



General Social Survey
Enquête sociale générale

Twenty Years of General Social Survey Data

Michael Wendt



Statistique
Canada

Statistics
Canada

Canada



Outline

- Background
- Historical data base
- Data model
- Examples
- Conclusion



Statistique
Canada

Statistics
Canada

Canada 



Background

- Have 20 complete cycles of data
- User guides, data dictionaries, methodology descriptions, weighting and estimation guides, etc. available
- Many research articles have been written over the years



Statistique
Canada

Statistics
Canada

Canada 



Historical Data Base

- Now have a wealth of data available
 - 330,000 respondent records (albeit over different years), thousands of variables on each respondent
 - Also, time use information, episodes of victimization, etc.
- Idea is to package the data in an easily accessed form and provide tools for use
- Provide a database containing the 20 years of data



Statistique
Canada

Statistics
Canada

Canada



Historical Data Base

- Not pooled database but separate data sets with harmonized variables and concepts in some coherent way
- Ambitious version of idea
 - “Glue” respondent files together by cycle
 - 330,000 rows x thousands of columns
 - Have a year (= cycle = file) identifier
 - Bootstrap weights
 - Provide links to other files extant like instances of time use, victimization episodes





Historical Data Base

- Unfortunately, cannot simply “stack” the 20 data sets on top of each other
- Issues
 - Variables can have different names, formats, code sets
 - Processing methods have changed over the years (fortunately, not that much)

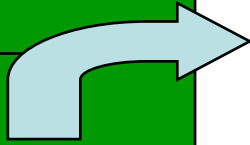
cycle	age	sex	...
1			
2			
3			
⋮			





Data Model

	Core
Cycle	
1	
2	
3	
4	
5	
...	



respondent information common to all or almost all cycles

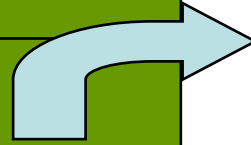
age, sex, marital status, living arrangements, province, etc.





Data Model

	Core	Control
Cycle		
1		
2		
3		
4		
5		
...		



unique identifiers, meta data, newly derived variables, methodological control

record id, weight, matching variables, etc.





Data Model

	Core	Control	Theme
Cycle			
1			
2			
3			
4			
5			
...			



themes that have been repeated

time spent doing yard work, etc.





Data Model

	Core	Control	Theme	Special
Cycle				
1				
2				
3				
4				
5				
...				

topics in only a few consec. cycles

'net use, etc.





Data Model

	Core	Control	Theme	Special	Other
Cycle					The rest ? or !
1					
2					
3					
4					
5					
...					





Data Model: Summary

	Core	Control	Theme	Special	Other
Cycle					
1					
2					
3					
4					
5					
...					





Historical Data Base

- Will allow users to
 - Follow estimates over time
 - Create time series of estimates
 - Follow “cohorts” over time (age 25 in 1990, age 30 in 1995, etc.)
 - Look at similar subpopulations at different time periods (age 25 in 1990 versus age 25 in 1995)
 - Combine small domains
 - Use different cycles to increase sample sizes
 - In general, cannot pool all the data and look at 160,000 females, say, because there is no real reference population





Data Model: Variable Uses

- Core
 - Estimates over long time series
 - Calibration
 - Cohorts (different year use different age) and age domains (different year yet same age)
- Control
 - Building the data base
 - Quality checks
 - Tools such as variance estimation (Bootstrapping)
 - Matching to other data sets





