Office for National Statistics
Swyddfa Ystadegau Gwladol

Using R for analysis and production of Price Indices for the Production and Services sector of the economy.

Matthew Mayhew
Index Numbers Methodology
Office for National Statistics

Contents

Brief description of the UK's PPI, SPPI, EPI and IPI.

Production – Sampling

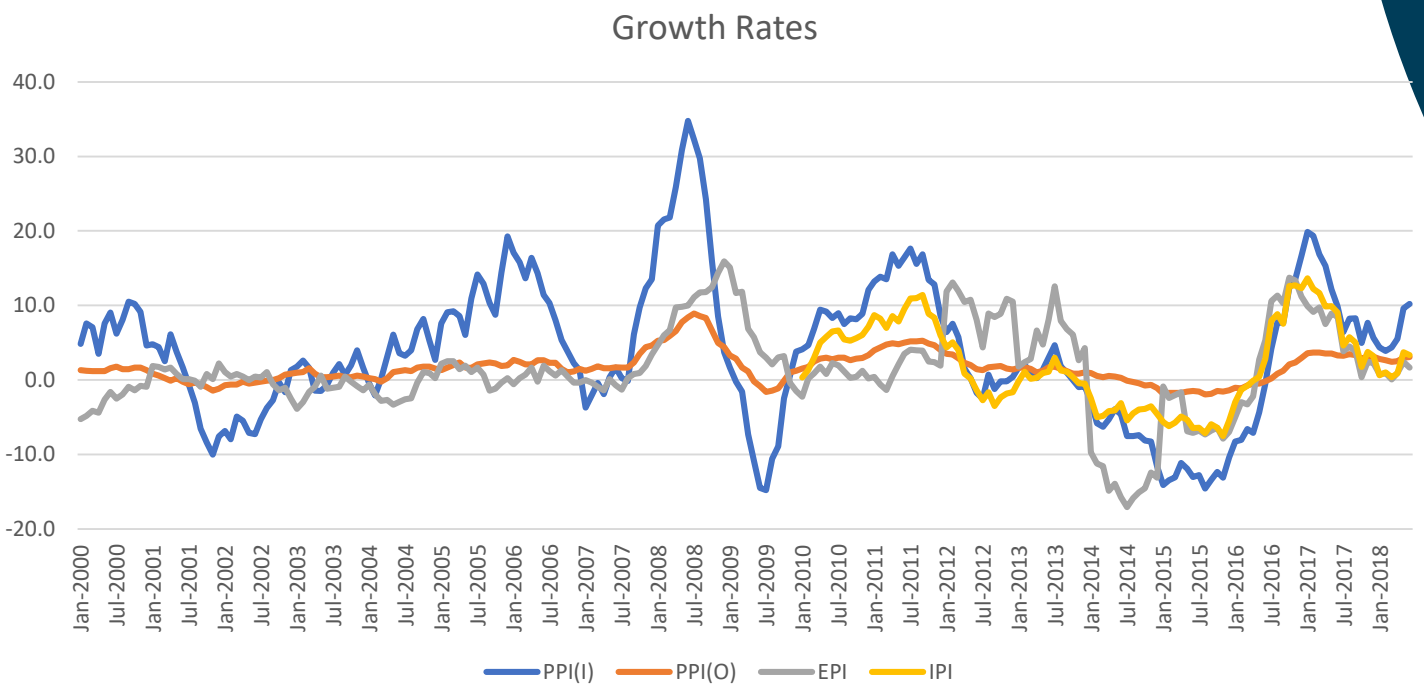
Analysis – Analysing the impact of Chain-Linking

UK PPI, SPPI, EPI and IPI

A brief introduction to these outputs

What are the price indices?

- The Producer Price Index (PPI) measures the price changes of products produces by UK manufacturers on the domestic market
- The Services Producer Price Index (SPPI) measures the prices of services provided by UK Businesses
- The Export Price Index (EPI) measures the price changes of products produces by UK manufacturers on the non-domestic market
- The Import Price Index (IPI) measures the price changes for inputs into the production sector bought from the non-domestic market.



