Methods for calculating price indices from alternative data sources: CLIP and other animals

Matt Mayhew
Senior Public Policy Analyst
Policy Evidence and Analysis Team

21 May 2019

Main Topics

- Product Definition
- Index Methods
- Expenditure Proxies

Product Definition

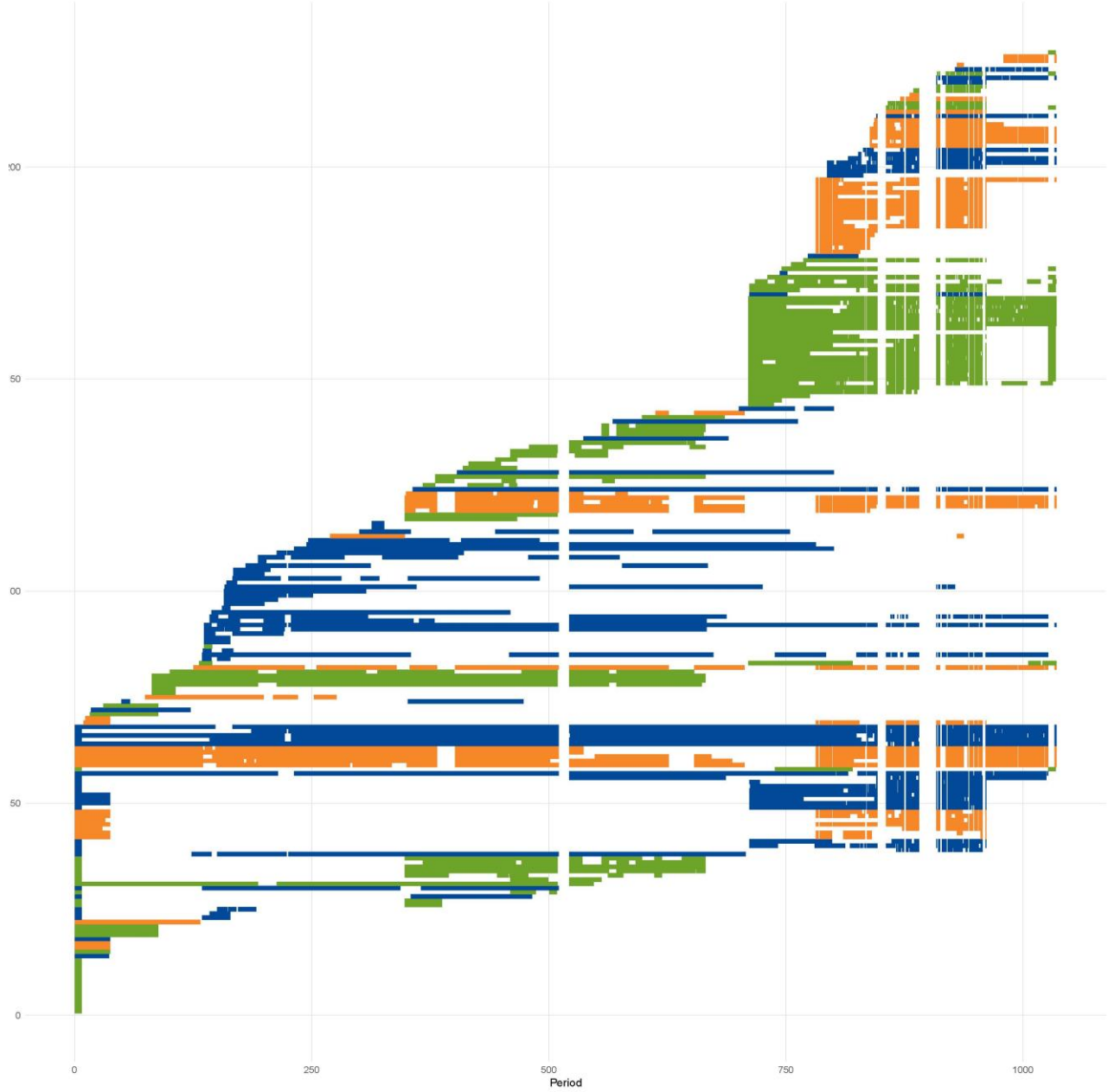
The problem of product churn

Product Churn

Product churn is the process of products leaving and/or entering the sample.

1. Product goes out of stock, temporally leaves the sample
2. Product is restocked, and re-enters the sample
3. Product is discontinued
4. Product is new to the market
5. Product is rebranded

Product Churn Apples



Source: ONS

Problems caused

We are unable to easily track a product across time

Traditional method then fails

Cannot identify comparable replacements easily when products go out of stock

Solution: Track groups of products

Instead of tracking individual products we can track a group of products with similar characteristics.

This allows for products to come in and out of the groups but the group itself will still exist.

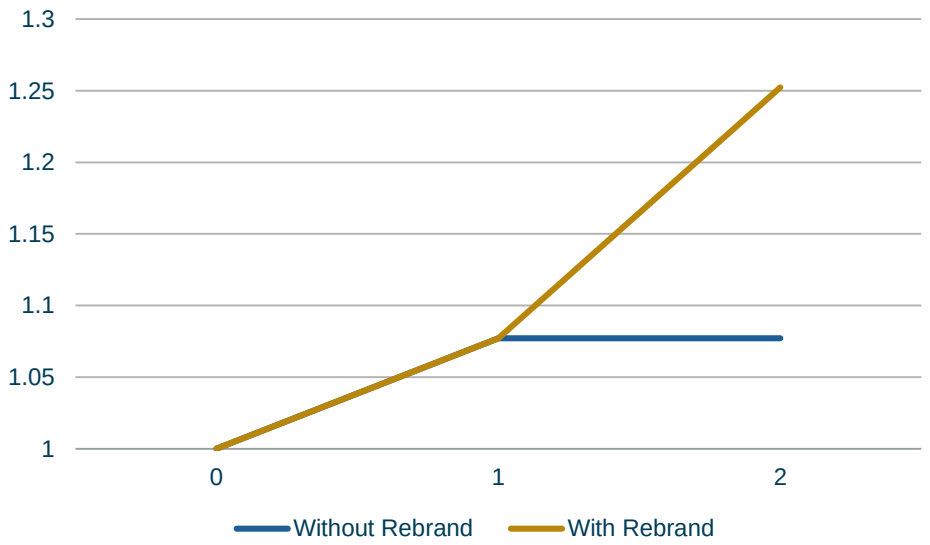
Why are rebrands a problem?

