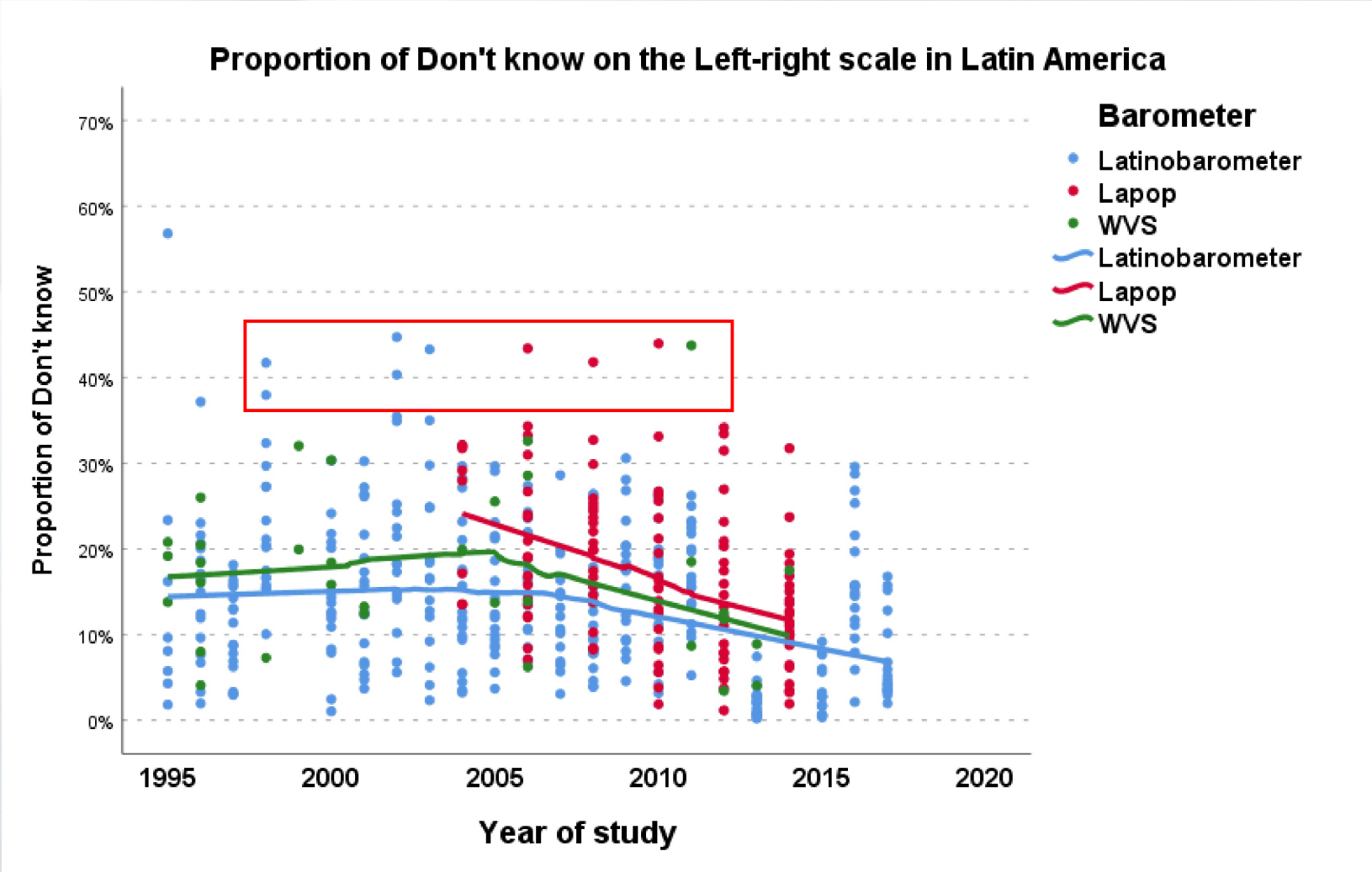
Issue no 1. The missing data/ missing question issue

- At the individual level:
  - No answer
  - Don’t knows
- At the survey level:
  - Questions are not asked to the respondents.
    - Not relevant or not possible in some countries
      - No president in parliamentary regimes
      - No “church” in muslim countries
      - Regional institutions and the likes.
      - Topic too sensitive in some country or year

The left-right scale : Proportion of item non-response in Latin America



The left-right scale: Proportion of Don’t know in Latin America



Two types of missing values

At the individual level

- The mean proportion of missing values (including Don't know and no answer) is around 20% until 2010, consistent across projects.
  - It is decreasing (towards 10% for LB) in recent years
- The mean proportion varies by country and year, from close to 0% to around 50%, and even 70% (Paraguay in 1995) in one case.
- The proportion of Don't knows constitutes the major part of missing values at the individual level.

