Outlier detection for grocery scanner data in Consumer Price Statistics

Mario Spina
On behalf of many

Office for National Statistics

Introduction

- Presentation based on outlier detection for grocery scanner data publication
- Background to data cleaning
  - Junk filters vs outlier detection
  - Methods explored
- Dump prices
- Analysis overview
  - Methods
  - Indices analysis
  - Consumption segment and seasonality
  - Results
- Future developments and conclusions

Background to data cleaning

- ONS are introducing new, bigger data sources in CPI
- ONS transformed rail fares and second-hand cars categories
- Next ONS are planning to introduce grocery scanner data in CPI from 2025
- Data cleaning selects observations relevant for index calculation
- Building on previous work on outlier detection for rail fares and second-hand cars
- Will explore price-quantity relative outlier detection, and the combination with price relative methods

4 Junk filters vs outlier detection

Data cleaning consists of two underlying components:

Junk filter

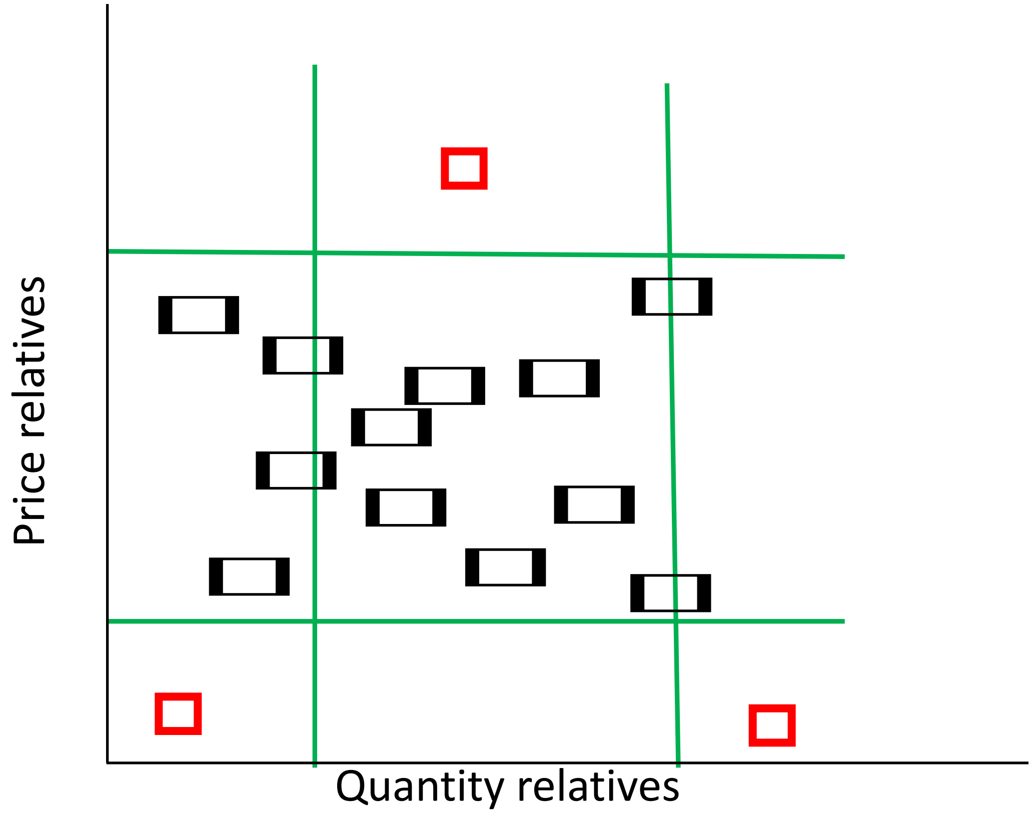
Determines observations out of scope

Example:

- Removing products sold by weight
- Removing transactions not linked to a UK region

Outlier detection

Identifies products with extreme and potentially erroneous or out-of-scope price or quantities movements



Office for National Statistics

5 Methods explored

Based on our previous analysis, ONS explored the following methods for grocery scanner data:

- Price relative fences (p-dump)
- Price-quantity relative fences (pq-dump)
- Price and price-quantity relative fences (combined)

Abbreviation	Keep row if…
p-dump	RP in [Lp, Up] (E1)
pq-dump	RP in [Lp, Up] OR RQ in [Lq, Uq] (E2)
combined	(E1) AND (E2)

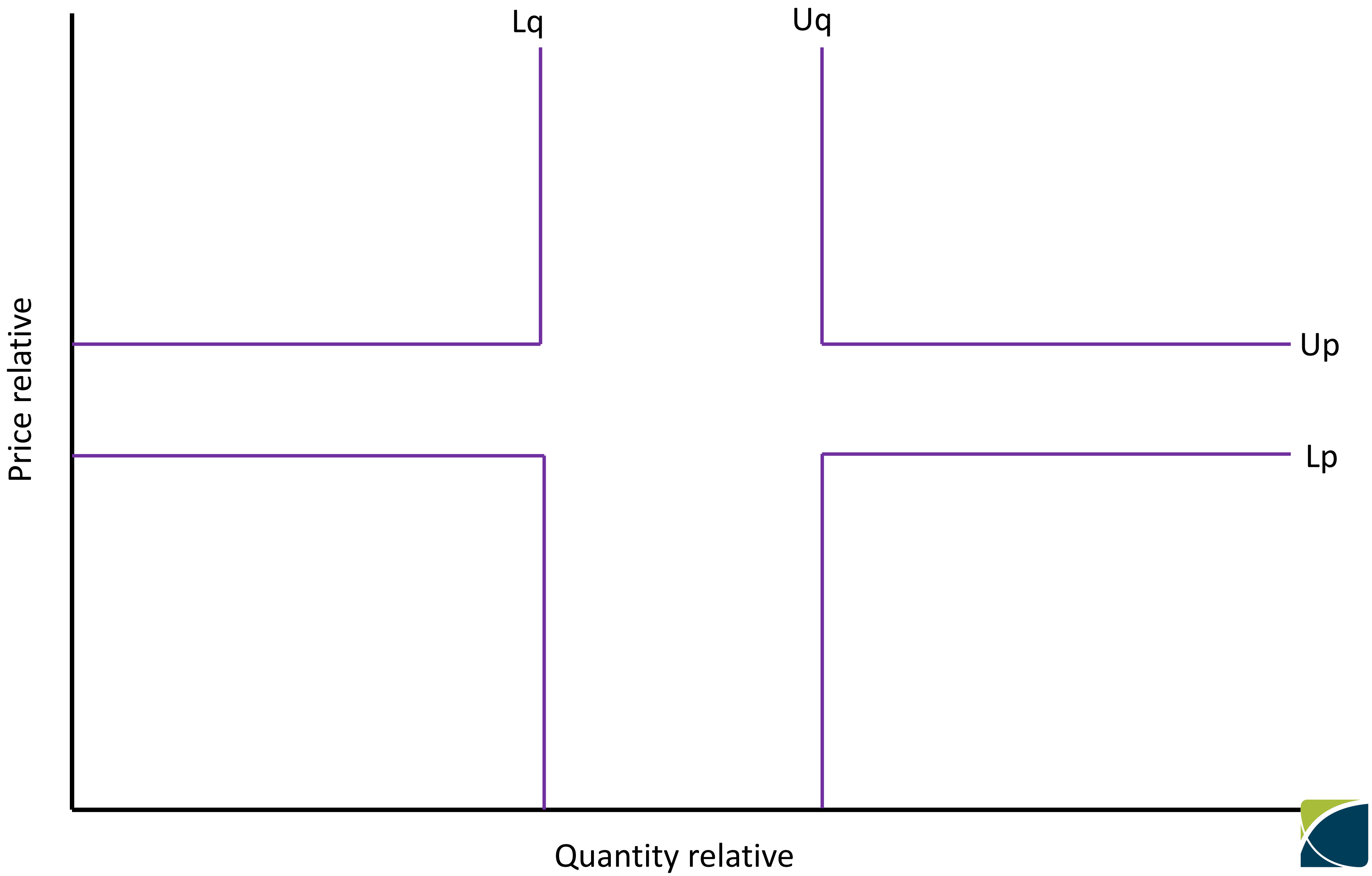
- Note: RP, RQ are price relative, quantity relative
- Lp(q) is the lower fence for price (quantity) relative
- Up(q) is the upper fence for price (quantity) relative

