



UN GWG on Big Data – Scanner Data

Michael Smedes, Task Team on Scanner Data



Task Team on Scanner Data

- ▶ **Aim: Increase the effective use of scanner data in official statistics...**
 1. ..through lowering the barriers of entry for countries by providing a library of methods, guidance and training
 2. ..via the sharing of experience, practice and learning between countries on the use of scanner data
 3. ..and through supporting Public-Private collaboration in the acquisition and use of scanner data



Task Team on Scanner Data

► Deliverables:

1. Delivery of a tool hosted on the UN Global Platform for analysis, monitoring and index estimation using historic scanner data from Nielsen
2. Accompanying training and instructional material on the use of the tool
3. Accompanying methodological guidance material summarising and referencing to literature, recommendations and cataloging good practice



New data = new methods?

- ▶ These new data sources allow for different index methods to be used
- ▶ E.g, there are index methods that allow for new products to enter the market during the year
- ▶ Scanner data also gives us an opportunity to change the expenditure weights over time



But how do we choose?

- ▶ Inflation statistics are often one of the most high-profile releases that an NSI produces
- ▶ Changing the methodology and data source requires a lot of research and analysis before an NSI can decide on a final plan of implementation



Instructional Guide

Scanner Data Methods: Instructional Guide

Table of contents

0.	Scope of this document.....	
1.	Introduction and literature.....	
2.	A concise overview of index methods.....	
2.1	Preliminary considerations.....	
2.2	Bilateral index methods.....	
2.3	Multilateral index methods.....	
2.3.1	The GEKS method.....	
2.3.2	The CCDI method.....	
2.3.3	The Geary-Khamis method.....	
2.3.4	Time Product Dummy method.....	
2.3.5	Hedonic method.....	
2.4	Updating methods.....	
2.5	Other choice aspects.....	
3.	Discussion: What methods can be used when?.....	
3.1	Index methods.....	
3.2	Updating methods.....	
4.	The need for monitoring and analysis.....	
5.	Case studies.....	
5.1	CPI production.....	
5.2	Research.....	
	References.....	



Case studies

- ▶ Many NSIs are starting to research the feasibility of using these data in a production environment
- ▶ The instructional guide is a useful summary of current literature and includes case studies of NSIs who have implemented the data in production



UN Global Platform

- ▶ The aim of this tool is to allow NSIs access to tested index method code, and to practise using the different methodologies on some training data



Nielsen data

- ▶ 5 Product Groups -milk, cauliflower, water, charcoal,
- ▶ Canada, weekly data June 2015 to June 2018
- ▶ Key variables: unique ID, product name, salesunits (quantity), salestonnage or saleslitres (quantity), salesCAD (expenditure)
- ▶ Derive a unit price for each product



Demo of using the UN Global Platform – initial analysis

Secure | <https://developers.officialstatistics.org/projects/detail/a0-b34c0c593ff04a66a48ebd001a92c08a/view>

Projects Deployments Channels

Projects / Scanner Data (Restricted Access) / View

jupyter

Files Running Clusters

Select items to perform actions on them. Upload New Refresh

<input type="checkbox"/>		Name	Last Modified
<input type="checkbox"/>	Folder	tmp	3 days ago
<input type="checkbox"/>	File	Analysis of cauliflower data.ipynb	Running 4 minutes ago
<input type="checkbox"/>	File	Analysis of milk data.ipynb	an hour ago
<input type="checkbox"/>	File	Data wrangling FEWS (cauliflower).ipynb	an hour ago
<input type="checkbox"/>	File	Data wrangling FEWS (milk).ipynb	an hour ago
<input type="checkbox"/>	File	R notebook FEWS (cauliflower).ipynb	an hour ago
<input type="checkbox"/>	File	R notebook FEWS (milk).ipynb	an hour ago
<input type="checkbox"/>	File	20180926_cauli.zip	3 days ago
<input type="checkbox"/>	File	anaconda-project.yml	3 days ago
<input type="checkbox"/>	File	CAULIFLOWER.csv	2 hours ago
<input type="checkbox"/>	File	data for FEWS cauli.csv	an hour ago