Two types of missing values

At the individual level

- What can we do about it?
  - The high proportion of missing values indicate that answering the question may be problematic for some respondents.
    - Either it does not mean anything to them or...
    - The question is considered too sensitive
  - Can we – should we replace missing values when the proportion is so high?
  - **Is there a relationship between the political situation and non response.**
    - Non-response starts to decline with the “giro a la izquierda” (turn to the left) in Latin America.
    - Which means that it may be related to respondents’ ideological stand.

Two types of missing values

At the individual level

- What can we do about it?
- Three solutions
  - 1. Drop cases with missing values → bias...
  - 2. Replace with mean of the scale by country and year.
    - Heuristic solution since it does not influence the analysis.
  - 3. Missing value imputation using the information from relationships over time and within countries (Wutchiett and Durand, submitted)
    - It is an advantage of combined data.
    - It may help taking into account the possible ideological bias in non-response.
    - At the same time, it "boosts" existing relationships.

The missing data/ missing question issue

At the survey level

- The question is not asked in some countries and years.
  - Because it is not really relevant (West Indies)
  - Because it may be too sensitive (Paraguay, Colombia, Bolivia) at certain periods.
- Can we impute at the survey level?
  - Using the information that we have from the same countries
    - In the same year but different survey projects.
    - In the same country but for different years.
    - In the same year for similar countries.
  - What about file grafting (Aluja-Banet & Thiô, 2001)?

Missing questions for scales

At the survey level

- Trust in institutions:
  - In the 17 survey projects from 1995 to 2017 that my team combined
    - 133 different institutions surveyed
    - Average of 12.5 institutions surveyed in each survey, from 3 to 23.
  - The “solution” usually applied:
    - Use “the question” which is present in the largest proportion of surveys.
    - Compute a scale with a restricted number of questions that are present in the highest proportion of surveys.
  - The consequence: analyses pertain almost exclusively to political trust, and use trust in parliament or a scale of 3-5 items.

Missing questions, what can we do?

- We can conceptualize items as samples of all the questions that could have been asked to cover a given topic.
- We can use either multivariate or repeated measures multilevel analysis.
  - Answers to questions are nested within respondents.
  - We group institutions that are close conceptually and assess whether empirical criteria hold (similar mean and std for grouped institutions for similar countries and projects).
- We need to control for the differences between projects.

Trust in institutions, repeated measures

	Model 0		Model 1	
Intercept	3.956	***	4.221	***
Level Institutions				
Media (REF)				
State/President			-0.048	***
Governments			-0.388	***
Parliament			-0.655	***
Elections			-0.323	***
Political Parties			-1.119	***
International Org.			-0.173	***
Army			0.314	***
Police			-0.214	***
Public Admin.			-0.202	***
Judiciary			-0.380	***
Church			0.619	***
Trade Unions			-0.619	***
ONG- Civil Society			-0.069	***
Financial Organizations			-0.173	***
Enterprises			-0.354	***

- Different means for usual items composing scales of political trust.
- Compared with trust in media, on a 7-point scale:
  - Political parties: -1.12
  - Parliament: -.66
  - Public administ.: -.20
  - **Army: +.31**

Trust in institutions, repeated measures

	Model 0		Model 1	
Intercept	3.956	***	4.221	***
Level Institutions				
Media (REF)				
State/President			-0.048	***
Governments			-0.388	***
Parliament			-0.655	***
Elections			-0.323	***
Political Parties			-1.119	***
International Org.			-0.173	***
Army			0.314	***
Police			-0.214	***
Public Admin.			-0.202	***
Judiciary			-0.380	***
Church			0.619	***
Trade Unions			-0.619	***
ONG- Civil Society			-0.069	***
Financial Organizations			-0.173	***
Enterprises			-0.354	***

Variance	Model 0		Model 1	
Measures	2.427	63.0%	2.257	61.3%
Respondents	1.051	27.3%	1.064	28.9%
Country-Year	0.088	2.3%	0.090	2.4%
Country-Source	0.284	7.4%	0.273	7.4%
Total	***3.849***		***3.682***	

- Within respondent variance in answers to trust questions: 63%
- Difference between institutions explains
  - 7% of the variance at the measurement level.
  - 4% of the variance at the country-source level.

Missing values: summary

- At the individual level:
  - We can impute based on all the information present in combined longitudinal data sets.
  - Should we differentiate bw DK and NA?
  - Beware when missing values may be related to the political situation.
- At the survey level:
  - Use a within- respondent level in multi-level analysis in order to keep and analyse all the information available.
  - Use IRT models (Van der Meer, 2019) to achieve equivalent scales between surveys.
  - Data fusion - survey grafting (Aluja-Banet and Thiô, 2001).

