Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

LAD
THE LONGITUDINAL ADMINISTRATIVE DATABANK

Presentation to Research Data Centre Conference

Paul Roberts

November, 2015

Outline of Presentation

1. History of and rationale for the LAD
2. Construction and contents of the LAD
3. T1FF construction
4. LAD construction
5. Accessing the LAD and confidentiality
6. Research applications

Section 1: The rationale for and the development of theT1FF and LAD – a short history

Original Mandate(s)

- Small Area and Administrative Data Division (SAAD)
  - Produce yearly migration data in order to support the Population Estimate Program of Demography division
    - Migration estimates
  - Supply personal and family income data for low levels of Geography
    - T1FF
- 2004 Strategic Streamlining Initiative
  - Cost recovery mandate

LAD History

- 1979: STC program to develop data from administrative records → T1FF
- 1989 first LAD completed (1982-1986)
  - Economic Council PSID inspired analysis - T1FF/SA files
- 1999 - LAD expands to 20% of T1FF
  - Some funding by HRSDC in the 1990's
  - 2%, 5%, 10%, 20%
- 2004 – SSI cost recoverable mandate
- 2010 – SAAD merges with ISD

Section 2:
T1FF Construction

Source of information:

T1 Family File (T1FF)

T1FF/LAD: Universe and Coverage

Target Population

- Persons who completed a T1 tax return for the year of reference or who received CCTB (Canada Child Tax Benefits)
  - Their non-filing spouses (including wage and salary information from the T4 file)
  - Their non-filing children identified from three sources (the CCTB file, the births files, and an historical file)
  - Filing children who reported the same address as their parent.

T1FF - Source of information

- Personal identifier (Social Insurance Number - SIN)
- Mailing address
- Birth date, gender and marital status
- Sources of income
- Deductions, exemptions and tax credits
- Information about family members included on individual tax forms

T1FF - Processing

- Coverage enhancement
- Creation of Census Families
- Geography components
- Imputation of Income variable

T1FF Strengths

- 100% of Canadian tax filers
- Not longitudinal, but good for cross-sectional, single year analysis
- More than 70% filing rate compared to population estimates of Canadians
- 96% coverage rate when including the dependents
- Exists since 1982

T1FF - Areas for improvement

- Definition for administrative purposes not necessarily related to concepts of interest
- Coverage of certain populations
- Mailing addresses versus place of residence
- Self reporting
- Subject to change over time due to legislative changes to tax laws

Section 3:
LAD Content and Construction

Longitudinal Administrative Databank

Coverage and content

- 20% longitudinal sample of the T1FF
- 1982 to most recent year of T1 data (2013)
- Selection of variables from the T1FF
- Information available at the individual, spouse/parent and family level
- The primary source for variable information is the LAD Data Dictionary.

Demographic Variables

- Individual Demographics
  - age, sex, marital status, language, etc.
- Family Demographics attached to each selected individual
  - type of family (Couple , Lone Parent, Person not in census family)
  - number & age of children
- Spouse or parent information
- Geography
  - province/territory, city, town
  - postal: FSA
  - census: CMA, CD, Census Tract

Income and other variables

- Employment Income
  - Wages, Salaries, Commissions, Tips
  - Self-employment
- Investment Income and other Income
  - Net rental income
  - Alimony
  - Other Pensions
  - RRSP
  - Limited Partnership
- Tax credit such as tuition fees
- Transfer Payments incl.
  - Old Age Security
  - Net Federal Supplement
  - Canada/Quebec Pension
  - Employment Insurance
  - Social Assistance
  - Workers' Compensation
  - Child Tax Benefits
- Other variables such as disability amount
- Two-digits NAICS (since 1999 – from the Business Register)

Immigration variables

- Since 2002, the LAD contains information on recent immigrants at time of landing – 1980 to 2012. Variables include:
  - Official languages ability indicator
  - Country of citizenship at landing
  - Country of last permanent residence
  - Country of birth
  - Level of education at landing
  - Landing year
  - Marital status at landing
  - Native language (or mother tongue)
  - Intended place of destination
  - Intended occupation

