Combining Data: Why Not Dream Big?

Claire Durand,
Department of sociology,
Université de Montréal

Presented at Computational Methods for Survey and Census Data in the Social Sciences, June 20th, 2014, Montréal, Qc, Canada

Outline

- Imagine...
- Example 1: Combining surveys results:
  - Evolution of support for sovereignty
  - Evolution of voting intention for Obama and Romney in 2012
- Example 2: Combining data files:
  - Evolution of trust in Canada
  - Aboriginals living outside FN communities and the communities they live in
- Conclusion

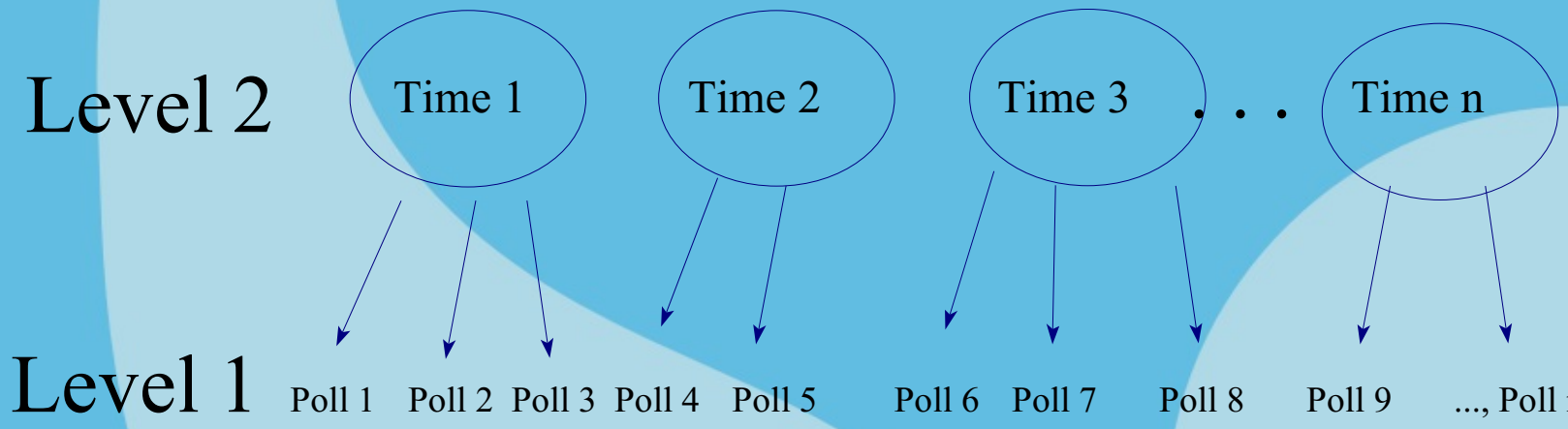
Imagine

- The actual situation is characterized by access to huge quantities of data from different sources
- We would like to be able to use the full potential of all these data
  - To trace evolution with time of different attitudes and behaviors
  - To compare across regions and groups and fully understand the differences occurring in time and space.
  - To understand how the context in which people live may influence their behaviors and attitudes
- But we are hindered by
  - The fact that measures of similar concepts are not always the same
  - The idea that this means that we cannot compare across studies, groups, time and space

Example 1: Combining survey results

Multilevel modelling

- **At Level 2: evolution with time and its predictors.**



- **At level 1: variation between polls and its predictors.**

Example 1: Combining survey results

A) Evolution of support for sovereignty in Quebec

- Close to 700 polls between 1976 and 2008
- Questions differ in:
  - Wording, i.e. whether the question pertains to an opinion or to voting intention
  - Constitutional option, i.e. whether the question refers to sovereignty with an association/partnership with the rest of Canada, to sovereignty per se, to independence or to separation
  - Mode of administration, prop. of undecideds and sample size
- The research questions are:
  - What is the likely evolution of support for sovereignty?
  - Which events, if any, influenced this evolution?
  - Is the evolution the same whatever the question asked -- voting intention or attitude, constitutional option?

Example 1: Combining survey results

A) Support for Quebec Sovereignty 1976-2008 (Yale & Durand, 2011)

- 696 polls, 7 periods, 3 under study
- At level 1:
  - Question wording (constitutional option):
    - Separation
    - Independence
    - Sovereignty
    - Sovereignty- association or sovereignty-partnership
  - Type of question: voter intent vs favorability, mandate
  - Proportion of undecideds
  - Sample size
- At level 2:
  - Time, time squared, 3rd power;
  - Elections;
  - Events: Accords -- Meech Lake, Charlottetown -- and sponsorship scandal.