Issue no. 2: Time

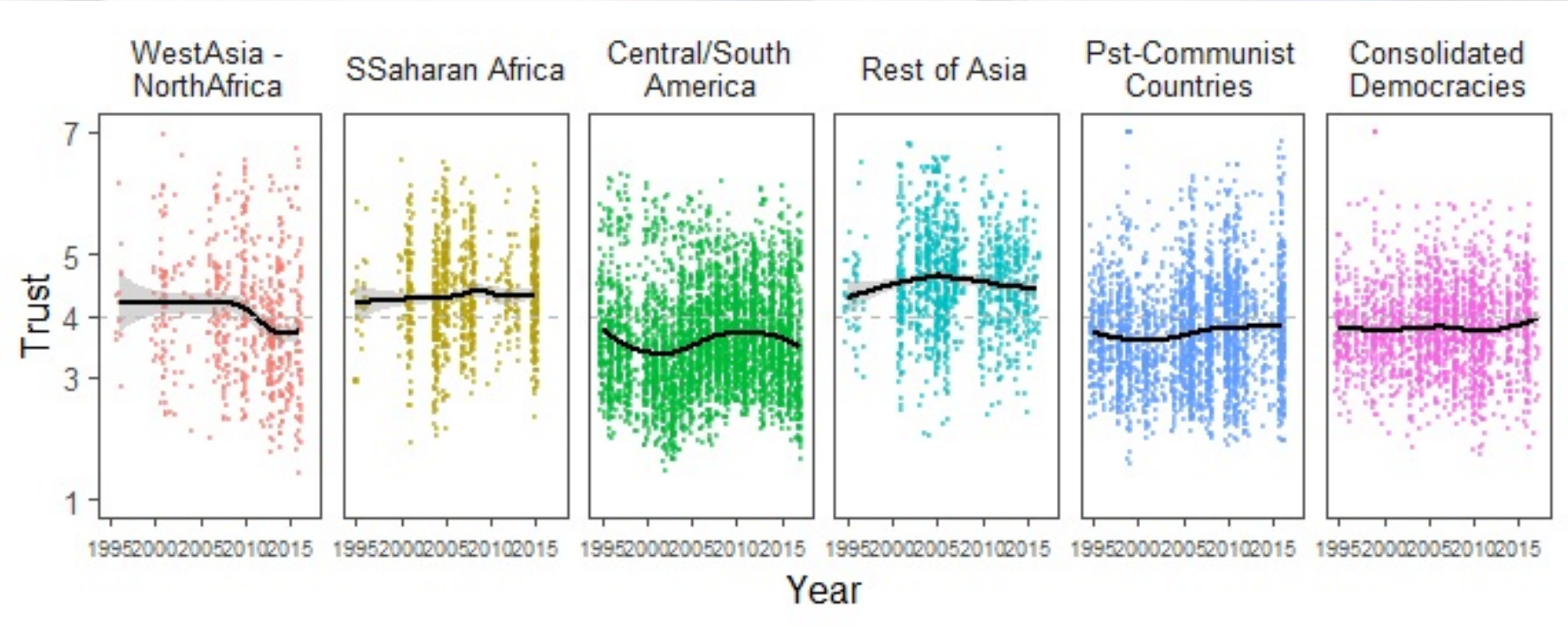
- One of the goals when we combine survey data is to analyse change over time.
  - It is particularly relevant when we have data over a long period.
- We need to be able:
  - To assess the form that trends may take – linear, quadratic, cubic...
  - To assess the possible impact of some events
    - In the country where they occur
    - In the region that may or may not be affected.
  - To assess whether trends are similar
    - In different countries or regions
    - For different measures.

How to analyze change over time

- V**isualize** the trends.
- Introduce time in our analyses
  - As a variable with different components -- linear, quadratic, cubic.
  - As variables indicating events.
- Assess possible interaction effects, that is,
  - Trends may differ between measures and between country groupings.

Visualize trends

- There is a tendency to present line graphs of average values.
- Local regression allows for visualizing variation between surveys & measures as well as mean change.



Visualizing leads subsequent analyses

- Each point represents mean trust in one institution - survey conducted in one country and year by one international survey project. We notice much variability around the mean.
- Trust is stable in consolidated democracies, in Post-communist countries and in Sub-Saharan Africa.
- There is a drop in overall institutional trust in North Africa and West Asia (WANA) since 2011.
- Trust in Latin America follows a fish-like (cubic) trend. Recent increases may be attributed to the turn to the left (Pena Ibarra and Durand, submitted).

The visualized longitudinal trends do not take into account...

- The fact that some sources (survey projects) are present only in some countries and years.
- Methodological differences between sources.

