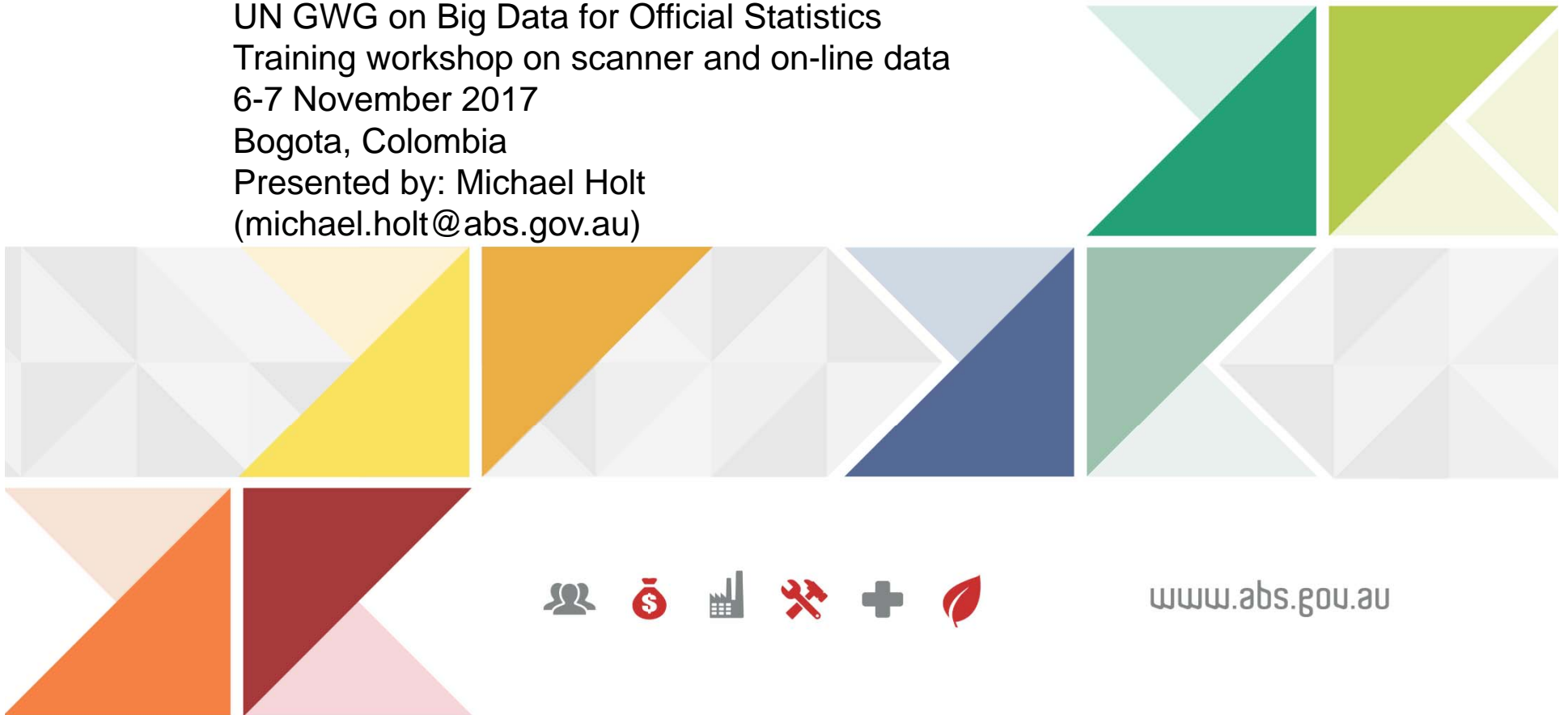




# An introduction to scanner data methods

UN GWG on Big Data for Official Statistics  
Training workshop on scanner and on-line data  
6-7 November 2017  
Bogota, Colombia  
Presented by: Michael Holt  
([michael.holt@abs.gov.au](mailto:michael.holt@abs.gov.au))



[www.abs.gov.au](http://www.abs.gov.au)



1. Traditional index methods
2. Characteristics of scanner data
3. Introduction to multilateral index methods
4. Introduction to extension methods
5. Monitoring systems for new methods
6. Implementation plan of new methods – ABS case study
7. Conclusion



- “A consumer price index measures the change in the prices paid by households for goods and services consumed” [ABS \(2016\)](#)
- In practice, the aggregation of many prices into a single number (price index)
- Basic information required to compile a price index: prices, expenditure shares and classifications (e.g. product, product groupings)
- Focus of this session is on methods that utilise expenditure weights, but they can be adapted to unweighted data (e.g. web scraped data)
- Let us first cover some traditional price index methods



- Laspeyres

$$I_L^{0,t} = \left( \frac{\sum_i p_i^t q_i^0}{\sum_i p_i^0 q_i^0} \right)$$

*With weight:*

$$I_L^{0,t} = \sum_i w_i^0 \left( \frac{p_i^t}{p_i^0} \right)$$

- Paasche

$$I_P^{0,t} = \left( \frac{\sum_i p_i^t q_i^t}{\sum_i p_i^0 q_i^t} \right)$$

*With weight:*

$$I_P^{0,t} = \frac{1}{\sum_i w_i^t \left( \frac{p_i^0}{p_i^t} \right)}$$

# Traditional index methods



Item		Price (\$)	Quantity	Expenditure (\$)	Expenditure shares	Price relatives
<b>Period 0</b>						
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
<b>Total</b>				<b>14 750</b>	<b>1.0000</b>	
<b>Period t</b>						
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
<b>Total</b>				<b>14 217</b>	<b>1.0000</b>	

Laspeyres

$$= (0.3932 \times 1.0345) + (0.1864 \times 0.8182) + (0.1085 \times 1.0500) + (0.1627 \times 0.9167) + (0.1492 \times 1.0909) \times 100$$

$$= 98.51$$

Paasche

$$= 1 / ((0.4220 / 1.0345) + (0.1424 / 0.8182) + (0.0768 / 1.0500) + (0.2321 / 0.9167) + (0.1266 / 1.0909)) \times 100$$

$$= 97.62$$





- Fisher

$$I_F^{0,t} = \left( \frac{\sum_i p_i^t q_i^0}{\sum_i p_i^0 q_i^0} \right)^{\frac{1}{2}} \left( \frac{\sum_i p_i^t q_i^t}{\sum_i p_i^0 q_i^t} \right)^{\frac{1}{2}}$$

- Törnqvist

$$I_T^{0,t} = \prod_i \left( \frac{p_i^t}{p_i^0} \right)^{\frac{1}{2}(s_i^0 + s_i^t)}$$

# Traditional index methods



Item		Price (\$)	Quantity	Expenditure (\$)	Expenditure shares	Price relatives
<b>Period 0</b>						
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
<b>Total</b>				<b>14 750</b>	<b>1.0000</b>	
<b>Period t</b>						
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
<b>Total</b>				<b>14 217</b>	<b>1.0000</b>	

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

Fisher

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

$$= 98.06$$

$$= 1/2 \times (0.3932 + 0.4220) \times \ln(1.0345)$$

$$+ 1/2 \times (0.1864 + 0.1424) \times \ln(0.8182)$$

$$+ 1/2 \times (0.1085 + 0.0768) \times \ln(1.0500)$$

$$+ 1/2 \times (0.1627 + 0.2321) \times \ln(0.9167)$$

$$+ 1/2 \times (0.1492 + 0.1266) \times \ln(1.0909)$$

$$= -0.0199$$

and then taking the exponent multiplied by 100

$$= e^{-0.0199} \times 100$$

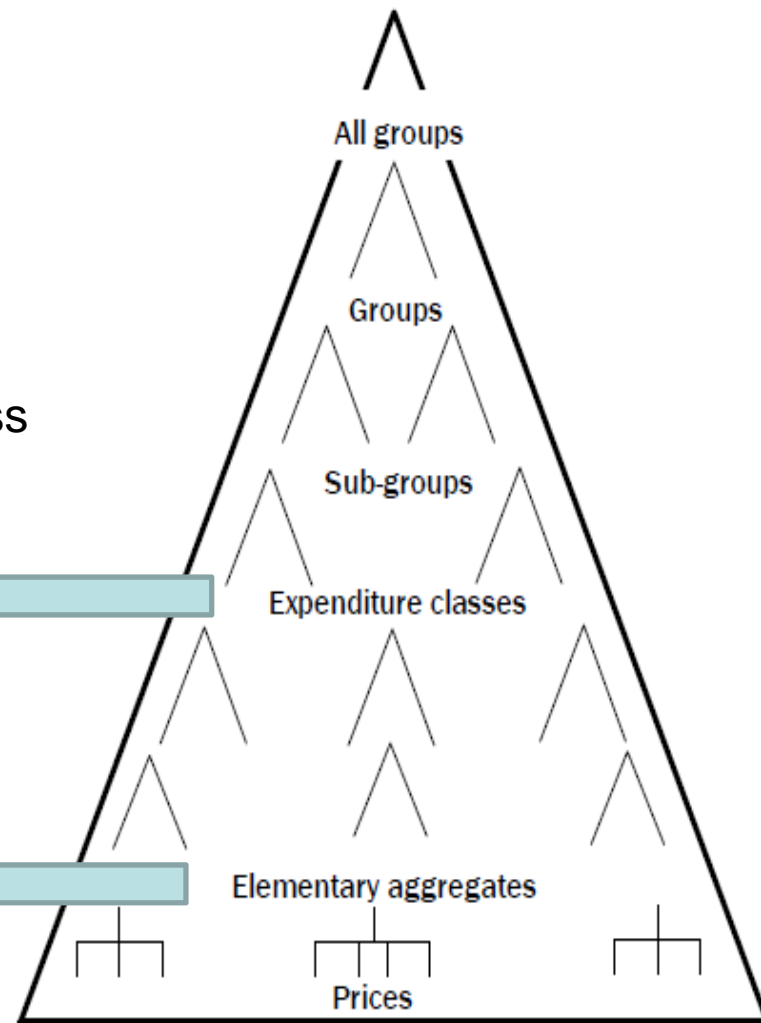
$$= 98.04$$

## Traditional index methods

- Choice of price index formula dictated by available data at the ABS
- Jevons (geometric mean) formula used at elementary level
- Lowe Index formula used at expenditure class level (and above)

