Canada's National Household Survey: Frustration and Compromise

Thomas G. Exter, chief demographer

December 2013

Some bad press for the voluntary NHS…

- "Canada's voluntary census is worthless…here's why…"
  - Globe and Mail (October 4, 2013)

- "Canadian income data 'is garbage' without census, experts say…"
  - Globe and Mail (October 7, 2013)

PBS Data Team Objectives

- Understand the potential for non-response bias
- Develop a methodology for making reasonable DA-level estimates with NHS data
- Produce a set of estimates for:
  - Base Year (2011)
  - Current Year (2014)
- For selected high-value NHS variables

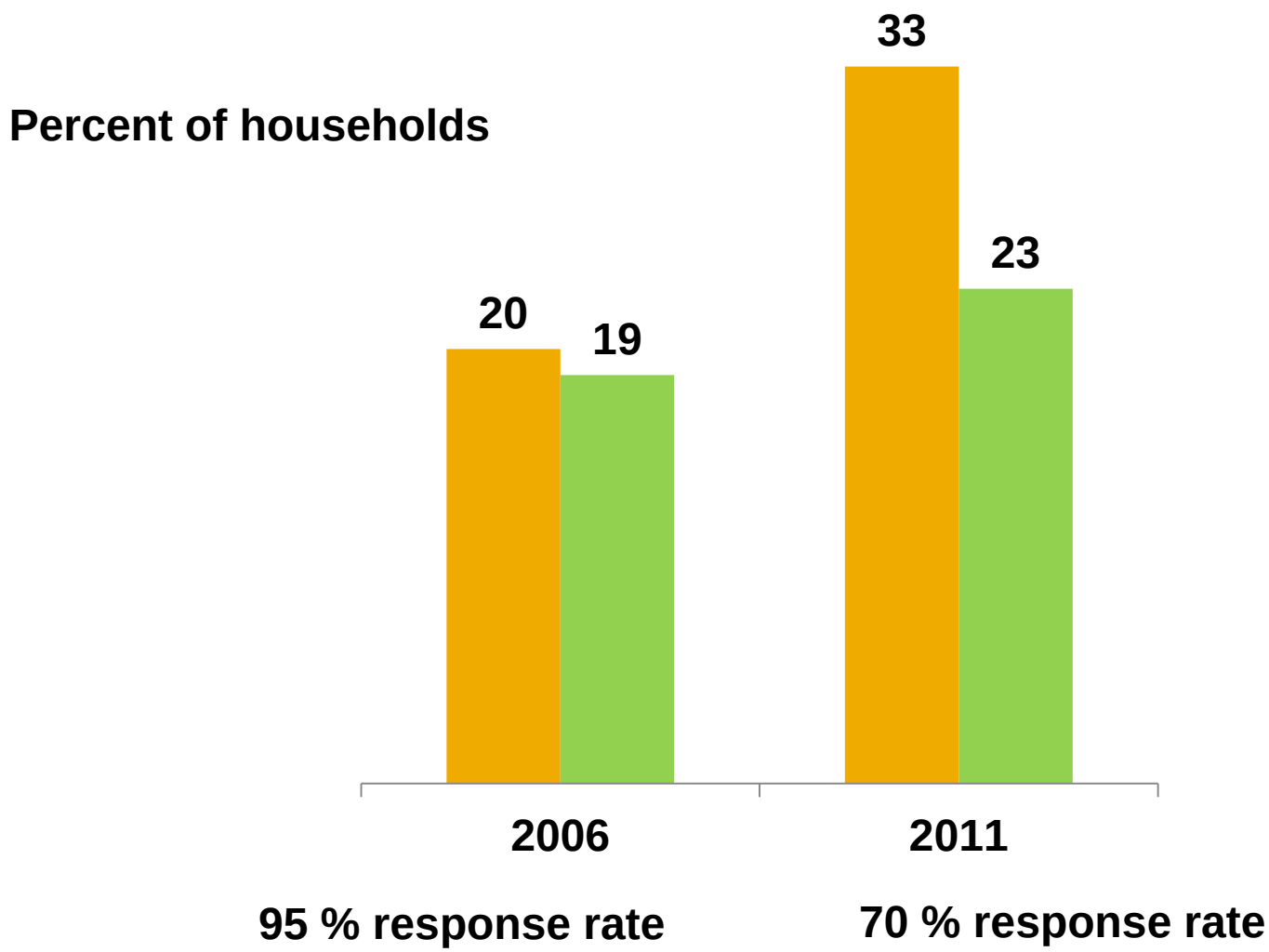
Assumptions

- Global non-response rates (GNR) indicate but don't measure non-response bias.
- Increased uncertainties require mitigation.
- Mitigation through imputation and modeling can build user confidence in the data.