At level 1

Effects related to polls and questions

Table 1 – Summary of Average Effects Linked to Measure

		1976–1979	1989–1995	1995–2008
Fixed effects				
	Intercept	39.84*** (1.51)	60.71*** (1.05)	48.98*** (1.57)
	Voterint	*n.s.*	-3.18*** (0.76)	-4.04** (1.35)
	Sovereignty	—	-7.63*** (1.09)	-6.75*** (0.87)
	Independence	—	-13.46 (0.90)	-8.95*** (1.31)
	Separation	—	-16.84*** (1.08)	-11.38*** (1.52)
	Mandate	14.01*** (0.85)	—	—
	Extreme	-20.27*** (1.67)	—	—
	Size	*n.s.*	*n.s.*	*n.s.*
	Non-disclosers	*n.s.*	*n.s.*	0.21* (0.0941)
Variance component				
Level-1	*R*	20.04	24.99	19.93
	(%)	66	49	61
Level-2	*Intercept*	10.10***	25.68***	12.53***
	(%)	34	51	39
Deviation		371.89	1762.09	1700.45
	Parameters	5	7	8
	DL	21	58	121

* P< 0.05
** P<0.01
*** P<0.001
n.s. not significant. The variable was tested in one previous model and removed from the model.

Compared to attitudes re: Sov-association.

Voter intent: -3 pts to -4 pts

Sovereignty: -7 pts to -8 pts

Independence: -9 pts to -14 pts

Separation: -12 pts to -17 pts

Mandate: + 14 pts

Extreme: -20 pts

+ item NR →+ support

49%-66% of variance btw polls, the rest, between time units

At level 2

Effects related to time and events

Table 2 – Final Models of Change for 1989–1995 and 1995–2008.

	Sovereignty-partnership	Sovereignty	Independence	Separation
1989–1995				
intercept	45.07***	45.07	29.55***	37.74
month	1.79***	1.79	2.18*	0.73***
meech1	-2.30***	-2.51***	-2.79**	-1.10***
charlot1	0.48***	0.48	0.48	0.48
1995–2008				
intercept	55.30***	50.92***	47.10***	55.30
month	-0.31***	-0.31	-0.31	-0.87***
month²	0.002***	0.002	0.002	0.007***
sponsorship	4.76**	4.76	4.76	-9.46***
gomery1	-0.55***	-0.55	-0.55	-0.55

* P< 0.05
** P<0.01
*** P<0.001

Support for various options...

1989-1995

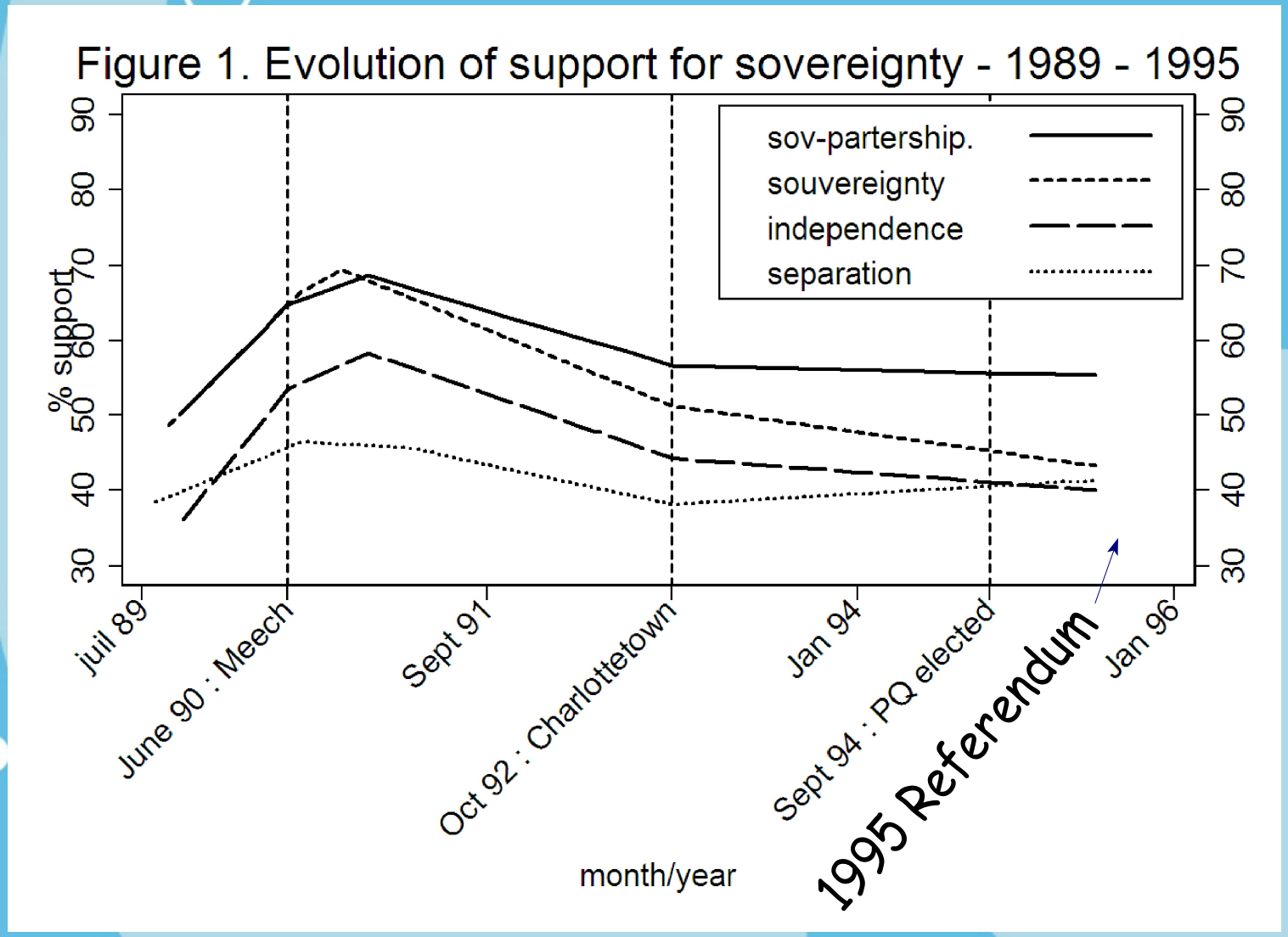
↑ with time
↓ after Meech failure
↑ after Charlottetown failure