Dump prices

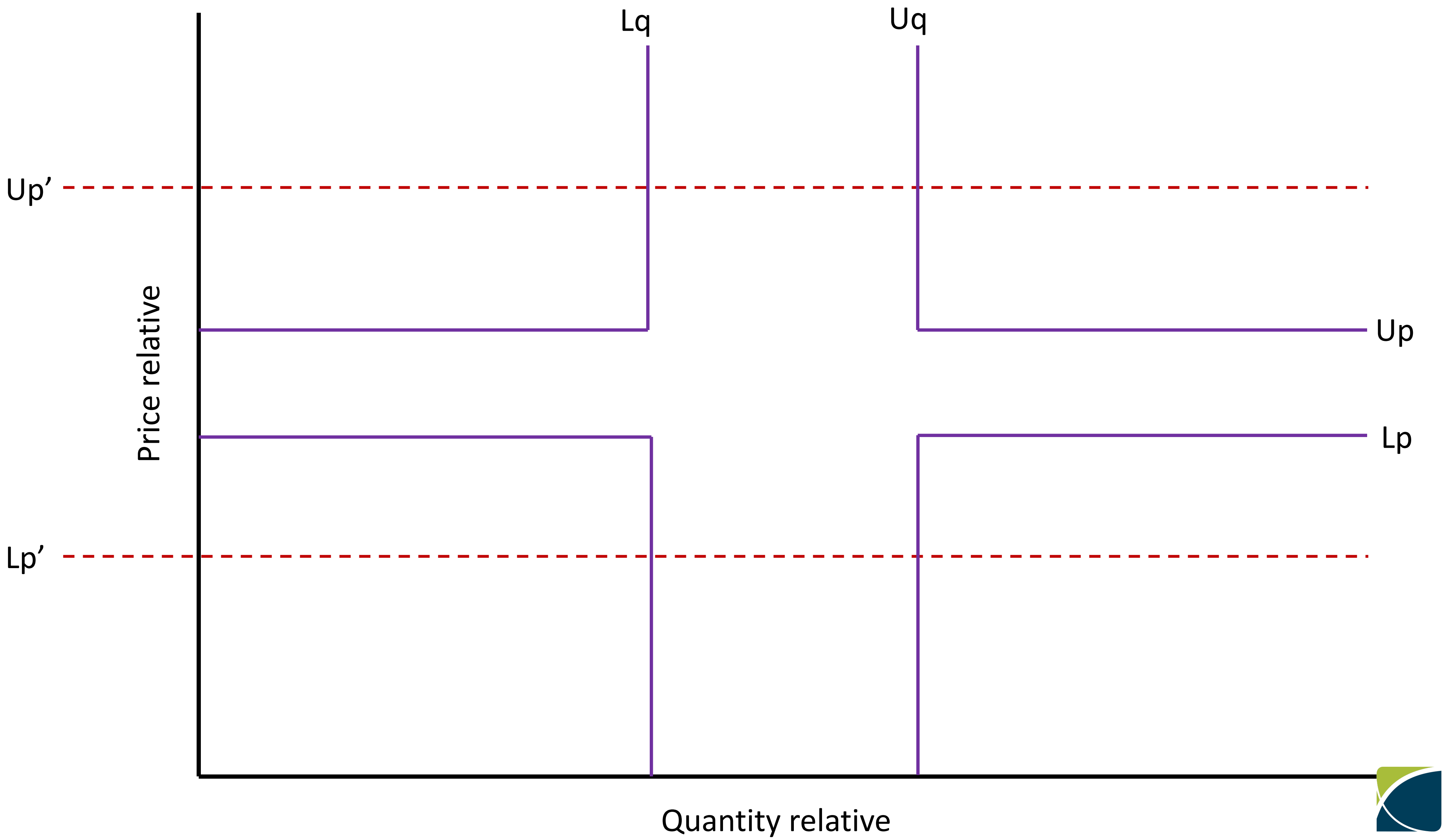
- Occur at the end of a product's life cycle, particularly common in grocery.
- Characterised by a large price and quantity drop. Can be observed using scanner data.
- Different from "clearance sticker products" as the quality is different, often due to nearing expiry date.
- International guidance recommends to remove dump prices, as might bias the index.
- GEKS-T might be biased by dump prices.

Product	Price, Jan	Price, Feb	Quantity, Jan	Quantity, Feb
1	3	3	10000	10000
2	3	0.5	10000	1
			Törnqvist	0.6389