Introducing time in analysis

Trust in institutions - basic models										
	Model 0		Model 3		Model 4a		Model 4b		Model 4c	
Intercept	3.956	***	4.180	***	4.459402	***	4.16134	***	4.175306	***
Level Institutions										
Level Country-Year-Source										
Time			0.0104	**	0.0122	**	0.0109	**	0.0012	ns
Time^2			0.0001	ns	0.0001	ns	0.0002	ns	0.0001	ns
Time^3			-0.0001	**	-0.0002	***	0.0001	***	0.0000	ns
Level Country-Source										
LAPOP					0.185		0.383	**	0.490	***
WVS-EVS					-0.277	***	-0.072		-0.102	
medium scale (5-7)					-0.399	***	-0.058		-0.147	
Long scale (10-11 pts)					-0.665	***	-0.332	***	-0.388	***
Consolidated Dem. (REF)										
West Asia N. Africa							0.038		0.302	
Time									*-0.022*	
Time^2									*-0.004*	**
Sub-Saharan Africa							0.472	***	0.451	***
Central/South America							-0.313	***	-0.364	**
Time									*0.025*	**
Time^2									*-0.001*	
Time^3									*0.000*	***
Rest of Asia							0.682	***	*0.662*	***
Post-Communist Countries							-0.089		-0.086	
Variance	Model 0		Model 3		Model 4a		Model 4b		Model 4c	
Measures	2.427	63.0%	2.257	61.2%	2.257	61.9%	2.257	63.1%	2.257	63.1%
Respondents	1.051	27.3%	1.061	28.8%	1.061	29.1%	1.061	29.7%	1.061	29.7%
Country-Year	0.088	2.3%	0.088	2.4%	0.087	2.4%	0.089	2.5%	0.088	2.5%
Country-Source	0.284	7.4%	0.281	7.6%	0.243	6.7%	0.172	4.8%	0.170	4.7%

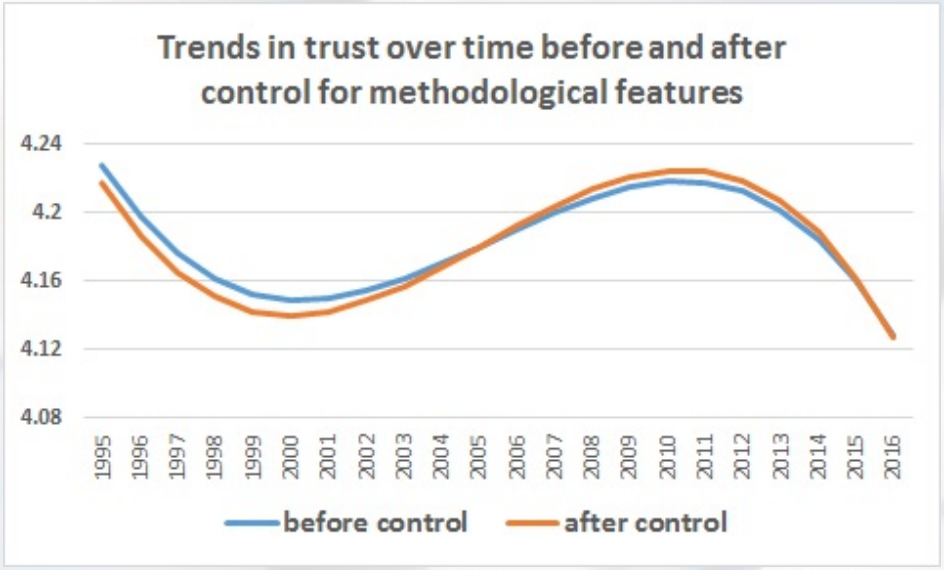
- 1. Introduce time, time^2 time^3
- 2. Introduce methodological features
- 3. Introduce country grouping
- 4. Introduce time * country grouping
- 5. Introduce time*institution