1995-2008

U shaped with time
↑ after spons. Scandal except for separation
↓ after Gomery report

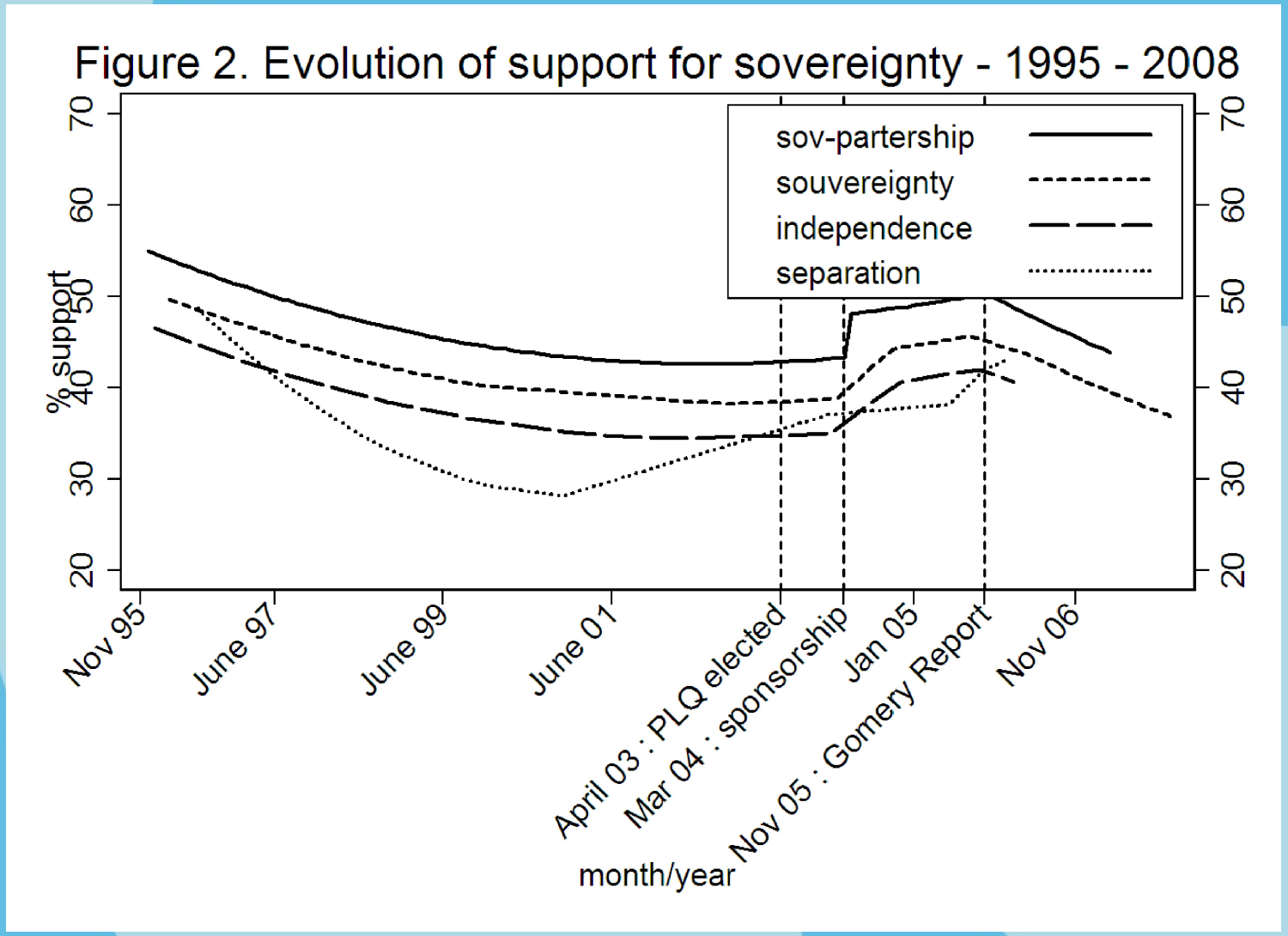
Combined model 1989-1995

Evolution according to constitutional option - voter intent



Combined model 1995-2008

Evolution according to constitutional option - voter intent



Example 1: Combining survey results

B) Evolution of voting intentions for Obama and Romney, U.S. 2012

- The question:
  - What is the likely evolution of voting intentions for the 2012 US presidential election?
  - Is this evolution the same whatever the survey mode of administration?
  - What is the impact of using a likely voter model?