Rebrands are where an item is taken off the market and brought back in a slightly different format.

This is purely cosmetic, but might be introduced at a different price.

Example: Rebranding

Product	0	1	2
A	8	10	10
B	7	5	5
C	5	7	-
D	-	-	11



Product D is a rebrand of product C

Methods of finding groups

CLIP

MARS

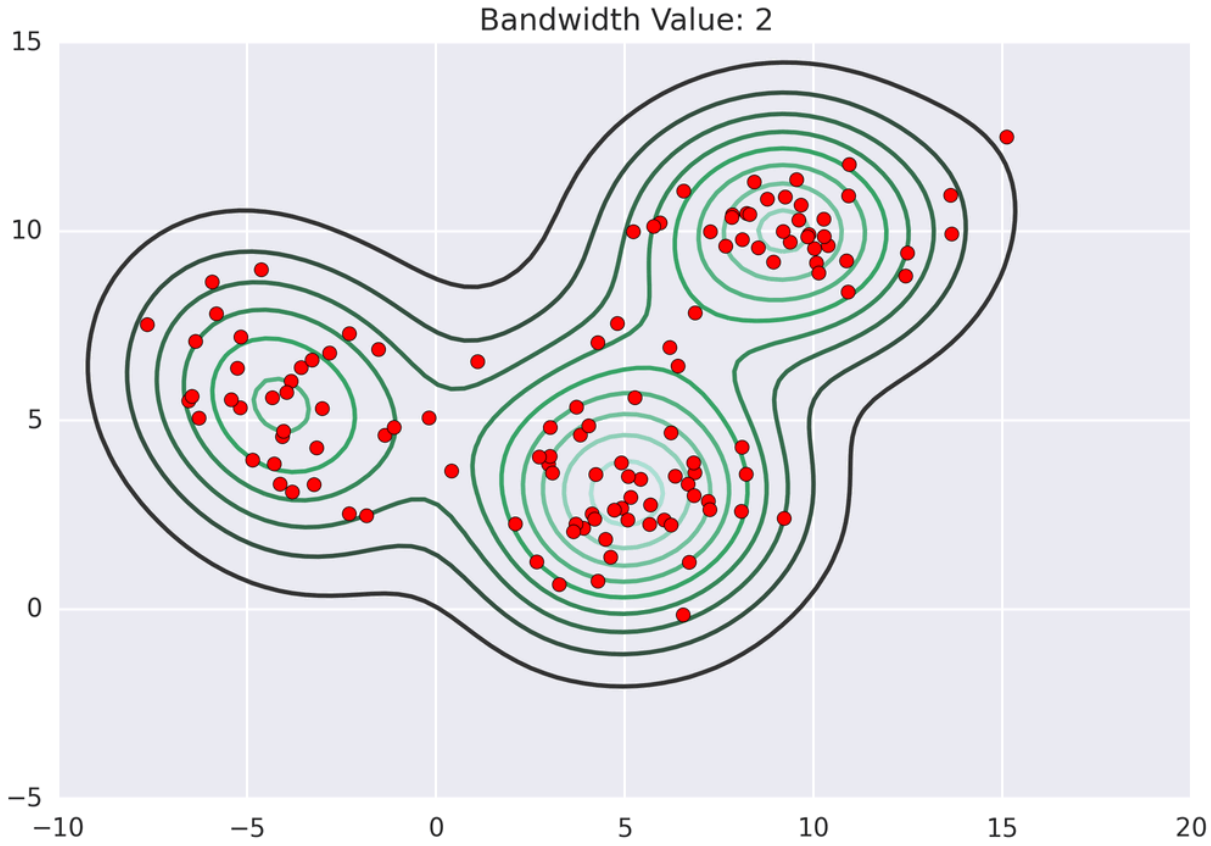
CLIP

Clustering Large datasets Into Price indices (CLIP) finds these groups by density based clustering, (the Mean Shift Algorithm)

Products with similar characteristics will be grouped together.

Mean Shift Clustering

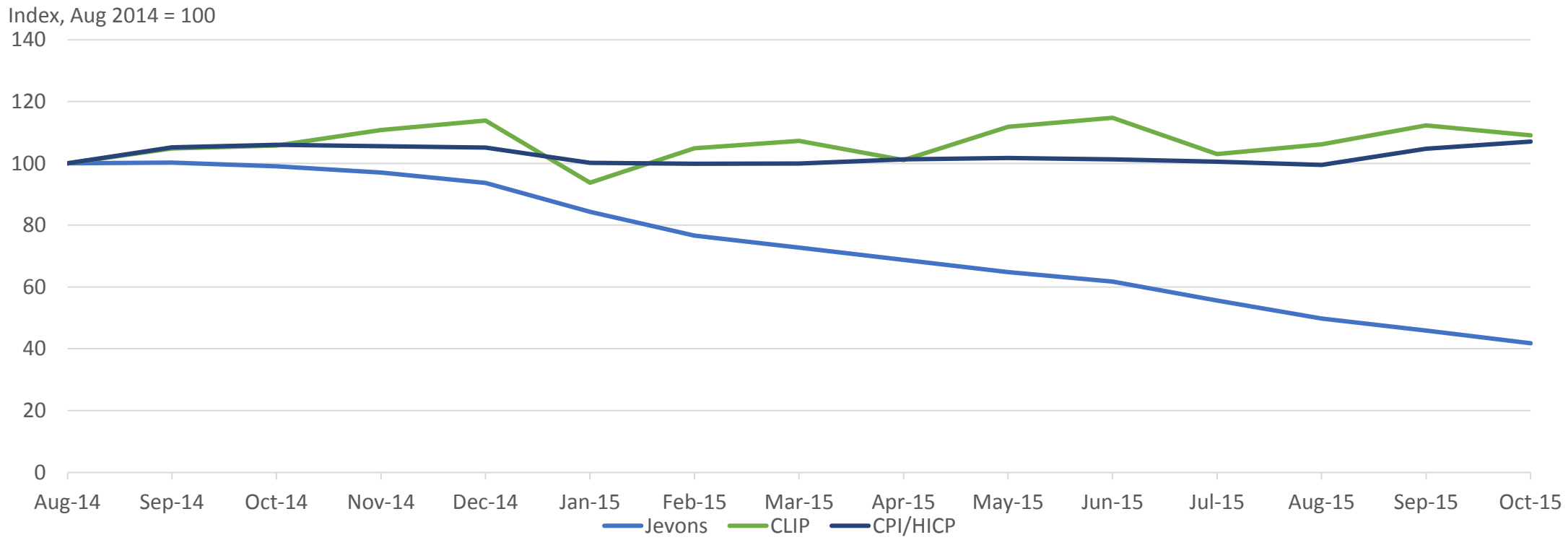
Mean Shift Clustering associates individual products with their nearest local maxima, in the KDE for the joint distribution.