Quantity-price relative plane

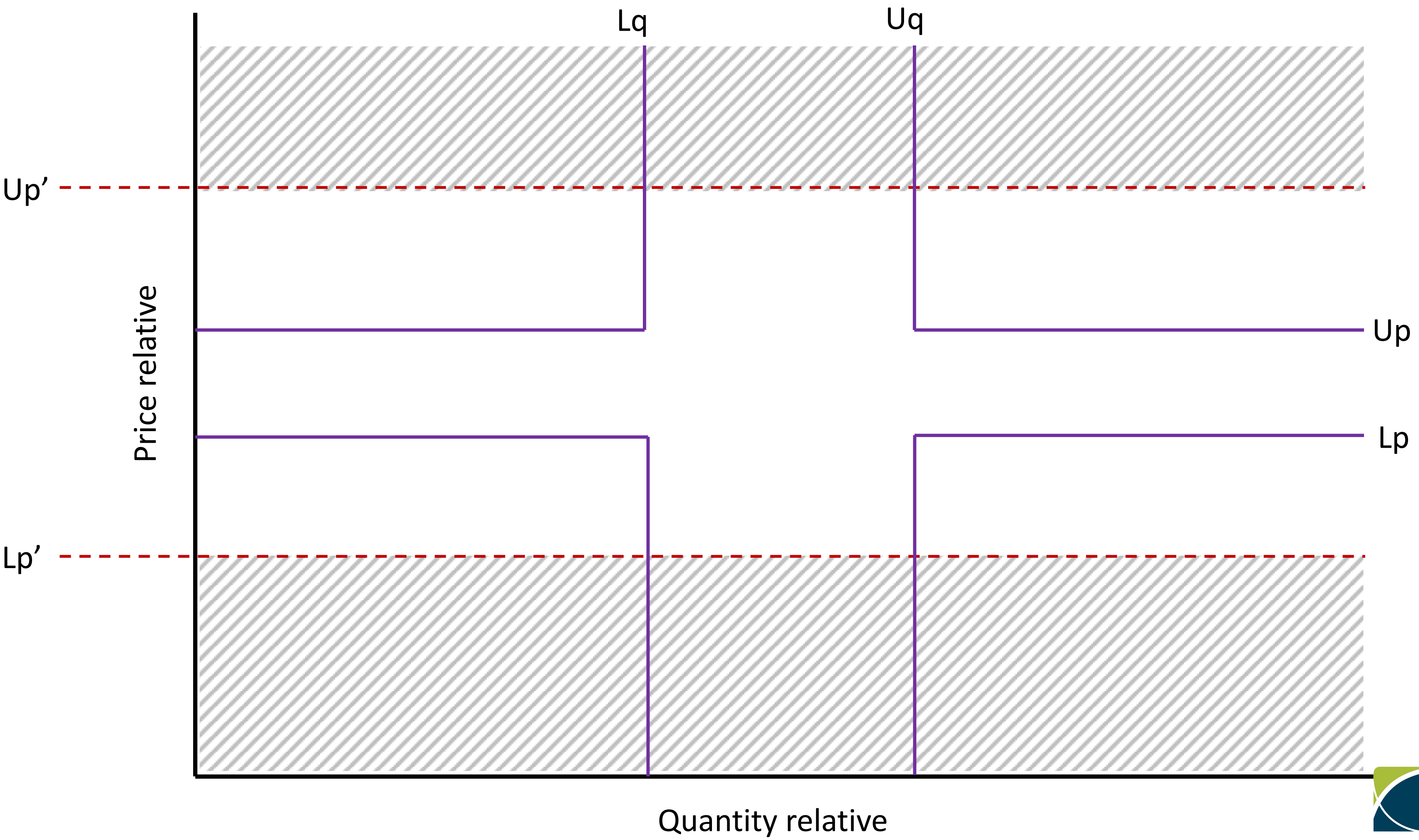


Quantity-price relative plane



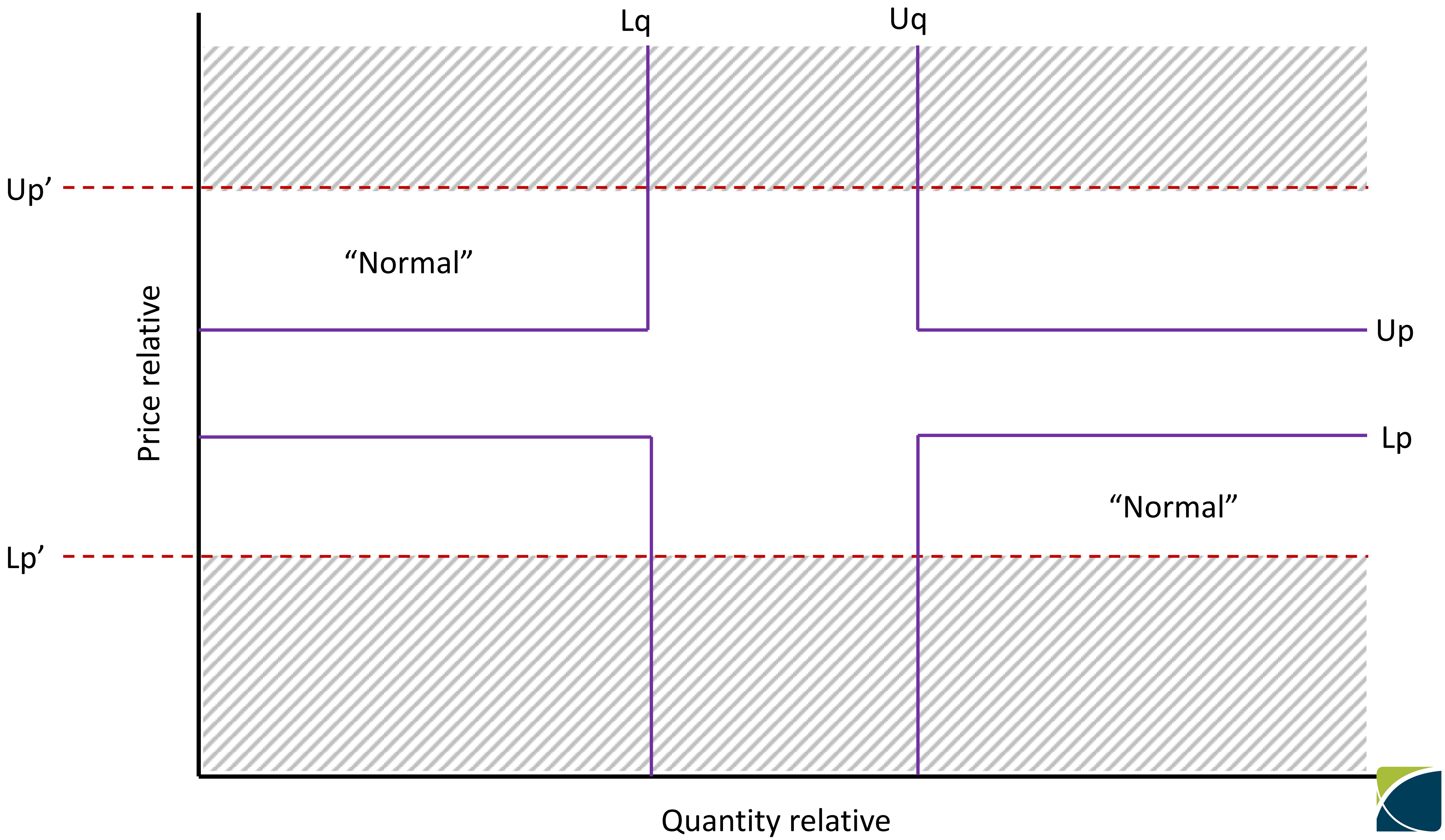
Office for National Statistics

7 Quantity-price relative plane

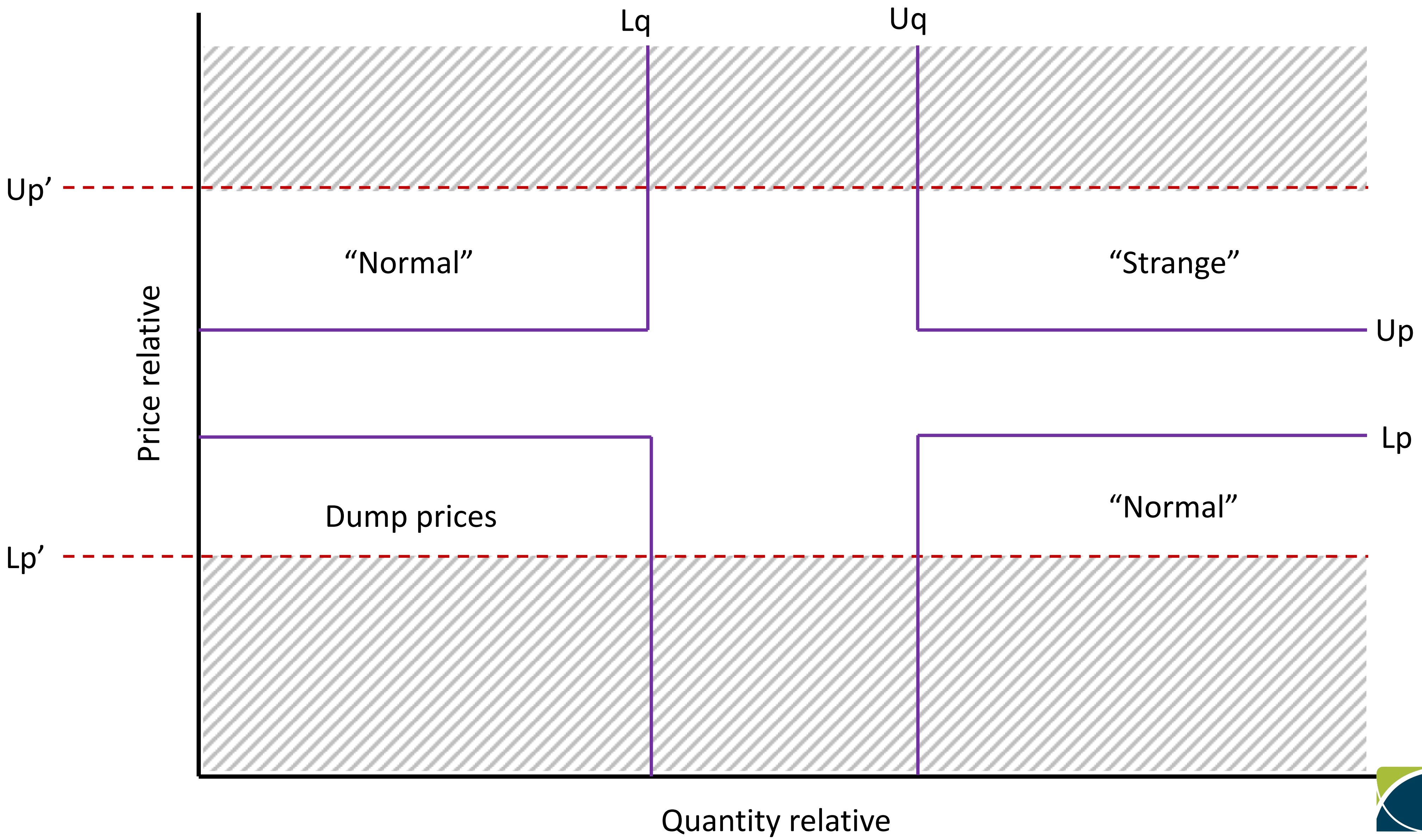


Office for National Statistics