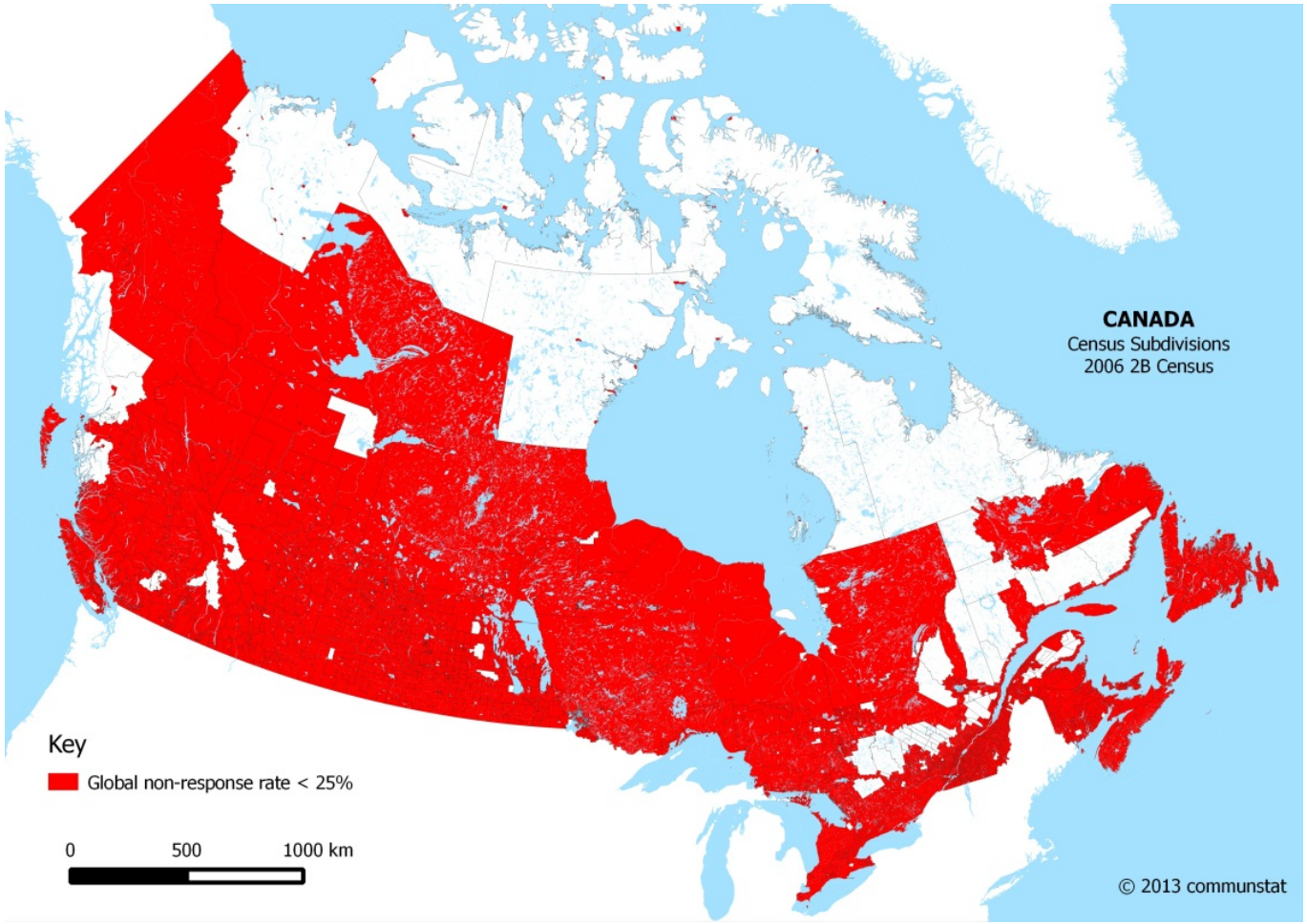
Suppression for “Quality’s Sake”

- For 2006 long-form sample data, Statistics Canada suppressed data where non-response was 25 percent or more.
- For the 2011 NHS data, the GNR suppression threshold for “quality” was 50 percent or more.
- Official criteria for “quality” suppression relaxed from 25 percent to 50 percent.