Data Model: Variable Uses

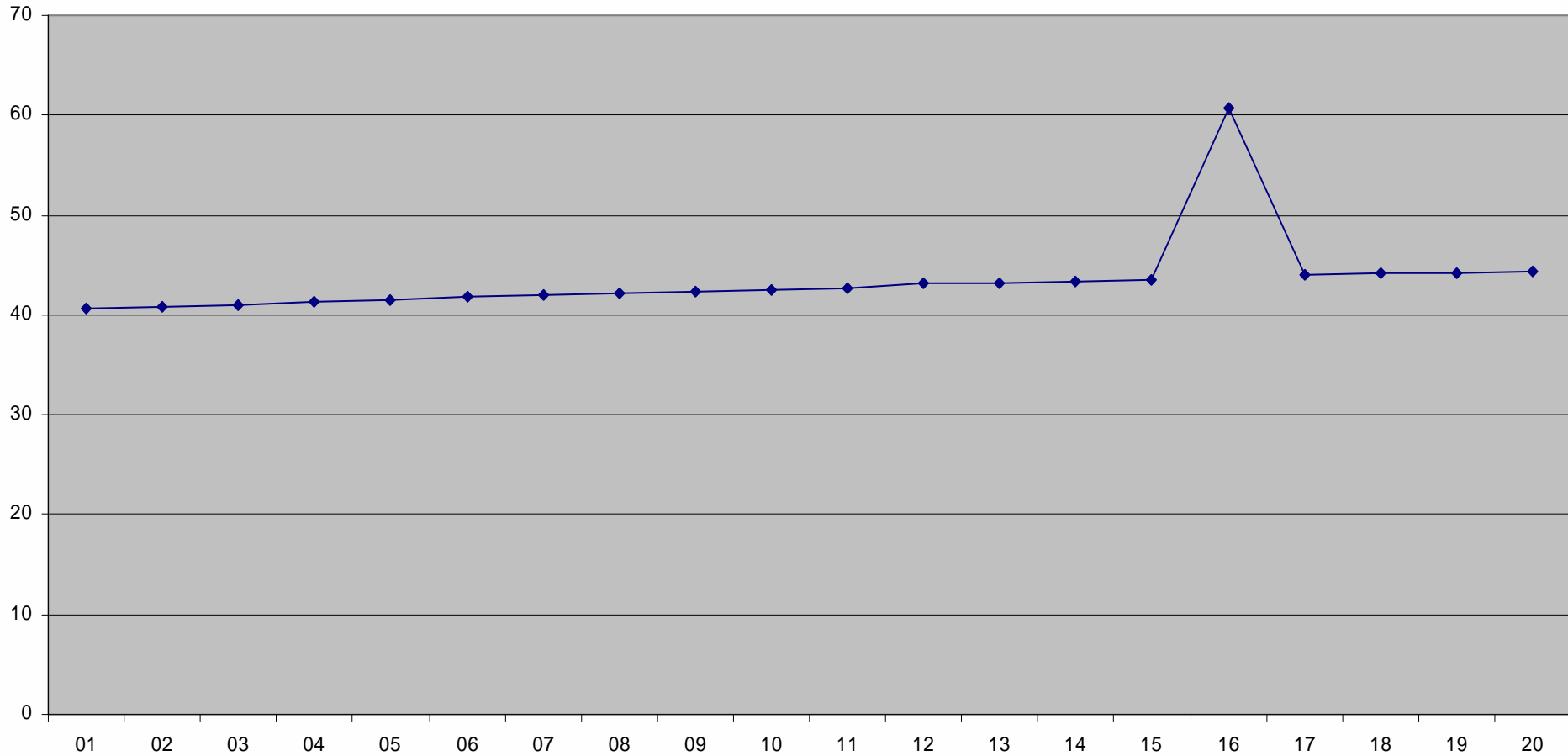
- Theme
 - Cohorts
 - Time series with fewer points but more detail
- Special
 - Small domains
 - Very short time series
- Other
 - Depends on type, quantity, etc.