CLIP Method

1. Cluster the products within an item in the base period
2. Use a decision tree to find the “rules” that make up a cluster
3. When new products appear assign them to a cluster using these rules.

All clothing



MARS

Match Adjusted R-squared creates groups by maximising the homogeneity of the groups whilst minimising product churn, subject to the characteristic.

MARS – Product Homogeneity

Product homogeneity is measured by the proportion of the total variance explained by a grouping, an R^2, i.e.

$$\rho^2_{K,t} = \frac{\mathrm{Var}_K(p^t)}{\mathrm{Var}_A(p^t)}$$

K is a specific group, A is all products.

MARS

Product churn is minimised by maximising the overlap of the partition between the two periods. This is measured as follows:

$$\mu_{K,t} = \frac{\text{Sum of quantities of products in K in both 0 and } t}{\text{Sum of quantities of all products in sold in } t}$$

MARS

Each group then gets a MARS score:

$$MARS_{k,t} = \rho_{k,t}^2 \mu_{k,t}$$

Averaging prices

Once a grouping of clusters is found the price of cluster is needed

A geometric mean is taken of the products to calculate this price.

Index Methods

Towards an index framework

Office for National Statistics

Index methods

There are multiple methods that are used to calculate price indices, but it is difficult to say which is the “best”.

This depends on the application.

Criteria that could be used to assess

1. Resources
2. Theoretical Properties
3. Transitivity
4. Characteristicity
5. Flexibility
6. Interpretability
7. Cohesion

The ONS' statistical quality framework will also be used as part of the assessment.

Price Indices for alternative data sources

- Jevons
- GEKS-J
- Geary-Khamis
- Time Product Dummy/Time Dummy Hedonics

Jevons

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

GEKS - J

$$P_{GEKS-J}^{0,t} = \left(\prod_{i=1}^{T} P_J^{0,i} \, P_J^{i,t} \right)^{\frac{1}{T+1}}$$

Geary-Khamis

$$P_{GK}^{0,t} = \frac{\sum_{i=1}^{n} p_i^t q_i^t}{\sum_{i=1}^{n} v_i q_i^t}$$

$$v_i = \sum_{z \in T} \phi_i^z \frac{p_i^z}{P_{GK}^z}$$

$$\phi_i^z = \frac{q_i^z}{\sum_{s \in T} q_i^s}$$

Hedonic Indices

Time Product Dummy hedonic model

$$\ln p_i^t = \alpha + \sum_{j=1}^{t} \delta^j D_i^j + \sum_{k=1}^{n-1} \gamma_k D_{k,i} + \varepsilon_i^t$$

Time Dummy hedonic model

$$\ln p_i^t = \alpha + \sum_{j=1}^{t} \delta^j D_i^j + \sum_{k=1}^{K} \beta_k z_{k,i} + \varepsilon_i^t$$

The index is

$$P_H^{0,t} = \exp \delta^t$$

Extension Methods

The previous methods are estimated on a window.

If this window changes then revisions occur.

We don’t want to revise the CPIH.

Extension methods allow us to extend the series without revisions.

Direct Extension

This method is like the current CPIH methodology:

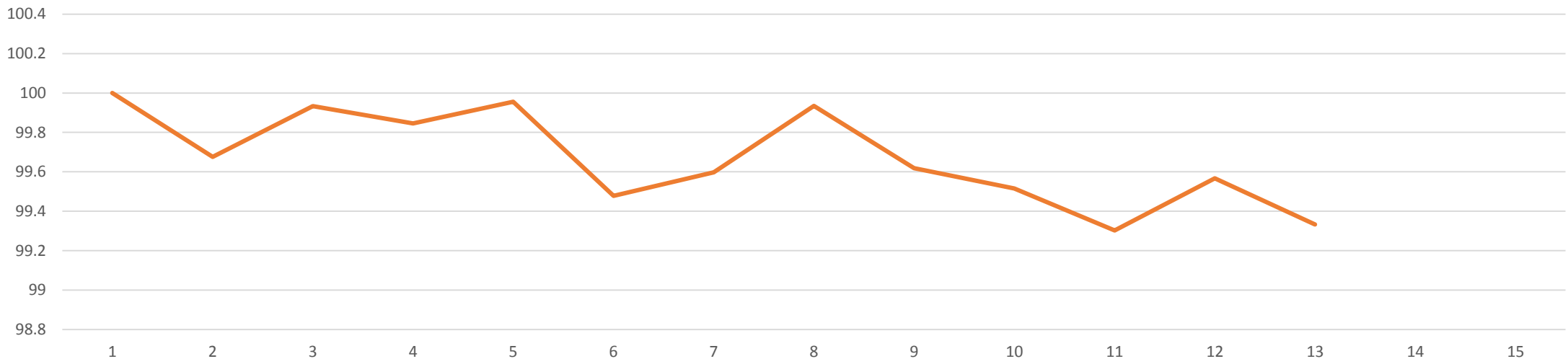
$$P_D^t = P^l \times P^{l,t}$$

Movement Splice

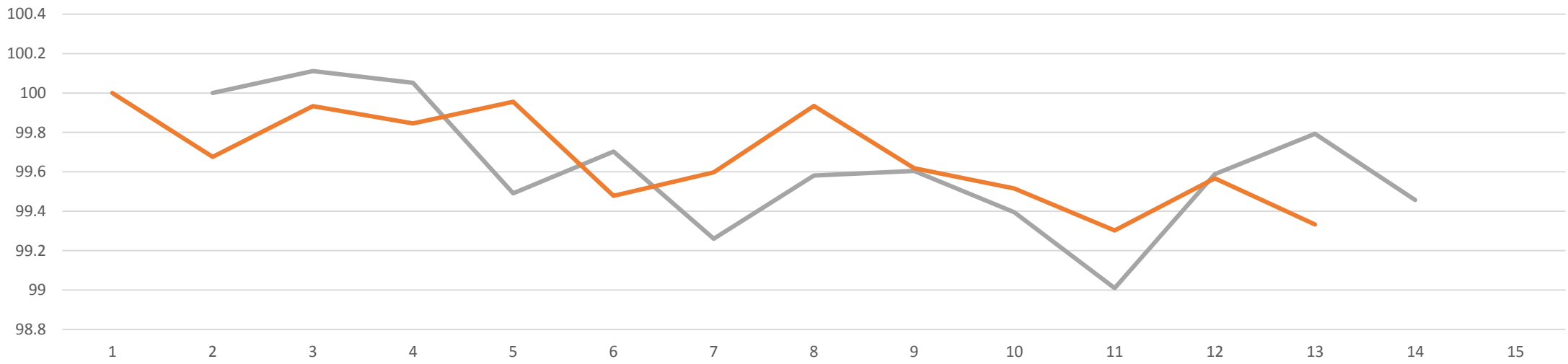
$$P_{MS}^{t} = P^{t-1} \times P_{t-T}^{t-1,t}$$

$P_{t-T}^{t-1,t}$ price movement between *t-1* and *t* using the latest multilateral window between *t-1* and *t.*