$$P_L^t = \sum_{i=1}^n \left( \frac{p_i^t}{p_i^0} \right) s_i^{0b}$$

$$P_J^t = \prod_{i=1}^n \left( \frac{p_i^t}{p_i^0} \right)^{\frac{1}{n}}$$





## Characteristics of scanner data

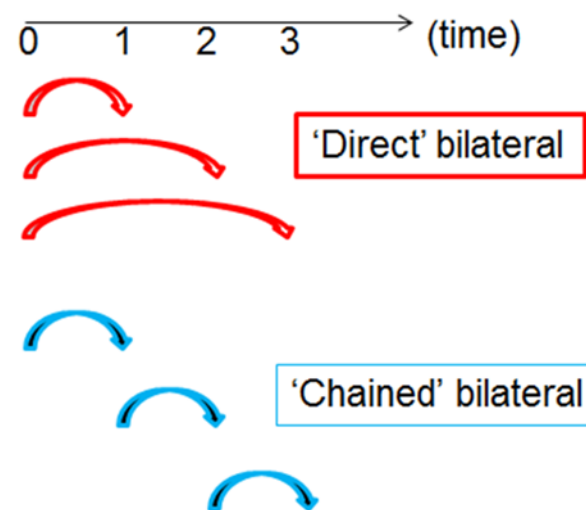


- Scanner data contains detailed information about transactions, dates, quantities, product descriptions, and values of products sold
- Ideally, we would use a method that:
  - Uses census of products
  - Weights prices at the product (and product group) level
  - Automated processes (less resources)
- [ILO/IMF Consumer Price Index manual](#) recommends 'superlative' indexes (e.g. Fisher, Törnqvist) as the ideal CPI target

Product ID	Store Location	Product Description	Time Period	Revenue (\$)	Units sold	Unit value (\$/unit)
U0001	Sydney CBD	CARROTS PREPACKED 1KG	Jan-16	5000	2500	2.00
U0001	Sydney CBD	CARROTS PREPACKED 1KG	Feb-16	7000	4000	1.75
U0001	Sydney CBD	CARROTS PREPACKED 1KG	Mar-16	4100	2000	2.05
U0001	Sydney CBD	CARROTS PREPACKED 1KG	Q1-16	16100	8500	1.89
U0002...						

## Characteristics of scanner data

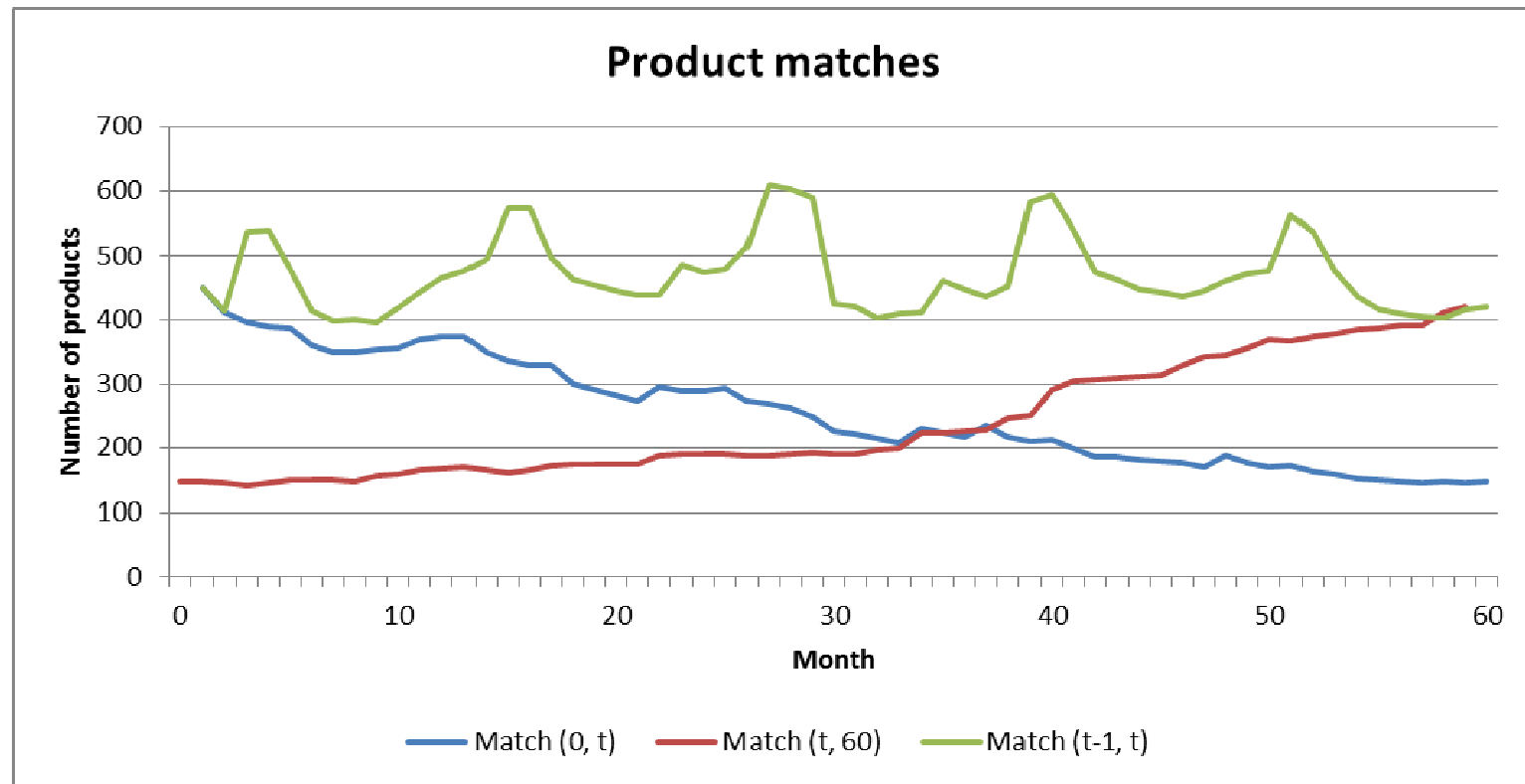
- Can we just apply these methods (e.g. Törnqvist, Fisher) directly to scanner data?
- Could use 'direct' or 'chained' weighted bilateral indexes
- However, dynamic nature of transactions data can make these methods perform poorly (i.e. traditional price index methods break down when applied to this new data source)



