Traditional price index methods in the context of scanner data

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

- Definition and uses of a consumer price index
- Review of traditional price index formula
- Construction of a national consumer price index
- Preprocessing of scanner data for CPI calculation
- Scanner data and traditional price index formula
- Multilateral price index methods: general principle

Consumer price index: definition and uses

- What is the Consumer Price Index (CPI):
  - An indicator of the changes in consumer prices that are experienced by a target population. It measures *average price changes* by comparing, *through time*, the cost of a *fixed basket* of goods and services.
  - The goods and services in the basket must be of unchanging or equivalent quantities and qualities
  - Reflects a *pure price change,* NOT a Cost-Of-Living Index
  - Price movements of CPI product categories are **weighted** according to their relative importance in the total expenditures of consumers

Consumer price index: definition and uses

- Uses of the CPI:
  - Central banks: as a measure of inflation, for monetary policy
  - Benefits recipients, workers and unions: to index, escalate or adjust nominal values
  - Governments: budget projection, deflation of nominal values to obtain constant dollar figures
  - Other uses: financial markets and traders
- Not always easy to accommodate all specific requirements of all uses

Review of traditional price index formula

- Index formula for elementary price indices

$$I_{Carli}^{0:t} = \frac{1}{n}\sum_{i}^{n}\left(\frac{p_i^t}{p_i^0}\right)$$

$$I_{Dutot}^{0:t} = \frac{\frac{1}{n}\sum_i^n p_i^t}{\frac{1}{n}\sum_i^n p_i^0}$$

$$I_{Jevons}^{0:t} = \prod_{1}^{n}\left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{n}}$$

	t=0	t=1	t=2	t=3	t=4	t=5	t=6
	Prices (standardized)						
Candy 1	$ 6.00	$ 6.00	$ 7.00	$ 6.00	$ 6.00	$ 6.00	$ 6.60
Candy 2	$ 7.00	$ 7.00	$ 6.00	$ 7.00	$ 7.00	$ 7.20	$ 7.70
Candy 3	$ 2.00	$ 3.00	$ 4.00	$ 5.00	$ 2.00	$ 3.00	$ 2.20
Candy 4	$ 5.00	$ 5.00	$ 5.00	$ 4.00	$ 5.00	$ 5.00	$ 5.50
Arithmetic Mean prices	$ 5.00	$ 5.25	$ 5.50	$ 5.50	$ 5.00	$ 5.30	$ 5.50
Geometric Mean Prices	$ 4.53	$ 5.01	$ 5.38	$ 5.38	$ 4.53	$ 5.05	$ 4.98
Carli Month-to-Month index	100.0	112.5	108.9	101.8	91.2	113.2	100.0
Carli Chained month-to-month index	100.0	112.5	122.5	124.8	113.9	128.9	129.0
Carli Direct index on t=0	100.0	112.5	125.6	132.5	100.0	113.2	110.0
Dutot Month-to-Month index	100.0	105.0	104.8	100.0	90.9	106.0	103.8
Dutot Chained month-to-month index	100.0	105.0	110.0	110.0	100.0	106.0	110.0
Dutot Direct index on t=0	100.0	105.0	110.0	110.0	100.0	106.0	110.0
Jevons Month-to-Month index	100.0	110.7	107.5	100.0	84.1	111.4	98.7
Jevons Chained month-to-month index	100.0	110.7	118.9	118.9	100.0	111.4	110.0
Jevons Direct index on t=0	100.0	110.7	118.9	118.9	100.0	111.4	110.0

Review of traditional price index formula

- Index formula for elementary price indices

- **Are chained and direct indexes equal?**
  - Carli index: Chained **≠** Direct
  - Dutot and Jevons index: Chained **=** Direct
  - So avoid chained Carli index!
- **When prices go back to base level (t=4 and t=0), index level should go back to 100**
  - Chained Carli index fails this property!
- **By default, G20 countries – except Japan, use, generally a Chained Jevons index rather than a Chained Dutot**
  - Dutot only works well for very homogeneous products, which means for EAs that are very narrowly defined and products that have the same unit of measure