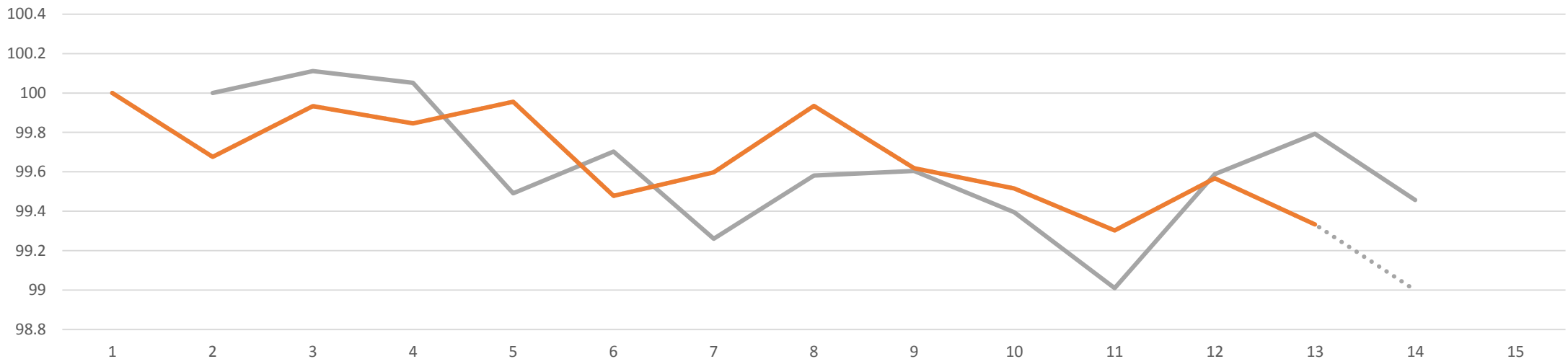
Movement Splice



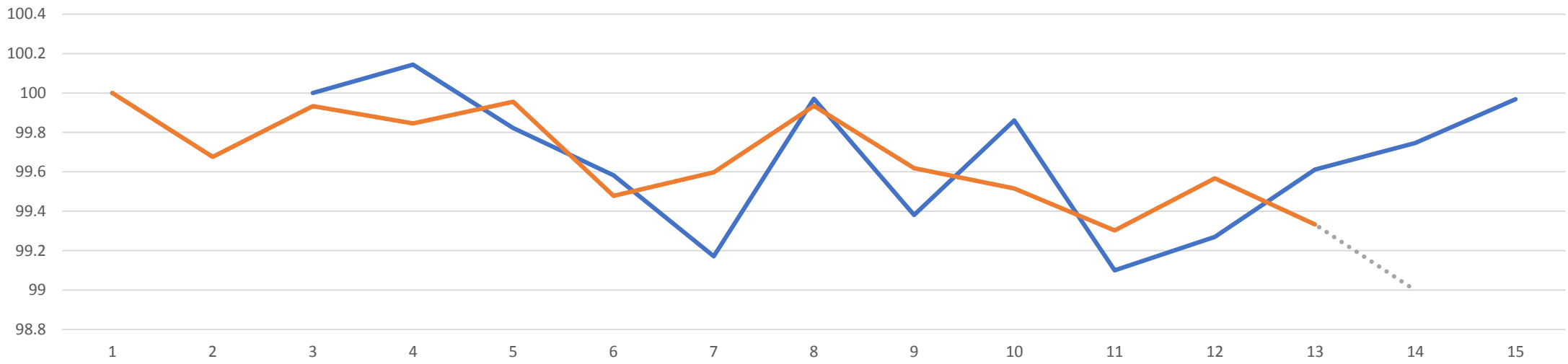
Movement Splice



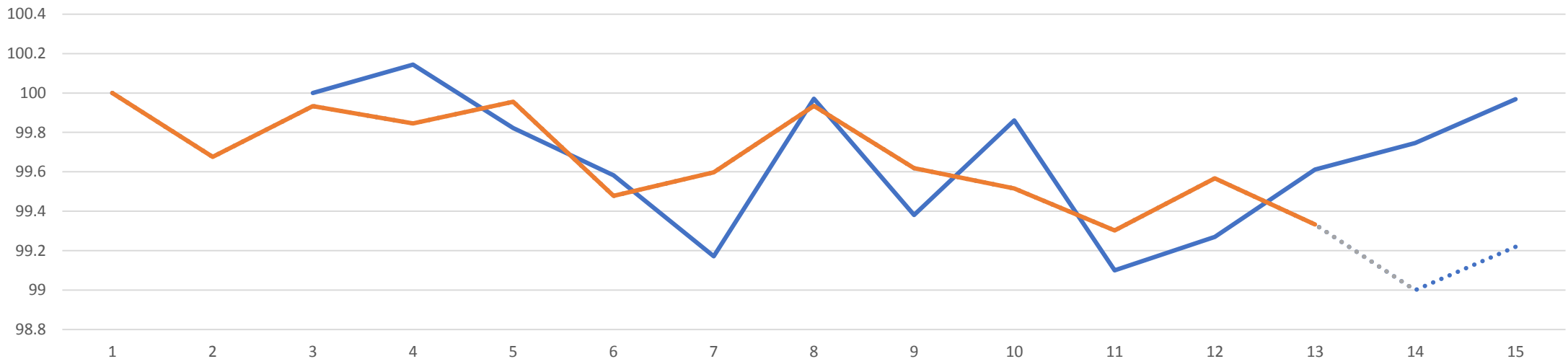
Movement Splice



Movement Splice



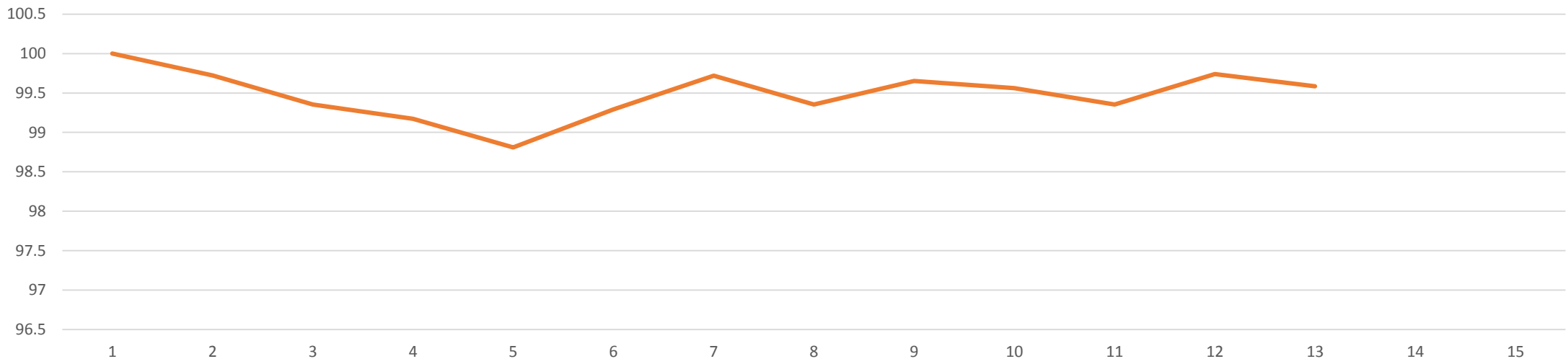
Movement Splice



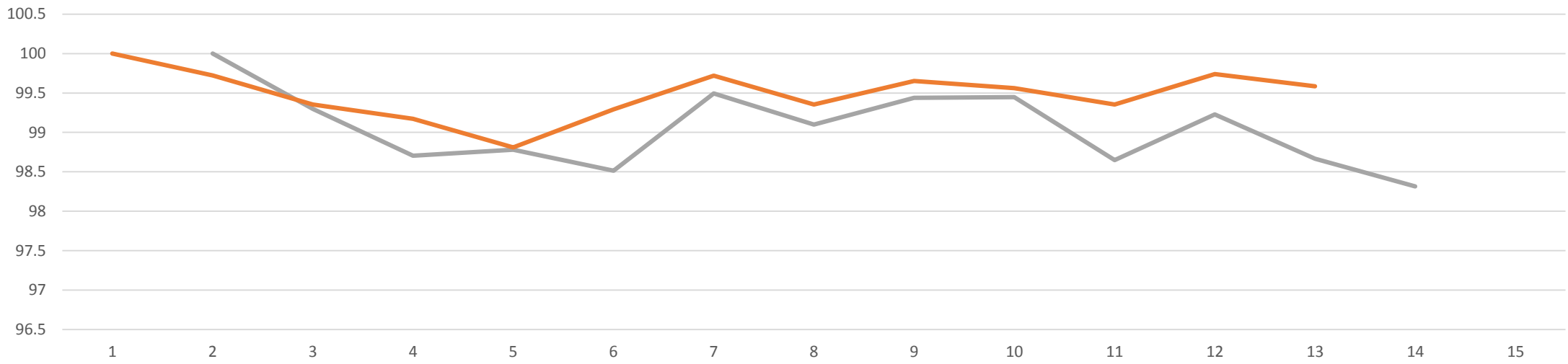
Window Splice

$$P_{WS}^{t} = P^{t-1} \times \frac{P_{t-T}^{t-T,t}}{P_{t-T-1}^{t-T,t-1}}$$

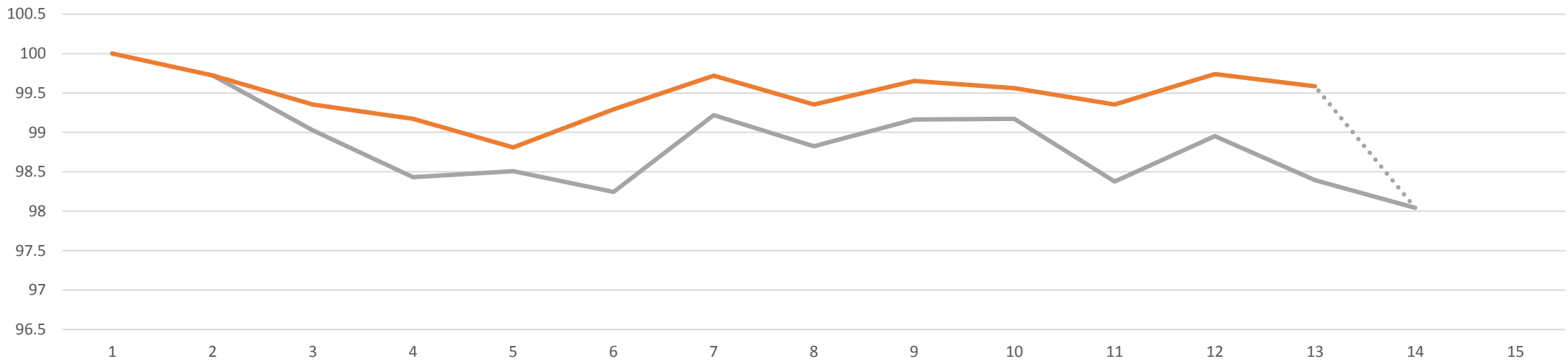
Window Splice



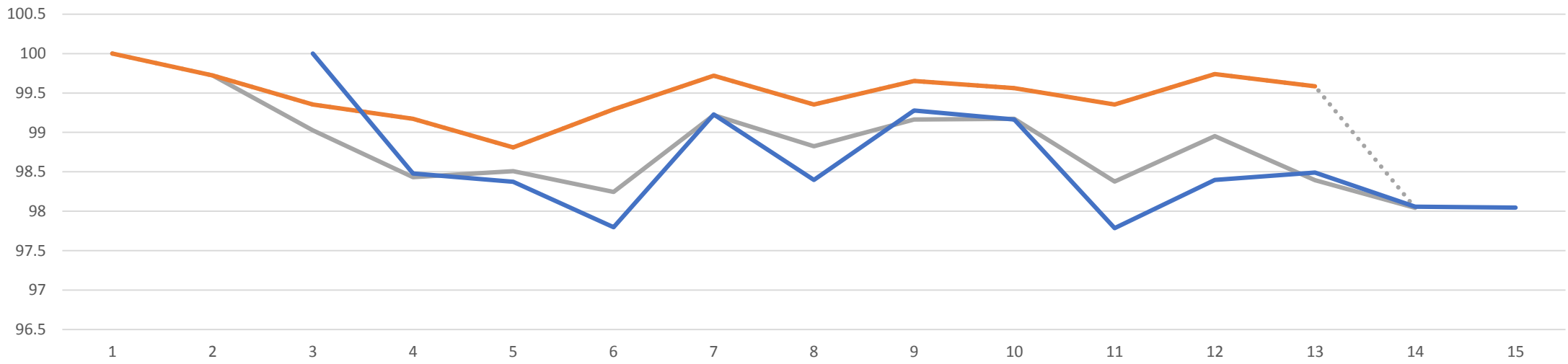
Window Splice



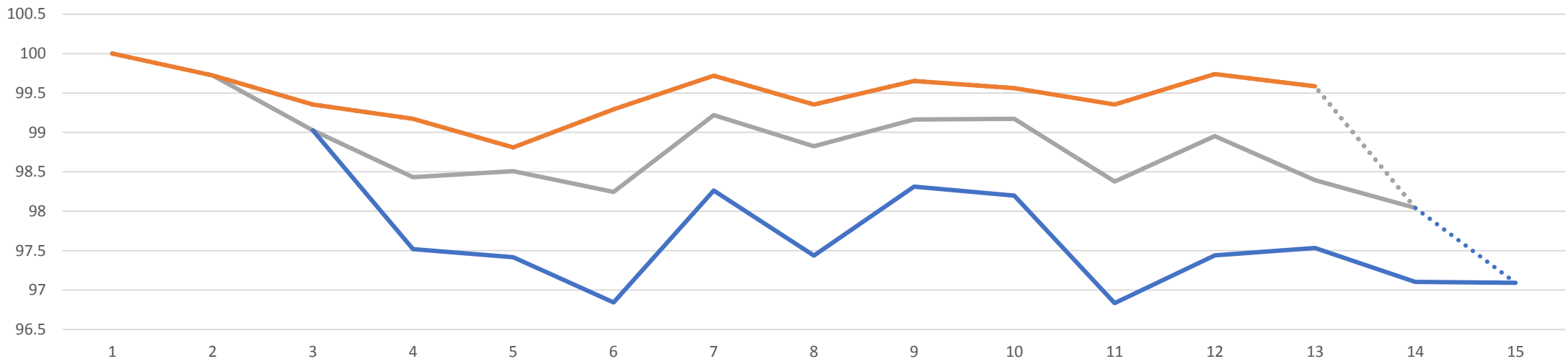
Window Splice



Window Splice



Window Splice



Half Splice

$$P_{HS}^{t} = P^{t-1} \times \frac{P_{t-T}^{t-\frac{T}{2},t}}{P_{t-T-1}^{t-\frac{T}{2},t-1}}$$

Fixed Base Monthly Expanding

$$P_{FBME}^{t} = P^{0,t}$$

Expenditure Proxies