Tax Free Savings Account Variables

The LAD currently has three TFSA variables:

- TFSA contributions
- TFSA calendar year end
- TFSA withdrawals

Annual information for these variables exists from 2009 to 2013

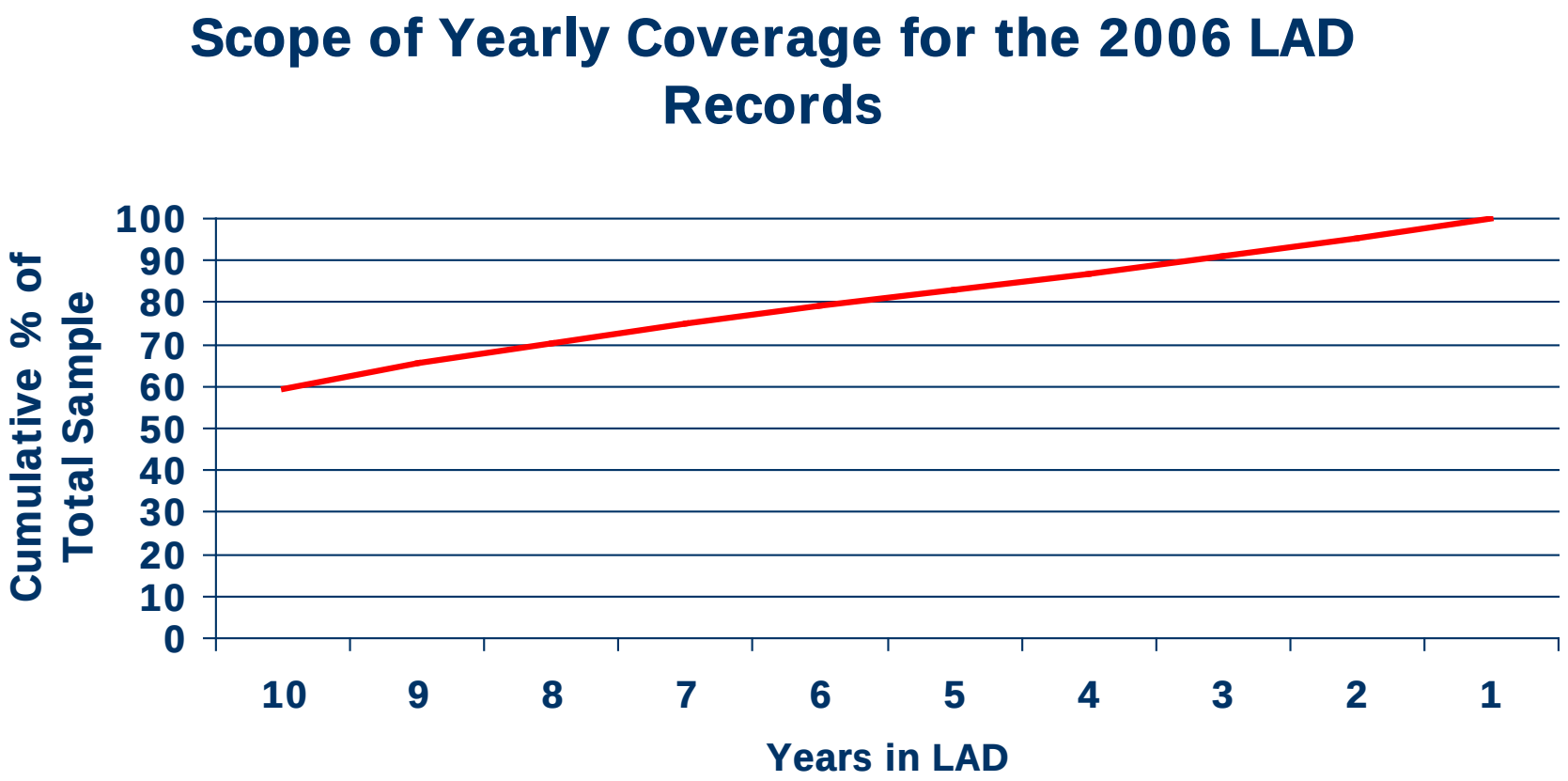
LAD Coverage

Taxfilers and Dependents by Age Group for Canada, 2005
Comparison T1FF, LAD and Population Estimates