Example: Estimates as a Time Series

Average Age of Respondent



Statistique
Canada

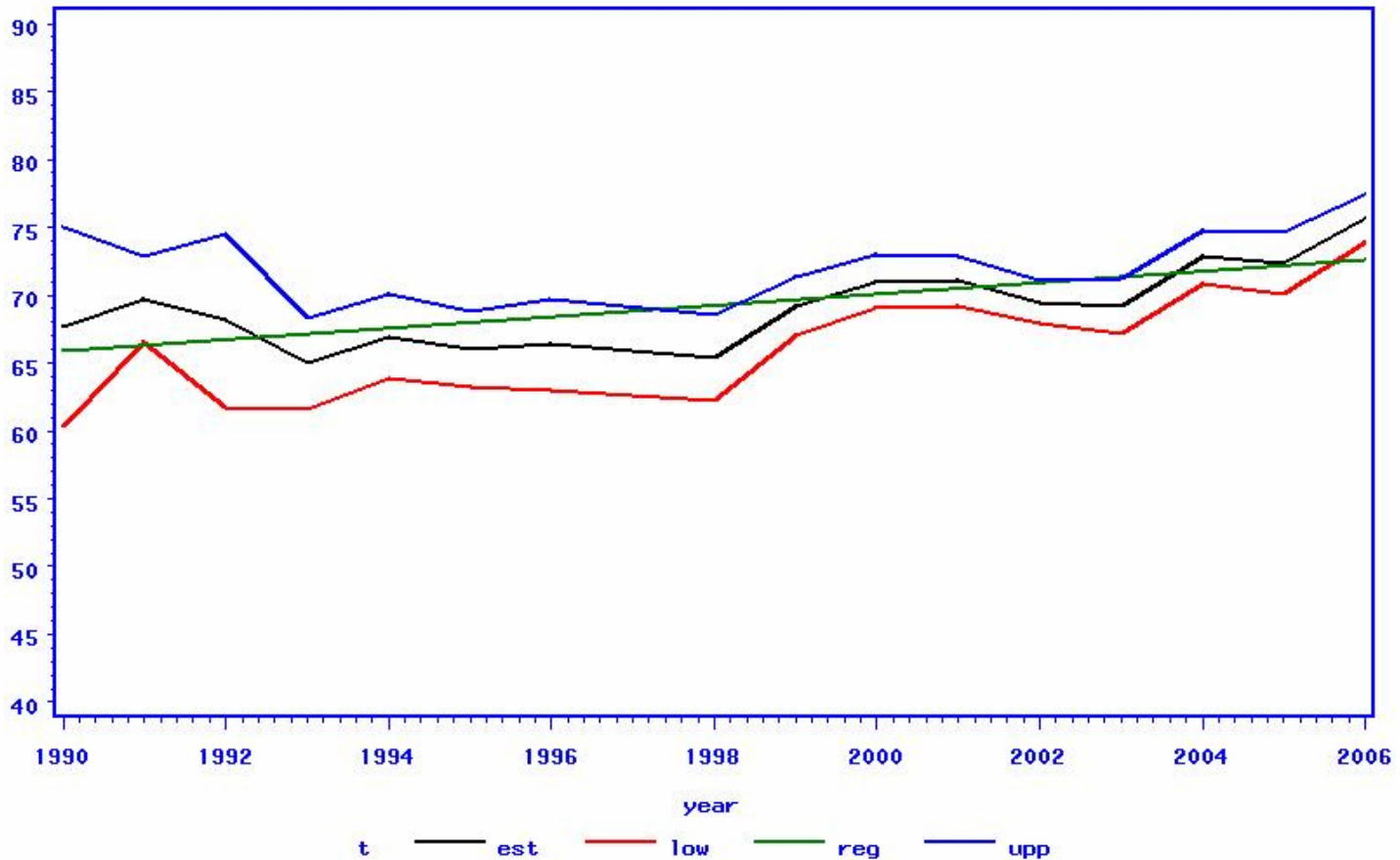
Statistics
Canada

Canada



Example: Estimates as a Time Series

Percentage of dwellown = 1 for 45+ in Quebec



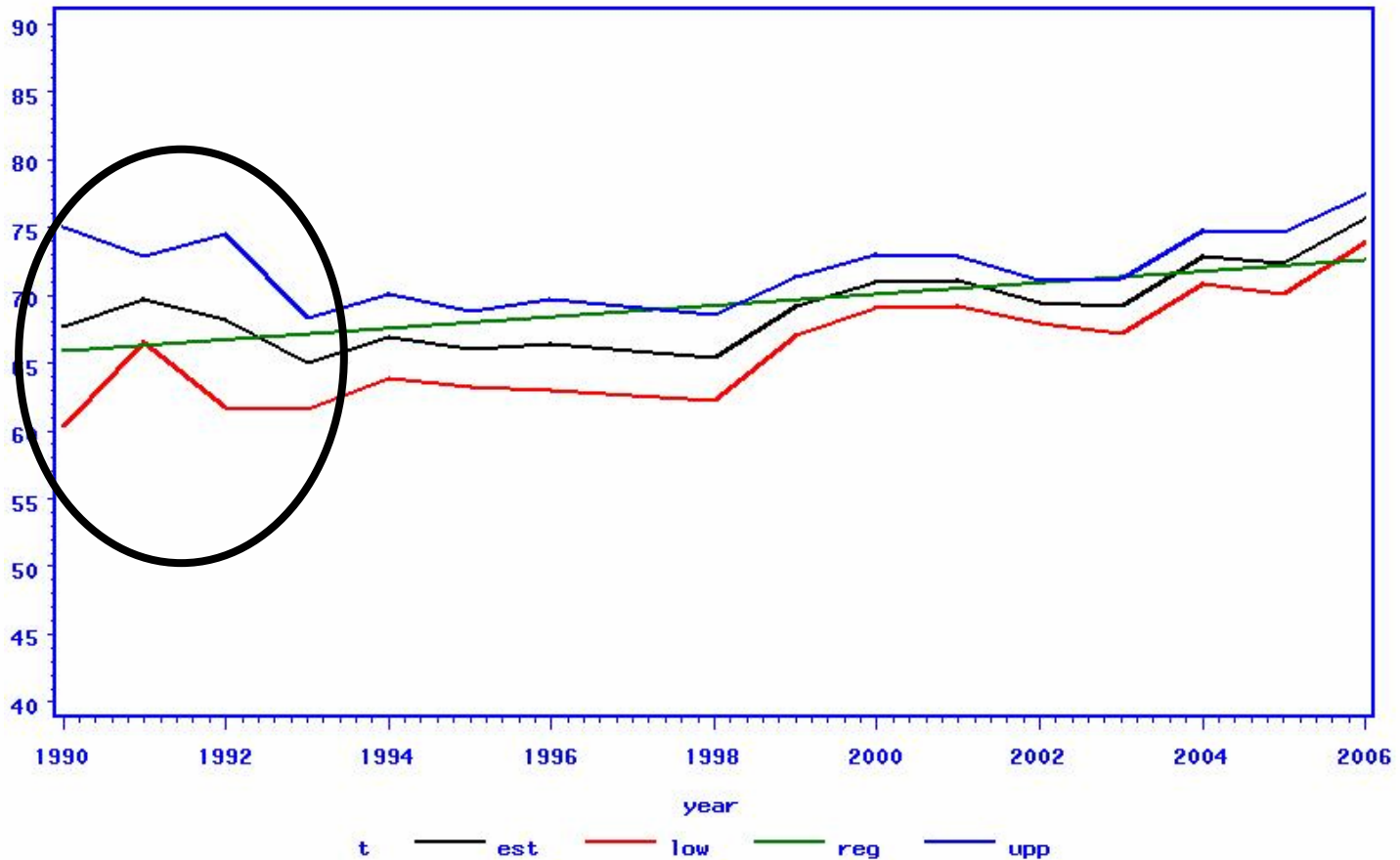
note: for illustrative purposes only
source: Statistics Canada, General Social Survey





Example: Estimates as a Time Series

Percentage of dwellown = 1 for 45+ in Quebec



note: for illustrative purposes only
source: Statistics Canada, General Social Survey