How to perform analysis

B) Evolution of voting intentions for Obama and Romney, U.S. 2012

- Level 2 model: Defining time: week (vs day)
  - Time, time squared, time cubic, power 4, power 5
- Level 1:
  - Dependent variable:
    - Estimate of voting intention for Obama or Romney
  - Independent variables:
    - Mode of administration (not significant)
    - Number of days poll is in the field
    - Sample size
    - Proportion of non-disclosers
    - Use of a likely voter model

Equations…Final model

LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering)

$$OBAMA = \beta_0 + \beta_1(UNDEC2) + \beta_2(NBJOURS) + \beta_3(LIKELY\_V) + \beta_4(SAMPLESQ) + r$$

LEVEL 2 MODEL (bold italic: grand-mean centering)

$$\beta_0 = \gamma_{00} + \gamma_{01}(TEMPS) + \gamma_{02}(TEMPS2) + \gamma_{03}(TEMPS3) + \gamma_{04}(TEMPS4) + u_0$$

$$\beta_1 = \gamma_{10}$$

$$\beta_2 = \gamma_{20}$$

$$\beta_3 = \gamma_{30} + \gamma_{31}(TEMPS) + \gamma_{32}(TEMPS2)$$

$$\beta_4 = \gamma_{40}$$

Only the intercept is allowed to vary per week.

- At level 1: support for Obama is influenced by
  - The proportion of undecideds in the poll, the number of days the poll was in the field, the use of a Likely voter model and the sample size ($1/\sqrt{n}$)
- At level 2,
  - The intercept is influenced by time (linear, quadratic, cubic and power 4).
  - The influence of the likely voter model varies with time linear and quadratic.