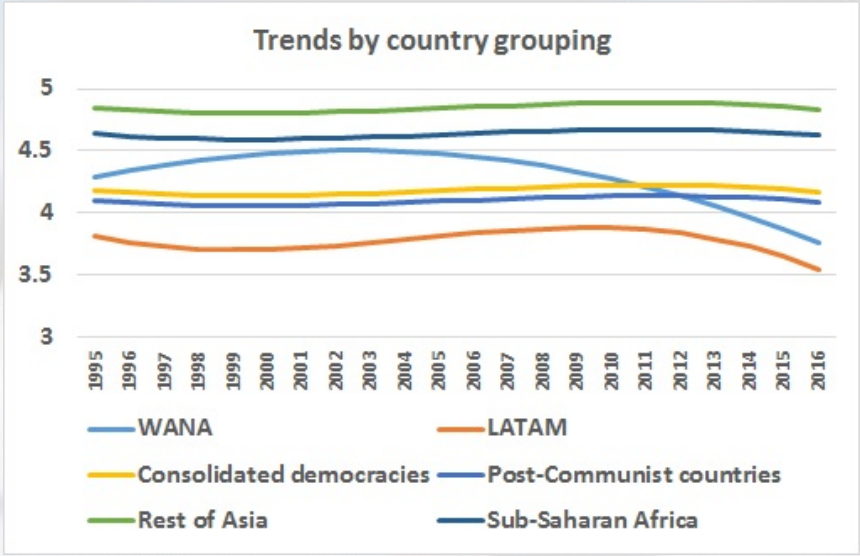
Focus on trends, sources & countries

	Model 3		Model 4a		Model 4c	
Intercept	4.180	***	4.459402	***	4.175306	***
Level Country-Year-Source						
Time	0.0104	**	0.0122	**	0.0012	ns
Time^2	0.0001	ns	0.0001	ns	0.0001	ns
Time^3	-0.0001	**	-0.0002	***	0.0000	ns
Level Country-Source						
LAPOP			0.185		0.490	***
WVS-EVS			-0.277	***	-0.102	
medium scale (5-7)			-0.399	***	-0.147	
Long scale (10-11 pts)			-0.665	***	-0.388	***
Consolidated Dem. (REF)						
West Asia N. Africa					0.302	**
Time					*-0.022*	
Time^2					*-0.004*	**
Sub-Saharan Africa					0.451	***
Central/South America					-0.364	**
Time					*0.025*	**
Time^2					*-0.001*	
Time^3					*0.000*	***
Rest of Asia					*0.662*	***
Post-Communist Countries					-0.086	

- Introducing methodological features slightly modifies the estimated trend.
- Introducing time in country groupings leads to overall trends non significant.

Trends in trust in the media after controls





- Impact of methodological features is small.
- Different trends in WANA countries and Latin America compared with other country-groupings.

Trends, sources and countries

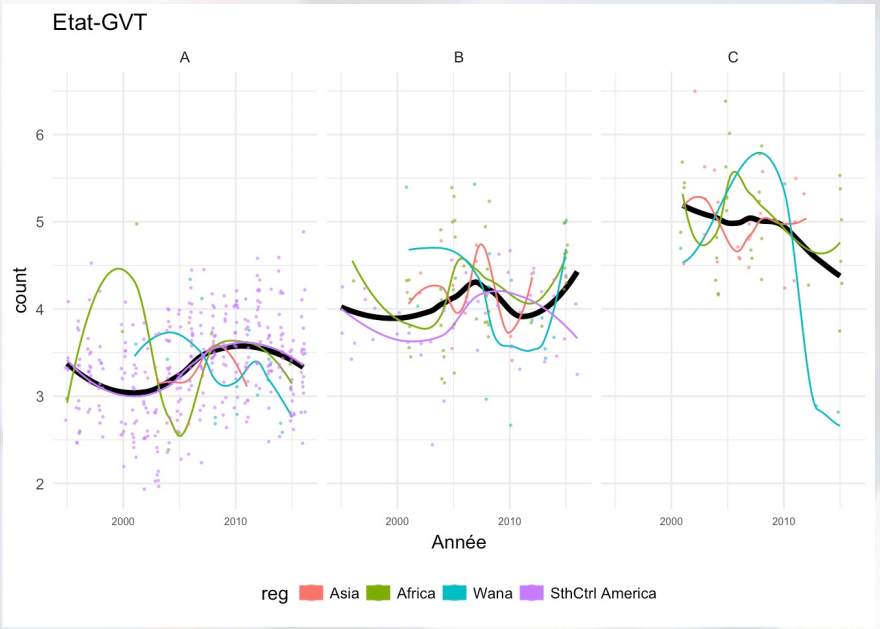
- 1. Before controlling for possible differences due to sources of data and trends in the country groupings,
  - Time and time^3 are significant overall.
- 2. After controlling for the presence and the methodological features of the survey projects
  - Slight change in trends.
- When we introduce interactions of time variables on country groupings
  - Time variables become non significant overall.
  - Time and time^3 are significant in Latin America only.
  - WANA starts with more trust than consolidated democracies but time^2 significant (sharp decline).