Growth Rates

Production – Sampling

Sample Design

- Subsamples of two existing ONS surveys, PRODCOM (PPI) and Annual Survey Goods and Services (SPPI)
- EPI and IPI are samples from HRMC customs declarations data.
- Stratified sample by product classification and business size.

Adjusting Customs Data

- Customs data isn't in the correct product classification and the business definition is different.
- For certain businesses the ONS splits them up into different smaller units (known as suppliers), e.g. its retail arm and its production. HMRC records data for the larger businesses (known as traders)
- Trader sales are apportioned between suppliers by employment as follows:

```
HMRC <- HMRC %>%
  group_by(CN_Code,TRADER)%>%
  mutate(total_employment  = sum(employment),
         apportioned_sales =
sales*employment/total_employment)
```

- There are three size bands; small businesses, medium business and large businesses.
- Defined by businesses turnover
- The stratum boundaries are found by the cumulative root f rule.

```
library(stratification)

strata <- strata.cumrootf(x =
product$sales , n = n_product , Ls = 3,
alloc = c(q1,q2,q3) , model = NULL)
```

Size bands

Iteration 1

Size Band	frequency	$\sqrt{f}$	Size Band	$\frac{1}{3}\sum\sqrt{f}$
[0,100]	30	5.48	1	3.93
(100,200]	40	6.32		
(200,300]	22	4.69	2	5.59
(300,400]	50	7.07		
(400,500]	25	5		
(500,600]	10	3.16	3	2.26
(600,700)	13	3.61		

Iteration 3

Size Band	frequency	$\sqrt{f}$	Size Band	$\frac{1}{3}\sum\sqrt{f}$
[0,100]	30	5.48	1	3.93
(100,200]	40	6.32		
(200,300]	22	4.69	2	3.92
(300,400]	50	7.07		
(400,500]	25	5	3	3.92
(500,600]	10	3.16		
(600,700)	13	3.61		

Cumulative Root F method

Sample Sizes for businesses

- The sample sizes for product and size band is calculated using a Neyman allocation:

$$n_{h_k,P} = \frac{nN_{h_k,P}s_{h,p}}{\sum_{j,\ell} N_{j_\ell,P}s_{j,p}}$$

- These sample sizes are then inflated in order to oversample because of non-response as follows:

$$\tilde{n}_{h_k,P} = n_{h_k,P}\left(1 + \frac{d_{c_h}}{100}\right)$$

- For disclosure control reasons it is then checked to see if $\sum_k \tilde{n}_{h_k,P} \geq 3$ per product, if not they are set to 3.

Sample Sizes for price quotes

- A second sample size is calculated per product for the amount of price quotes to request, this is done using a Neyman allocation but using the variances of prices.
- These are then apportioned to the size bands depending on the total sales in each band.
- quotes are allocated to each size band so that each business is asked for a price, and that a business is either asked for 1, 2 or 3 quotes.

Drawing the sample

- The sample is taken using stratified simple random sampling.
- This is done as follows:

```
library(purrr)
Sampled_business <- sample_frame %>%
      group_by(product , size_band ) %>%
      nest()%>%
      full_join(sample_sizes , by = c("product" ,
"size_band") %>%
      mutate(sample = map2(data, nh, sample_n )%>%
      select(product, size_band, nh, Nh, sample) %>%
      unnest()
```

- Easy to update code than previous SAS code
- Can implement methodological changes better
- More variety in sampling and stratification methods
- Faster and more reliable.
- Easier to integrate

Why was R used?

Analysis – Impact of Chain-linking

Problem Definition

- As part of a Eurostat regulation change governing Short Term Statistics, NSIs were required to move their producer price indices to an annual chain-link due to updated weights.
- What impact would occur on the indices given these changes?
- Reusable code needed.

What is Linking?