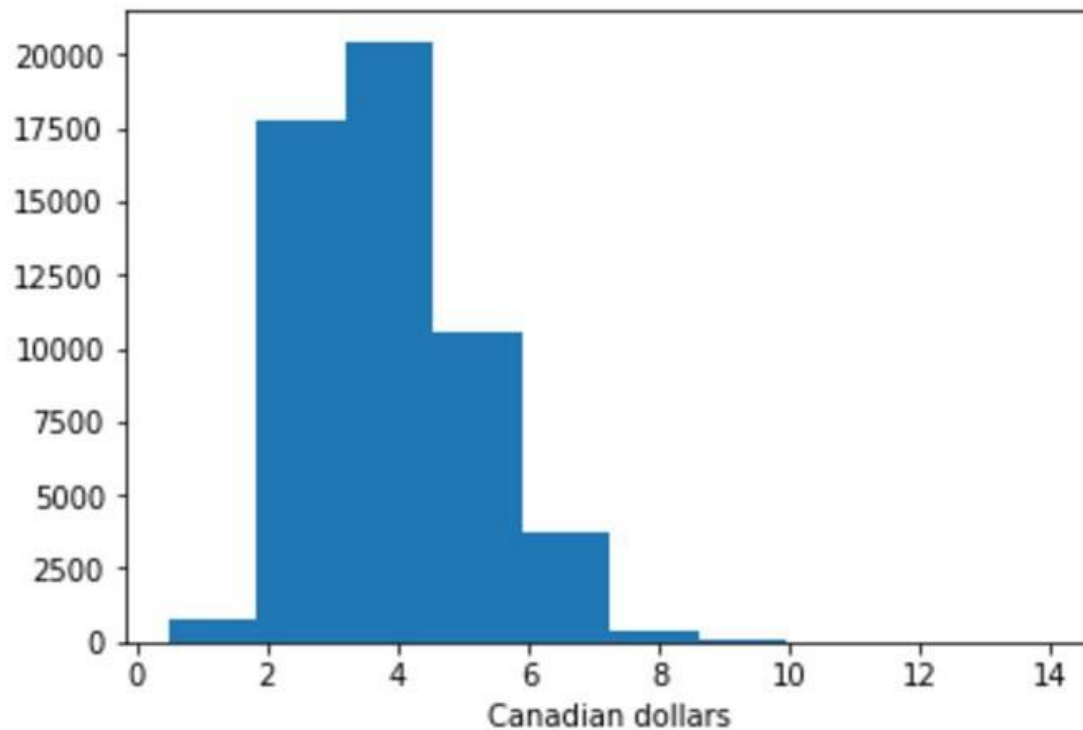
Nielsen data

- ▶ Churn (cauliflower) -
 - ▶ 40 unique products at the beginning of the period
 - ▶ 46 unique products at the end
 - ▶ 35 products remained in the sample over the 3 years