Review of traditional price index formula

- Index formula for aggregate price indices

Index name	Laspeyres	Paasche	Fischer	Törnqvist	Lowe
Formula	$I_{L,A}^{0:t}=\frac{\sum_{i=1}^{n}p_i^t q_i^0}{\sum_{i=1}^{n}p_i^0 q_i^0}$ $I_{L,A}^{0:t}=\sum_{i=1}^{n}s_i^0\left(\frac{p_i^t}{p_i^0}\right)$ $s_i^0\equiv\frac{p_i^0 q_i^0}{\sum_{i=1}^{n}p_i^0 q_i^0}$	$I_{P,A}^{0:t}=\frac{\sum_{i=1}^{n}p_i^t q_i^t}{\sum_{i=1}^{n}p_i^0 q_i^t}$ $I_{P,A}^{0:t}=\frac{1}{\sum_{i=1}^{n}s_i^t\left(\frac{p_i^t}{p_i^0}\right)^{-1}}$ $s_i^t\equiv\frac{p_i^t q_i^t}{\sum_{i=1}^{n}p_i^t q_i^t}$	$I_{F,A}^{0:t}=\left(I_{L,A}^{0:t}\times I_{P,A}^{0:t}\right)^{\frac{1}{2}}$	$I_{T,A}^{0:t}=\prod_{i=1}^{n}\left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{2}\left(s_i^0+s_i^t\right)}$	$I_{Lo,A}^{0:t}=\frac{\sum_{i=1}^{n}p_i^t q_i^b}{\sum_{i=1}^{n}p_i^0 q_i^b}$ $I_{Lo,A}^{0:t}=\sum_{i=1}^{n}s_i^{0b}\left(\frac{p_i^t}{p_i^0}\right)$ $s_i^{0b}\equiv\frac{p_i^0 q_i^b}{\sum_{i=1}^{n}p_i^0 q_i^b}$

Review of traditional price index formula

Item		Price ($)	Quantity	Expenditure ($)	Expenditure shares	Price relatives
Period 0						
White fresh bread	loaves	2.90	2 000	5 800	0.3932	1.0000
Apples	kg	5.50	500	2 750	0.1864	1.0000
Beer	litres	8.00	200	1 600	0.1085	1.0000
LCD TV	units	1 200.00	2	2 400	0.1627	1.0000
Jeans	units	55.00	40	2 200	0.1492	1.0000
Total				**14 750**	**1.0000**	
Period t						
White fresh bread	loaves	3.00	2 000	6 000	0.4220	1.0345
Apples	kg	4.50	450	2 025	0.1424	0.8182
Beer	litres	8.40	130	1 092	0.0768	1.0500
LCD TV	units	1 100.00	3	3 300	0.2321	0.9167
Jeans	units	60.00	30	1 800	0.1266	1.0909
Total				**14 217**	**1.0000**	

Laspeyres

= (0.3932 x 1.0345) + (0.1864 x 0.8182) + (0.1085 x 1.0500) + (0.1627 x 0.9167) + (0.1492 x 1.0909) x 100
= 98.51

Paasche

= 1 / ((0.4220 / 1.0345) + (0.1424 / 0.8182) + (0.0768 / 1.0500) + (0.2321 / 0.9167) + (0.1266 / 1.0909)) x 100
= 97.62

Review of traditional price index formula

Item		Price ($)	Quantity	Expenditure ($)	Expenditure shares	Price relatives
Period 0						
White fresh bread	loaves	2.90	2 000	5 800	0.3932	1.0000
Apples	kg	5.50	500	2 750	0.1864	1.0000
Beer	litres	8.00	200	1 600	0.1085	1.0000
LCD TV	units	1 200.00	2	2 400	0.1627	1.0000
Jeans	units	55.00	40	2 200	0.1492	1.0000
Total				**14 750**	**1.0000**	
Period t						
White fresh bread	loaves	3.00	2 000	6 000	0.4220	1.0345
Apples	kg	4.50	450	2 025	0.1424	0.8182
Beer	litres	8.40	130	1 092	0.0768	1.0500
LCD TV	units	1 100.00	3	3 300	0.2321	0.9167
Jeans	units	60.00	30	1 800	0.1266	1.0909
Total				**14 217**	**1.0000**	