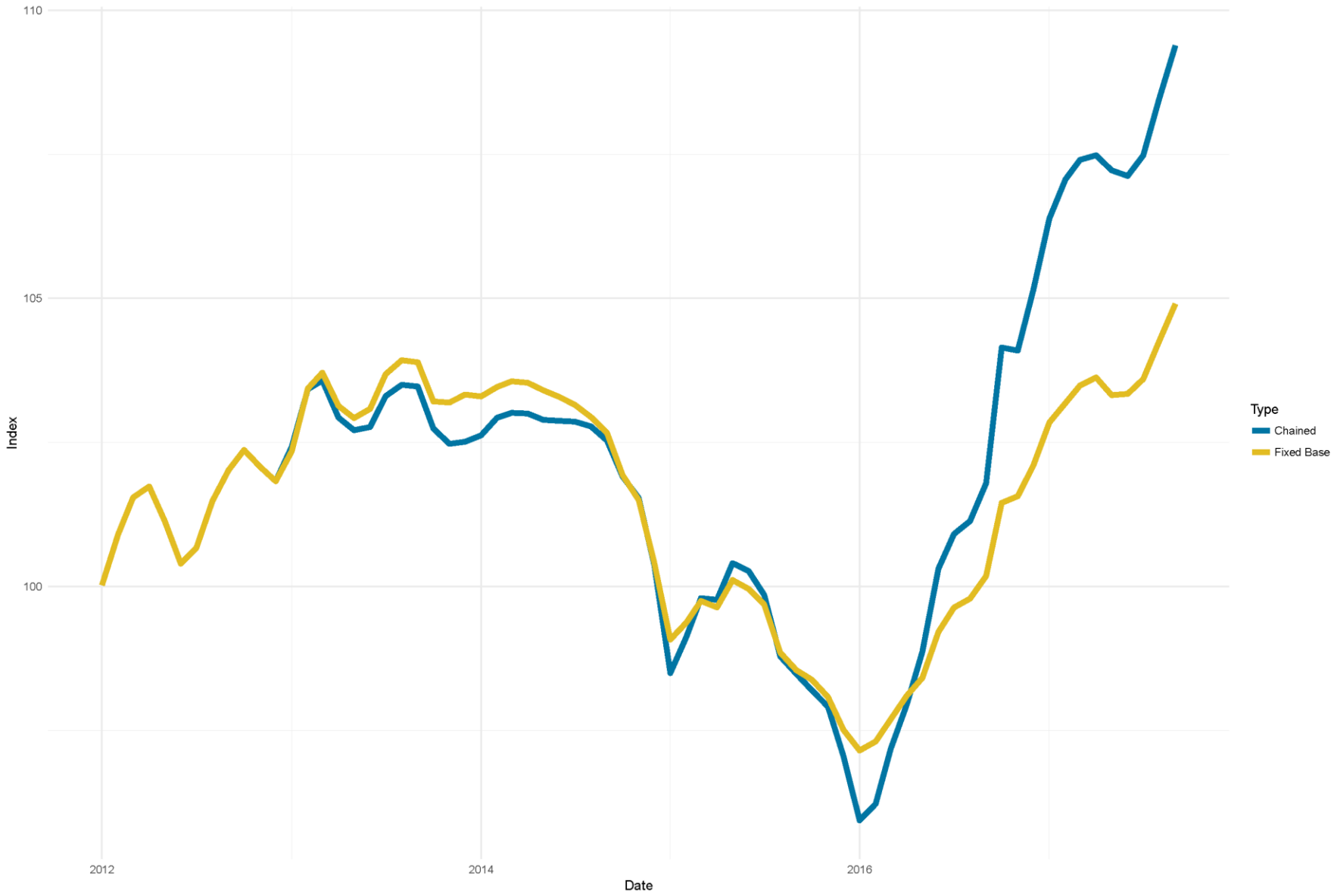
- Linking is the smoothing out of discontinuities in a index series.
- Discontinuities occur because of sample update, weight updates, change of base period or classification changes.
- Mathematically it is done by:

$$I^t_{LINKED} = \begin{cases} I^t_{OLD} & t < L \\ I^t_{NEW} \times \dfrac{I^L_{OLD}}{I^L_{NEW}} & t \geq L \end{cases}$$