## Characteristics of scanner data

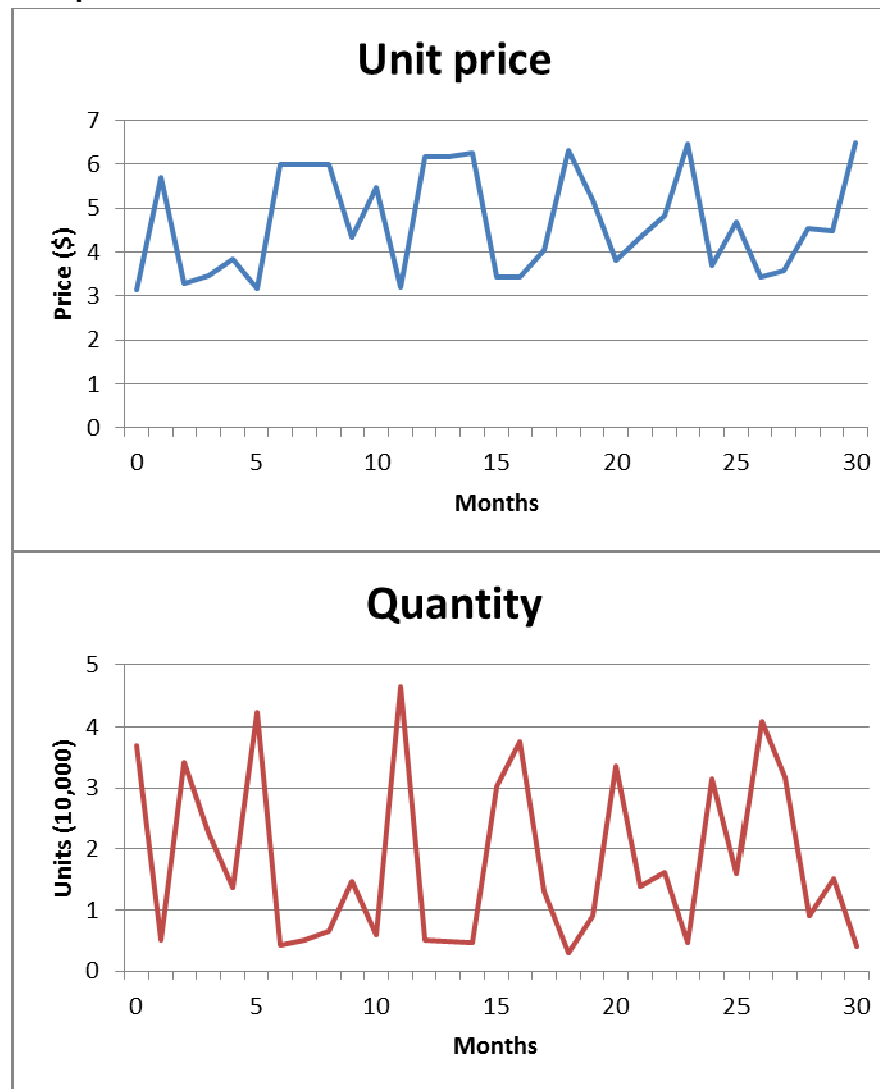


- 'Direct' bilateral indexes suffer from a 'matching' problem



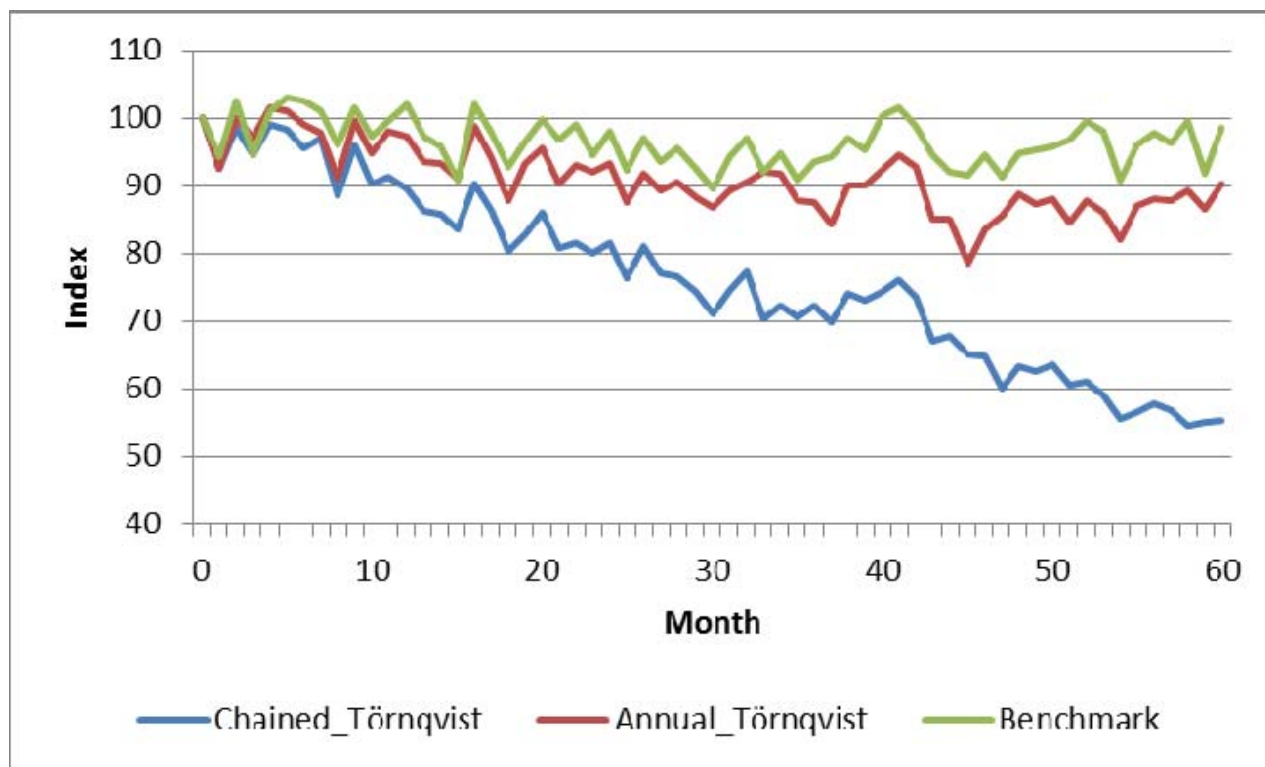
## Characteristics of scanner data

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



## Characteristics of scanner data

- Chained bilateral indexes suffer from a 'chain drift' problem



## Characteristics of scanner data



- An interim solution for National Statistical Institutions (NSIs) – continue to use a geometric mean at the elementary level
- Statistics Netherlands (CBS) and Statistics New Zealand (SNZ) have implemented multilateral methods for some components of their CPIs
- ABS will implement a multilateral method in December 2017

<i>Country</i>	<i>Transactions data items</i>	<i>Elementary aggregation formula</i>
Belgium	Supermarket items	Geometric mean
Denmark	Supermarket items	Geometric mean
Iceland	Supermarket items	Geometric mean
Netherlands	Supermarket items	Geometric mean
	Mobile phone and department store items	GK method with a direct annual extension
New Zealand	Audio visual and household appliance items	Imputation Törnqvist Rolling Year GEKS (ITRYGEKS)
Norway	Food medical, retail, petrol and pharmacy items	Geometric mean
Sweden	Supermarket items	Geometric mean
Switzerland	Supermarket items	Geometric mean

## Characteristics of scanner data

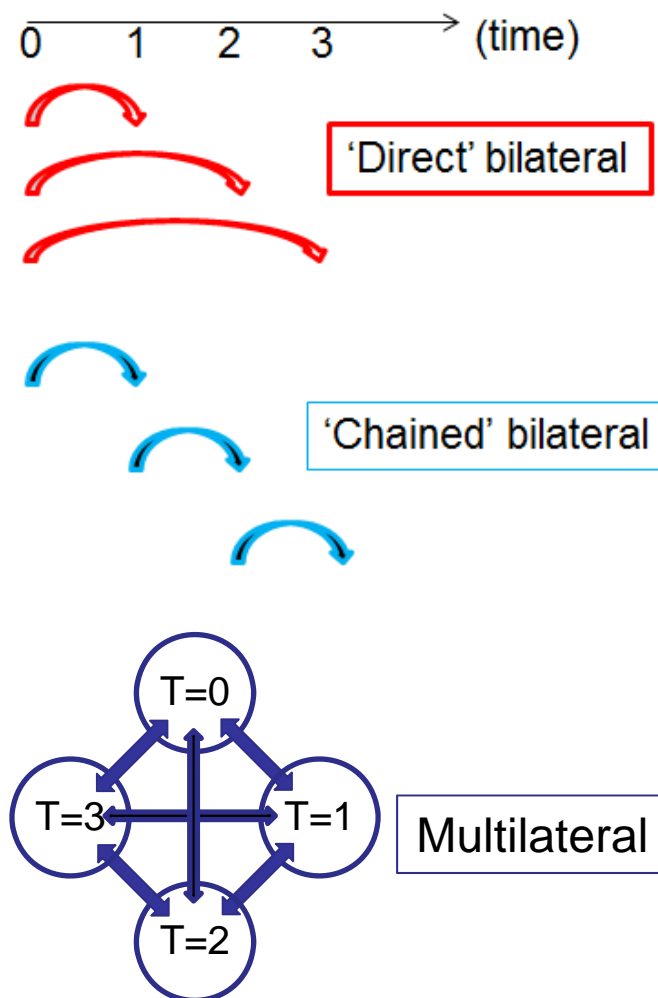


- Summary
- Dynamic nature of scanner data (e.g. item churn, specialling) cause problems for traditional index number formula
- As a result, most NSIs have initially implemented unweighted (geometric mean) methods into production
- NSIs conducting lots of research into using different index (multilateral) methods that maximise the information available on scanner datasets
- Some NSIs (e.g. CBS, SNZ, ABS) moving to implementing multilateral methods
- Questions?

## An introduction to multilateral methods



- Bilateral index methods compare prices across two time periods
- Multilateral index methods make price comparisons across multiple (three or more) time periods
- Historically used in constructing spatial price indexes
- Multilateral methods use all matched products, weight products by economic importance and are free of 'chain drift'



$$Index_{jl} = Index_{jk} \times Index_{kl}$$



## An introduction to multilateral methods



- NSIs have conducted and supported research on scanner data for over twenty years
- Ivancic, L., Fox, K. J. & Diewert, E. W. 2011. Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics*, 161, 24-35.



- The international price statistics community has reached a consensus that multilateral methods are the most appropriate approach to compile price indexes
- International CPI manual currently being [revised](#) to recommend multilateral methods

## An introduction to multilateral methods



- This session will introduce some ‘well-known’ multilateral methods – but this is not an exhaustive list of available multilateral methods
- For a more exhaustive list of methods for temporal aggregation of scanner data see [de Haan, Willenborg and Chessa \(2016\)](#)
- This introduction will cover three ‘families’ of multilateral methods:
  - Dummy variable regression (e.g. TD, TPD)
  - Gini, Eltetö and Köves, and Szulc (GEKS)
  - Geary-Khamis (GK)
- The types on methods we could use for each family depends on the data available (available information on product characteristics)



### Dummy variable regression

- Dummy (0,1 Indicative variables): takes the value of 0 or 1 to indicate the presence of some characteristic (category)

$$\ln p_i^t = \delta^0 + \sum_{k=1}^K \beta_k z_{ik} + \varepsilon_i^t$$

where :  $p_i^t$  is the price of item  $i$  in period  $t$

$z_{ik}$  is the quantity/indicative of the  $k$ -characteristic ( $k = 0, \dots, K$ ) for item  $i$

$\beta_k$  is the corresponding regression parameter



## Dummy variable regression

Product	Period	Price	Characteristics	
			Organic	Imported
pear	0	5	0	0
pear	0	6	1	0
pear	0	5	1	1
banana	0	3	0	1
banana	0	5	1	0
banana	0	4	0	0
pear	1	8	0	0
pear	1	10	1	0
pear	1	7	0	1
apple	1	5	0	1
apple	1	7	1	0
apple	1	4	0	0
banana	2	4	0	1
banana	2	6	1	0
banana	2	5	0	0
apple	2	7	0	1
apple	2	9	1	0
apple	2	8	0	0



Model results				
Estimate parameters	Intercept	Organic	Imported	Predicted
		1.862	0.16338	-0.162
pear	0	1	1	6.445486
apple	-0.00183	1	1	6.433701
banana	-0.38224752	1	1	4.397923
pear	0	0	0	6.436597
apple	-0.00183	0	0	6.424829
banana	-0.38224752	0	0	4.391858
pear	0	0	1	5.473947
apple	-0.00183	0	1	5.463939
banana	-0.38224752	0	1	3.735017