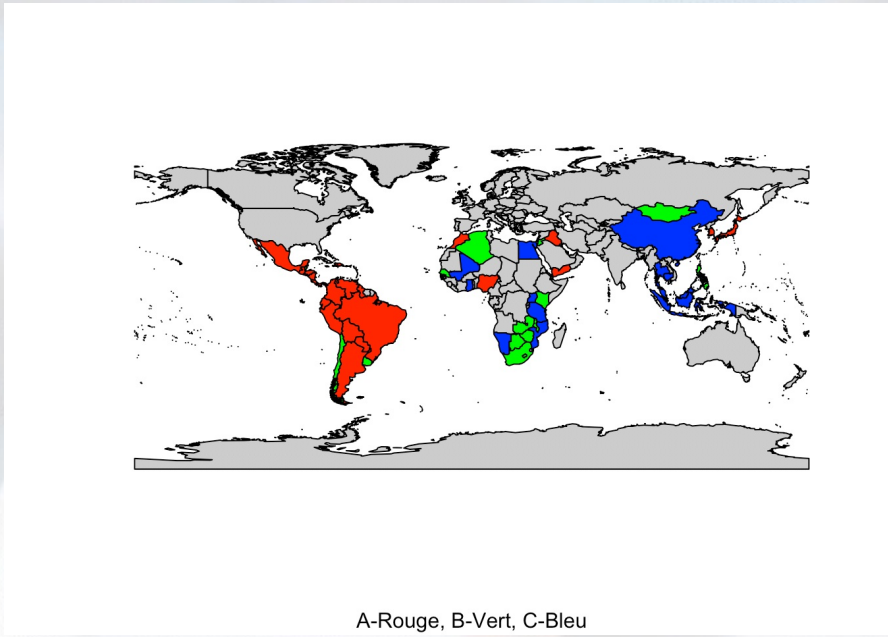
Focus on trends and institutions

	Model 3		M4e + t sur religion & Army	
Intercept	4.180	***	4.138	***
Level Institutions				
Media (REF)				
State/President	-0.048	***	-0.048	***
Governments	-0.388	***	-0.389	***
Parliament	-0.655	***	-0.655	***
Elections	-0.323	***	-0.323	
Political Parties	-1.119	***	-1.119	***
International Org.	-0.173	***	-0.171	***
Army	0.314	***	0.271	***
- Time			*0.026*	***
Police	-0.214	***	-0.214	***
Public Admin.	-0.202	***	-0.200	***
Judiciary	-0.380	***	-0.380	***
Church	0.619	***	0.643	***
- Time			*-0.030*	***
Trade Unions	-0.619	***	-0.619	***
ONG- Civil Society	-0.069	***	-0.069	***
Financial Organizations	-0.173	***	-0.173	
Enterprises	-0.354	***	-0.354	***

- Time trends may also differ by institution surveyed
- Army:
  - From .314
  - To .271 +.026 by year (=**+.52** over 20 years on a 7-point scale)
- Church:
  - From .619
  - to .643 -.03 by year (=**-.6** over 20 years)

What if trajectories are important: Trust in government.





A-Rouge, B-Vert, C-Bleu

- Trajectory analysis allows for classifying trends.
- Most of Latin America is in the low trust cluster (red); most of Asia is in the high trust cluster (Blue); Africa & Wana are mixed.
- Problem of predicting the past with the future...

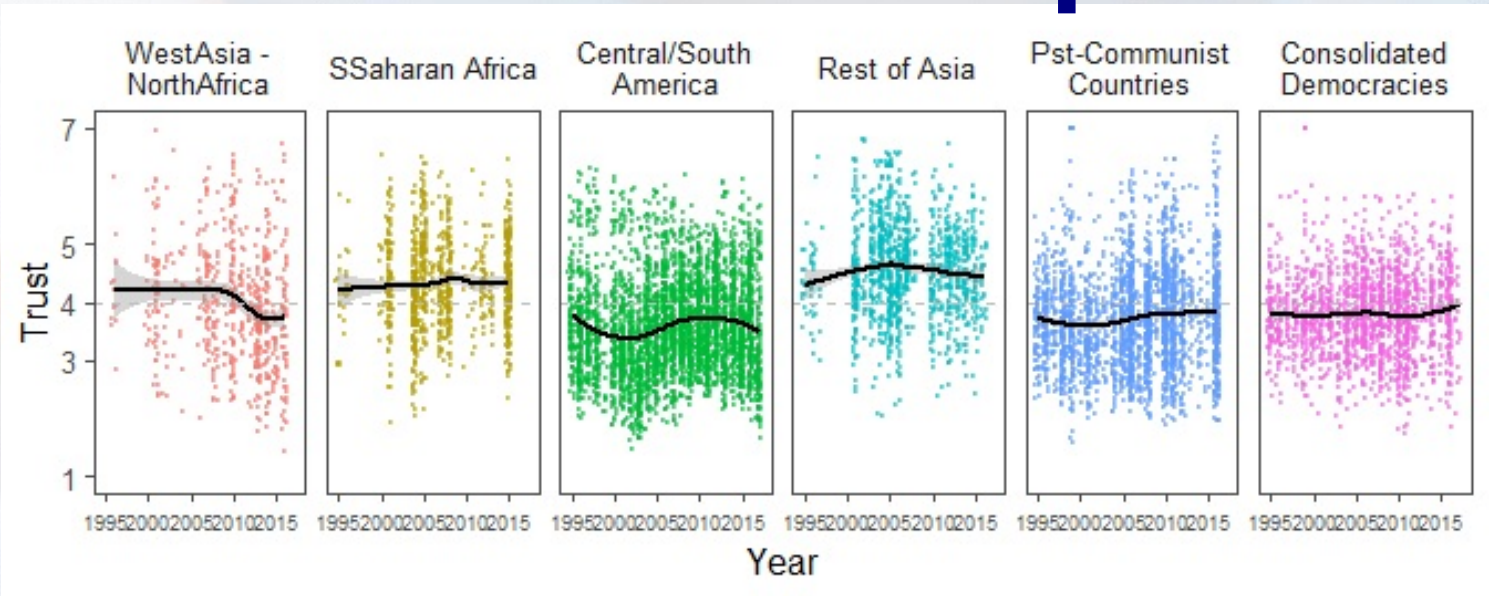
Issue no. 2: summary

- First, visualize data.
- Second, introduce time in analyses:
  - Assess trends taking into account,
    - Methodological features of the source of data.
- And assess specific trends
  - In different country groupings.
  - For different measures.
- In order to understand specific explanations for these trends.
- Third: Examine whether trajectories could be a fruitful avenue.