- Chain-linking is the repeated use of linking

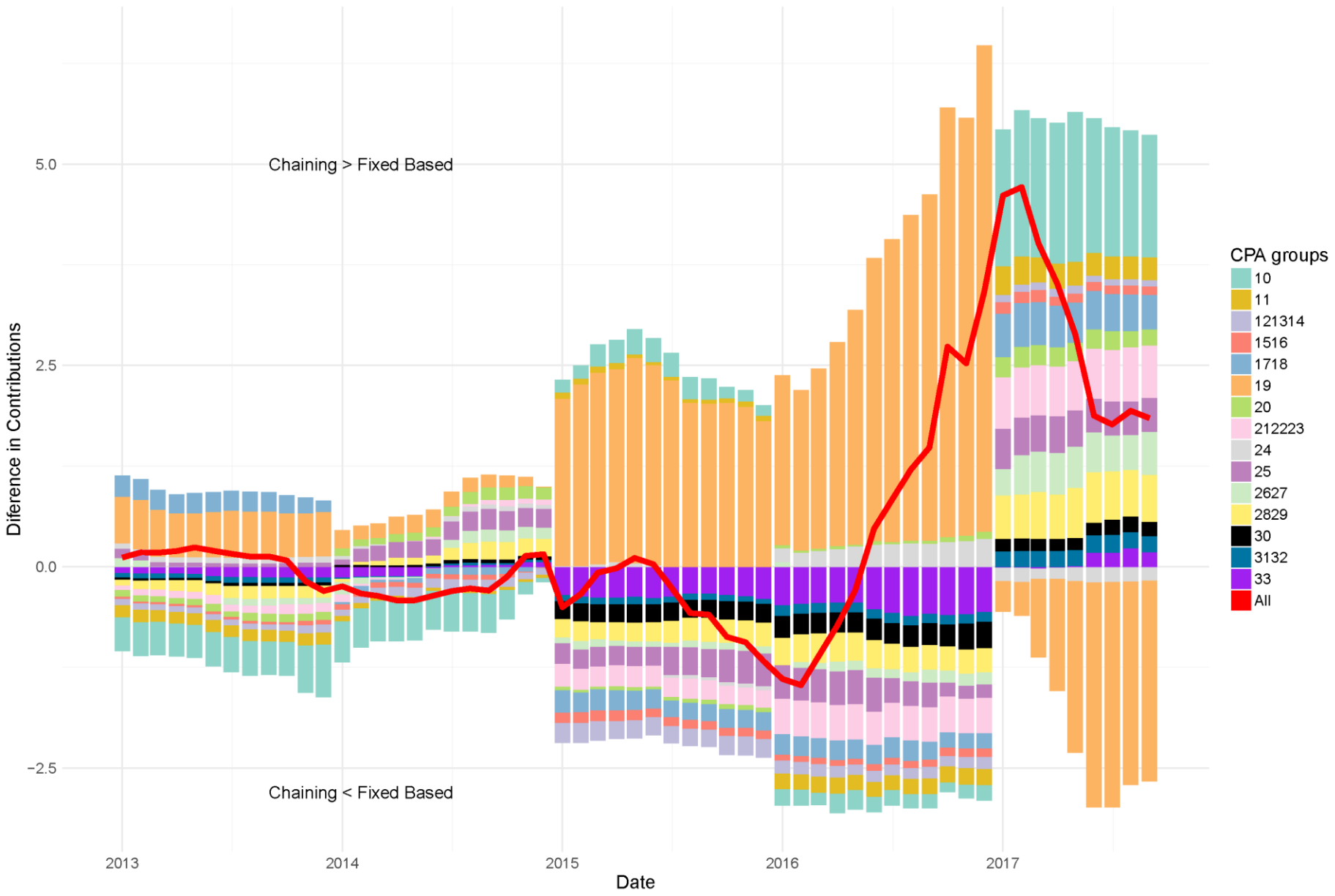
Method

- Rebase the PPI each December and the SPPI each Quarter 4
- Aggregate to item level using the Carli formula and using the Jevons (where appropriate)
- Aggregate to division and all items levels
- Calculate growth rates and contributions