AGE GROUP	Taxfilers 2005	Taxfilers & Dependents 2005		DEMOGRAPHY Population Estimates	COVERAGE		
	T1FF (#)	T1FF (#)	LAD (#)	2006 (PR) - July 1st (#)	T1FF (col. b/e) (%)	T1FF (col. c/e) (%)	LAD (col. d/e) (%)
column a	column b	column c	column d	column e	column f	column g	column h
Under 20	1,163,220	7,928,440	1,220,090	7,823,056	14.87	101.35	15.60
15 +	23,899,640	25,478,590	24,198,680	26,997,972	88.52	94.37	89.63
15 - 64	19,802,780	21,307,300	20,035,460	22,675,362	87.33	93.97	88.36
65 - 74	2,184,220	2,231,410	2,225,720	2,276,066	95.96	98.04	97.79
75 +	1,912,640	1,939,880	1,937,500	2,046,544	93.46	94.79	94.67
65 +	4,096,860	4,171,290	4,163,220	4,322,610	94.78	96.50	96.31
All Ages	23,951,820	31,099,150	24,271,380	32,649,482	73.36	95.25	74.34

Sources:
Statistics Canada, Special tabulation from T1 Family File
Statistics Canada, Longitudinal Administrative Databank
Statistics Canada, CANSIM, Table 051-0001