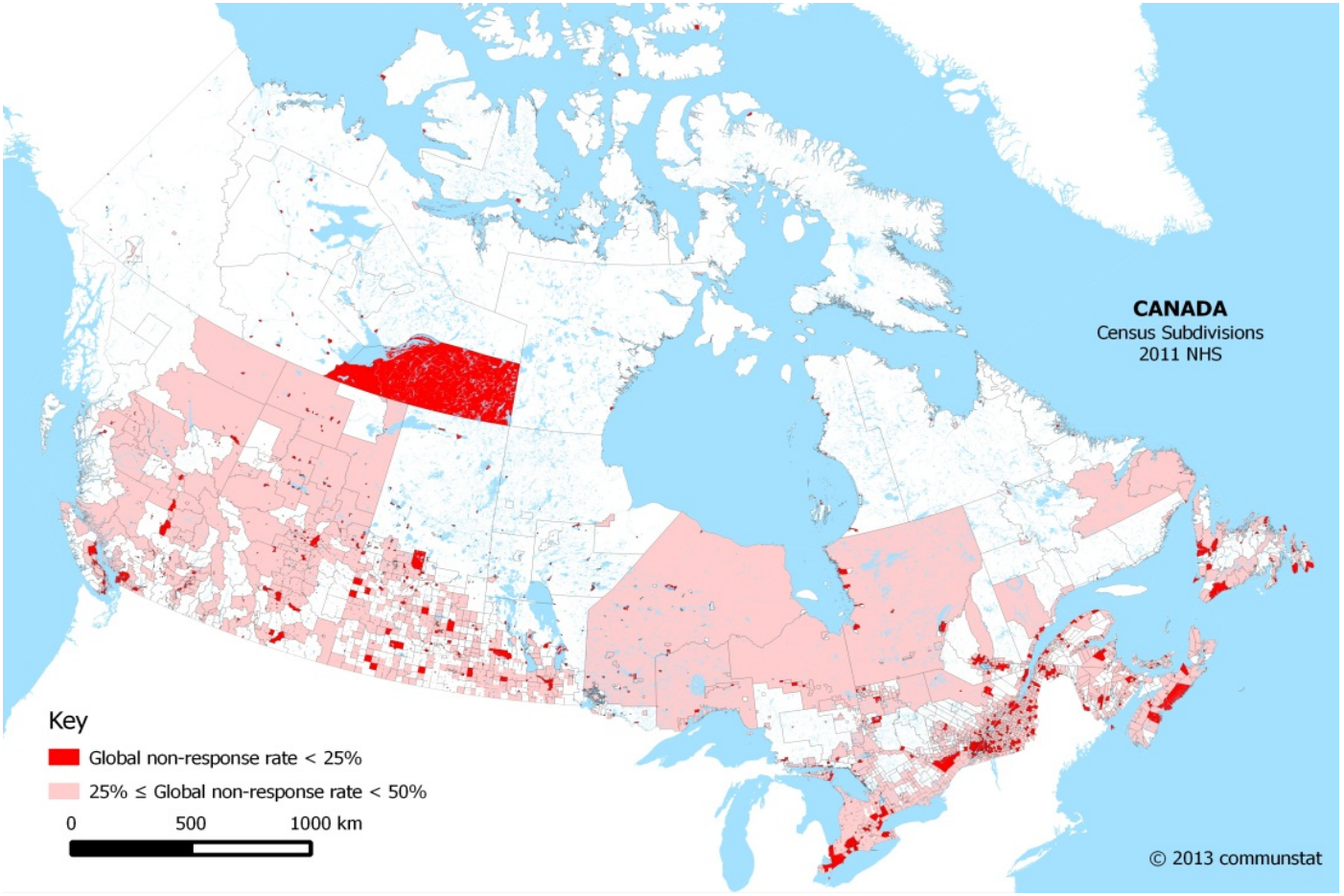
Sample frame and response: 2006 and 2011



2006 Census long-form: >75% response



2011 NHS: at least 75 percent response



Accepting the 50 percent threshold…

- For example, among CSD's (N = 5253):
  - Reasonably good data (< 50% GNR)
    - (N = 3439, 65 percent of CSD's)
  - Compromised data (>=50% GNR)
    - (N = 1814, 35 percent of CSD's)

Consolation Prize

- "Good data" areas contain nearly 90 percent of the population.
- Among Dissemination Areas, 90 percent of the population live in areas with at least a 50 percent response rate.
- Reality Check: Only 25 % of the population live in areas that meet the 2006 "quality" threshold.

Normal Suppression for confidentiality…

- Data not published: < 40 pop areas
- Income suppression:
  - <250 pop, or
  - < 40 households

The PBS Strategy

- Use conservative imputation methods where data is unreliable (GNR 50+%)…
- And, where data are suppressed for confidentiality reasons…
- No “black box” methods to adjust for non-response bias … (e.g. multiple imputation in R)

- Essence of the method:
  - 1. substitute a good mean in lieu of bad data
  - 2. derive mean from higher level geographies
  - 3. model income distributions from means

Base Year Estimates (2011)

- Process:
  1. Select most important, high-value NHS variables.
  2. And, all short-form census variables.
  3. Filter out “bad” or suppressed data and run imputation routine to substitute better data.
  4. Control to undercount-adjusted total.

Current Year Estimates (2014)

- Update all Base Year Estimates to Current Year Estimates. (constant DA distribution)