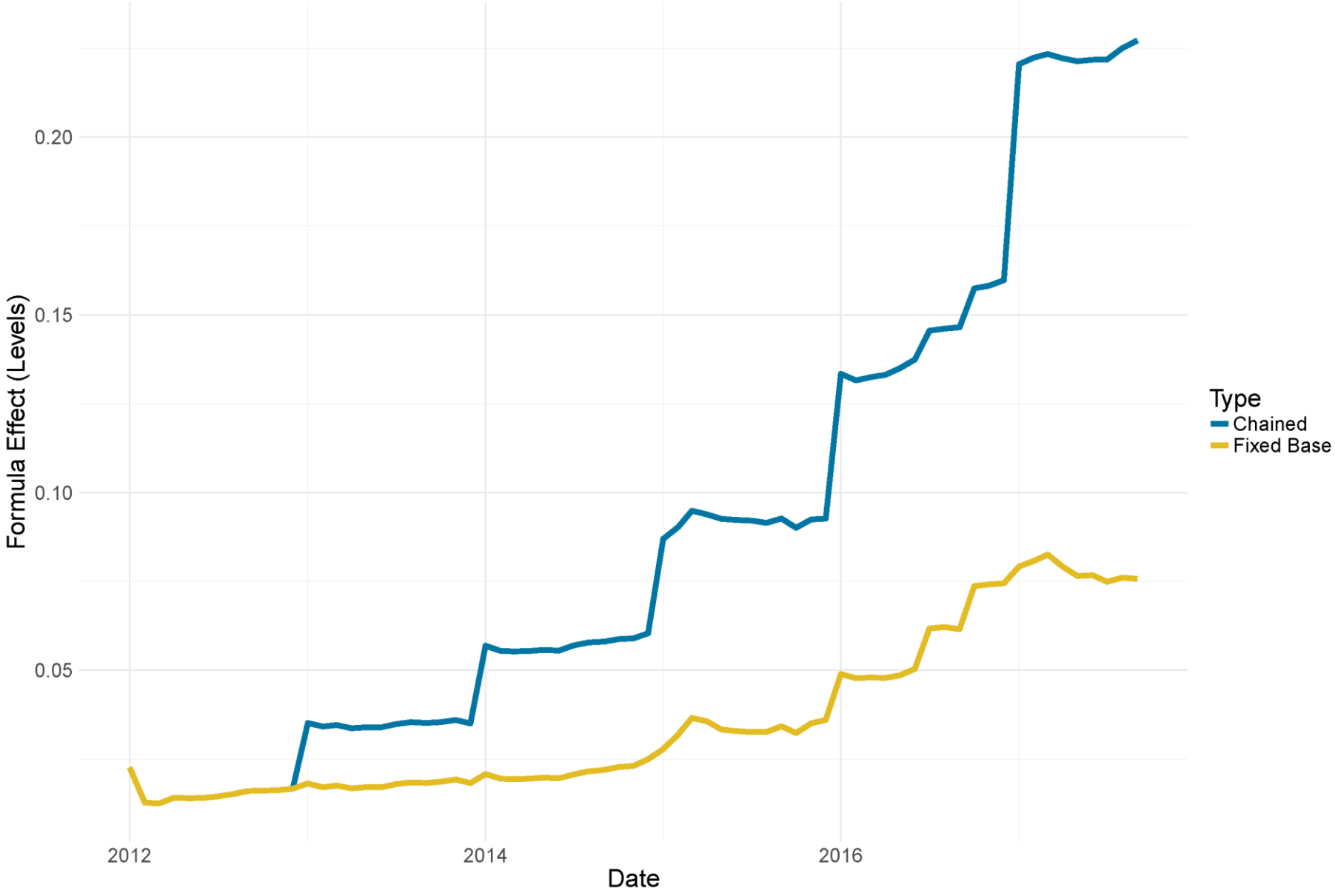
Effect on PPI Levels

Source: Own calculations only indicative of methods change and not of the final output