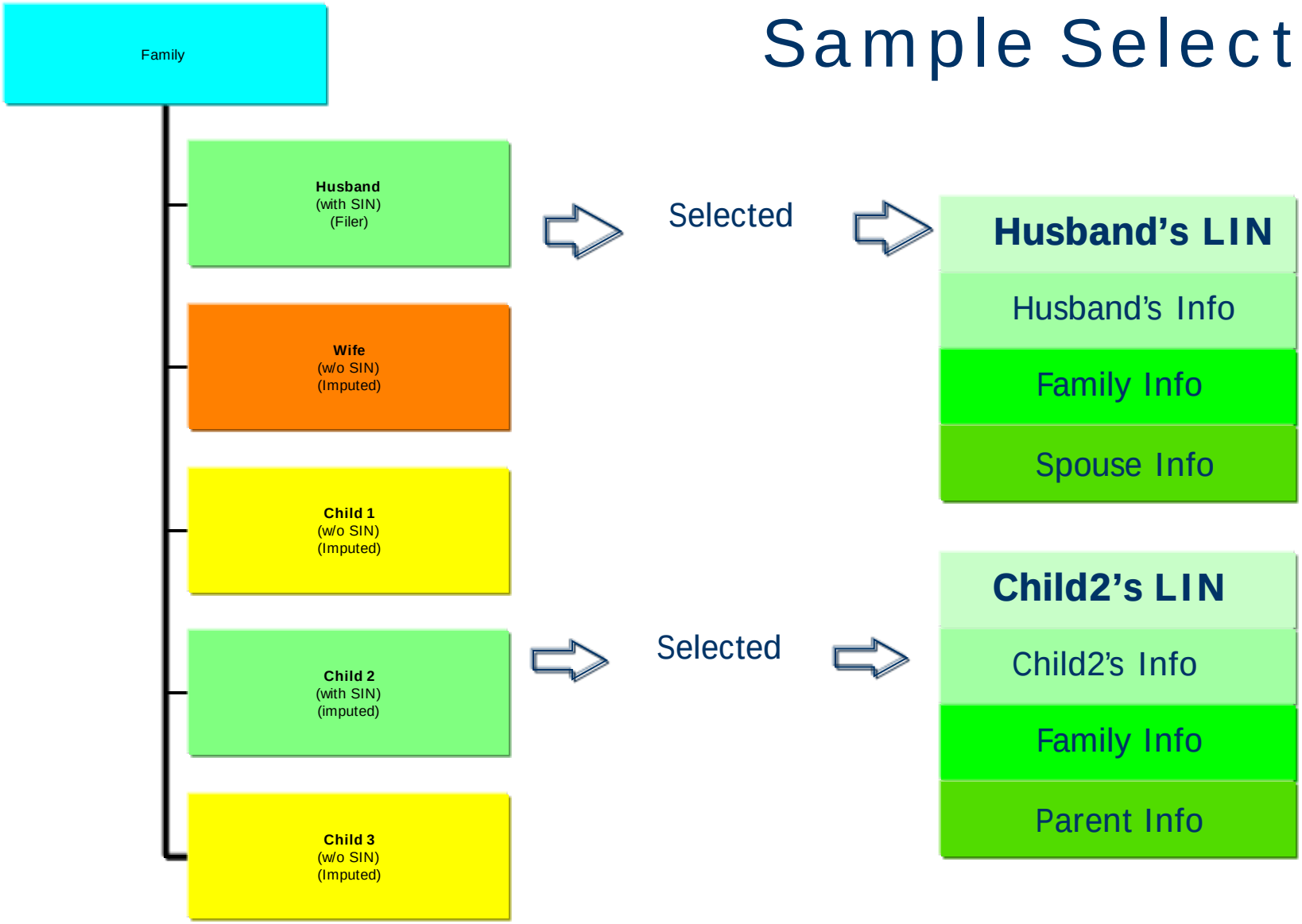
Quality indicator: Persistency



Sampling Scheme

- Sampling frame: T1FF (T1 Family File)
- SIN is used as the seed for sampling
- Constant unique SIN via SIN cross-referencing
- Sampling fraction: 20%
- No imputation for missing records

Sample Selection



Sample Dynamics

- Entry
  - Young adults
  - Immigrants
- Exit
  - Deceased
  - Emigrants
- Missing / Sporadic presence
  - Late filers
  - Non-filers
  - Non-residents

Techniques to deal with inconsistencies across years data

- SIN cross-referencing.

- Register file: keep the first year of valid data and use as constant for years following for Gender, Year of Birth, Year of Death, Year of Landing variables.

Section 4:
LAD Access and Confidentiality

LAD Data Access & Sharing

- Very controlled access
  - Access to micro-data limited to only Statcan or RDC
  - No release of micro-data (no pumfs)
  - Confidential aggregate data must remain on-site
  - Secure physical environment
- Research assistant service on a cost recovery basis
- Record linkage requests
- Custom cost-recovery research assistance for clients

LAD in RDCs

- Up until the last two years, access to the LAD has been quite restricted for researchers outside of Statistics Canada.
- Consultations were held between CRA and Statistics Canada, in 2012, to investigate the potential of providing access to the LAD via the Research Data Centre network.
- In 2013 and 2014, a successful pilot project was completed that allowed Federal government researchers to access the LAD from within the FRDC located within Statistics Canada.
- Based on this, the LAD is now being gradually rolled-out to the larger RDC network

Disclosure Control Techniques

- Rules to prevent disclosure
  - Addition of noise
  - Suppression of cells with low counts
  - Dominance tests
  - Residual disclosure avoidance
  - Rounding

Section 5: LAD Research

Returning to LAD's Research Roots

- Broaden the base of experienced LAD research analysts
- Strengthen the links between subject matter expertise and internal/external LAD research projects
- Central development of the base LAD in ISD