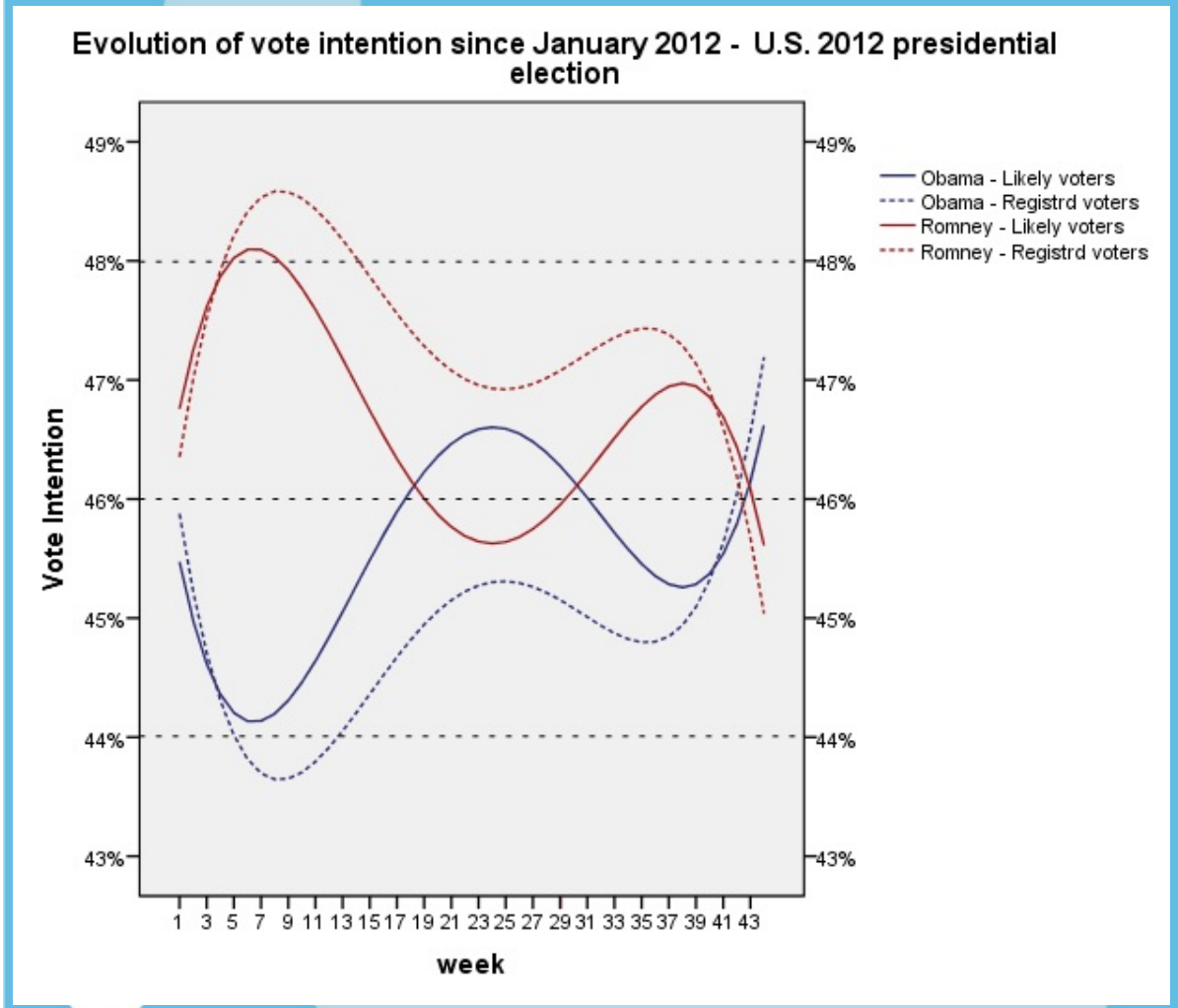
Results: Obama

Fixed effects		Coefficient	Std error	T-ratio	d.f.	P-Value
INTERCEPT1: B0						
INTRCPT2	G00	50.9356	0.7750	65.7220	39	0.000
TEMPS	G01	0.0615	0.0214	2.8690	39	0.007
TEMPS2	G02	-0.0110	0.0025	-4.4370	39	0.000
TEMPS3	G03	-0.0001	0.0001	-1.4560	39	0.153
TEMPS4	G04	0.0000	0.0000	5.2930	39	0.000
UNDEC2 SLOPE: B1						
INTRCPT2	G10	-0.5028	0.0280	-17.9280	381	0.000
NBJOURS SLOPE: B2						
INTRCPT2	G20	-0.1516	0.0390	-3.8900	381	0.000
LIKELY_V SLOPE: B3						
INTRCPT2	G30	1.3170	0.4382	3.0060	381	0.003
TEMPS	G31	0.0001	0.0148	0.0060	381	0.995
TEMPS2	G32	-0.0039	0.0014	-2.8660	381	0.005
SAMPLESQ SLOPE: B4						
INTRCPT2	G40	-89.6405	14.5841	-6.1460	381	0.000

Note: Events could have been added but there was no cue that some important events had influenced voting intentions substantially.

Evolution of voting intention for Obama and Romney, US election 2012

Graph generated using SPSS (or Stata, or…)



- Likely Voter Model: 59% of the polls.
- Registered voters or adults: 41%
- All the other variables have been put at the mean -- number of days (4.22), sample size (1268), proportion of non disclosers (7.77).

Final results: variance explained

Prediction of voting intention for Obama

	Model 0	Model Niv1	Full model
Var. Niv. 2: weeks	.52	.52	.19
Var. Niv. 1: polls	4.19	1.82	1.76
Prop. var btw weeks	11.0%	22.2%	9.7%
Prop var. explained btw polls	-	56.6%	58.0%
Prop. var. explained btw we>		-	63.5%

- At the beginning, 11% of the variance is between weeks, 89% between polls.
- Variables at level one -- number of days in the field, sample size, proport. of undecideds and use of a likely voter model -- explain 57% of the variance between polls.
- Evolution with time -- including the effect of the varying impact of the likely voter model -- explains 63.5% of the variance with time.

Example 2: Combining data bases (i.e., individual records)

A) Evolution of trust towards institutions in Canada

- 50 surveys with questions pertaining to trust in institutions from 1976 to 2008
- n=127,500 respondents.
- Measures vary according to:
  - The object of trust: religion, schools, Unions, media, etc.
  - Whether the object is the institution itself or the people within the institution, i.e. religion vs preasts, schools vs teachers, unions vs union leaders, etc.
  - The wording and the number of response categories
- For each survey, it is necessary to figure out how to modify the data files so that each data base is on a common basis, including socio-demographics: looking for the smallest common denominator.