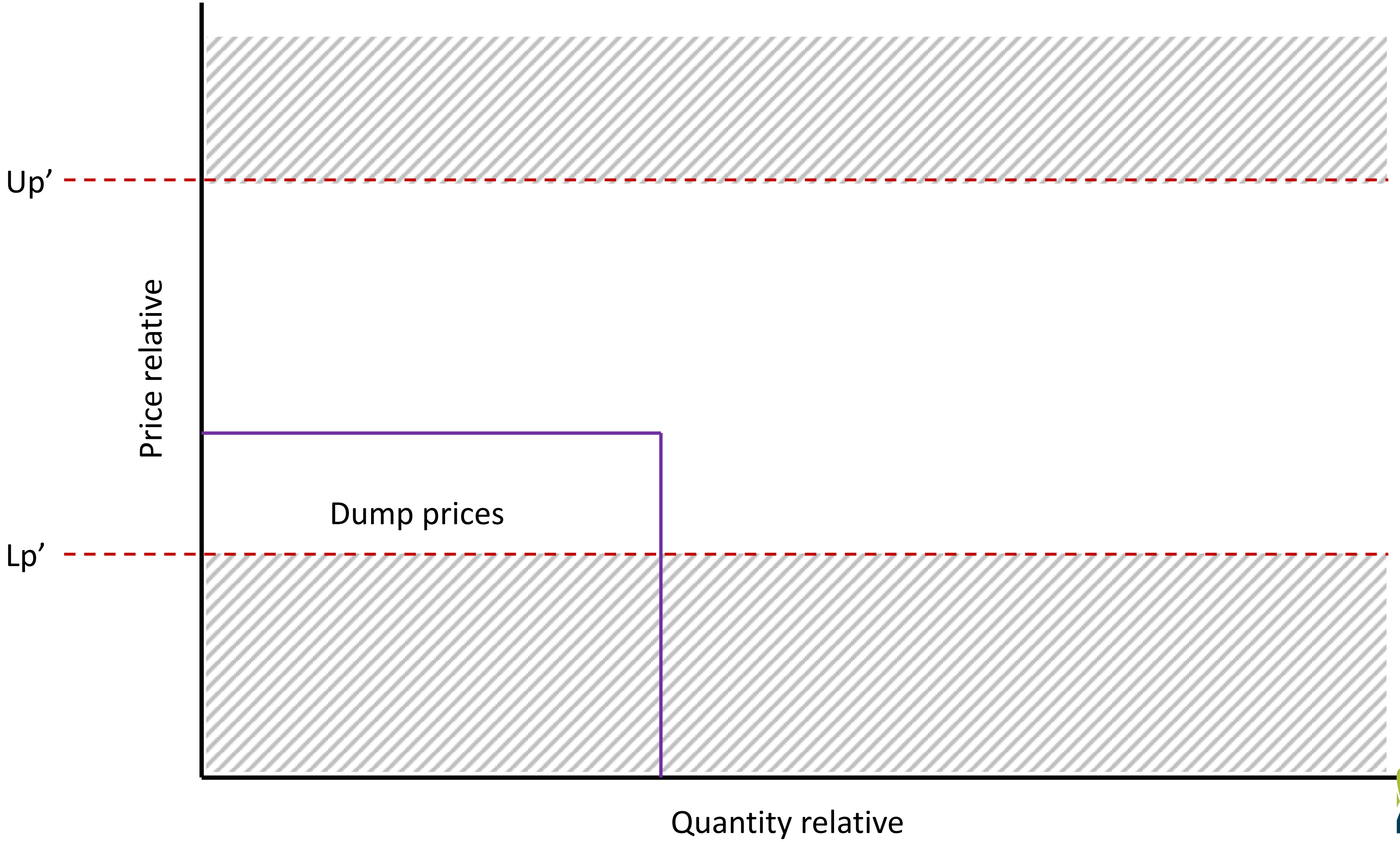
7 Quantity-price relative plane



Quantity-price relative plane

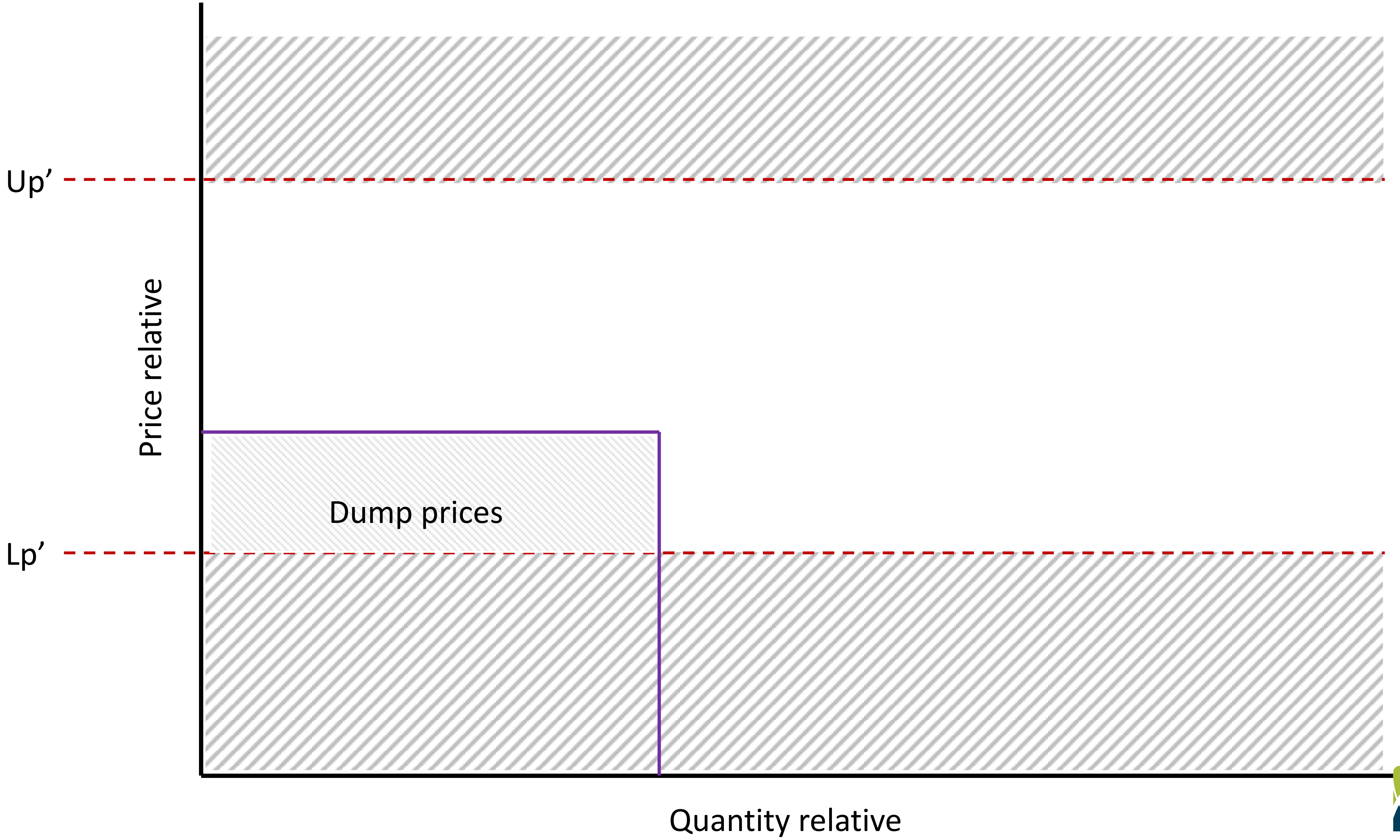


7 Quantity-price relative plane

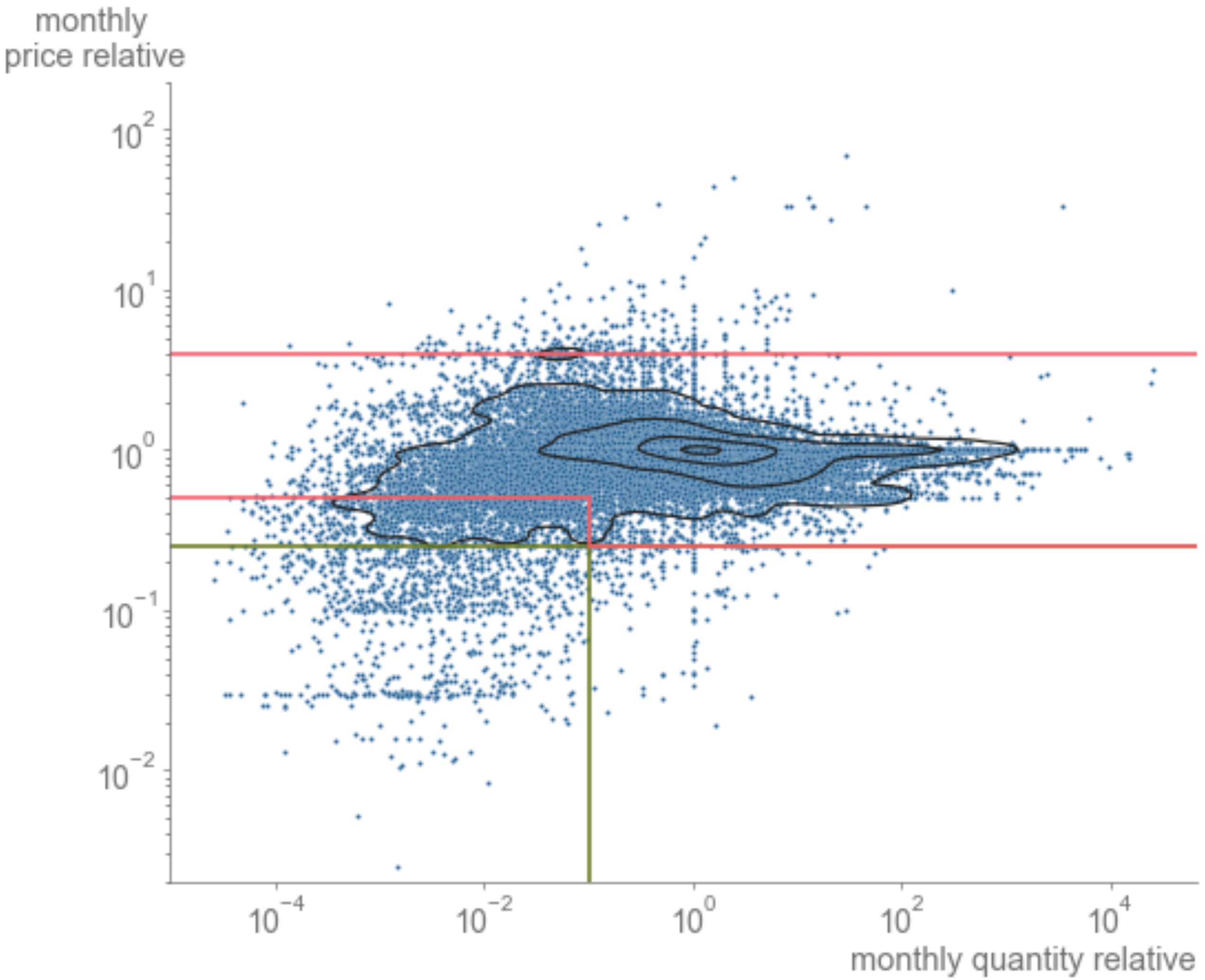


Office for National Statistics

Quantity-price relative plane



Quantity-price relative plane



Office for National Statistics

Analysis overview

The analysis was broken down as:

- Outlier detection methods explored
- High level indices analysis at various levels of aggregation
- Consumption segment analysis and seasonality

