Simple implementation of scanner data in the Canadian CPI

UN GWG on Big Data for Official Statistics

Workshop on Scanner Data and Official Statistics

Kigali, Rwanda, 29 April – 1 May 2019

Delivering insight through data for a better Canada

Outline

- Context
- Overview of scanner data acquisition
- Adopting an implementation approach
- Incremental implementation strategy
- Simple implementation
- Scanner data challenges

Delivering insight through data for a better Canada

Canada

Context

Statistics Canada has undertaken a modernization initiative that aims to:

- Ensure more timely and responsive statistics
- Develop and release more granular statistics
- Ensure cost effective products and service delivery

Overview of scanner data acquisition

- CPI price collection relies on ~ 90 000 prices collected each month

- Need to modernize CPI program:
  - New data sources and IT developments
  - Growth in third-party data providers

- 2015: Scanner data received from one major Canadian grocery retailer
  - Weekly files: very large size (storage and # data points)

- 2017 & 2018: Two additional Canadian grocery retailers
- Advantages of scanner data for CPI
  - Reduce response burden
  - Improve CPI input data quality (prices and weights)
  - Better quality CPI data
  - Potential use of the universe of all consumer products

Adopting an implementation approach

- Key considerations

  - Timeliness of data acquisition for CPI processing
  - Continuity of data acquisition in the long term
  - New data sources required appropriate methodologies
  - Infrastructure availability
  - Integration into existing systems
  - Low risk tolerance for the CPI (non-revisable)
  - Need to develop the required skill set

Statistics Canada / Statistique Canada

Delivering insight through data for a better Canada

Canada

Incremental implementation strategy

- Long term: Full implementation
    - A micro-data processing system to use with large volumes and varieties of administrative data
    - Full use of all scanner data
    - Multilateral index methods

- Short term: Simple implementation
    - One-for-one approach
    - Replace prices collected in the field with scanner data, in a "1-for-1" way
    - One retailer (up to now; 2 more retailers planned for July 2019)
    - Food and non-food products sold in grocery stores
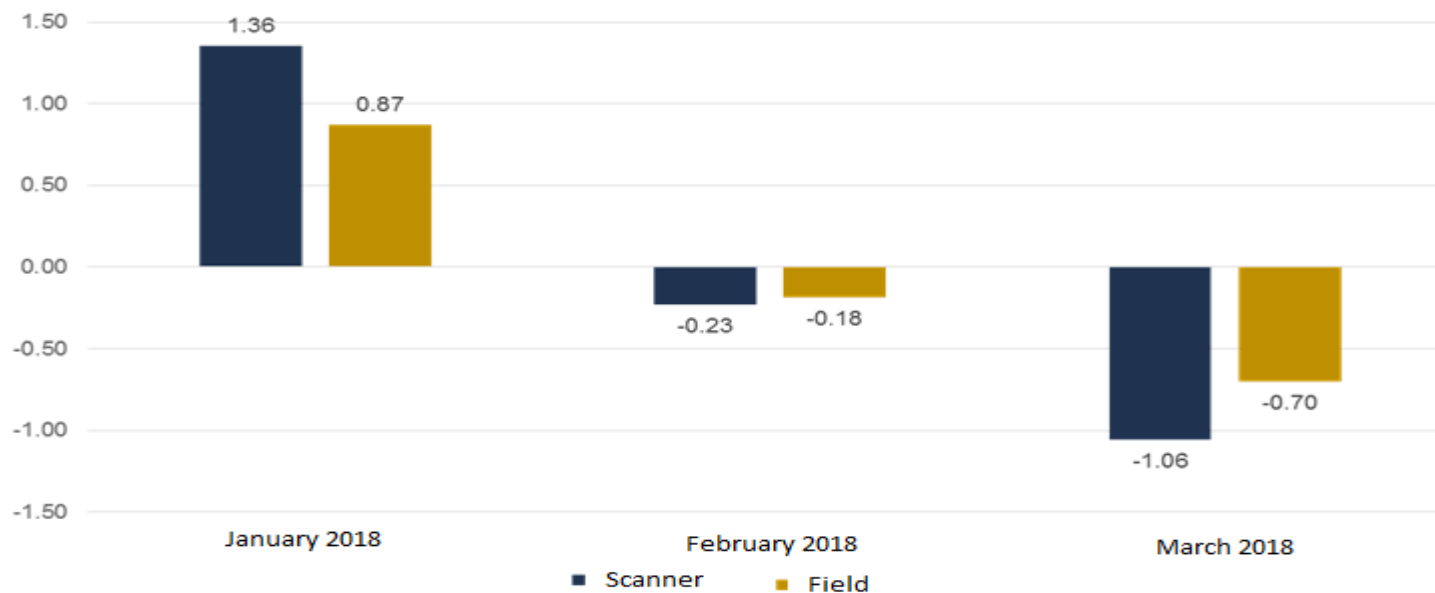
Simple implementation: Motivation