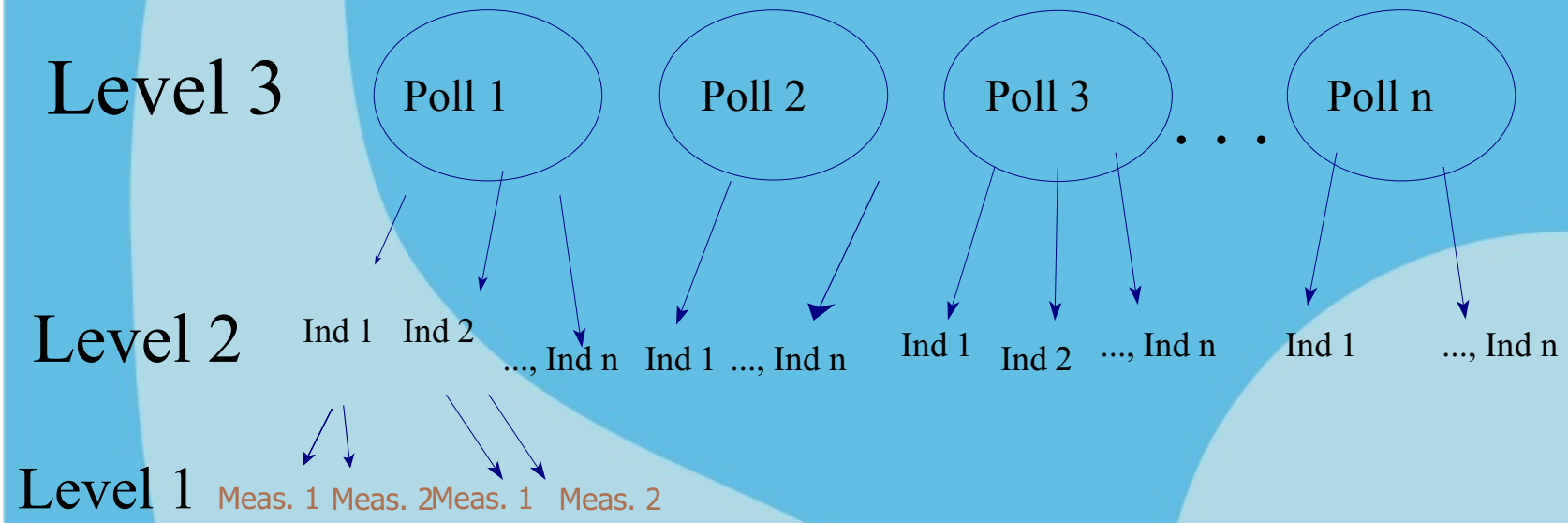
How to proceed

- In each file, variable names changed to common names:
  - AnswerTrustReligion, object TrustReligion (people vs institutions),...
  - AnswerTrustSchools, object TrustSchools (people vs institutions),...
  - AnswerTrustUnions, object TrustUnions (people vs institutions),...
  - For the whole file, year of survey, wording of trust questions,...
- Data from all the surveys are combined into one file.
- Then, restructure the file so that there are as many lines per respondent as the number of Trust questions asked to each respondent.
  - Person 1,
    - Line1: objectTrust (religion), Trust (religion), etc.
    - Line2: objectTrust (school), Trust (school), etc.
    - Line3: objectTrust (unions), Trust (unions), etc.

Example 2: Combining survey data

Multilevel modelling

- **At Level 3: evolution with time and its predictors.**



- **At level 2: individuals and their characteristics.**
- **At level 1: Trust and its objects and characteristics.**

Equations… Final model

- At level 1: Trust at the question level
  - Trust= π_0+ π_1(religion) +π_2(Unions) + e
- At level 2: Trust at the individual level
  - π_0= β_{00}+ β_{01}(Maritimes) +β_{02} (Quebec) + β_{03}(Ontario) + β_{04} (old)+r_0
  - π_1= β_{10}
  - π_2= β_{20}
- At level 3: Trust at the survey level
  - β_{00}= γ_{000}+μ_{00}
  - β_{01}= γ_{010}+γ_{011} (Year)
  - β_{02}= γ_{020}+γ_{021} (Year)
  - β_{03}= γ_{030}+γ_{031} (Year)
  - β_{04}= γ_{040}
  - β_{10}= γ_{100}+γ_{101} (Year)
  - β_{20}= γ_{200}+γ_{201} (Year)

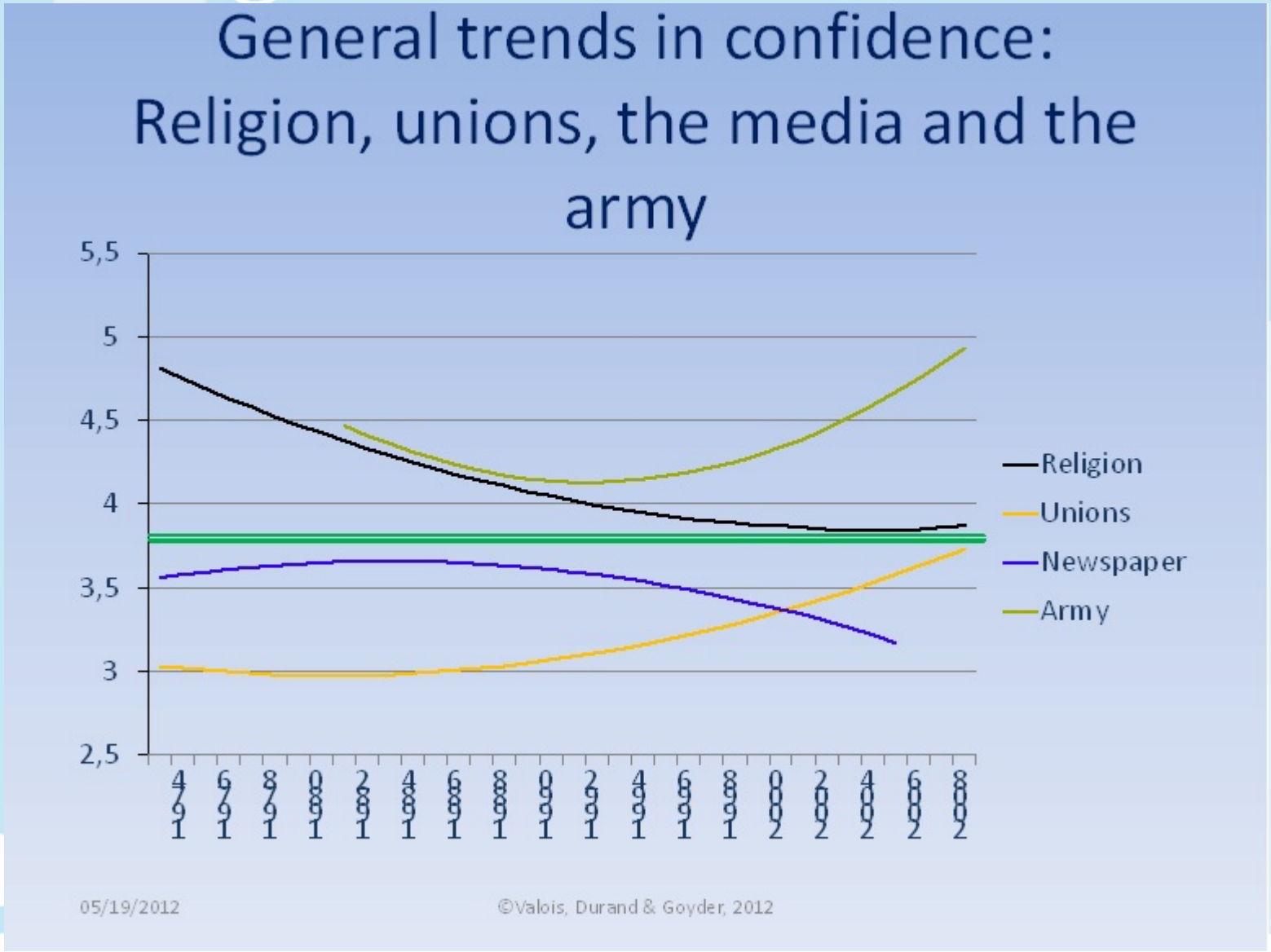
*Trust may evolve differently with time in different provinces and according to the object of confidence.