Fisher

$= (98.51 \times 97.62)^{1/2}$
$= 98.06$

Törnqvist is best calculated by first taking the logs of the index formula

$= 1/2 \times (0.3932 + 0.4220) \times \ln(1.0345)$
$+ 1/2 \times (0.1864 + 0.1424) \times \ln(0.8182)$
$+ 1/2 \times (0.1085 + 0.0768) \times \ln(1.0500)$
$+ 1/2 \times (0.1627 + 0.2321) \times \ln(0.9167)$
$+ 1/2 \times (0.1492 + 0.1266) \times \ln(1.0909)$
$= -0.0199$

and then taking the exponent multiplied by 100

$= e^{-0.0199} * 100$
$= 98.04$

Construction of a national CPI

- **Define the scope of the index**
  - Product coverage
  - Target population
- **Classifications**
  - Product classification
  - Geography classification
- **Source of expenditure weights and the frequency of their update**
  - Survey of Household Spending
  - CPI basket
  - Frequency of basket update
- **Sampling strategy**
  - Outlet sample
  - Product sample
  - Collection pattern for each product
- **Price collection**
  - In the field, by price interviewers
  - Using administrative data
  - Internet, online
  - Scanner data, etc.
- **Data editing and quality control of micro-data**

Construction of a national CPI

- CPI is built up from price indices for **elementary aggregates** (EA)
- Elementary aggregates are pairings of **lowest level product classes** and **lowest level geography classes**
  - Banana in geo strata 1, banana in geo strata 2 are two different EAs
  - Banana in Canada is not an EA
  - Some product classes may not be available in some geo classes
  - Canada: 695 lowest level product classes, 19 lowest level geo classes
- EAs price indices estimated by direct price observation or by imputation
- **Elementary price indices** :
  - In general, **Jevons index** formula is used in most countries
  - Prices for goods and services of **same quantity and same quality observed over time**
- **Aggregate level indices**:
  - Consumption expenditures, to give relative importance to product/geo classes
  - Expenditures must be **price-updated**: the expenditure value of each category is multiplied by its monthly price change. **Same quantities valued at current month's prices**
  - In general, **Lowe index** aggregation formula is used in most countries

Preprocessing of scanner data for CPI calculation

- **Typical structure of scanner data**

Variable	Format	Example
Date	Numeric or date	20180104, 01/04/2018
Store ID	Text or numeric	Store_0001
Store address	Text	123 ABC street, Region, postal code
Province/Region	Text or numeric	Ontario
Product identifier (UPC, SKU)	Numeric or text	ABC_0001
Retailer classification	Text or numeric	Grocery – Dairy – Cheese – Entertainment Cheese
Description	Text	‘Tasty’ Brand Brie Cheese 200g
Quantity	Numeric	61 units
Turnover	Numeric or currency	$501

- **Potential advantages of scanner data:**
  - Full enumeration of products sold during a given time period
    - ⇒ Universe of products purchased by consumers
  - Quantity and turnover information
    - ⇒ Actual average transaction price paid by consumers

⇒ Large scale increase in CPI product sample size
Lower amount of resources for price collection

- **Possible drawback: Risk of over-coverage**
  - Purchases made by businesses are included!
  - Is that really an issue?

Preprocessing of scanner data for CPI calculation

- **How is a product defined?**
  - A set of homogeneous items
  - Associated with: Global Trade Item Number (GTIN), Universal Product Code (UPC) or retailer assigned codes, Stock Keeping Unit (SKU)?
  - What if GTINs change frequently with very small or no changes in product characteristics: relaunches?

- **Data aggregation**
  - Individual transactions during a time period need to be aggregated
    - Total quantities, total turnover and average prices
  - Doing this aggregation at outlet (location) level or at retailer level?
    - Do prices differ across stores of the same retailer?
    - No, in general; but a market intelligence research recommended
  - Doing this aggregation at regional level?
    - National or regional pricing?
    - Do purchasing patterns vary by region? Regional weights may be needed!
  - For monthly CPI, how many weeks to include in the aggregation?
    - Depends on the CPI production cycle
    - Ideally, all days of the month
    - Practically, 2 or 3 weeks.

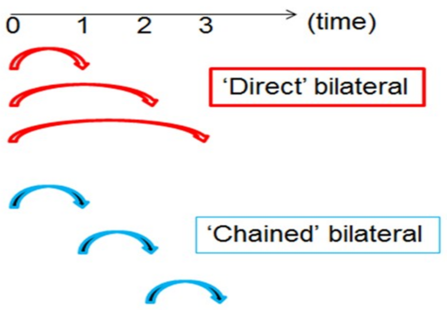
Scanner data and traditional price index formula

- Ideally, we would use a method that:
  - Uses census of products
  - Weights prices at the product (and product group) level
  - and automated processes (less resources)

- ILO/IMF Consumer Price Index manual recommends ‘superlative’ indexes (e.g. Fisher, Törnqvist) as the ideal CPI target

- Can we apply these methods directly to scanner data?