Research Projects

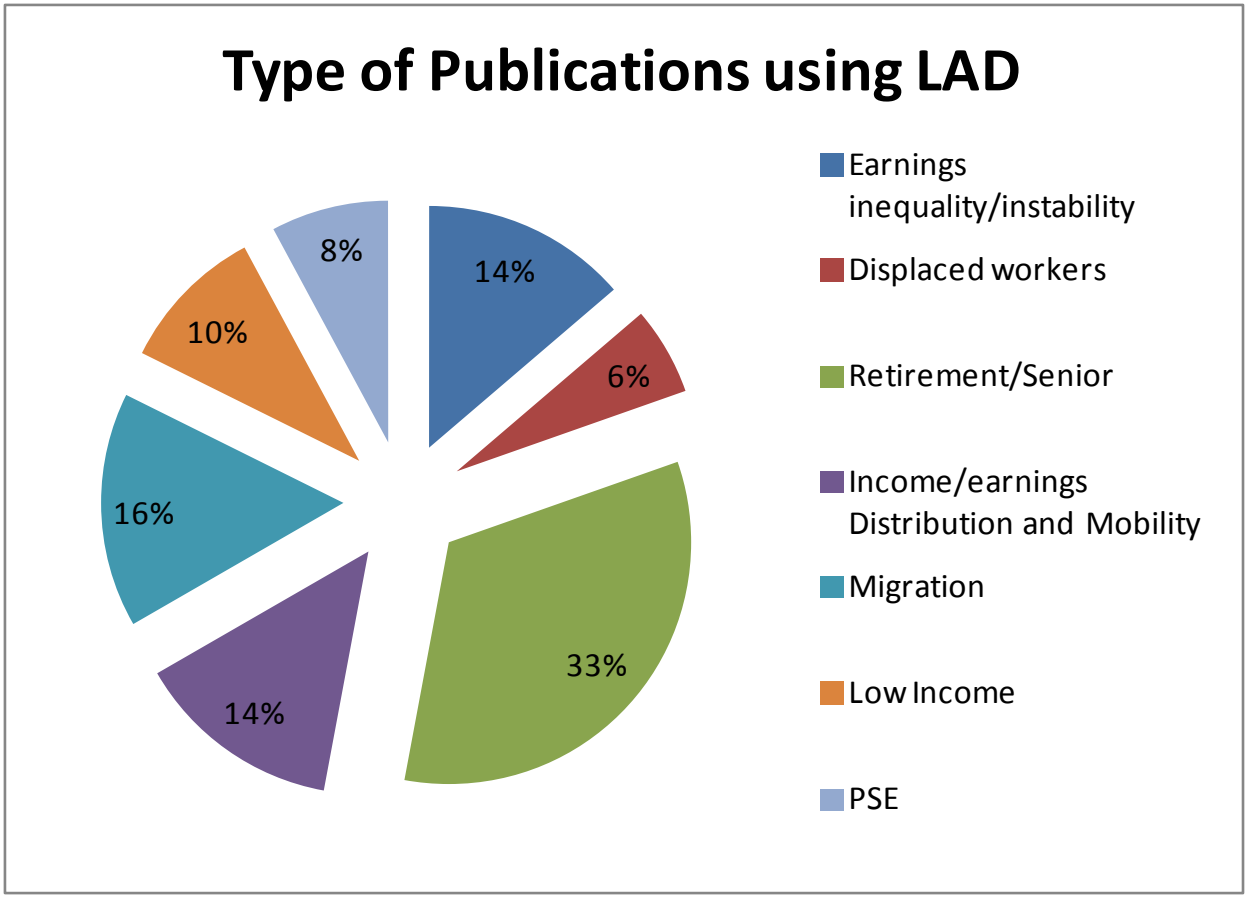
T1FF (Cross-Sectional)

- Refined geography (Census or Postal)
- Refined geographic areas (Client definitions)
- Multiple economic variables

LAD (Longitudinal)

- Record linkage
- Lifetime dynamics
- Event impacts

Labour & Income Related Publications with the LAD



LAD Development

- LAD Development
  - Improving documentation & guides
  - Production of CANSIM tables (High income)
  - Data quality research (fitness for use)
- Infrastructure
  - Expansion to RDC network (SAS and STATA data)
  - Stable pool of knowledgeable researchers and analysts

Conclusion

- The LAD is a good tool for studying many longitudinal socio-economic dynamics
- There are currently 31 years of reliable, comprehensive income data
- The LAD, as a 20% sample of the T1FF, can describe very small regions
- Having customizable geography and many economic variables makes the LAD an extremely versatile research tool.
- The result is a very useful research databank gradually being rolled-out to the RDCs