- Discuss 3 reasons suggesting outliers are mainly dump prices

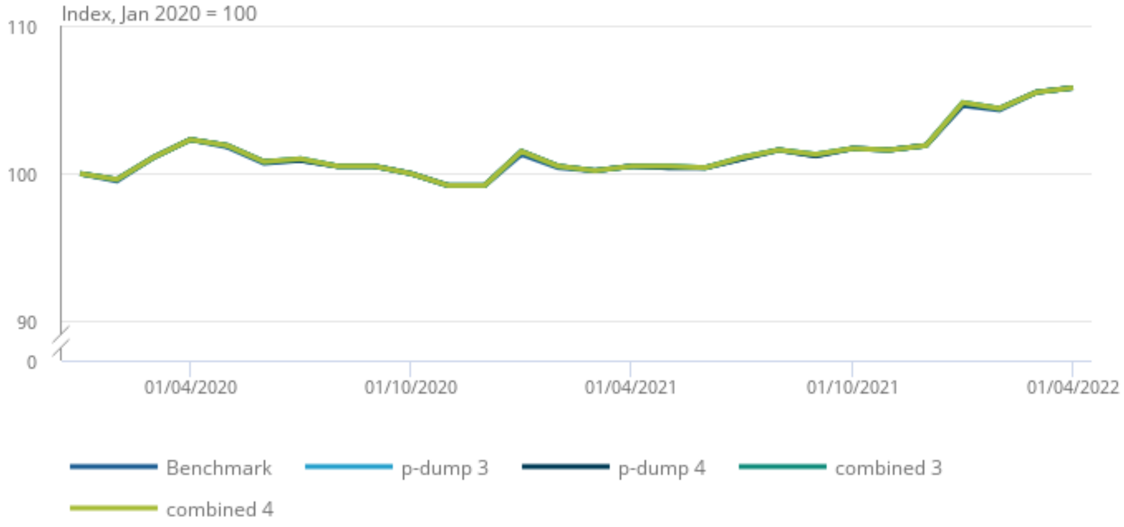
Outlier detection methods

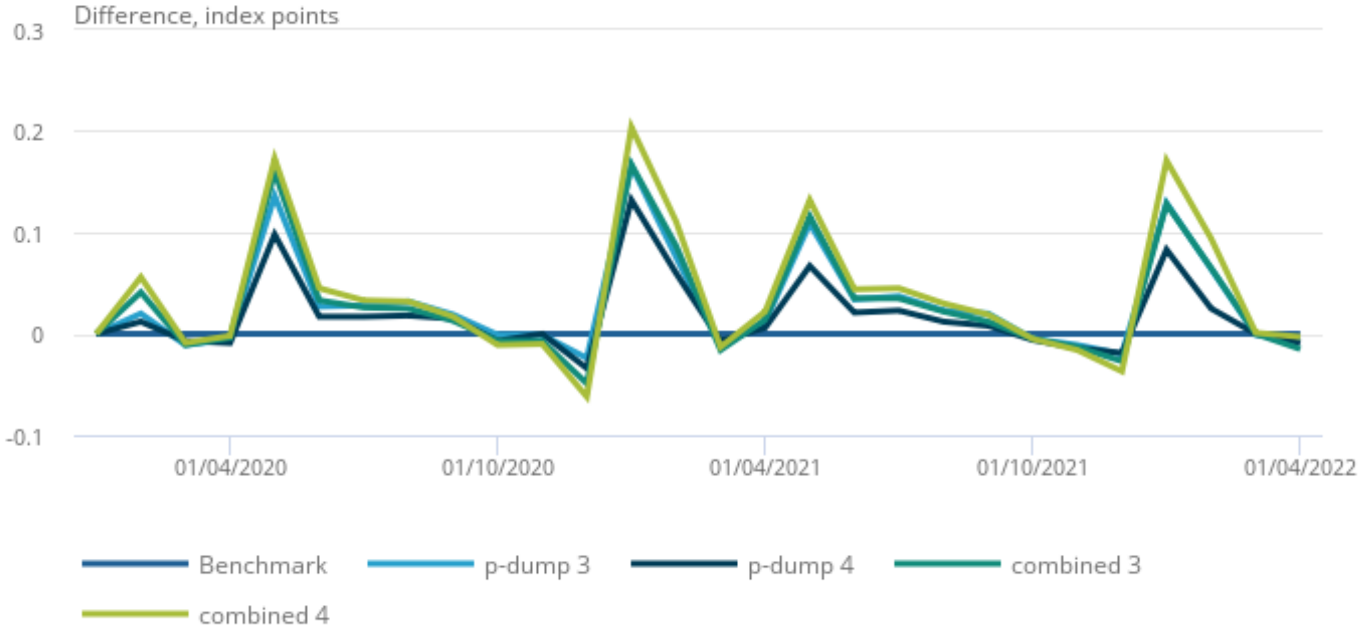
Definition of methods explored and percentage of data removed

Fencing method	Abbreviation	Keep row if…	% removed:	
			expenditure	rows
No outlier detection	benchmark	All rows kept	NA	NA
Price	p-dump 3	$0.3334 \leq rp_{t-1,t} \leq 3$	0.00852%	0.01671%
	p-dump 4	$0.25 \leq rp_{t-1,t} \leq 4$	0.00295%	0.00729%
Price-quantity	pq-dump 0.01	$0.5 \leq rp_{t-1,t}$ OR $0.01 \leq rq_{t-1,t}$	0.00015%	0.00082%
	pq-dump 0.1	$0.5 \leq rp_{t-1,t}$ OR $0.1 \leq rq_{t-1,t}$	0.00131%	0.00577%
Price and price-quantity	combined 3	p-dump 4 AND pq-dump 0.01	0.00308%	0.00781%
	combined 4	p-dump 4 AND pq-dump 0.1	0.00414%	0.01194%

Indices analysis at COICOP1

Exploring time window from January 2020 to April 2022
Over 130 consumption segments grouped into 4 COICOP3 categories