- Why a "1-for-1" use of scanner data?
  - Easy to control scope of the change
  - Manual validation of final product selection
  - Integrate into existing systems and processes with as little change as possible
  - Demonstrate to the retailer that we make official use of the data set

- Benefits
  - **Cost Savings:** Ability to decrease field collection earlier
  - **Reduced response burden:** No collection burden on this retailer's outlets
  - **Accuracy:** Two weeks of data being used instead of a single day price
  - **Representativeness:** Ability to use quantity data to initiate the product sample as well as for substitution product selection

Statistics Canada / Statistique Canada

Delivering insight through data for a better Canada

Canada

Simple implementation: How it works

- ## How does the "1-for-1" approach work?
  - The existing outlet sample from the retailer is used
  - Price data are calculated from the scanner database
    - No field collection from the retailer's outlets
    - Transactions prices as opposed to list prices
    - Unit value prices calculated for sample products at the outlet level
  - Sample of products re-initiated using the universe of products and sales information
  - Represents 20% of the total CPI product sample (~ 18 000 products out of 90 000)

- ## Implementation timeline
  - Parallel run (experimental): Food products, January-March 2018
  - April 2018: Food scanner data from 1 retailer
  - August 2018: Non-food scanner data from 1 retailer
  - July 2019: 2 more retailers implemented, food and non-food

Simple implementation: How it works

- Monthly movements based on scanner vs field collection data differ
  - Methodogical differences: price definition, products not exactly the same

Simple implementation: Product coverage

- Product coverage

CPI major component	Proportion of scanner data RPs
1. Food and non-alcoholic beverages	98.92
3. Household operations, furnishings and equipment	73.33
6. Health and personal care	33.33
8. Alcoholic beverages, tobacco products and narcotics	13.48

- Types of products not covered by scanner data

CPI major component	Types of products not in scanner data
1. Food and non-alcoholic beverages	Restaurant and fast food meals
3. Household operations, furnishings and equipment	Veterinary services, flowers, horticultural services, electric light bulbs
6. Health and personal care	Prescription drugs, eye care products, dental care products
8. Alcoholic beverages, tobacco products and narcotics	various types of beer, wine, whisky, liqueurs and dry gin