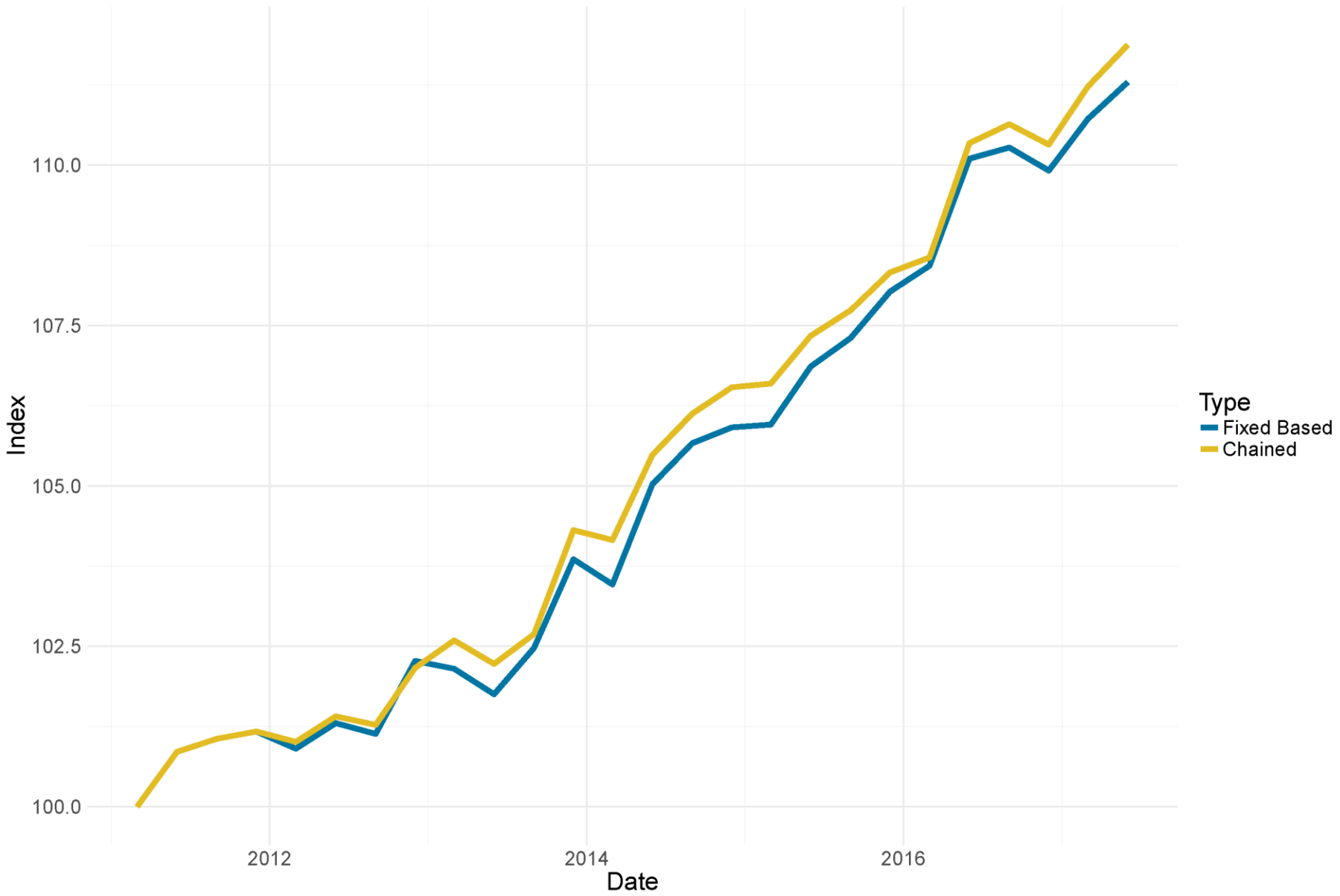
PPI Contributions

Source: Own calculations only indicative of methods change and not of the final output