### Time dummy (TD) model

- Extended from dummy variable regression (adding purchasing period into the model)

$$\ln p_i^t = \delta^0 + \sum_{t=1}^T \delta^t D_i^t + \sum_{k=1}^K \beta_k z_{ik} + \varepsilon_i^t$$

where (cont):  $D_i^t$  is a dummy variable (0,1) indicating item  $i$  purchased in period  $t$

$\delta^t$  is the time dummy regression parameter

- Then,

$$\hat{p}_i^0 = \exp(\hat{\delta}^0) \exp \left[ \sum_{k=1}^K \hat{\beta}_k z_{ik} \right]$$

$$\hat{p}_i^t = \exp(\hat{\delta}^0) \exp(\hat{\delta}^t) \exp \left[ \sum_{k=1}^K \hat{\beta}_k z_{ik} \right]$$

$$P_{TD}^{0,t} = \hat{p}_i^t / \hat{p}_i^0 = \exp(\hat{\delta}^t)$$





## Time dummy (TD) model

Product	Period	Price	Characteristics	
			Organic	Imported
pear	0	5	0	0
pear	0	6	1	0
pear	0	5	1	1
banana	0	3	0	1
banana	0	5	1	0
banana	0	4	0	0
pear	1	8	0	0
pear	1	10	1	0
pear	1	7	0	1
apple	1	5	0	1
apple	1	7	1	0
apple	1	4	0	0
banana	2	4	0	1
banana	2	6	1	0
banana	2	5	0	0
apple	2	7	0	1
apple	2	9	1	0
apple	2	8	0	0



## Time Dummy Regression

pear	0.00
banana	-0.47
apple	-0.23
Period 0	0.00
Period 1	0.27
Period 2	0.47

$$P_{TD}^{0,1} = \exp(\hat{\delta}^1) = \exp(0.27) = 1.31$$



Time product dummy (TPD) model

- Further extended from time dummy regression

$$\ln p_i^t = \delta^0 + \sum_{t=1}^T \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t$$

- Then,

$$\hat{p}_i^0 = \exp(\hat{\delta}^0) \exp(\hat{\gamma}_i)$$

$$\hat{p}_i^t = \exp(\hat{\delta}^0) \exp(\hat{\delta}^t) \exp(\hat{\gamma}_i)$$

$$P_{TPD}^{0,t} = \hat{p}_i^t / \hat{p}_i^0 = \exp(\hat{\delta}^t)$$



## Time product dummy (TPD) model

Product	Period	Price
pear	0	5
pear	0	6
pear	0	5
banana	0	3
banana	0	5
banana	0	4
pear	1	8
pear	1	10
pear	1	7
apple	1	5
apple	1	7
apple	1	4
banana	2	4
banana	2	6
banana	2	5
apple	2	7
apple	2	9
apple	2	8



Time Dummy Estimates	
pear	0.00
banana	-0.52
apple	-0.25
0	0.00
1	0.23
2	0.44

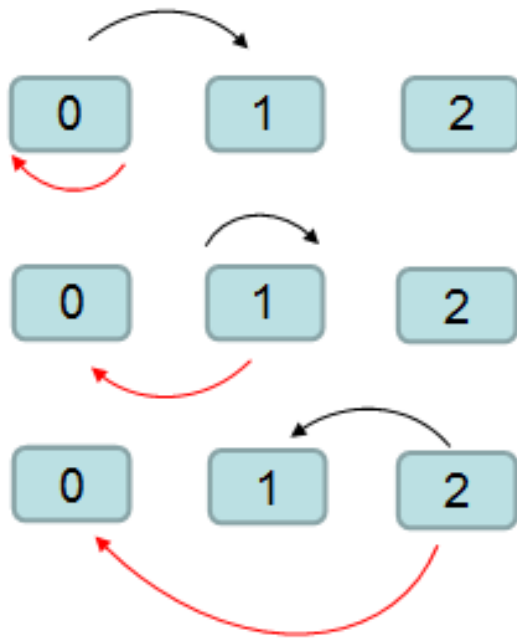
$$I_{TPD}^{0,1} = \exp(\delta^1) = 1.26$$





### Gini, Eltetö and Köves, and Szulc (GEKS)

- Geometric mean of all ratios of bilateral indexes where each entity is taken in turn as the base



$$P_{GEKS}^{0,1} = \left( \frac{P^{0,1}}{P^{0,0}} \times \frac{P^{1,1}}{P^{1,0}} \times \frac{P^{2,1}}{P^{2,0}} \right)^{1/3}$$



Gini, Eltetö and Köves, and Szulc (GEKS)

- If the dataset has limited/no characteristics but we have expenditure information, we can use a superlative bilateral index formula (e.g. Törnqvist, Fisher)
- If the dataset has characteristics and expenditure information, we can use a weighted time dummy regression for the bilateral links
- If the dataset has no expenditure information, we can use an unweighted bilateral formula (e.g. Jevons) or unweighted time dummy regression for bilateral links



### Geary-Khamis (GK)

- Can be expressed in terms of a ‘quality adjusted unit value index’
- The “Unit Value” of a set of (homogeneous) products is calculated as

$$\frac{\text{Total value of purchases or sales}}{\text{Sum of quantities}} = \frac{\sum_{i=1}^I p_i q_i}{\sum_{i=1}^I q_i}$$



### Geary-Khamis (GK)

- A Unit Value Index is a price index which measures the change in the unit values

$$P_{UV}^{0,t} = \frac{\text{Average price at time } t}{\text{Average price at time } 0}$$
$$= \left[ \frac{TotExp_t}{TotQuantity_t} \right] / \left[ \frac{TotExp_0}{TotQuantity_0} \right]$$



### Geary-Khamis (GK)

- Adding up quantities of dissimilar goods to form the unit value index isn't necessarily meaningful
- Some authors ([de Haan 2015](#), [Chessa 2015](#)) suggest using standardised or quality-adjusted quantities
- The idea is to apply quality adjustment factors to the various item quantities to express them in terms of a “base” item, and then simply add them up



### Geary-Khamis (GK)

- GK method expressed as a QAUUV, where adjustment factors ( $v_{i/b}$ ) are expressed as a quantity weighted average of deflated prices

$$P_{GK}^{0t} = \frac{V^{0t}}{Q_{GK}^{0t}} = \frac{\sum_{i \in U^t} p_i^t q_i^t / \sum_{i \in U^t} v_{i/b} q_i^t}{\sum_{i \in U^0} p_i^0 q_i^0 / \sum_{i \in U^0} v_{i/b} q_i^0}$$

$$v_{i/b} = \sum_{z \in T} \varphi_{i,z} \frac{p_{i,z}}{P_{GK}^z}$$

← Deflated price of product  $i$  in period  $z$

$$\varphi_{i,z} = \frac{q_{i,z}}{\sum_{s \in T} q_{i,s}}$$

← Quantity share of item  $i$  in period  $z$

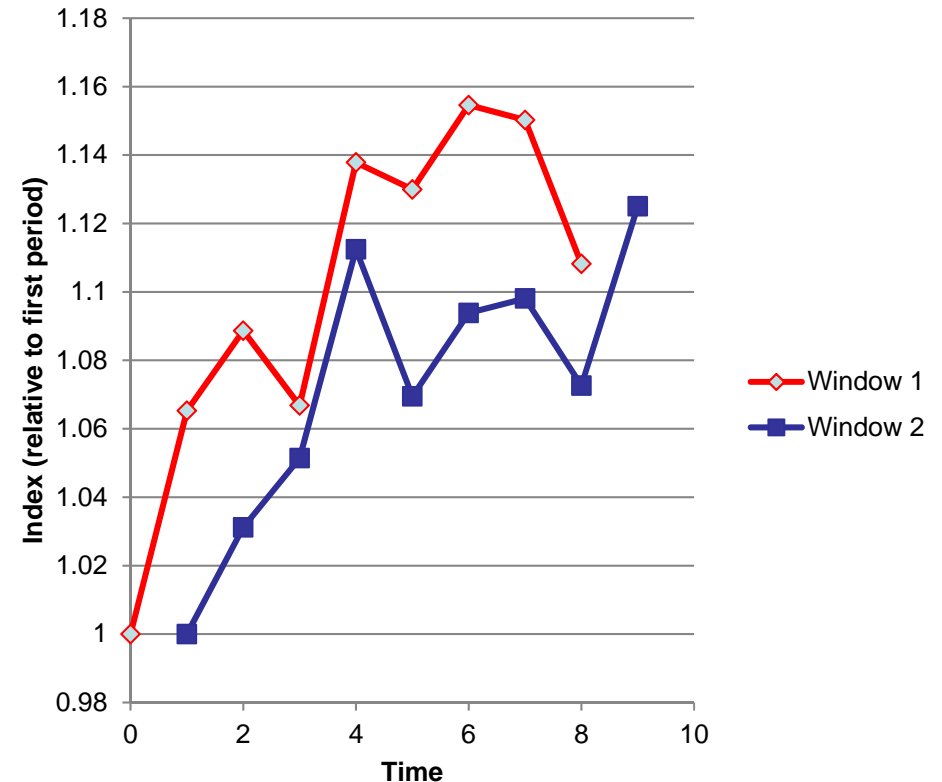
## An introduction to multilateral methods



- Summary
- Introduction to three types of multilateral methods
  - Regression based methods (e.g. TD, TPD)
  - GEKS based methods
  - GK/QAUV based methods
- The type of dataset (availability of product characteristics) will usually dictate the method – each multilateral method can be adapted accordingly
- Questions?



- Need to extend index each period
- Data from current period can alter comparisons between earlier periods
- Can't revise index in normal circumstances
- Two questions to address:
  1. How do we form a multilateral "window" incorporating the current period?
  2. How do we splice the results onto previous index levels?