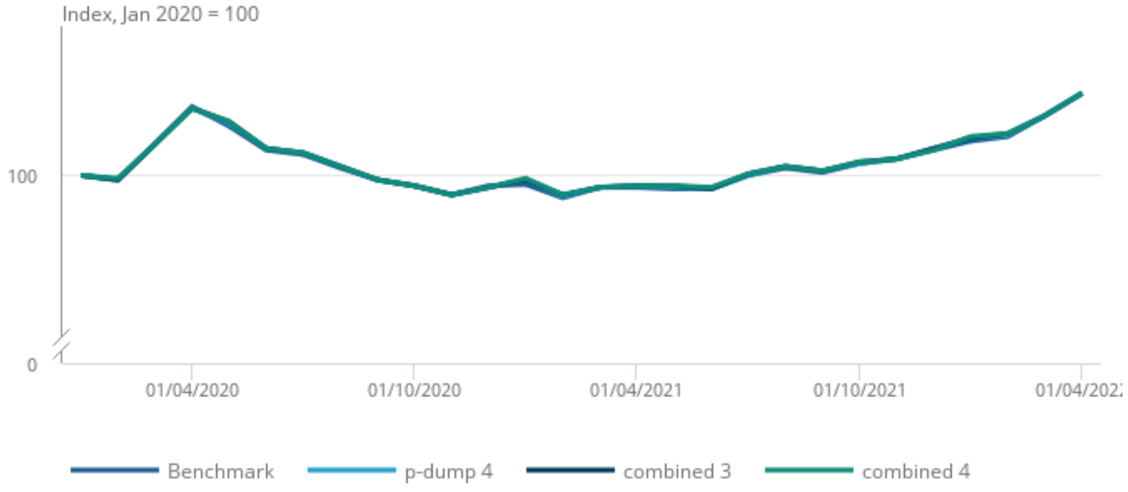


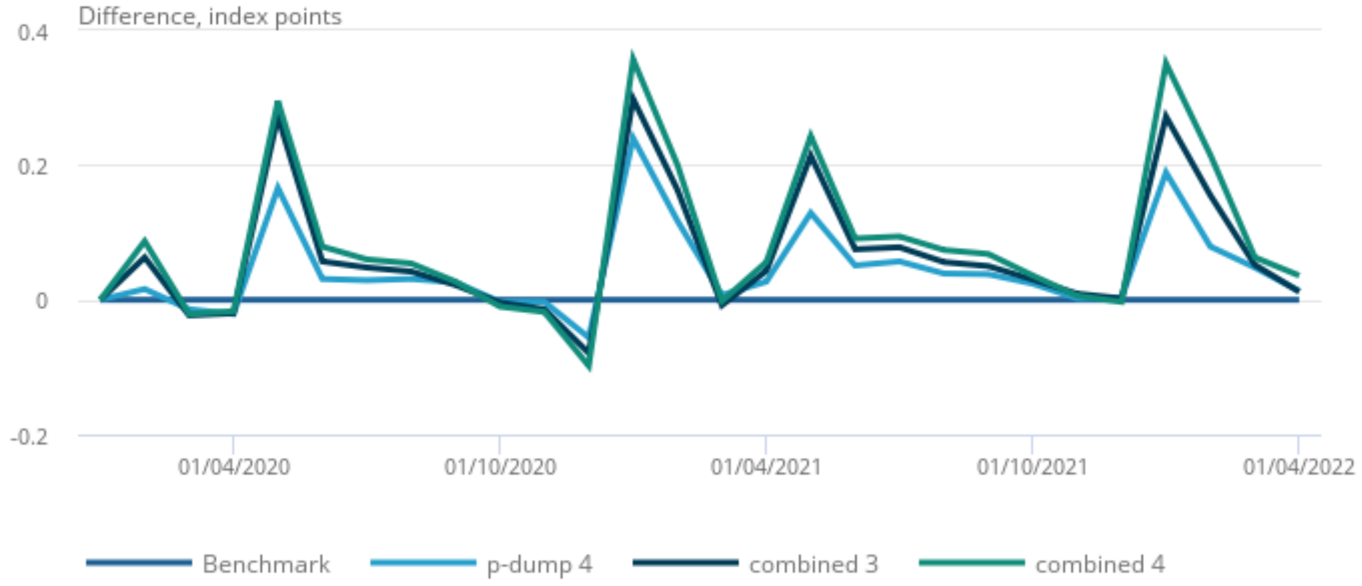
Similar trend observed, removal of price relatives >1, explore combinations of p-dump 4
Reason 1: indices mostly have a difference >0

Office for National Statistics

Indices analysis at COICOP3

102 CS in the Food COICOP3 category
80 CS show a difference larger than 0.1



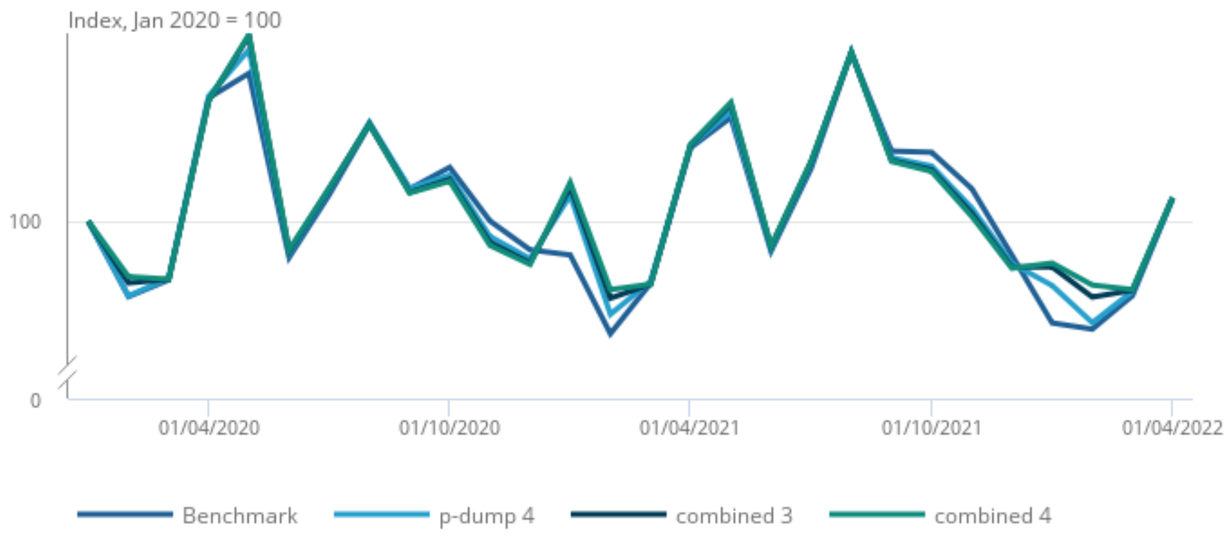


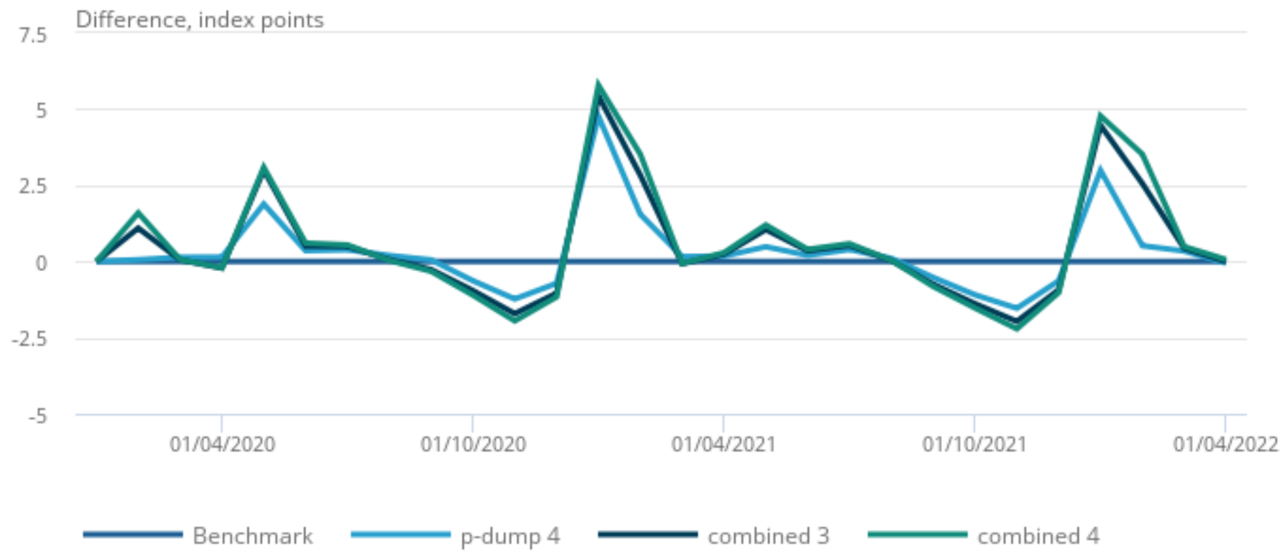
Difference increases, trend similar but price-quantity filters become more important

Office for National Statistics

13 Indices analysis at consumption segment

“Chocolate, assortment” consumption segment
Largest difference observed (about 6 index points) in January 2021