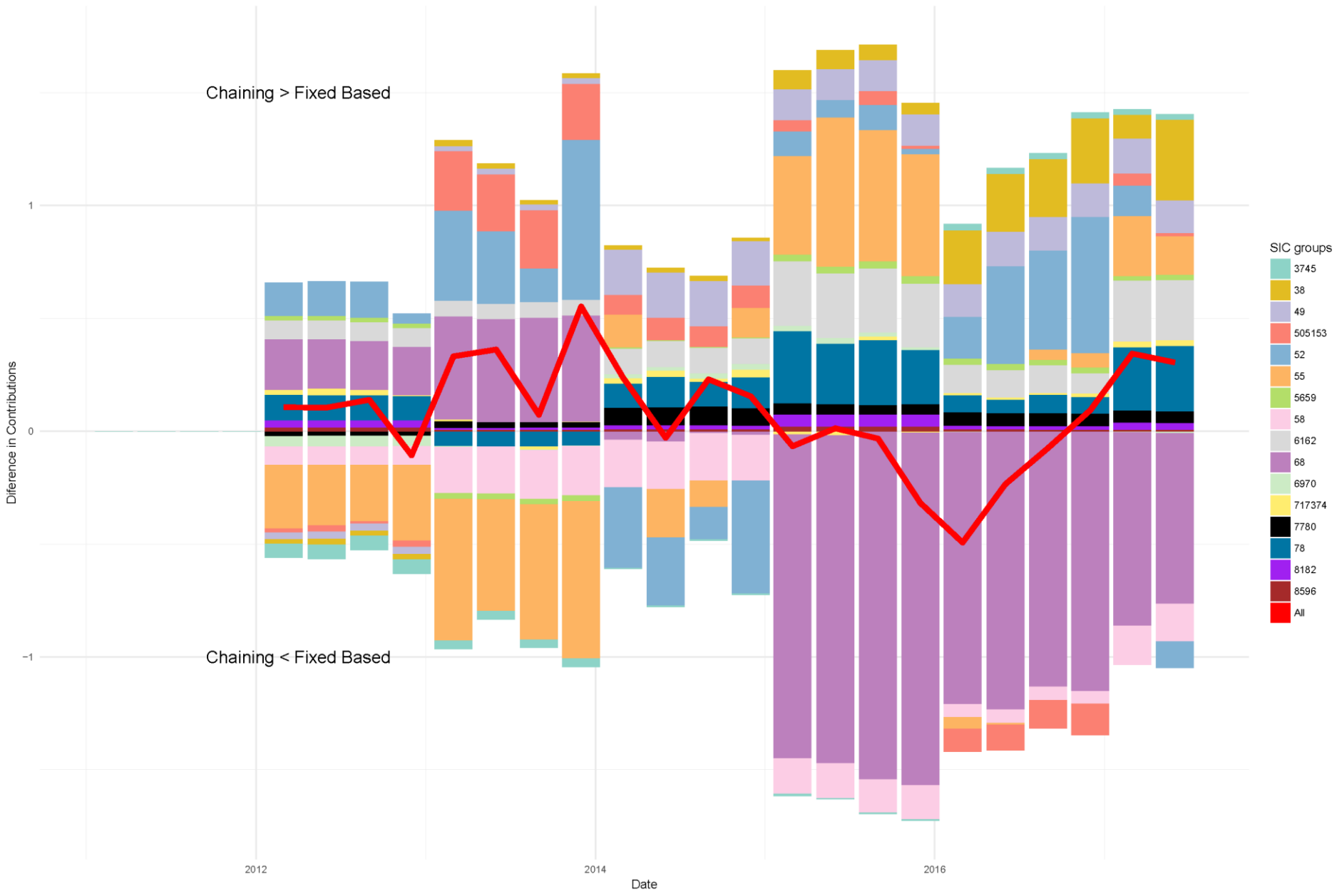
PPI Formula Effect

Source: Own calculations only indicative of methods change and not of the final output