Monitoring the data

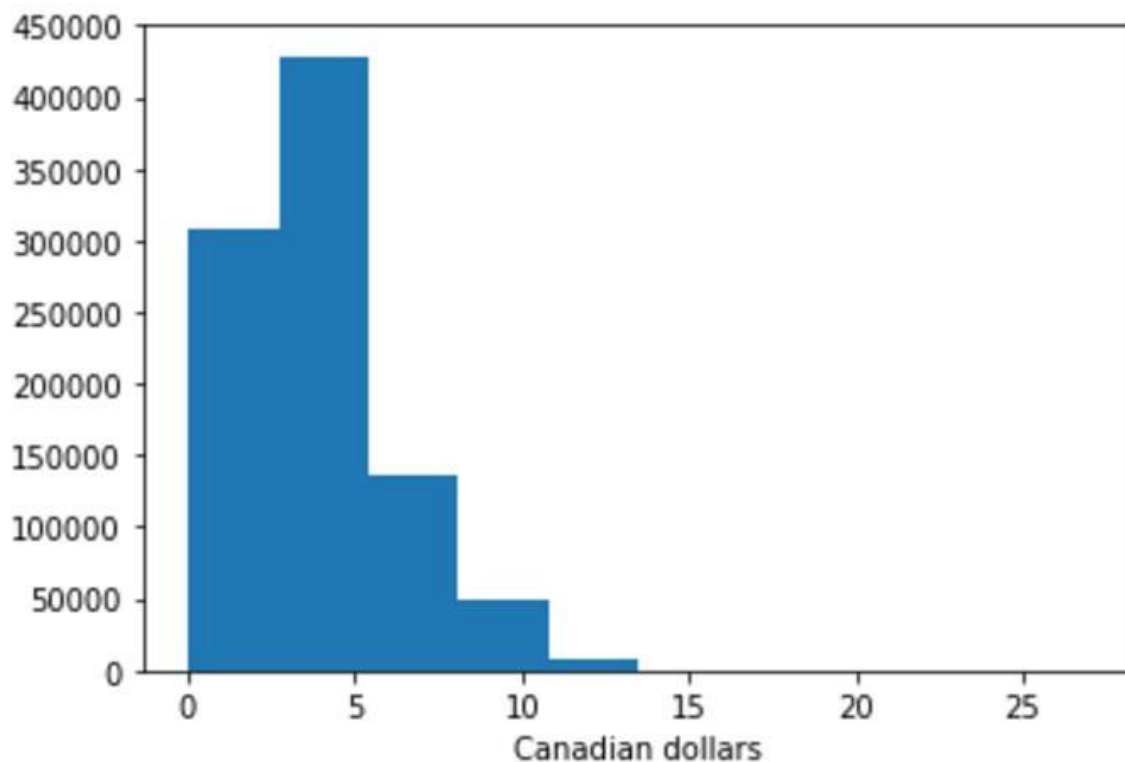
- ▶ Histogram of unit values (cauliflower)





Monitoring the data

► Histogram of unit values (milk)





Demo of using the UN Global Platform – FEWS index

Secure | <https://developers.officialstatistics.org/projects/detail/a0-b34-c0c593ffd4a66a48ebd001a92c08a/view>

Projects Deployments Channels

Projects / Scanner Data (Restricted Access) / View

jupyter

Select items to perform actions on them. Upload New Refresh

<input type="checkbox"/>	Name	Last Modified
<input type="checkbox"/>	tmp	3 days ago
<input type="checkbox"/>	Analysis of cauliflower data.ipynb	42 minutes ago
<input type="checkbox"/>	Analysis of milk data.ipynb	2 hours ago
<input type="checkbox"/>	Data wrangling FEWS (cauliflower).ipynb	2 hours ago
<input type="checkbox"/>	Data wrangling FEWS (milk).ipynb	2 hours ago
<input type="checkbox"/>	R notebook FEWS (cauliflower).ipynb	Running 2 hours ago
<input type="checkbox"/>	R notebook FEWS (milk).ipynb	2 hours ago
<input type="checkbox"/>	20180926_cauli.zip	3 days ago
<input type="checkbox"/>	anaconda-project.yml	3 days ago
<input type="checkbox"/>	CAULIFLOWER.csv	44 minutes ago
<input type="checkbox"/>	data for FEWS cauli.csv	2 hours ago
<input type="checkbox"/>	data for FEWS milk.csv	2 hours ago