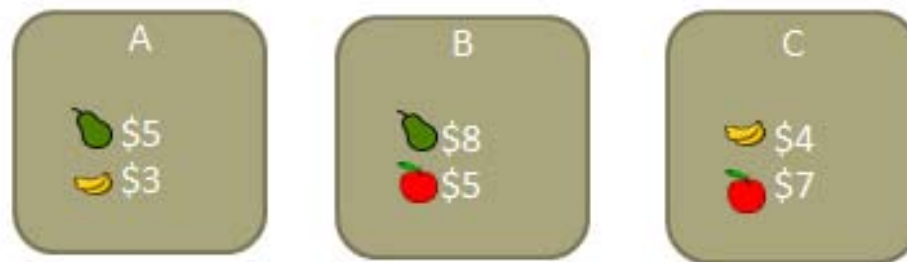







## An introduction to extension methods



Three period TPD



Fixed Effect Estimates	
	0.00
	-0.68
	-0.30
A	0.00
B	0.30
C	0.46




$$I_{TPD}^{A,B} = \frac{\exp(\delta^B)}{\exp(\delta^A)} = 1.35$$

## An introduction to extension methods



Four period TPD

A	B	C	D
 \$5  \$3	 \$8  \$5	 \$4  \$7	 \$6  \$6

Fixed Effect Estimates	
	0.00
	-0.61
	-0.37
A	0.00
B	0.37
C	0.50
D	0.62

$$I_{TPD}^{A,B} = \frac{\exp(\delta^B)}{\exp(\delta^A)} = 1.45$$

## An introduction to extension methods



- *Rolling or expanding window approaches*

Rolling window							
time	0	1	2	...	t	t+1	t+2
		-	-	-			
			-	-	-		
				-	-	-	

- Fixed length
- Variable start point

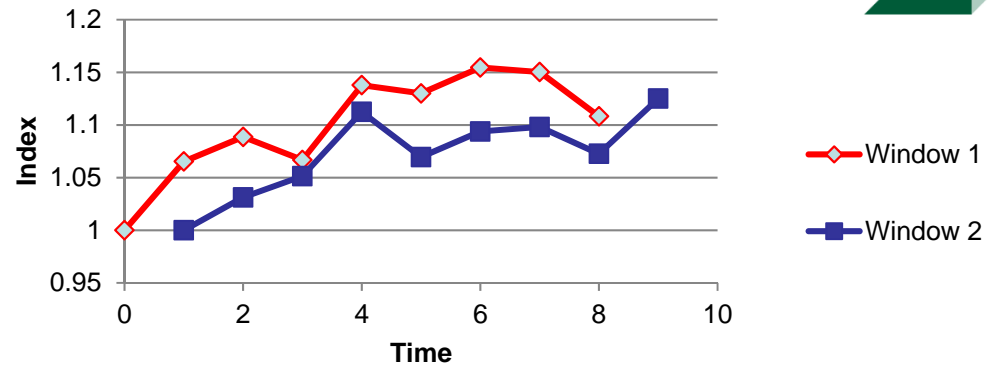
Expanding window								
time	0	1	2	...	t-1	t	t+1	t+2
		-						
		-	-	-				
		-	-	-	-			
							-	

- Variable length
- Fixed start point (can be updated from time to time)

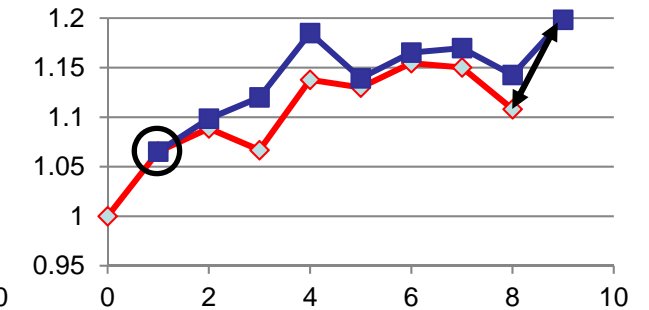
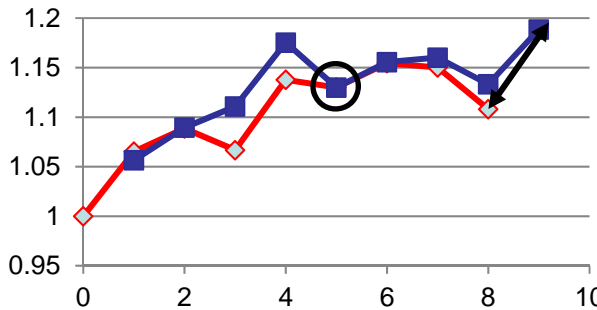
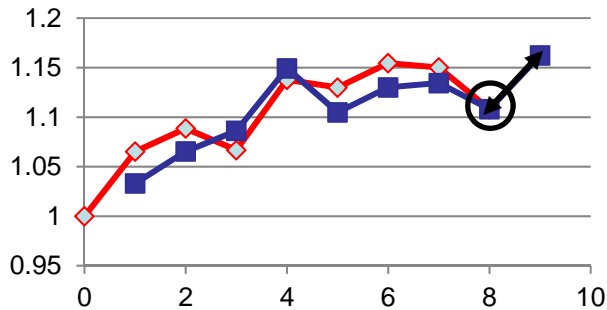
# An introduction to extension methods



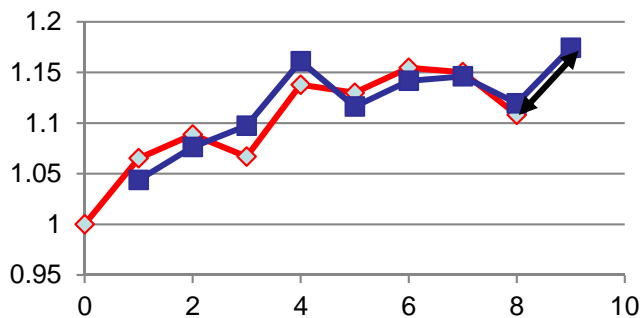
How should we link together successive windows?



Link in one period (which?)



Or take the (geo)mean over all possible links



## An introduction to extension methods



- Length of a multilateral window
- General consensus that the size of a multilateral window should be at least one year – but no consensus on optimal window size
- Considerations:
  - Seasonal products
  - Sensitivity to changes (preferences, utility/cost functions)
  - Product life cycles

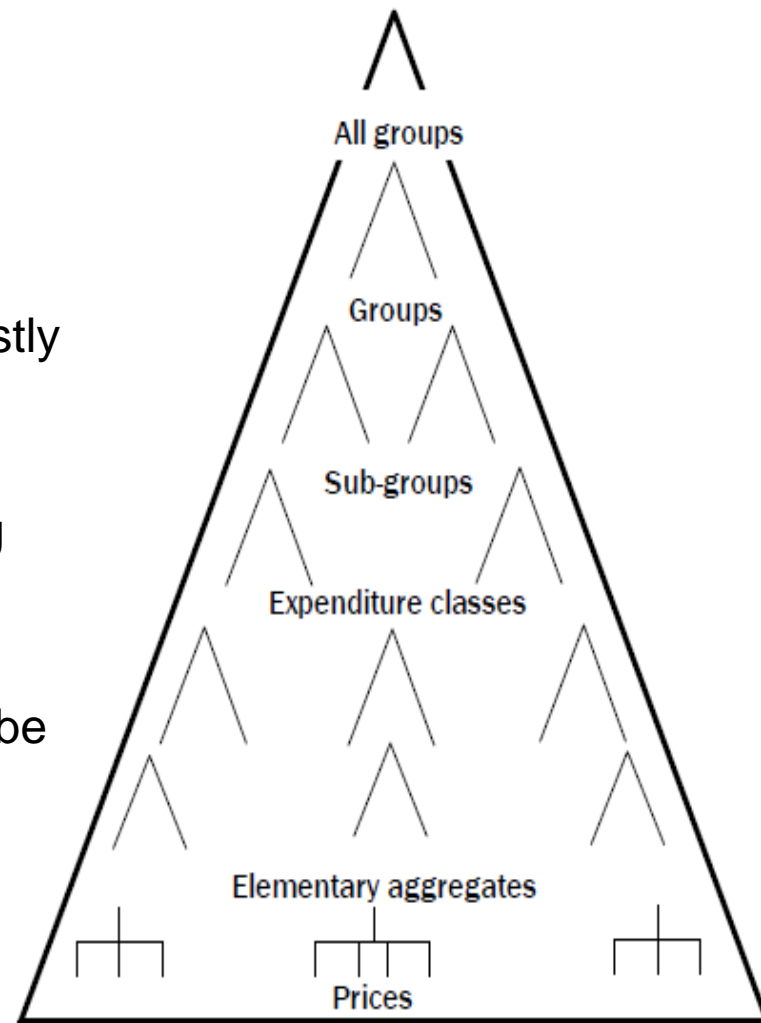




- Summary
- Additional periods to the multilateral window revise previous price movements
- Introduction to two families of extension methods:
  - Rolling window approach
  - Expanding window approach
- No recommendations on optimal window length – at least one year for rolling window approaches recommended. A matter for empirical testing
- Questions?