Statistique
Canada

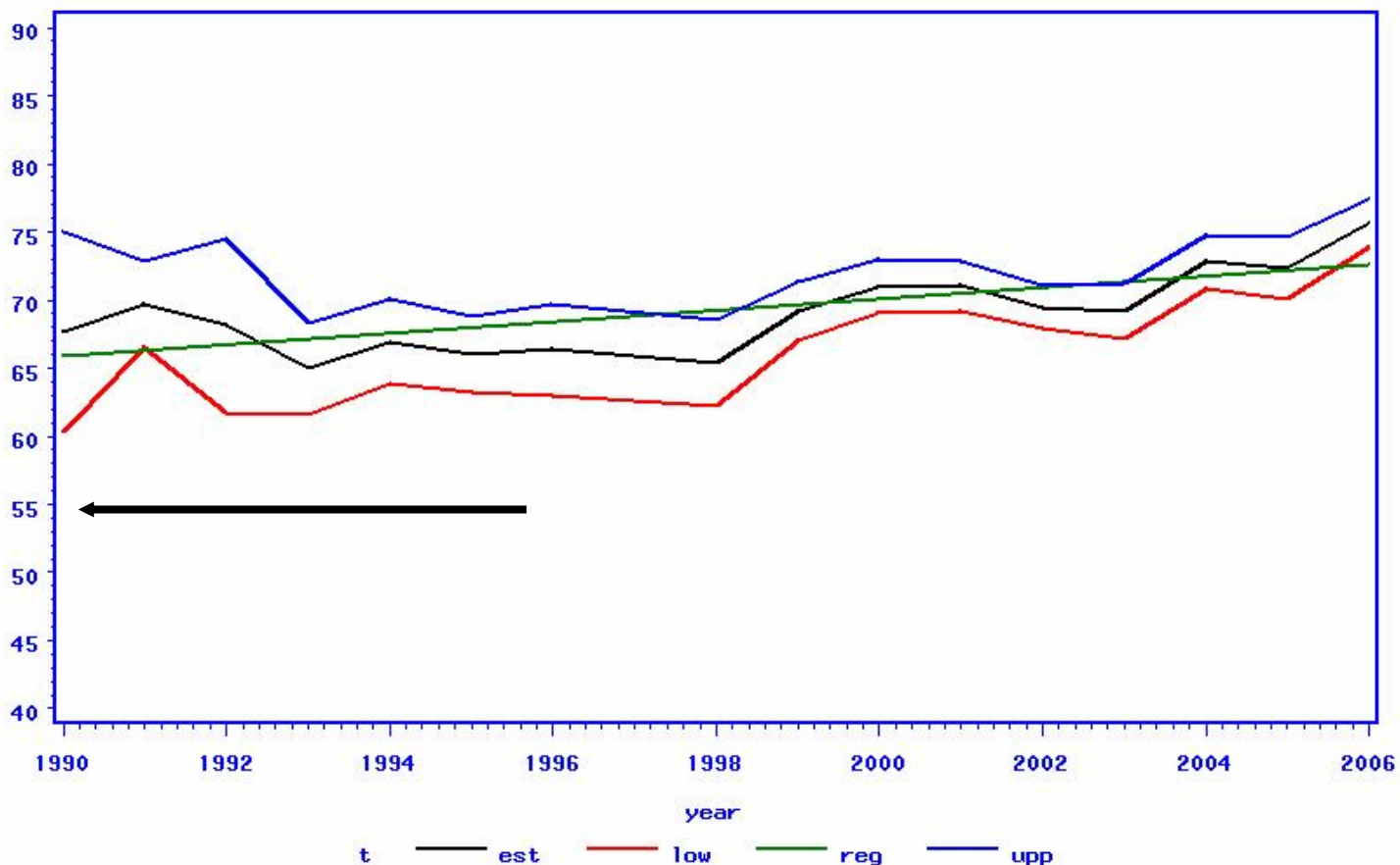
Statistics
Canada

Canada



Example: Estimates as a Time Series

Percentage of dwellown = 1 for 45+ in Quebec



note: for illustrative purposes only
source: Statistics Canada, General Social Survey





Example: Combining Small Domains

- Combine data across two or more cycles
- Advantage: more data (confidentiality concerns, variability concerns)
- Requires non-time-dependent analyses or few cycles with only mild time dependency
- Need to interpret estimates properly in the context of a target population (depends on type of estimate: descriptive versus model parameters)



Statistique
Canada

Statistics
Canada

Canada



Example: Combining Small Domains

- 15 year olds whose main activity last year was working at a paid job or business or looking for work

Cycle	15yo... etc.
13	under 10
14	under 10
15	under 10
total	over 10 yay!





Conclusion: Product

- Plan: “clean” product in RDCs by Spring, 2008
- Updates: GSS-21 and more variables
- Already have “beta” version
 - Have 896 variables
 - Some examples: age, cycle, dwellown, edu5, hsdsize, lanch, marstat, prov, recid_local, relig16, sex, wght_per
 - Team in place to create variables, procedures developed for populating the data base





Conclusion: Product

- Working on tools
 - Bootstrap weights, adding variables of interest to researchers, quality assurance checks, etc.
- Documentation
 - Data dictionary
 - User guides in readable format
 - Meta-data information
- Have started to do some analysis (Fact Sheets)





Conclusion: Contact

- Questions?
- Contact:
 - Michael Wendt
 - (613)-951-7314
 - michael.wendt@statcan.ca



Statistique
Canada

Statistics
Canada

Canada