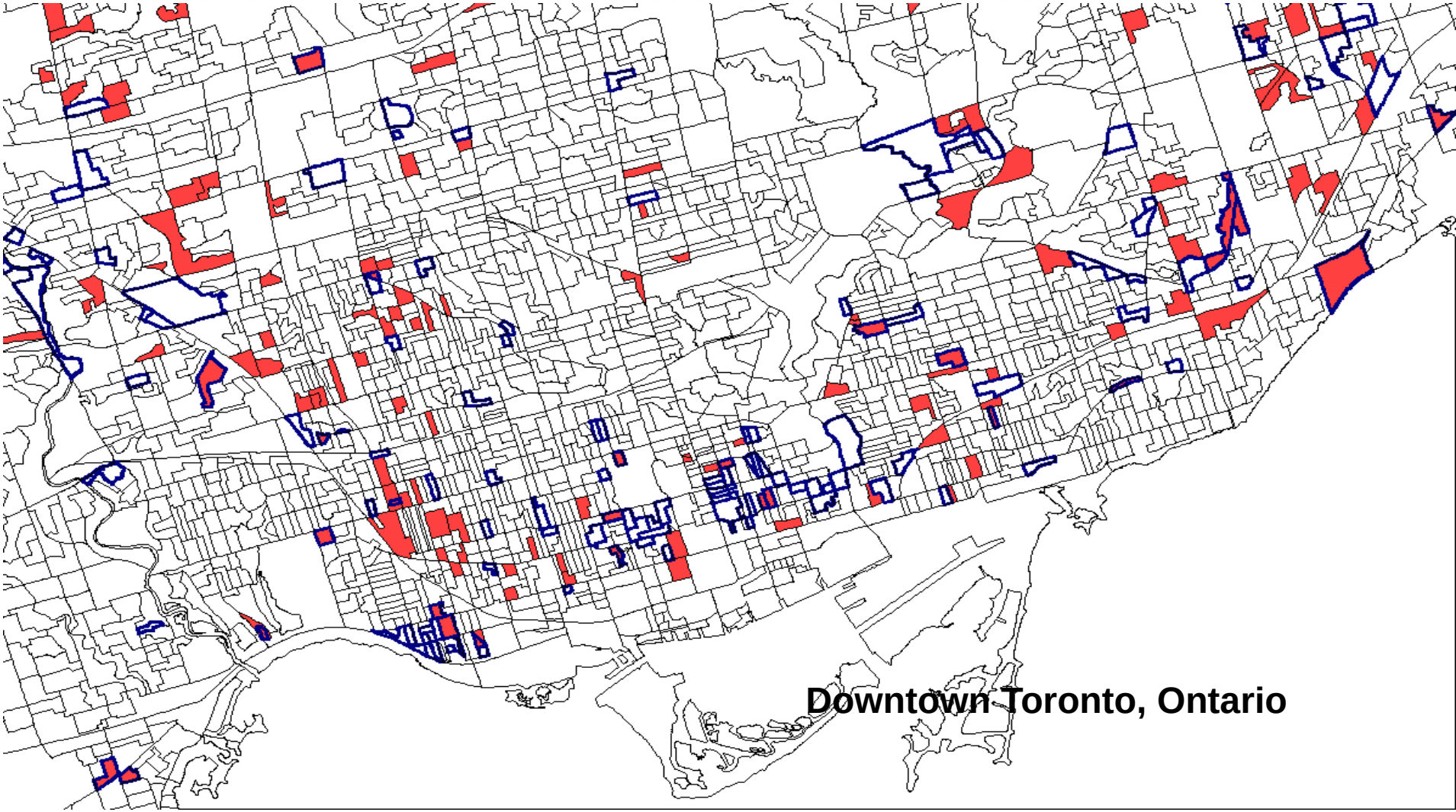
- Advantages:
  - Provides for reasonable trending from Base Year to Current Year.
  - NHS-based estimates updated to 2014 for selected variables.

household income issues …

- Issues with income:
  - More item non-response (CRA link, however)
  - More suppression, < 40 households
  - Random rounding affects sparse distribution data

- Client question:
  - Are we serving more or fewer youths from low-income areas?
  - LICO (2006) vs. LIM_AT (2011)