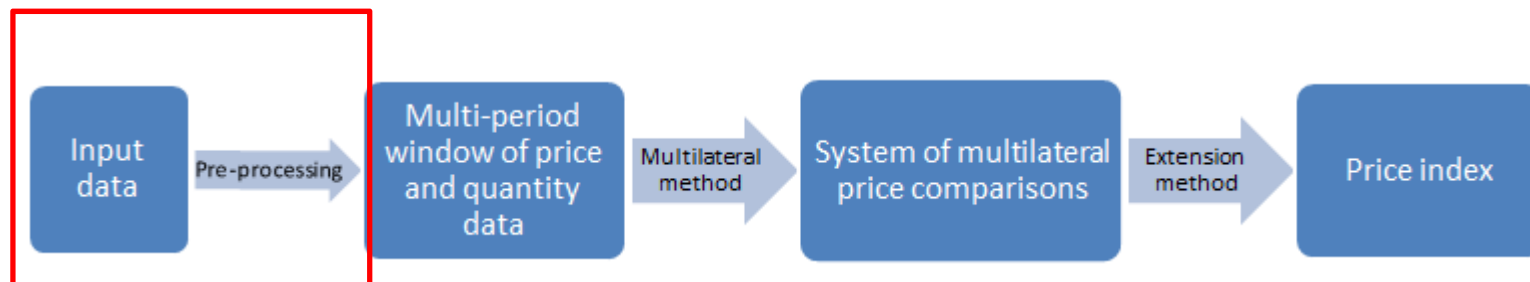
## Monitoring systems of new methods

- NSIs have traditionally had small purposive samples at the elementary level – makes it possible to micro edit?
- Scaling up the number of price observations used makes a micro editing strategy very costly for an NSI
- This means NSIs should consider monitoring input and output metrics
- We will cover some basic metrics that could be considered





- Input metrics
- Useful summary of practices adopted by [Statistics New Zealand](#)
- Plot time series of aggregate expenditure, quantities and average prices
- Plot time series of matching statistics (average matched expenditure share)
- Plot time series of sample sizes
- Plot time series of 'outlier' prices (proportion of expenditure removed)





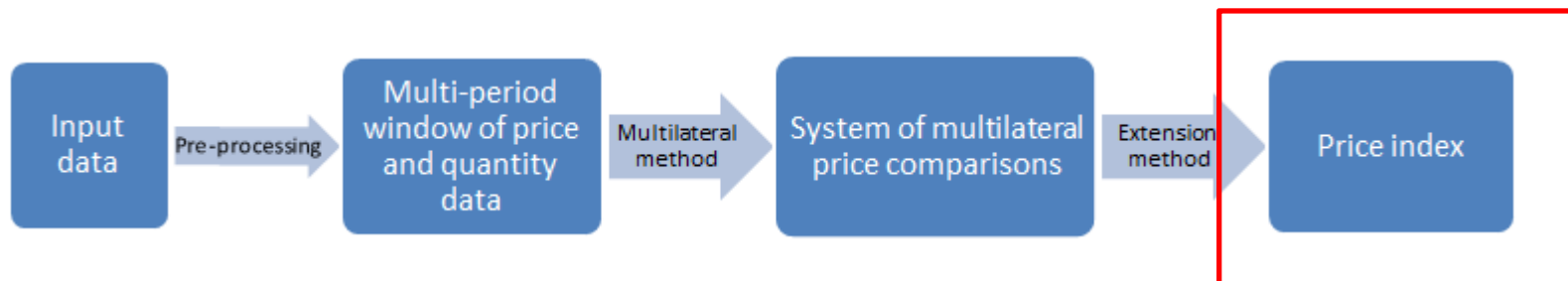
## Monitoring systems of new methods



- Input metrics
- Aim of the input checks is to detect abnormalities in the data that be caused by
  - Changes in how the respondent reports data to the NSI (e.g. different classifications, product codes, store reporting)
  - Weaknesses in pre-processing steps/methods used by NSIs



- Output metrics
- NSIs require the ability to explain price movements – the use of multilateral methods makes this more difficult?
- Decomposition tools are useful metrics to determine main contributors to price change for different respondents/cities
- ABS methodology team has recommended tools that help the editing strategy of the CPI

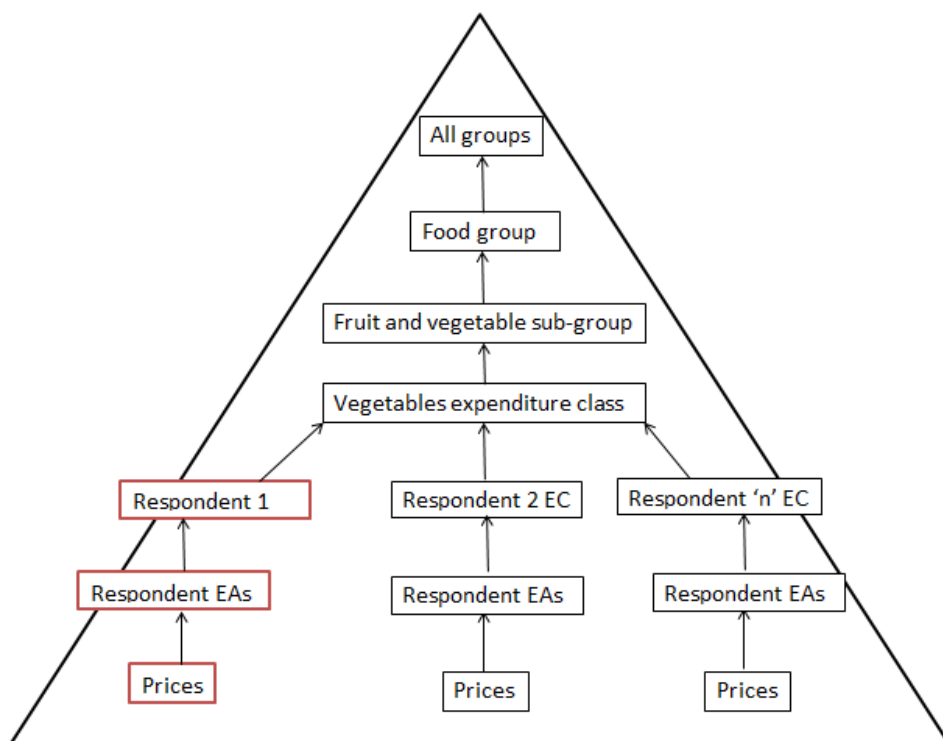




- Output metrics

- Modular approach to decomposition/editing

- Allows the ability to drill down from broad to fine levels of index classification



## Monitoring systems of new methods



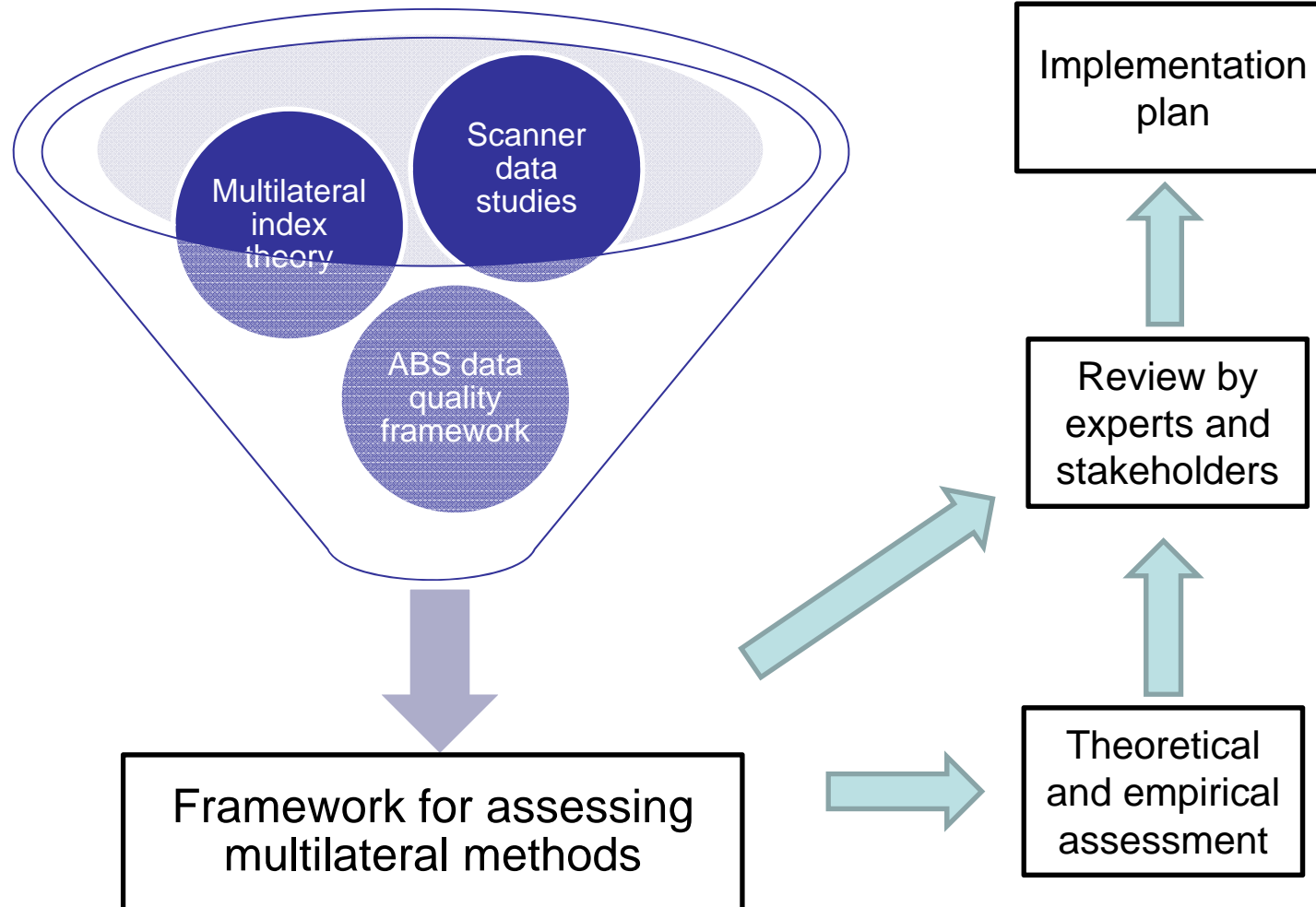
- Summary
- The use of large numbers of products for index compilation requires re-thinking traditional editing strategies
- NSIs (e.g. CBS, SNZ, ABS) have used dashboard type outputs to monitor the quality of input and output data
- Questions?