Issue no. 3: Disantangling variation due to country-level effects and survey-project coverage and methodological features

- When conducting analyses using multiple survey projects,
  - It is possible that what appears as change over time is due to the fact that different projects are conducted in different countries.
- Any analysis that combines data coming from different sources needs to:
  - Assess whether there are differences between sources.
  - And control for these differences.

For example



- The rise in trust in Latin America happens at the same time as LAPOP starts,... and LAPOP has a 7-point scale that leads to close to a half-point higher level of trust than the 4-point scales used by LatinoBarometro and WVS.
  - And LAPOP runs only in the Americas.
- Same problem for some projects in Eastern Europe.

Source vs country-level difference

- **After control for country-groupings**, compared with Barometers and 4-points scales, trust is
  - .38 points higher when using LAPOP rather than Barometers (no difference before control)
  - .33 points lower using a long scale (10-11 points) compared with .66 pts lower before control.
  - Other methodological features non-significant (significant differences in WVS-EVS & medium scales before control)

	Before controls		After controls	
LAPOP	0.185		0.383	**
WVS-EVS	-0.277	***	-0.072	
medium scale (5-7)	-0.399	***	-0.058	
Long scale (10-11 pts)	-0.665	***	-0.332	***

Source vs country-level difference

- Compared with consolidated democracies, after controlling for source, trust is
  - .47 points higher in Sub-Saharan Africa
  - .68 points higher in Asia
  - .31 points lower in Latin America (compared with no difference before control).
  - No difference in WANA and post-communist countries.

	Before controls		After controls	
West Asia N. Africa	0.070		0.038	
Sub-Saharan Africa	0.538	***	0.472	***
Central/South America	-0.082		-0.313	***
Rest of Asia	0.747	***	0.682	***
Post-Communist Countries	-0.112		-0.089	

Summary: Source vs country where survey is conducted

- It is not sufficient to harmonize survey data
- We need to control
  - For the presence-absence of a given source in a given year and country.
  - For the methodological differences that pre-existed before harmonization of the data.
  - See also Tomescu-Dubrow (2017) on the impact of indices of data quality.

Issue no. 4 : What about weights?

- Meta-analysis of individual data show differences in results using weights.
- Others (for example, Snijders and Boskers, 2012) suggest that if there is no substantial stratification,
  - It is not necessary to introduce design weights.
  - When we know the variables used to compile the design weights, we should examine the impact of introducing these variables as predictors.
- Weights are present for all the surveys that we have combined with an average of around 1.

What about weights at the survey level?

- "...a multilevel perspective takes into account the country's effect, and then we do not need population weights. This does not mean that weighting is not necessary in a multilevel perspective. As Pfeffermann et al. (1998) note, multilevel does not mean 'do not weight:' "When the sample selection probabilities are related to the response variable even after conditioning on covariates of interest, the conventional estimators of the model parameters maybe (asymptotically) biased" (p.24) Joye, Sapin & Wolf (2019).

What about weights at the survey level?

- Some countries are way larger than others.
- If we apply weights based on the population size, there will be too much variation in weights.
- Following Snijders and Boskers (2012) suggestion, we introduce
  - Ln (population size) and number of surveys per country as predictors in order to assess possible impact.

Introducing design weights variables

- The relationship between weight variables and trust is tenuous:
  - Small relationship before controls.
  - Virtually non significant after controls.
- However, while not changing the substantive conclusions,the presence of these variables has an impact on the coefficients for country groupings and methodological features.

		Before controls		After controls	
Intercept		**3.45**		**4.47**	
Ln(pop)		0.053	**	-0.015	n.s.
N surveys		-0.027	***	-0.010	*

Conclusion

- Challenges are numerous; we are just starting to explore solutions.
- Missing values at the survey level is a major challenge:
  - Use mutivariate-repeated measures, IRT models, imputation using all the information available.
- Taking time into account is essential
  - Visualizing data; introducing time variables; trajectories.
- The impact of methods and source
  - Can be controlled introducing variables.
- Weighting
  - Introduce weight variables but...we are only starting to explore this problem.