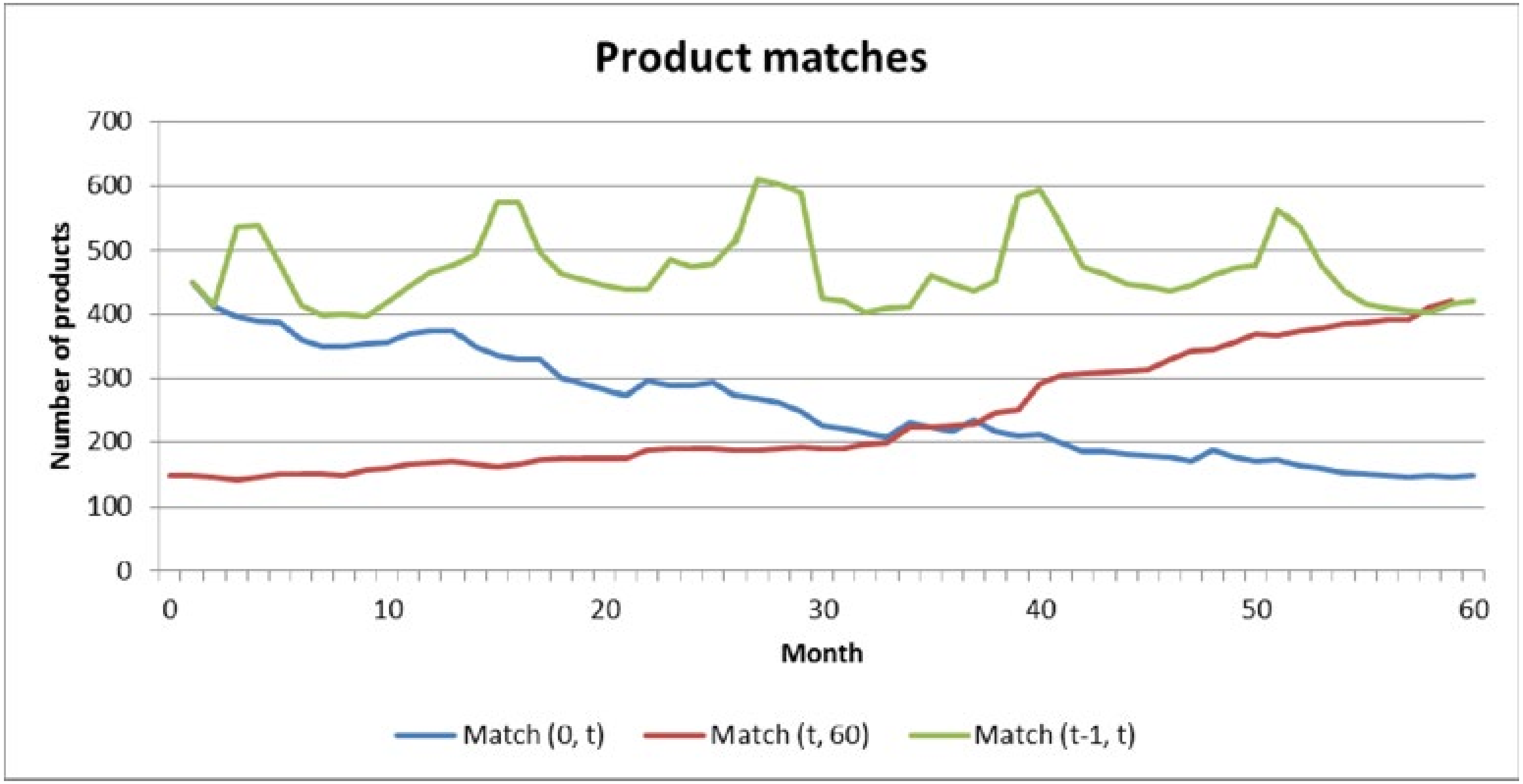
- Could use ‘direct’ or ‘chained’ weighted bilateral indexes



- However, dynamic nature of transactions data can make these methods perform poorly