- The implementation of new methods for CPI compilation is a major change for any NSI – this is something the ABS is currently experiencing with new methods being introduced from December 2017
- This sub-section provides a summary of some implementation considerations including:
  - Choice of method
  - Stakeholder consultation
  - Other implementation considerations



- Choice of method:
- Use of multilateral methods with new data sources is active field of international research
- Format of previous papers:
  - Present multilateral index as generalisation of preferred bilateral index
  - Refine to address issues observed / anticipated in real data
  - Present empirical comparisons (emphasise own innovations)
- Challenges for NSOs seeking a robust and future-proof method:
  - Assess importance/influence of each issue as best we can
  - Anticipate and/or respond positively to new developments





Aspect	Considerations
Resources	<ul style="list-style-type: none"> <li>• Can we scale up the amount of information used without scaling up manual effort?</li> </ul>
Theoretical justification	<ul style="list-style-type: none"> <li>• Axiomatic (test) approach</li> <li>• Economic approach</li> </ul>
Flexibility	<ul style="list-style-type: none"> <li>• How well can the method make use of datasets with more or less auxiliary information?</li> </ul>
Transitivity	<ul style="list-style-type: none"> <li>• Do direct and indirect price comparisons between two periods yield the same result?</li> </ul>
Characteristicity	<ul style="list-style-type: none"> <li>• How much is the price change between two periods influenced by prices in other periods?</li> </ul>
Interpretability	<ul style="list-style-type: none"> <li>• How easy is it to understand the methods conceptually?</li> <li>• How easy is it understand what is driving price movements?</li> </ul>



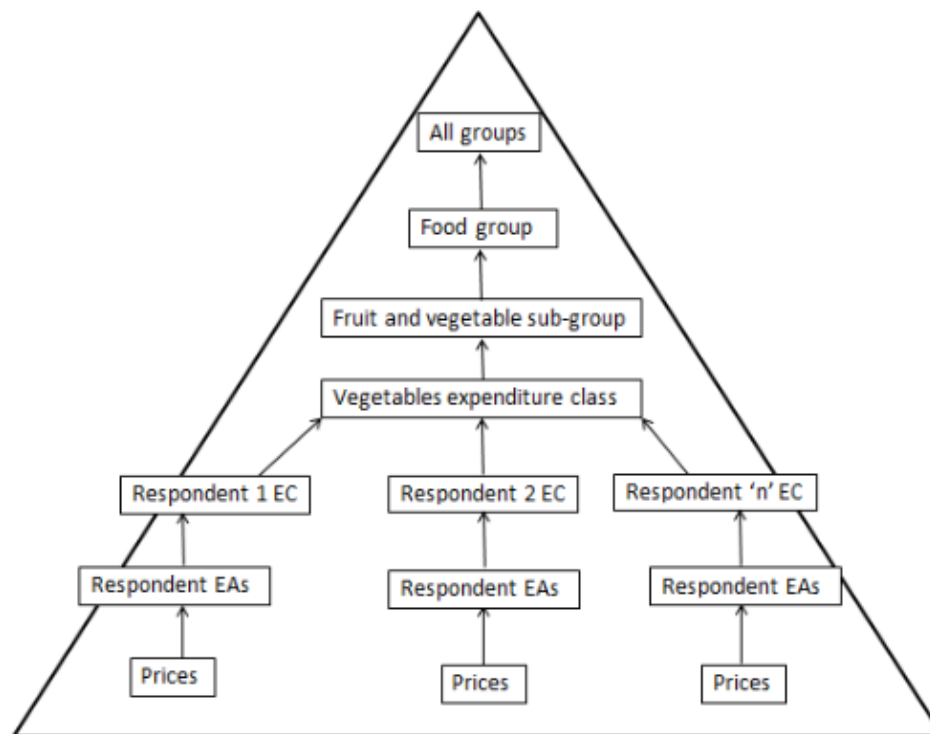


- Choice of method:
- Multilateral and extension methods assessed both theoretically and empirically
- Theoretical assessment helpful for
  - Contrasting bilateral and multilateral approaches
  - Understanding similarities and differences between multilateral methods
  - Anticipating and mitigating possible criticisms of specific choices
- Less helpful for
  - Assessing extension methods
  - Developing a clear ranking
  - Confirming which aspects are most influential (empirical assessment) and important (expert review)



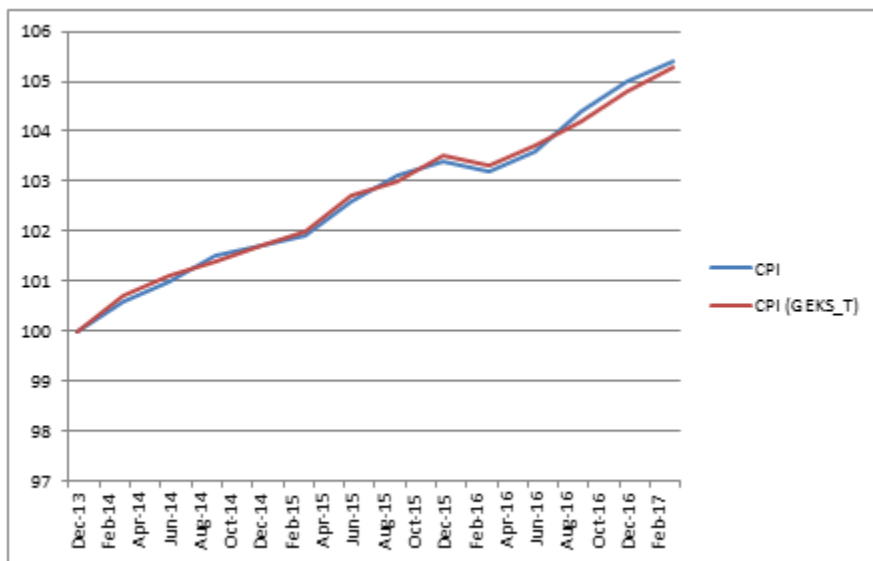
Modified aggregation structure compared for 28 ECs in the CPI

- Respondents weighted by market/expenditure share to produce published level indexes