Trust in institutions - basic models

	Model 0		Model 1		Model 2		Model 3		Model 4a		Model 4b		Model 4c	
Intercept	3.956	***	4.221	***	4.184997	***	4.180	***	4.459402	***	4.16134	***	4.175306	***
Level Institutions														
Media (REF)														
State/President			-0.048	***	-0.048	***	-0.048	***	-0.049	***	-0.049	***	-0.049	***
Governments			-0.388	***	-0.388	***	-0.388	***	-0.388	***	-0.388	***	-0.388	***
Parliament			-0.655	***	-0.655	***	-0.655	***	-0.655	***	-0.655	***	-0.655	***
Elections			-0.323	***	-0.323	***	-0.323	***	-0.323		-0.323		-0.323	
Political Parties			-1.119	***	-1.119	***	-1.119	***	-1.119	***	-1.119	***	-1.119	***
International Org.			-0.173	***	-0.173	***	-0.173	***	-0.173	***	-0.173	***	-0.173	***
Army			0.314	***	0.314	***	0.314	***	0.314	***	0.314	***	0.314	***
Police			-0.214	***	-0.214	***	-0.214	***	-0.214	***	-0.214	***	-0.214	***
Public Admin.			-0.202	***	-0.202	***	-0.202	***	-0.202	***	-0.202	***	-0.202	***
Judiciary			-0.380	***	-0.380	***	-0.380	***	-0.380	***	-0.380	***	-0.380	***
Church			0.619	***	0.619	***	0.619	***	0.619	***	0.619	***	0.619	***
Trade Unions			-0.619	***	-0.619	***	-0.619	***	-0.619	***	-0.619	***	-0.619	***
ONG- Civil Society			-0.069	***	-0.069	***	-0.069	***	-0.069	***	-0.069	***	-0.069	***
Financial Organizations			-0.173	***	-0.173	***	-0.173	***	-0.173	***	-0.173	***	-0.173	***
Enterprises			-0.354	***	-0.354	***	-0.354	***	-0.354	***	-0.354	***	-0.354	***
Level Respondent														
woman					0.028	***	0.028	***	0.028	***	0.028	***	0.028	***
age_cat					0.011	***	0.011	***	0.011	***	0.011	***	0.011	***
Nb questions asked					-0.005		-0.005		-0.009	*	-0.008	*	-0.007	
Prop_Non-resp.					0.274	***	0.274	***	0.274	***	0.274	***	0.274	***
Level Country-Year-Source														
Time							0.0104	**	0.0122	**	0.0109	**	0.0012	ns
Time^2							0.0001	ns	0.0001	ns	0.0002	ns	0.0001	ns
Time^3							-0.0001	**	-0.0002	***	0.0001	***	0.0000	ns
Level Country-Source														
LAPOP									0.185		0.383	**	0.490	***
WVS-EVS									-0.277	***	-0.072		-0.102	
medium scale (5-7)									-0.399	***	-0.058		-0.147	
Long scale (10-11 pts)									-0.665	***	-0.332	***	-0.388	***
Consolidated Dem. (REF)														
West Asia N. Africa											0.038		0.302	
Time													*-0.022*	
Time ^{2}													*-0.004*	**
Sub-Saharan Africa											0.472	***	0.451	***
Central/South America											-0.313	***	-0.364	**
Time													*0.025*	**
Time ^{2}													*-0.001*	
Time ^{3}													*0.000*	***
Rest of Asia											0.682	***	*0.662*	***
Post-Communist Countries											-0.089		-0.086	
Variance	**Model 0**		**Model 1**		**Model 2b**		**Model 3**		**Model 4a**		**Model 4b**		**Model 4c**	
Measures	2.427	63.0%	2.257	61.3%	2.257	61.2%	2.257	61.2%	2.257	61.9%	2.257	63.1%	2.257	63.1%
Respondents	1.051	27.3%	1.064	28.9%	1.061	28.8%	1.061	28.8%	1.061	29.1%	1.061	29.7%	1.061	29.7%
Country-Year	0.088	2.3%	0.090	2.4%	0.089	2.4%	0.088	2.4%	0.087	2.4%	0.089	2.5%	0.088	2.5%
Country-Source	0.284	7.4%	0.273	7.4%	0.278	7.6%	0.281	7.6%	0.243	6.7%	0.172	4.8%	0.170	4.7%
Total	***3.849***		***3.682***		***3.685***		***3.687***		***3.648***		***3.578***		***3.575***	
Deviance	80555452		80449736	*-105716*	80446129	*-3607*	80446119	*-10*	80446068	*-51*	80445959	*-109*	80445917	*-42*
dl	5		20	*15*	24	*4*	27	*3*	31	*4*	36	*5*	41	*5*
Year														
Country														

Mutiple questions, multiple issues, multiple solutions

- The use of one source over many sources for the same countries: effect of source vs country.
- The challenge of working with large, heterogenous, survey data.
- The challenge of weighting respondents, and countries?
- The use of multilevel models:
  - Taking time into account.
  - Taking within respondent variation into account.
  - Explained vs distributed variance.
  - Cross-level interactions.
  - The problem of the reference category
  - Residuals at the country-year level.
- Measurement equivalence and scales:
  - Use of IRT (van der Meer et al.), use of CFA (Marien, Zmerli?)?
- Classification of countries or of trajectories?
- Use of external data.