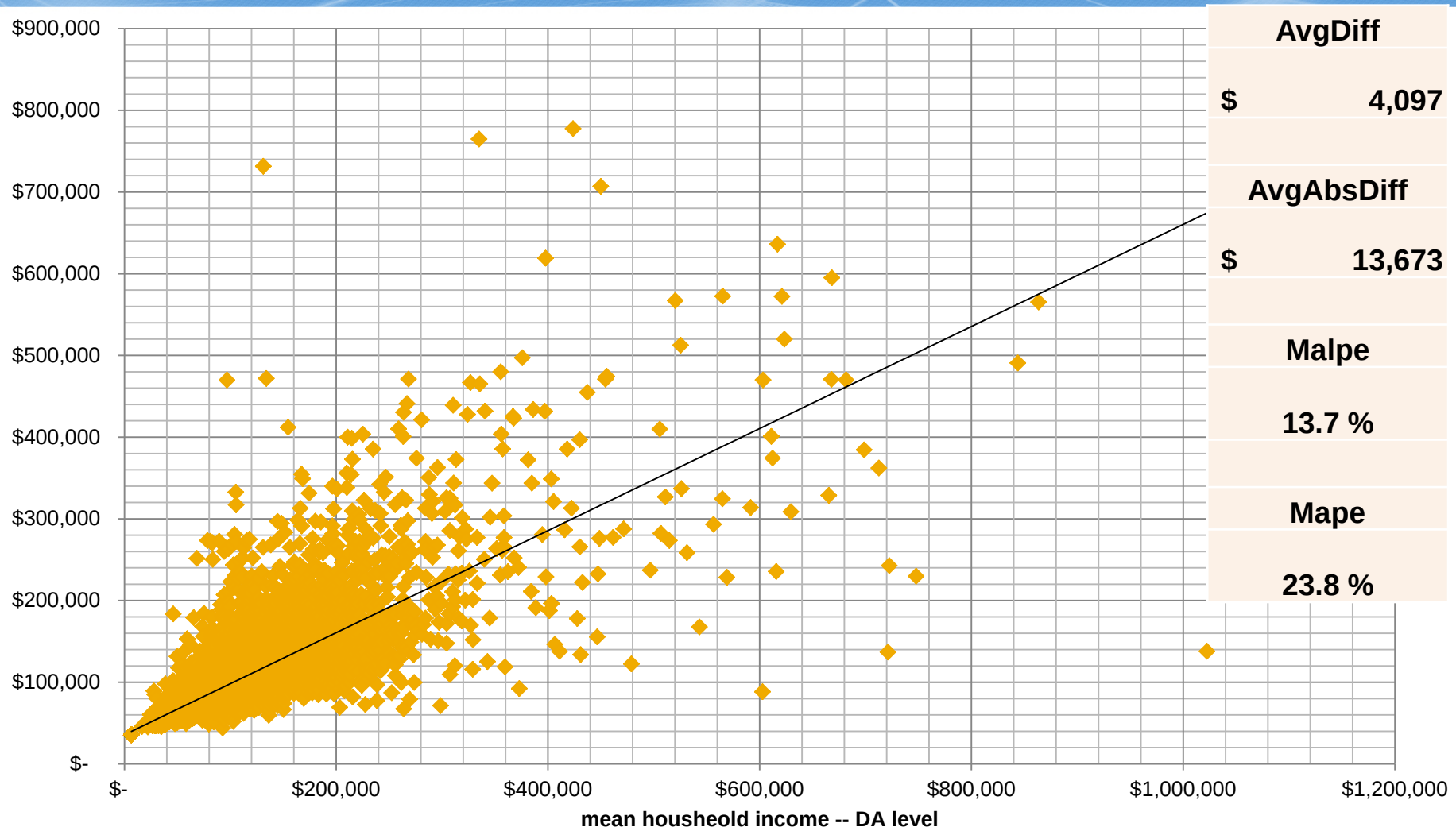
Only 37 % of low income DA's in common



Income Distribution Model, steps…

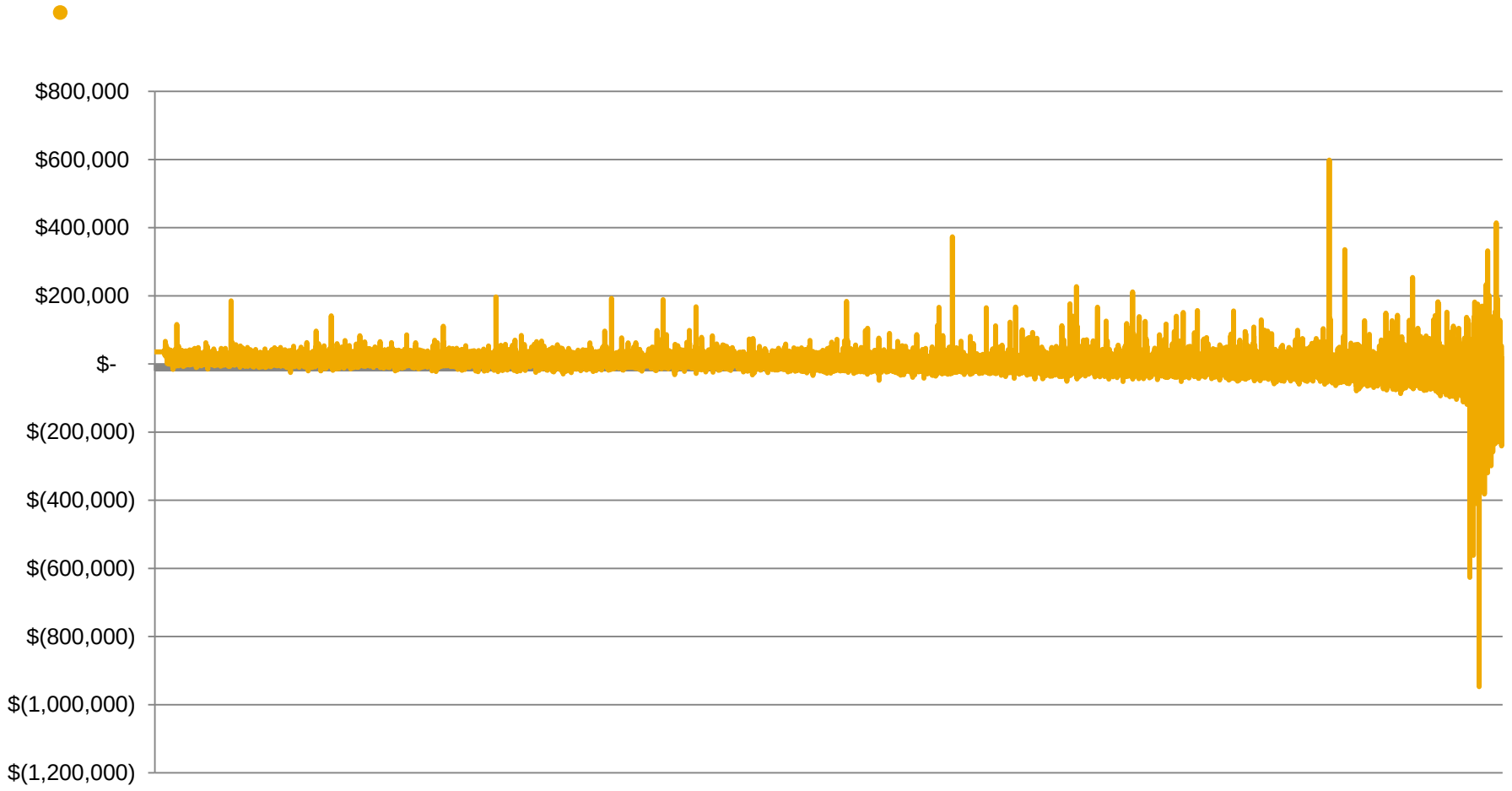
- 1. Assess mean income from NHS.
- 2. Apply imputation method for questionable or suppressed means.
- 3. Cap extreme differences versus 2006.
- 4. Compare various probability distribution functions based on mean income.
- 5. Solution: DA–level model based on derivative of Poisson distribution.
- 6. Assess result versus NHS at higher levels.