Scanner data and traditional price index formula

- 'Direct' bilateral indexes suffer from a 'matching' problem

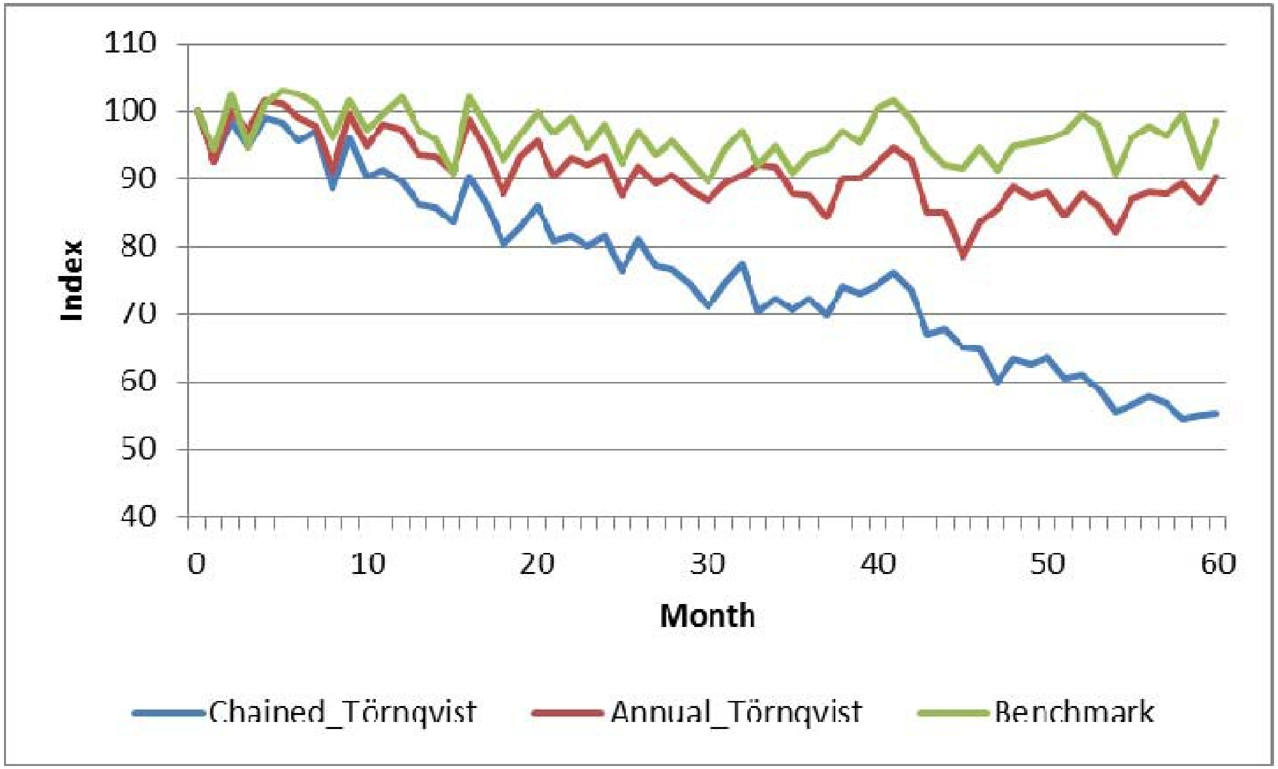


Scanner data and traditional price index formula

- Consumers responsive to sales: price and quantity bouncing can cause problems for chained indexes

Scanner data and traditional price index formula

- Chained bilateral indexes suffer from a ‘chain drift’ problem



Multilateral price index methods: general principle

- Many National Statistical Institutes (NSI) continue to use a geometric mean at elementary aggregate level
  - No weight used at this level
  - Mostly supermarket products
  - Belgium, Canada, Denmark, Iceland, Netherlands, Norway, Sweden, Switzerland
- A few NSI have implemented multilateral index methods for some of their CPI components:
  - Statistics Netherlands (CBS): Mobile phone and department store products
  - Statistics New Zealand (SNZ): Audio visual and household appliance products
  - Australia Bureau of Statistics (ABS): Food expenditure classes

Multilateral price index methods: general principle

- Bilateral index methods compare prices between two time periods
- Multilateral index methods:
  - Price comparisons across multiple (three or more) time periods
  - Historically used in constructing spatial price indexes (comparison of price levels between countries)
  - Use all matched products between any two months
  - Are “average” of multiple bilateral indexes:
    - Example: $$I_{jl} = \left(\prod_{k=0}^{T} I_{jk} I_{kl} \right)^{1/(T+1)}$$
  - Weight products by their economic importance (turnover)
  - Are free of ‘chain drift’

Questions?

Thank you!