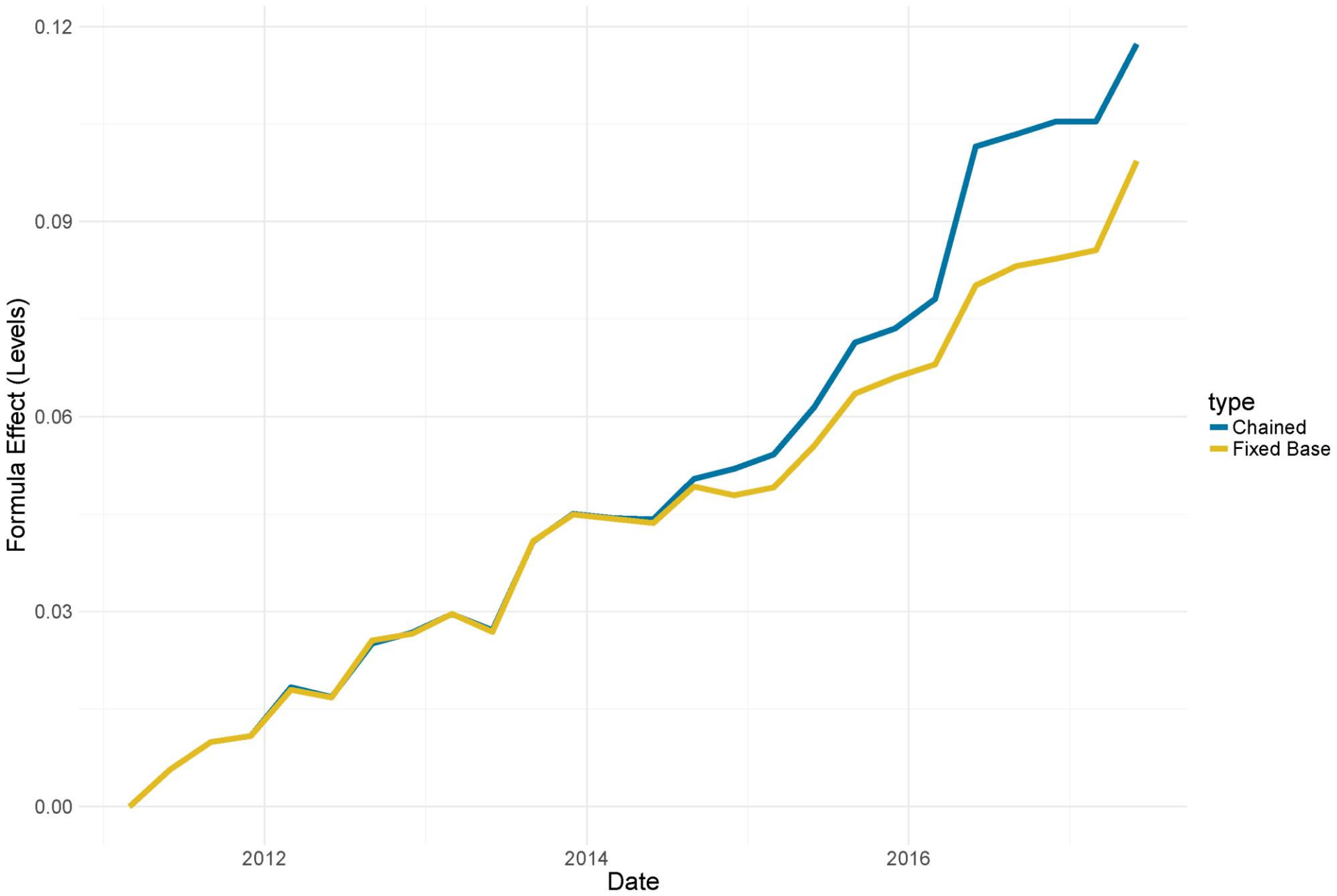
Effect on the SPPI Levels

Source: Own calculations only indicative of methods change and not of the final output



SPPI Contributions

Source: Own calculations only indicative of methods change and not of the final output



SPPI Formula Effect

Source: Own calculations only indicative of methods change and not of the final output

Results Summary

- Chain Linking has more of an effect on the PPI than the SPPI, average drift is 0.55pp for PPI and 0.31pp for SPPI.
- Chain-drift is always positive for the SPPI(except for Quarter 4 2012)
- Chain-Linking increases the contributions of Petroleum Products and Food Products in the PPI, due to large changes in prices in the short term and small in long term.
- In the SPPI, Retail Estate Agencies have the opposite effect, long term price changes are larger in the long term than in the short term, this is possibly down to the link to the asset price.
- The Formula effect is larger for the PPI than the SPPI, and is large enough to affect the published series.

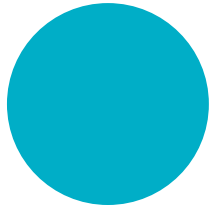
Conclusions

Why R was used?

- Reproducible code that is easy to edit
- General code was created so could be used on different dataset. Chain-linking was written for PPI and the used on SPPI data
- Visualisations of data were easily achieved and could be made more complex with a minimal amount of code.

- Production
  - Estimation of the sales
  - Price Updating
- Analysis
  - Does the PPI lead the CPI?
  - Time Based Methodology

Promoting R Use

Questions?