Mean Income: 2006 Census vs. 2011 NHS (Ontario DA's)

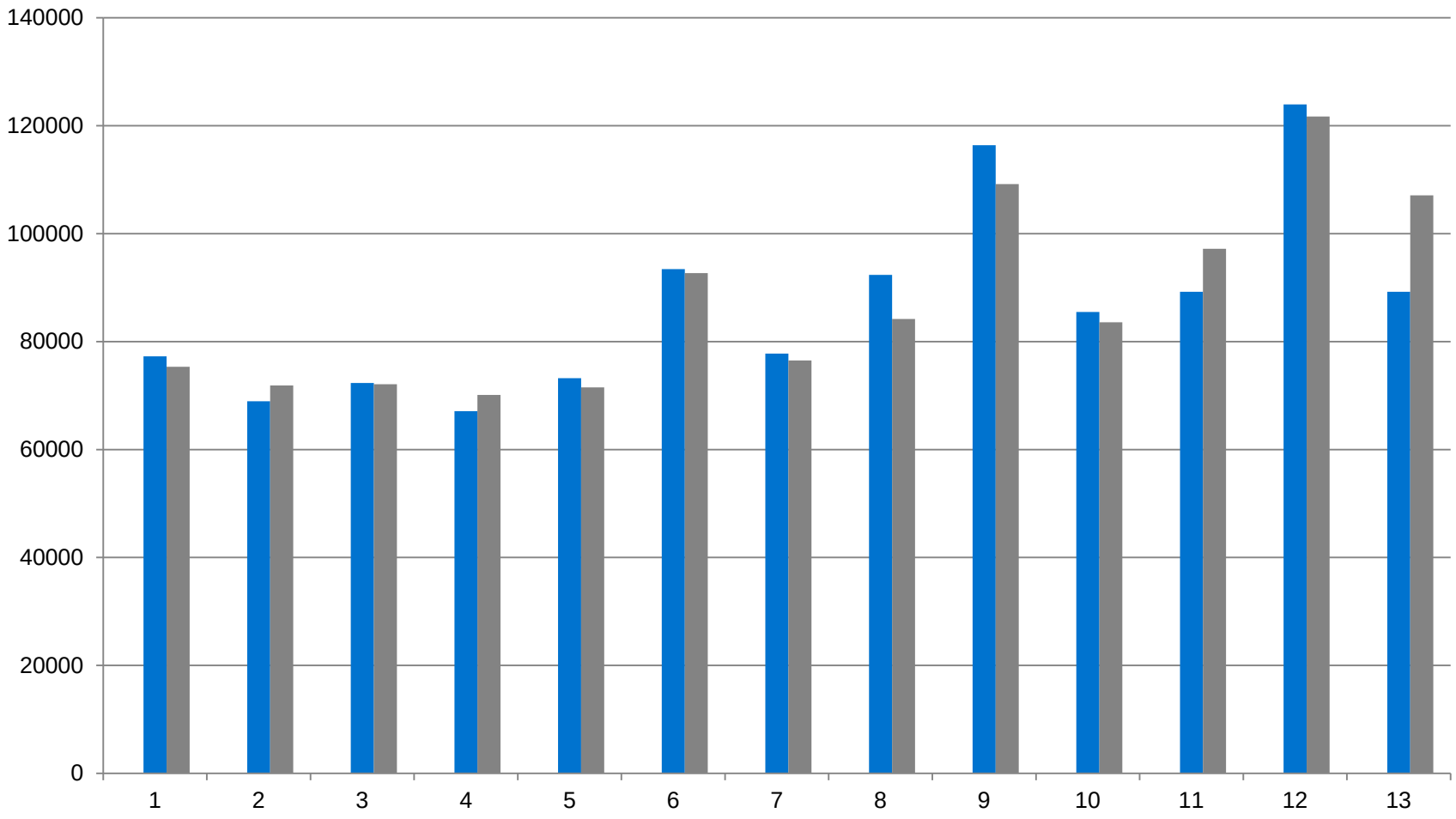


AvgDiff
$ 4,097
AvgAbsDiff
$ 13,673
Malpe
13.7 %
Mape
23.8 %

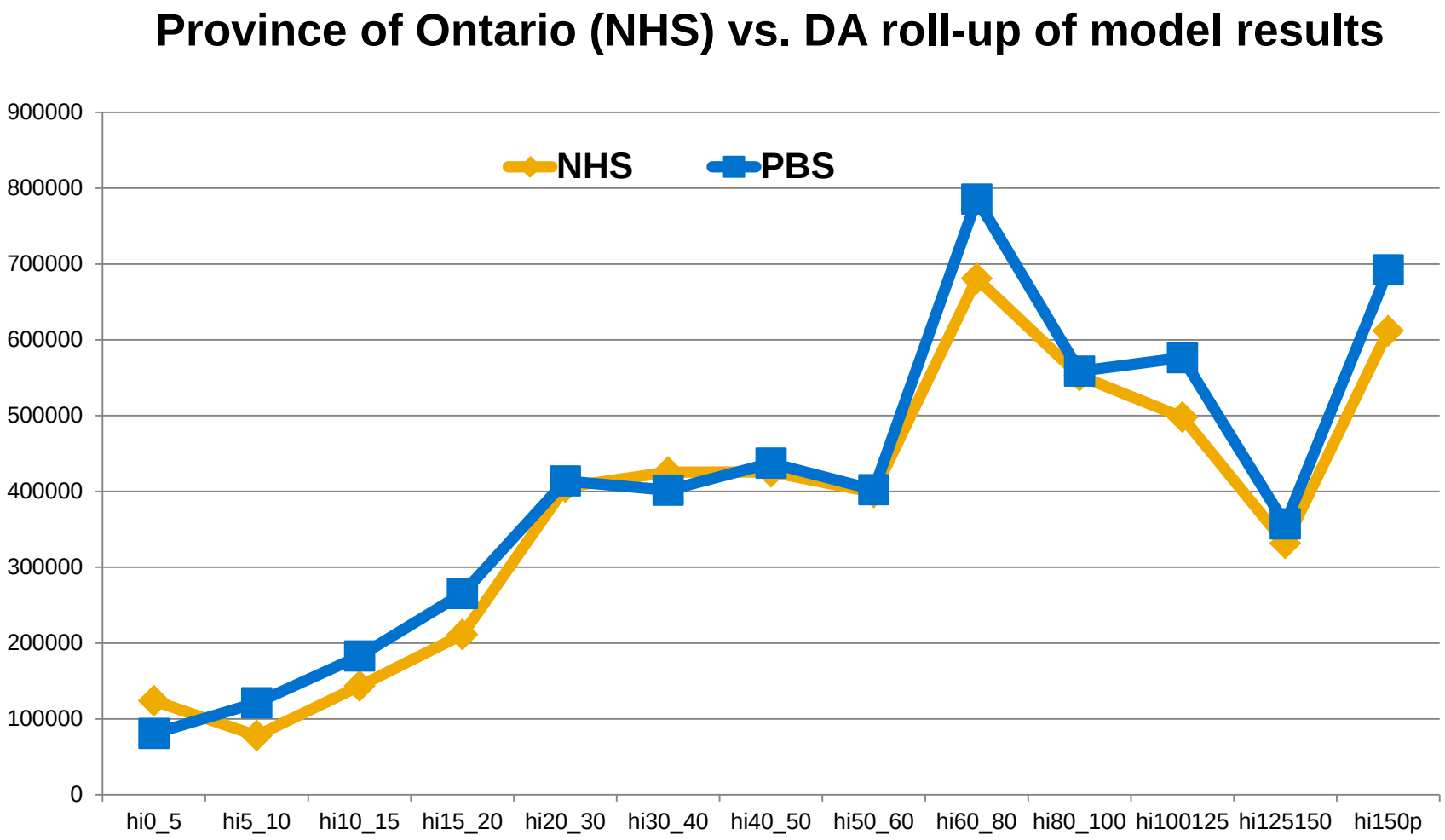
Plot of residuals ...