What to do with web scraped data

Office for National Statistics

Lack of quantities

Web scraped data are a richer source of data in terms of coverage and characteristics, but lack quantity information.

Scanner data have quantities, and the manual collection is designed to sample the most representative item.

Lack of quantities mean that products which are not bought have an equal influence on the index

Products ranking

The position of the product in the list of available products does give information about relative popularity.

Could this be used as a proxy for expenditure or quantities?

Transformation of ranks

Three options to convert ranks to expenditures shares

Option 1:

$$w_{i,1} = \frac{2}{n}\left(1 - \frac{r_i}{n+1}\right)$$

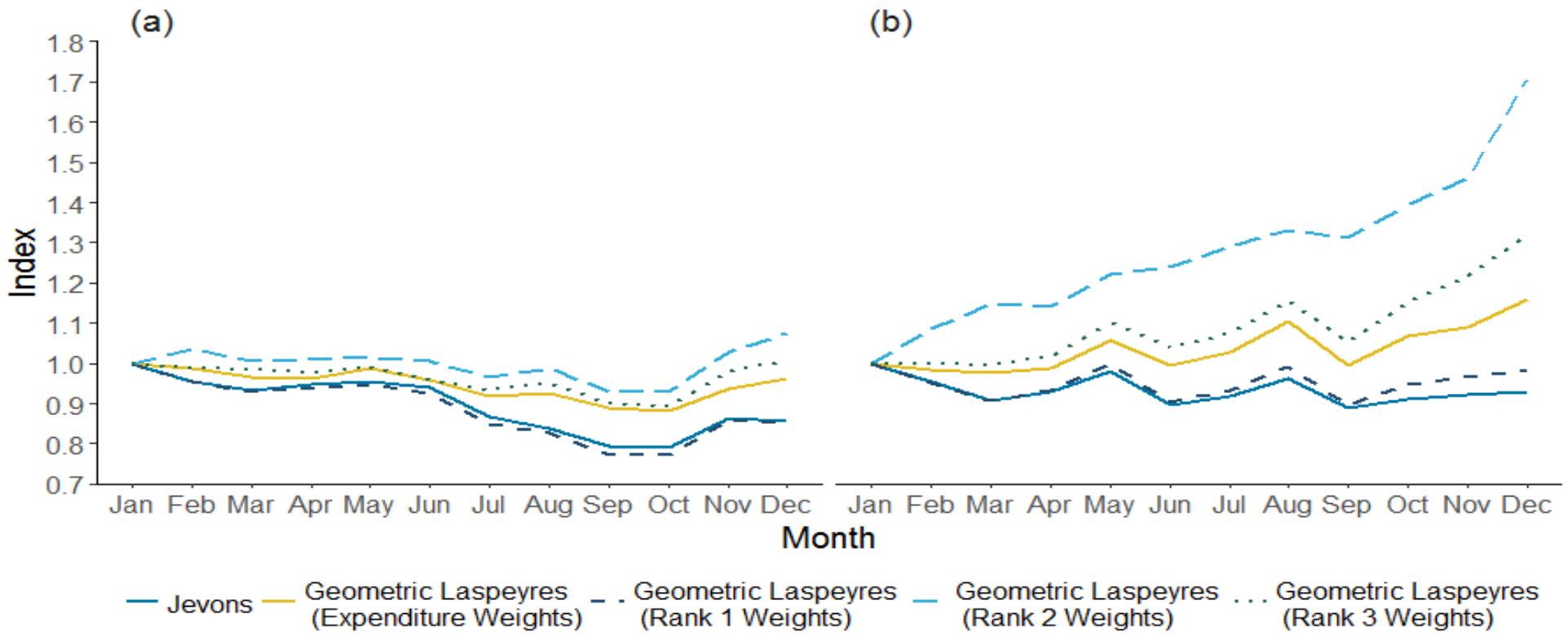
Option 2:

$$w_{i,2} = \frac{1}{\sum_{i=1}^{n} \frac{1}{r_i}}\left(\frac{1}{r_i}\right)$$

Option 3:

$$w_{i,3} = \frac{s_i^x}{\sum_{i=1}^{n} s_i^x} \qquad s_i = \frac{r_i}{\sum_{i=1}^{n} r_i}$$

Rank 3 Weights are closest to expenditure weights (x=6)



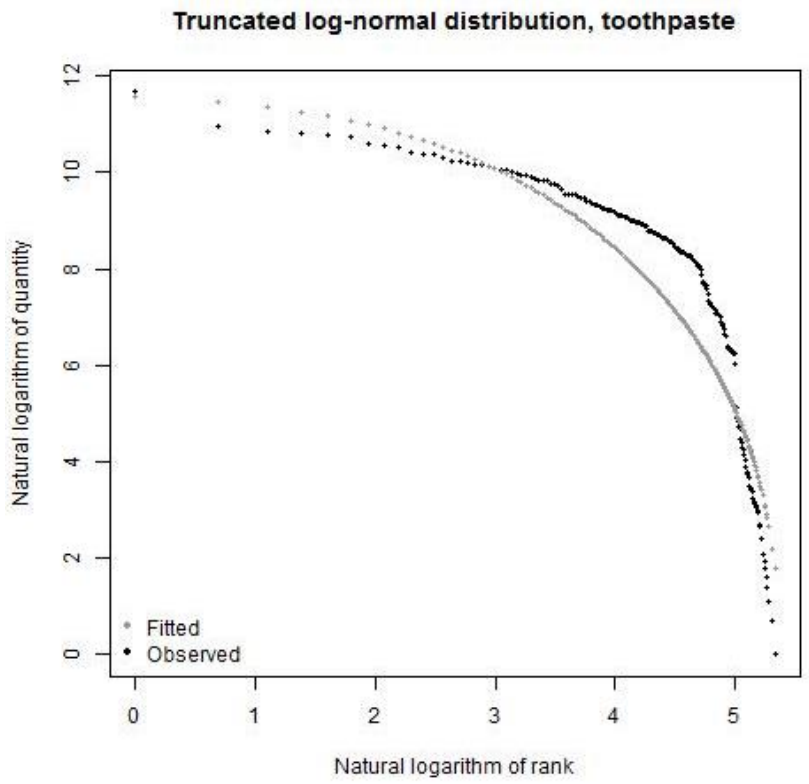
Distributions

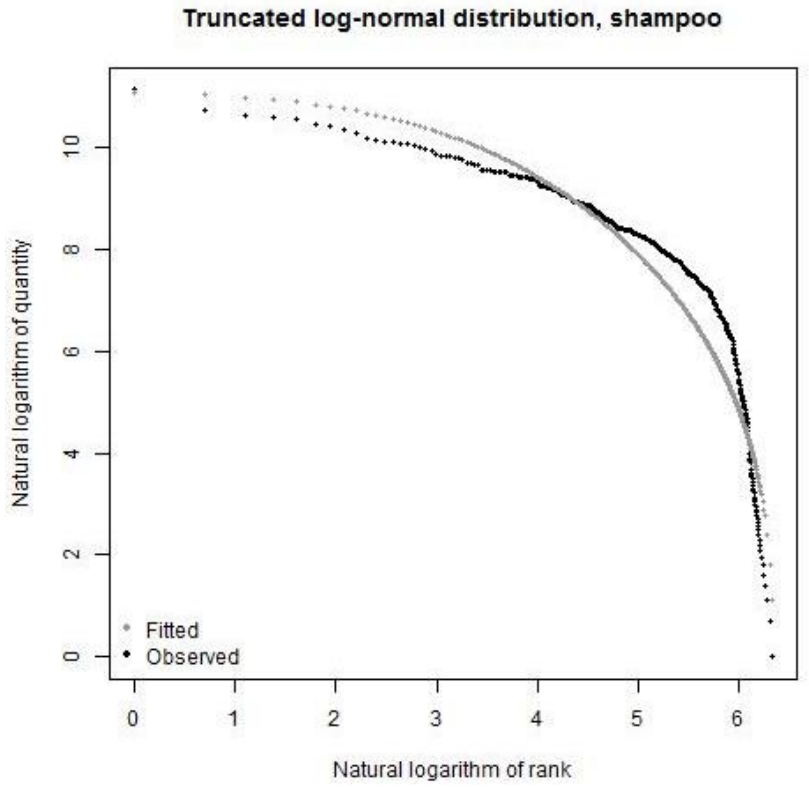
Estimate the distribution of the expenditure, and predict this from their ranks. This allows for retailers to provide summary stats.

$$q_i = F^{-1}\left(1 - \frac{r_i}{n}\right)$$

Possible candidates for F are, the log-normal, truncated log normal and the Pareto.