Statistics Canada / Statistique Canada

Canada

Simple implementation: Processing system

- Data processing system

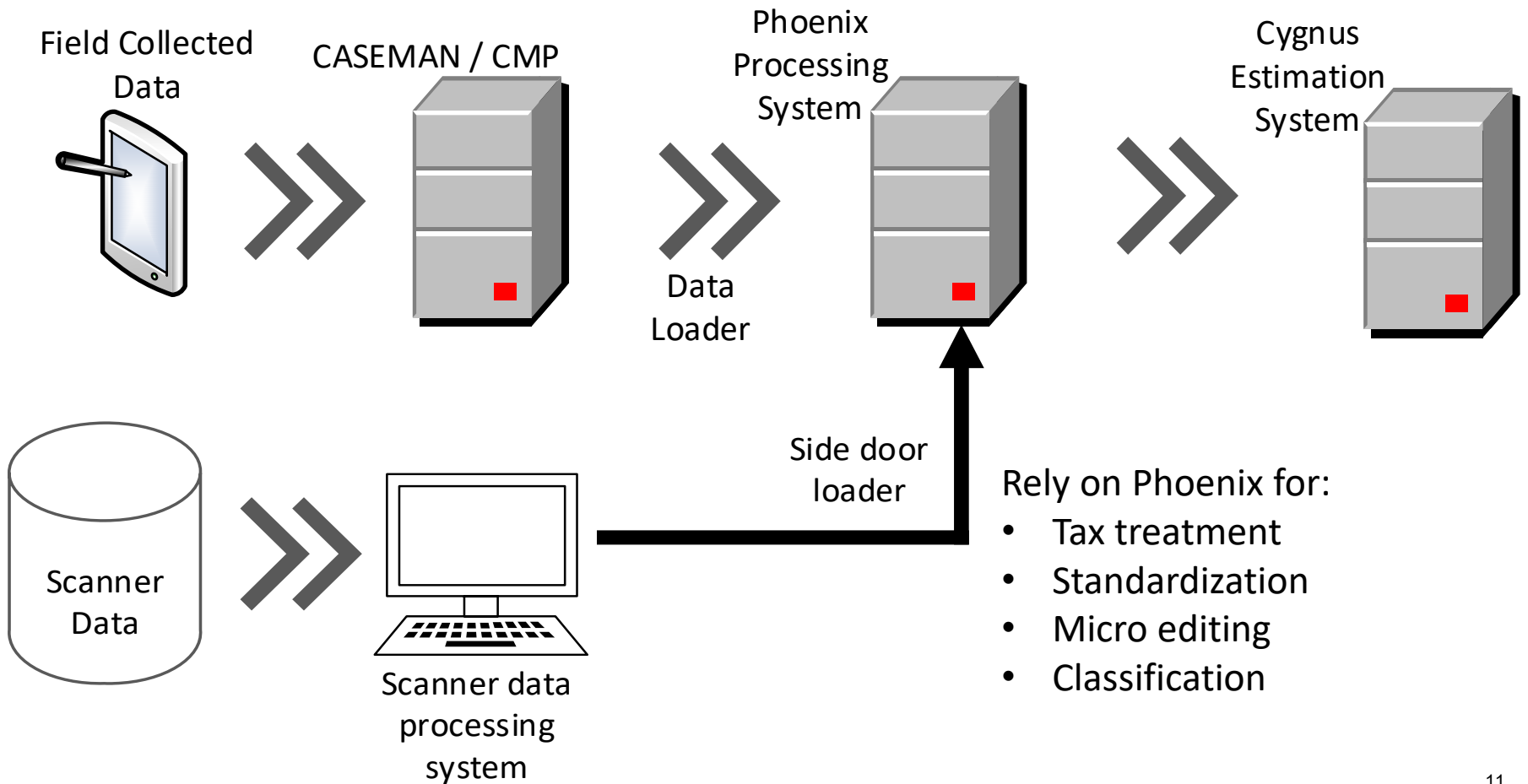| Load	Process	Substitution module	Prepare output

- Read in raw weekly scanner data files

- Aggregate weeks
- Clean variables
- Manage sample

- System proposes candidates for substitution
- User resolves missing UPCs

- Load final clean scanner data into micro-data processing system

Simple implementation: Processing system

- Data processing system

Field Collected Data

CASEMAN / CMP

Phoenix Processing System

Cygnus Estimation System

Data Loader

Scanner Data

Scanner data processing system

Side door loader

Rely on Phoenix for:
- Tax treatment
- Standardization
- Micro editing
- Classification

Scanner data challenges

- ## Acquisition
  - Data is sensitive and security is important

- ## "Big" nature of data
  - More significant IT storage and processing requirements

- ## Timeliness
  - How well does the availability of the data source correspond with the monthly production calendar?

- ## Methodology
  - Integration of scanner data with data collected in stores
  - Highly specialized skills required, e.g. data scientist
  - Quality assurance needs appropriate tools
  - With multilateral indexes, analysis of index results not standard

Questions?

Thank you!