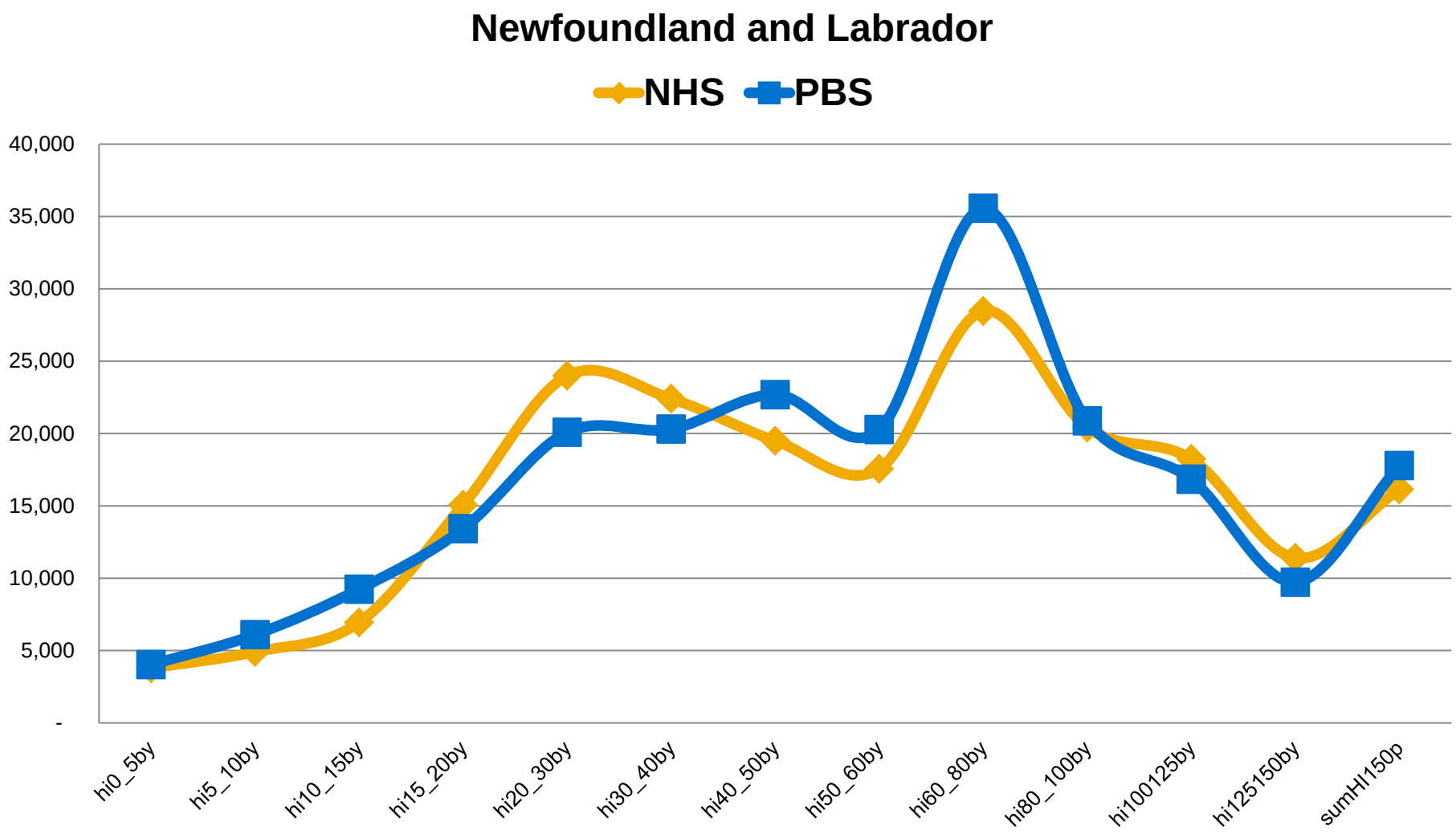
Mean Income by PR: model vs. 2013



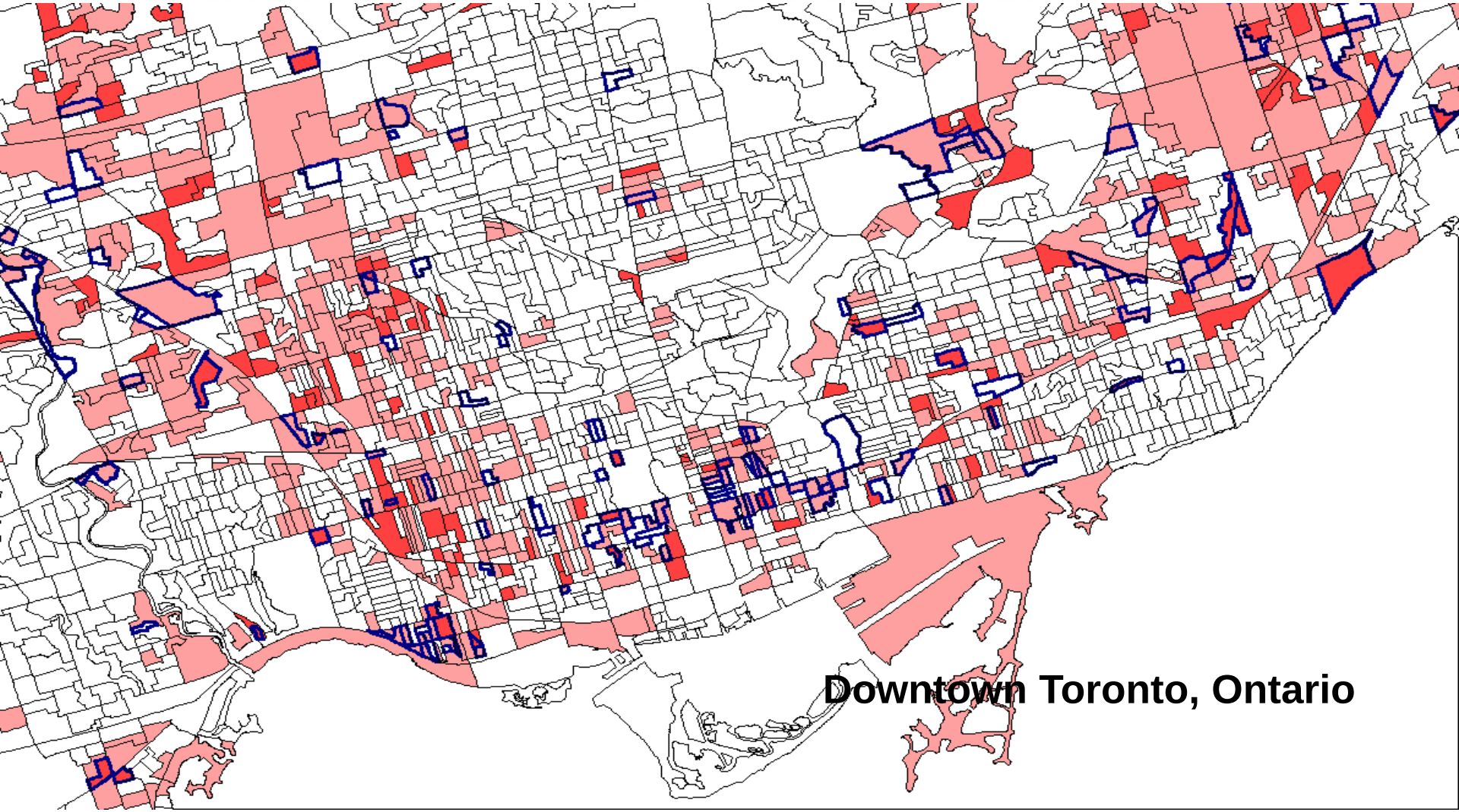
Income Distribution Model vs. NHS (2011)



Inequality Controversy



…a somewhat better match to low-income areas: 59% DA's in common at 25+% LIM_AT



Thank you!

…paper available from:

Tom.Exter@pb.com