Results: Trust

Fixed effects	MODEL 0		MODEL 1		MODEL 2		MODEL 3		MODEL 4	
π0: intcp3	3.852111	***	4.469601	***	4.393476	***	4.377160	***	4.371209	***
MARITIM: intcp3					0.336602	***	0.337172	***	0.375634	***
Year									-0.007682	*
QUEBEC: Intcp3					0.111681		0.113499		0.081894	
Year									0.012219	**
ONTARIO: intcp3					0.028412		0.029255		0.057478	*
Year									-0.008415	***
OLD: intcp3							0.079404	***	0.077850	**
RELIGION: intcp3			-0.393494	***	-0.393881	***	-0.394182	***	-0.353401	***
Year									-0.042896	***
UNION: intcp3			-1.356785	***	-1.356762	***	-1.356761	***	-1.334064	***
Year									-0.010732	
Variance Level1	2.75814		2.43858		2.44211		2.44282		2.40862	
Variance Level2	0.21294		0.35200		0.33575		0.33395		0.34674	
Variance Level3	0.50134		0.23865		0.24134		0.24037		0.25019	
Prop var niv2	6.1%		11.6%		11.1%		11.1%		11.5%	
Prop var niv3	14.4%		7.8%		7.9%		8.0%		8.3%	

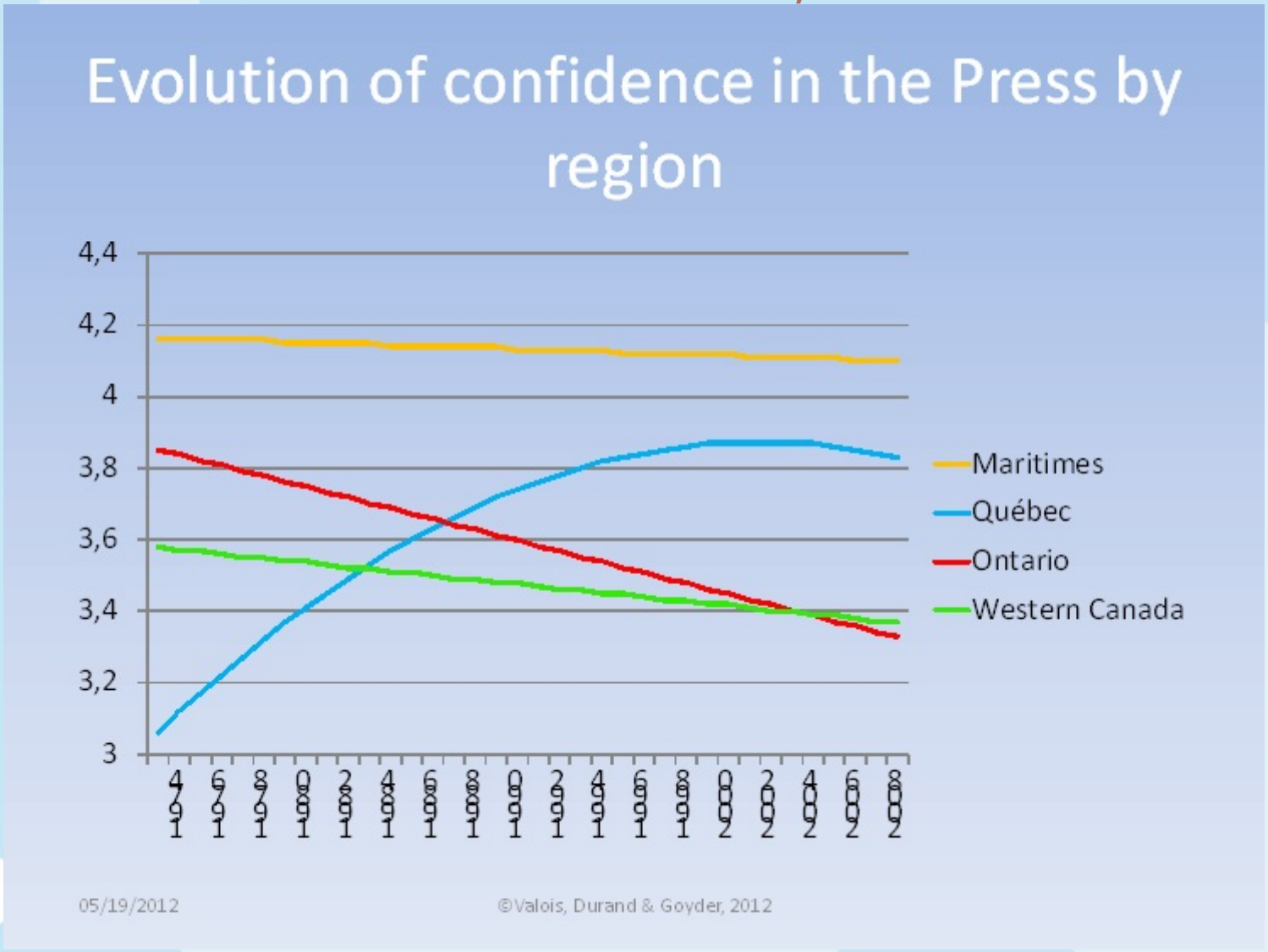
Evolution of trust in some institutions

With Isabelle Valois, 2012



Evolution of trust in media in different regions

With Isabelle Valois, 2012



Example 2: Combining data files

B) Individual and collective data

- 1) Aboriginal People Survey conducted among First Nation people living outside First Nation communities.
- 2) The question is whether First Nation people tend to fare better when they live in a community in better socio-economic condition.
  - We could hypothesize that FN people go to wealthier communities thinking that they will be able to improve their situation but do not manage well in these environments. It is the idea that urban FN people are the poor "urban Indians".
  - Or else, like non FN people, in a wealthier environment, they fare better.

Combining data: Example 2b

Individual and collective data

- For every FN respondent in APS, we have the identifier of the community where they live.
- For every community, we have an index of well-being, i.e. the IBC (index of community well-being based on income, education, activity and housing)
- It is easy to
  - Recuperate the information on the IBC with an appropriate software and make sure that the community identifier is entered in the same way as in the APS file;
  - Merge the two files;
  - Produce the level 2 files with one line per community, the IBC and other interesting information like the proportion of FN people living in the community, the mean level of education of FN people living in the community, etc.