Truncated Log-normal distribution was the best fit



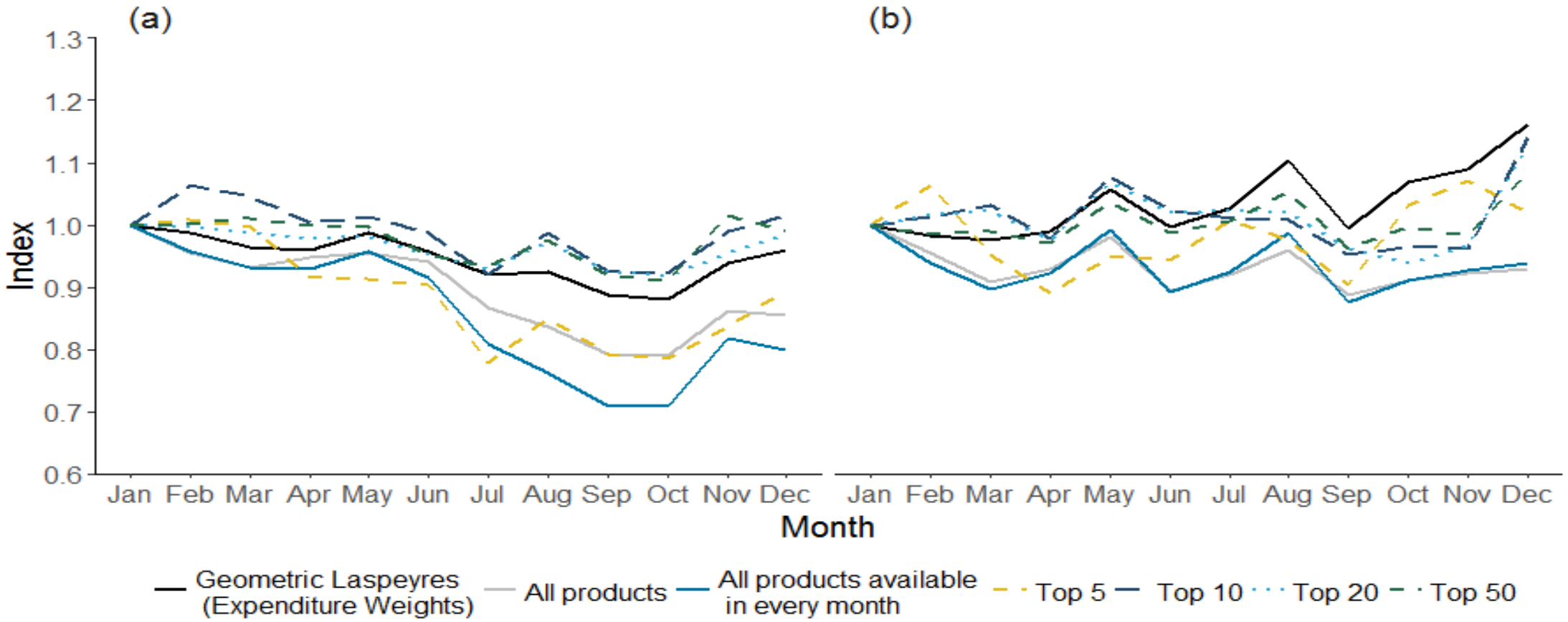


Subsets

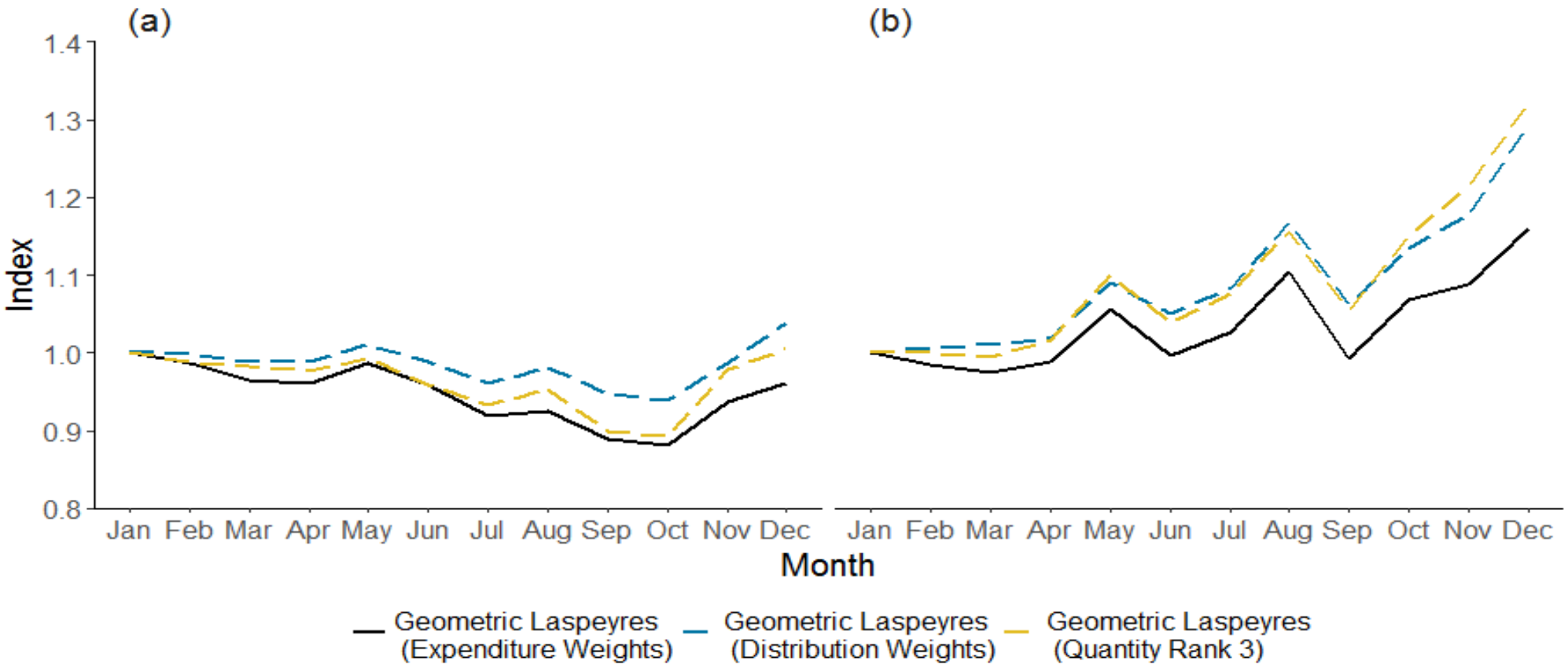
Taking the top 10/20/50 products by quantity sold and calculating a Jevons, how well does this perform against using weights?

A concern with Jevons is that less popular products have equal influence as the most popular. Taking these subsets emulate the “traditional” collection.

There is no best option



Rank 3 weights perform better than estimating distributions



Questions?