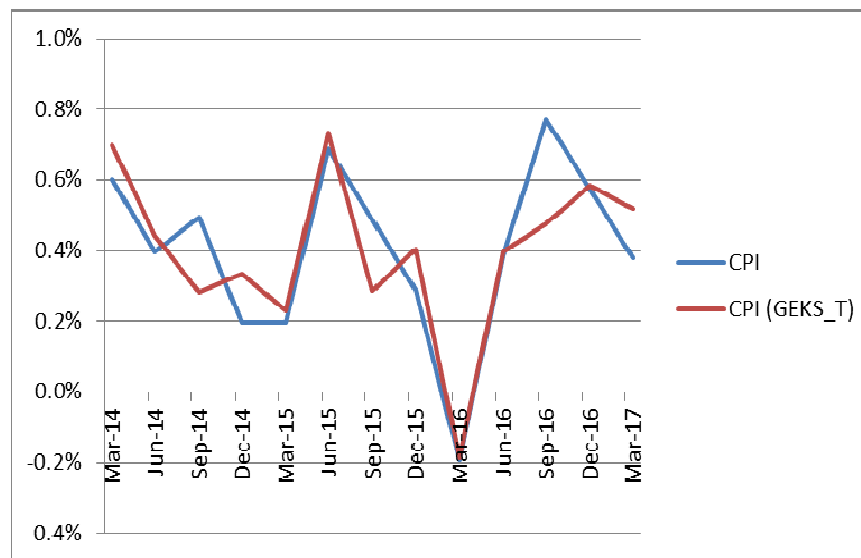




### All groups price index

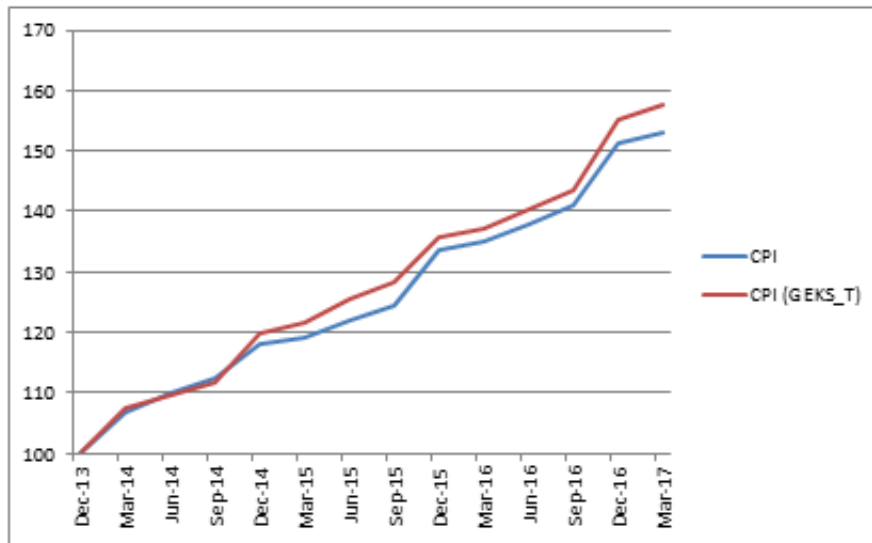


### All groups quarterly percentage change

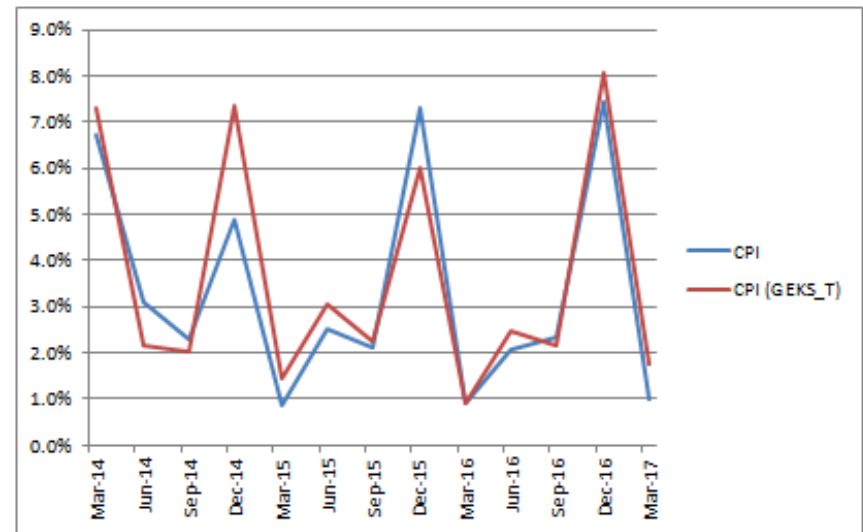




### Tobacco price index

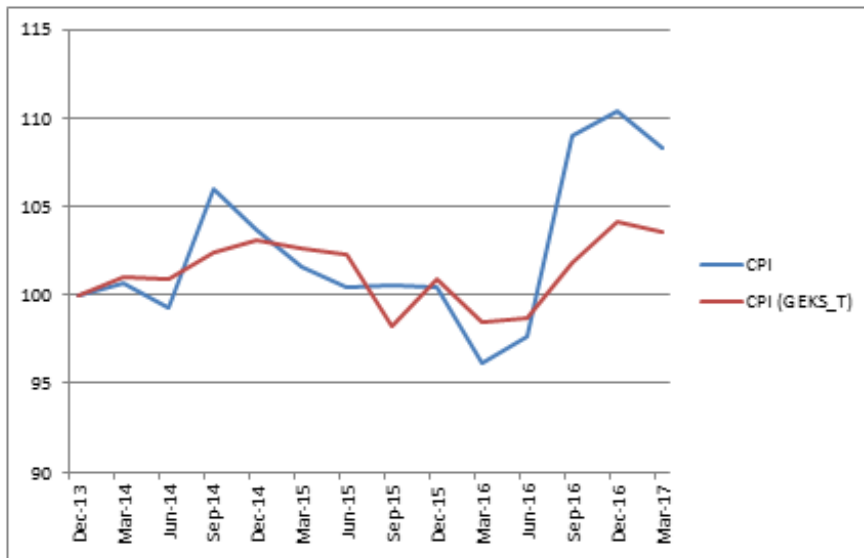


### Tobacco quarterly percentage change

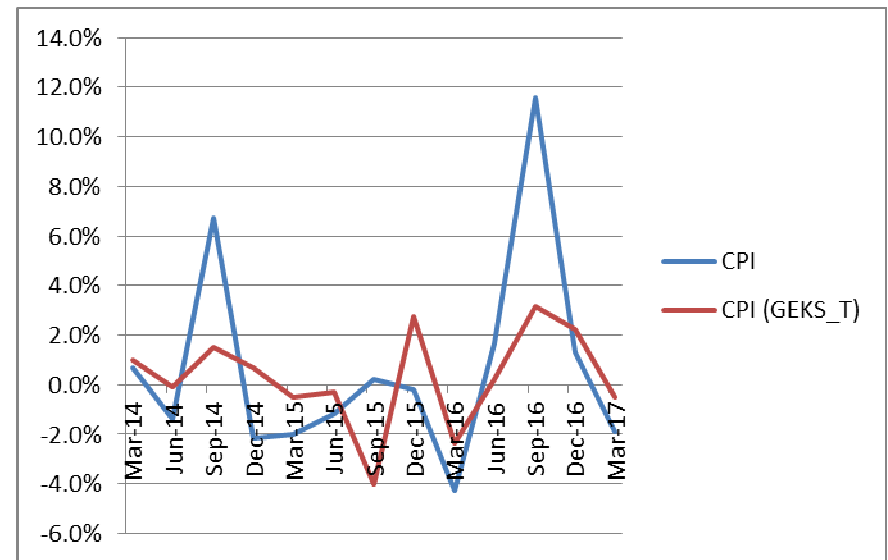




### Fruit and vegetables price index



### Fruit and vegetables quarterly percentage change





### Stakeholder consultation:

- Release of two information papers (November 2016, June 2017) seeking public submissions
- Peer review of index methods by two international price index experts
- Collaboration with other National Statistical Offices (e.g. New Zealand, Netherlands)
- Bilateral/multilateral workshops with government (e.g. central bank, Treasury)
- Presentations at relevant forums (domestically and internationally)





Based on stakeholder consultation and international expert peer reviews, [ABS \(2017\)](#) recommended the following methods for implementation

- GEKS-Törnqvist as preferred multilateral method
- Aggregate below the EC level using respondent classes
- Aggregate respondent classes together using Törnqvist index formula
- Mean splice with a rolling window of 9 quarters

Timetable for implementation:

- ABS to refine methods for June quarter 2017 and September quarter 2017 whilst parallel processing
- ABS to implement new methods in December quarter 2017



- Other implementation considerations:
- Data provided by respondents in raw format for their own reporting purposes
- ABS required to make practical decisions on using this data for price indexes. Some of these include:
  - Re-thinking of CPI aggregation structure, editing and reporting strategies
  - Cleaning of unusual/clearance prices (especially important for Törnqvist based indexes)
  - Strategies to deal with product churn (e.g. product entries, exits). Can different SKUs be considered the same item ('relaunched')?





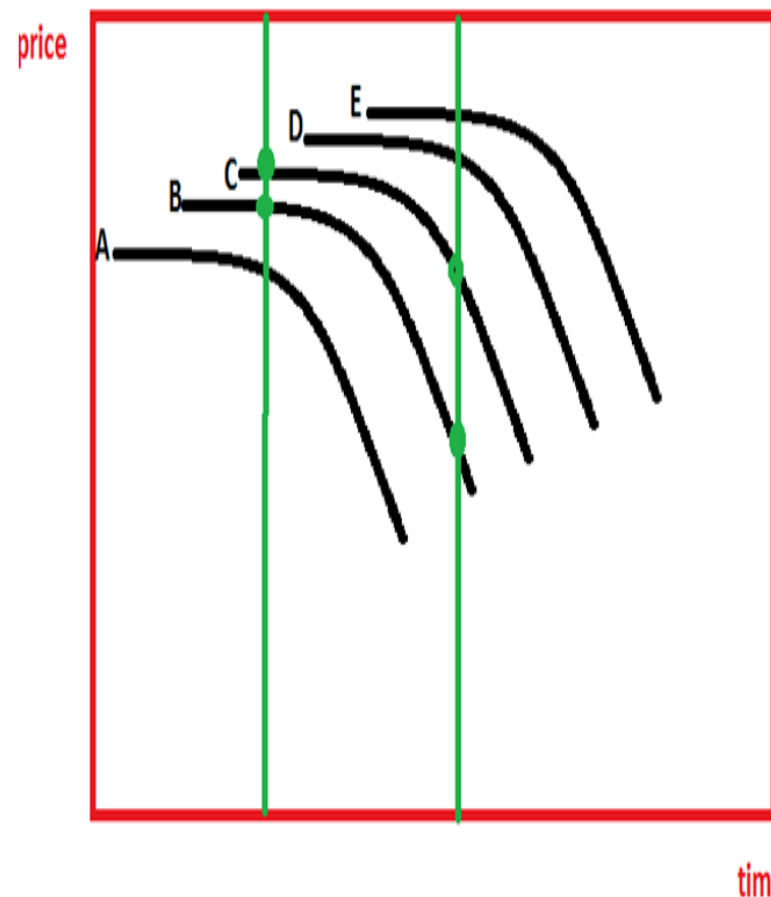
- Ongoing and future avenues of research:
- Methods for using new data sources in price indexes is an active area of research internationally
- ABS can continue to contribute and learn from others – we have a few avenues in mind:
  - Weighted GEKS
  - Hedonic indexes
  - Feature selection and extraction
  - Methods for aggregating datasets of online prices



- GEKS gives all bilateral comparisons equal weight
- In practice, some comparisons may be more important or reliable than others due to attrition/churn or seasonality
- Not yet a well established index method - needs more study before we'd recommend its use in official statistics



- We usually track the price changes of individual products
  - Sometimes within-product price changes may not reflect overall price changes
- Traditional approach: try to find a replacement for a disappearing product (not always feasible for dynamic sampling)
- Hedonic approach
  - represent products as bundles of price-determining characteristics (brand, type, flavour, size etc) and include these in a model
- Captures price changes when new products have features in common with existing products
- Feasibility depends on availability of these characteristics...





- Different datasets have different information about each product's characteristics (brand, type, flavour, size etc)
  - A. Some have separate data items for each characteristic
  - B. Others have a terse description in a text string
  
- Characteristic information useful for
  - Accounting for price and/or quality change, especially as product ranges are refreshed (e.g. via hedonics)
  - Product classification
  
- Interested in methods for
  - Extracting characteristic information (esp from text)
  - Deciding which characteristics are important
  - Some work has already been done on this at ABS and other agencies



Scanner data described so far are extracted from vendors' stock management systems

Alternatively, prices can be collected from websites:

- Manual collection
- Brought together by other companies (aggregators)
- Collected ourselves via automated tools - CPI are starting to develop “web crawlers”

Key differences include

- Doesn't require companies to actively supply data to the ABS (+)
- No information on volumes sold (-)

ABS is starting to venture into the collection space

- Interested in methods for aggregating large volumes of this data (if or when available)
- Other countries/researchers are making some progress



## Conclusion



- Traditional bilateral index formula can ‘break down’ when applied to scanner data – multilateral methods seen as the best solution
- Multilateral method requires an extension method for production (rolling or expanding window)
- Choice of window length – at least one year. Other options can be considered
- What is the preferred multilateral method? Depends on the characteristics of input data. Preferred multilateral method still debated internationally – useful to consider in local context
- Empirical results typically show little difference between different multilateral methods, more difference in extension method