Results

Dependent variable: Income in categories (Yves-Emmanuel Massé-François, 2013)

Fixed effects		Coefficient	Std error	T-ratio	d.f.	P-Value
INTERCEPT1: B0						
INTRCPT2	G00	3.637	0.257	14.160	507	0.000
IBC	G01	0.138	0.038	3.586	507	0.001
MALE SLOPE B1						
INTRCPT2	G10	1.382	0.248	5.578	2543	0.000
AGE_GROUP SLOPE B2						
INTRCPT2	G20	1.684	0.439	3.837	2543	0.000
HEALTH SLOPE B3						
INTRCPT2	G30	-0.911	0.398	-2.291	2543	0.022
EDUCATION SLOPE: B4						
INTRCPT2	G40	1.709	0.294	5.807	2543	0.000

- The higher the IBC of the community, the higher the income of FN people in the community.
- The model accounts for 10% of the level 1 variance (between individuals) and 29% of the variance at level 2 (between communities).

Limits

- When combining data, we need to have enough information at all levels, for example,
  - Variation between question wording (example1a) has to be spread on all time periods.
    - We had to perform analyses separately for the different periods in order to take this into account.
  - Use of likely voter model (example 1b) has to be spread also throughout the period.
- When combining data files,
  - We cannot take into account all the subtilities of question wording
  - It may be difficult to find a common denominator for response categories.
    - Use mean, put on a 7 point scale, use proportion of high trust or of low trust as dependent variables, etc.
  - It may be very difficult to find a common denominator for variables like age, income, etc.

Other possibilities

- Use of Item response theory (IRT) to put scales on the same standardized scale
  - Limit: When computing a composite scale of likert-type items, at least one question would have to be the same for all respondents.
- Use of local regression (loess) in order to estimate evolution with time for different groups or different wordings
  - Limit: mostly descriptive.
  - But analysis very easy to perform: Use for quick estimation of the evolution of voting intention.

Conclusion

Imagination in power

- There are incredible possibilities to combine data in order to get to the “big portrait”.
- It allows for a thorough use of the data already collected in order to better understand different phenomena and their evolution with time and within different contexts.