Price indices - cauliflower

Index June 2016 = 100





Price indices - milk

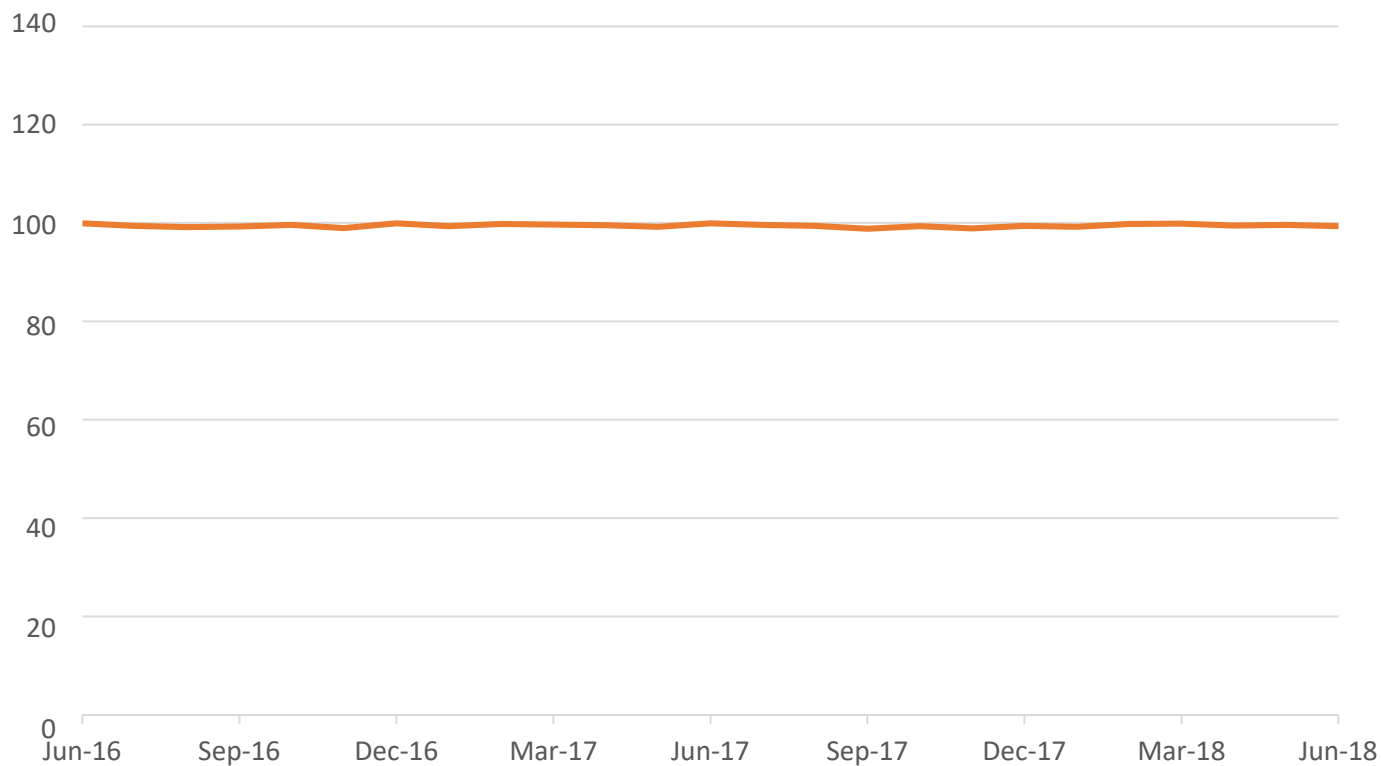
Index June 2016 = 100





Price indices - milk

Index June 2016 = 100





Lessons learnt so far - Platform

- ▶ Technology platform provides easy to use interface
- ▶ Availability of a range of trusted index methodology expedites learning
- ▶ The set-up of the platform requires a learning curve for the task team participants.



Lessons learnt so far – Nielsen

- ▶ Comprehensive clean data
- ▶ Value in establishing a partnership to establish best practices
- ▶ Adding a data dictionary would facilitate data understanding



Next steps

- ▶ Finalised Phase 1:
 - ▶ Instructional manual loaded to the platform
 - ▶ Expanded test data stored on UN Global Platform
 - ▶ Code up additional index methods to allow for testing
 - ▶ Training workshops conducted in Bogata, Colombia and Kigali, Rwanda



Next steps

▶ Commence Phase 2:

- ▶ Using Scanner Data to calculate CPI expenditure weights
- ▶ Data cleaning and sorting to provide analysis ready dataset
- ▶ Capacity Building - providing training material and courses
- ▶ Expanding membership to include new organisations: US Fed Reserve; Eurostat; Nielsen; Amazon