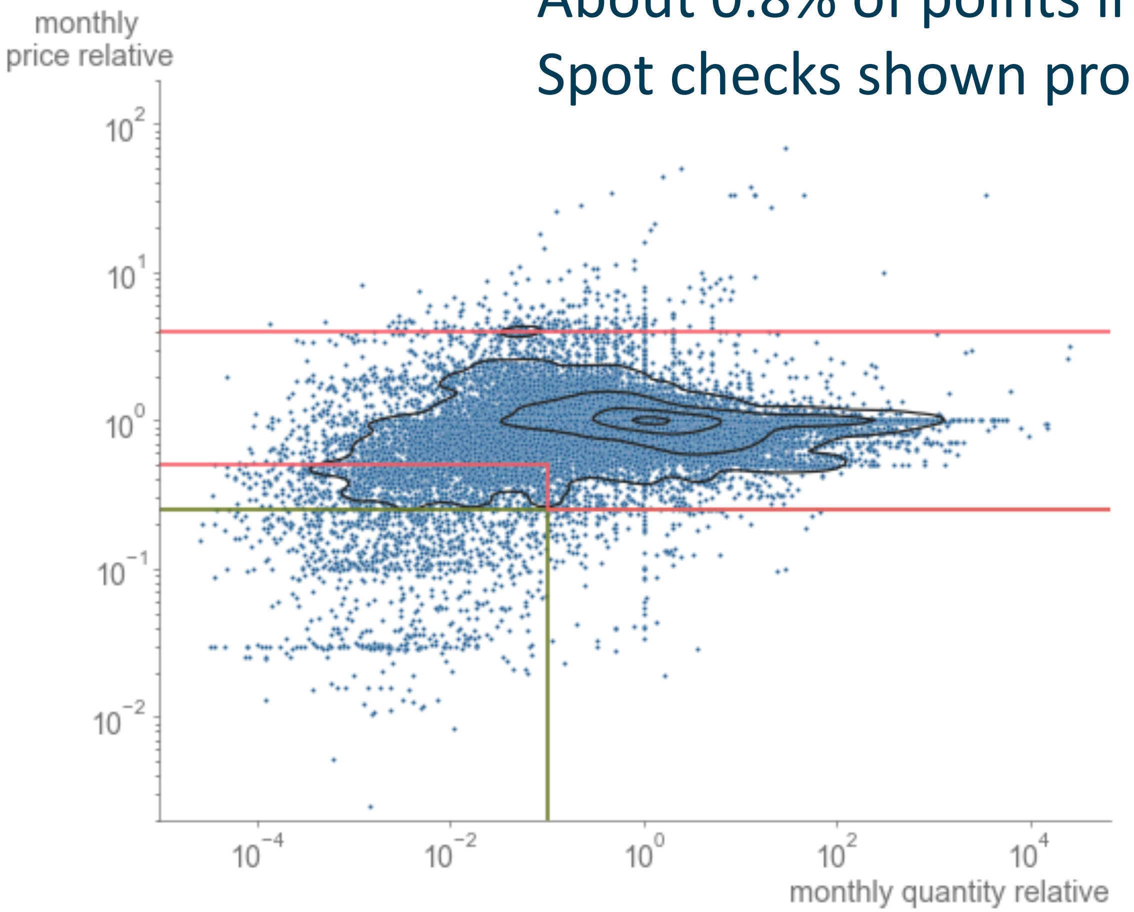
Pronounced seasonality, with difference peaks after Christmas and Easter due to heavy discounts

Consumption segment analysis

Over 60’000 points in the chart

About 0.8% of points in the bottom-left rectangle

Spot checks shown products being dumped from market

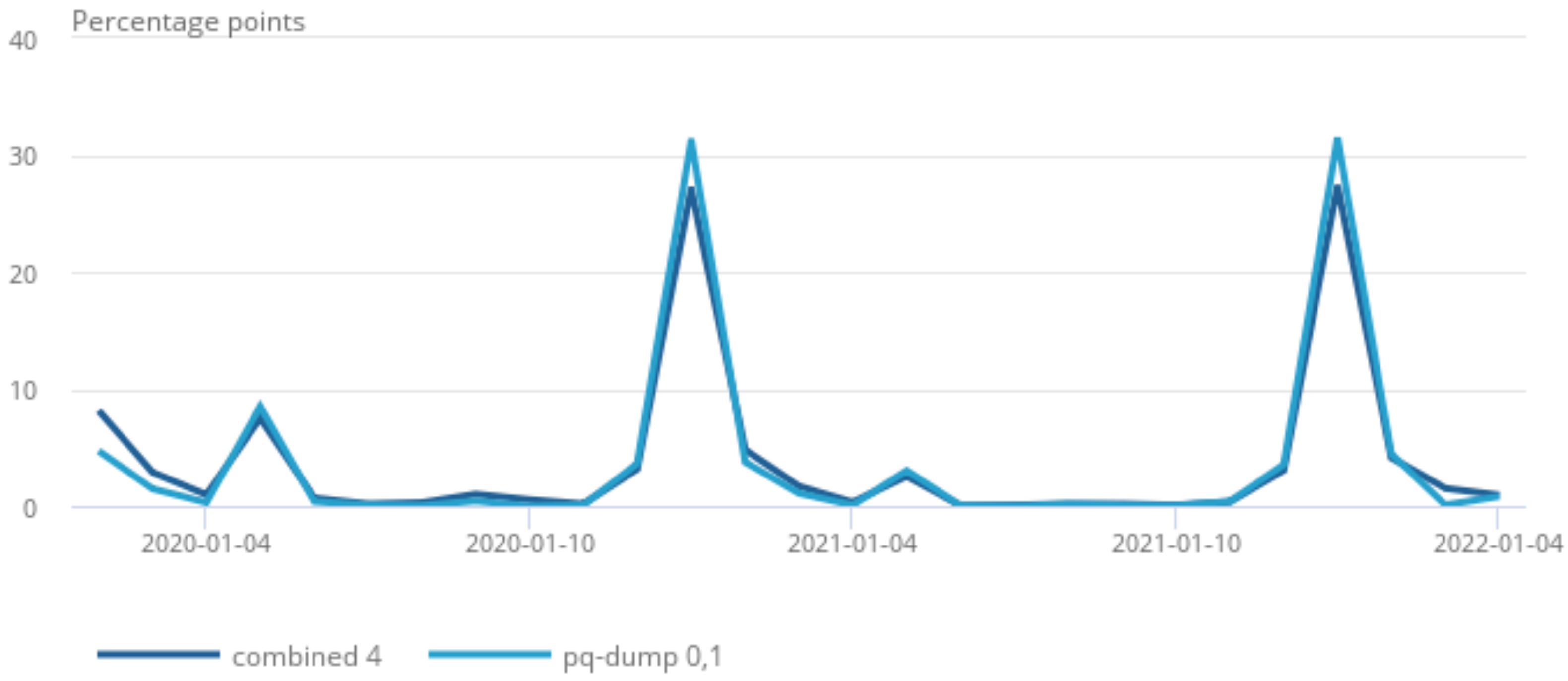


Fencing method	Abbreviation	Keep row if…	% removed:	
			expenditure	rows
Price	p-dump 4	$0.25 \le r^p_{t-1,t} \le 4$	0.0181%	1.9491%
Price-quantity	pq-dump 0.1	$0.5 \le r^p_{t-1,t}$ OR $0.1 \le r^q_{t-1,t}$	0.0121%	3.1202%
Price and price-quantity	combined 4	p-dump 4 AND pq-dump 0.1	0.029%	3.7516%

Reason 2: most of the outliers are in the “dump prices” quadrant

Consumption segment analysis - seasonality

We expect no explicit seasonality from price errors
Strong seasonality observed, with over 60% of outliers in January or May
Suggests that outlier are caused by dump prices



Reason 3: Strong seasonality

Office for National Statistics

Results: discussion

- ONS presented an analysis that suggests outliers detected in grocery scanner data are due to dump prices and suggested three reasons.

- The preferred approach combines price relative fences of [0.25,4] with a price-quantity filter with price relative fence of $r^{p}_{t-1,t} \leq 0.5$ and quantity relative fence of $r^{q}_{t-1,t} \leq 0.1$.

- This flags dump prices and removes the least amount of data (0.00344%) from index calculation, in line with previous studies

- The thresholds for the price filter have widened because of wider price distribution in grocery scanner data

Results: discussion

- The outlier detection strategy seems to (mostly) remove dump prices, correcting for a mild downward bias, with impacts in January and May.

- The strategy has larger impacts at lower levels of aggregation:
  - COICOP1 largest difference of 0.2 index points.
  - COICOP3 largest difference of 0.35 index points for food categories
  - Consumption segment largest difference of 6 index points for "Chocolate, assortment", and showing a strong seasonal structure.

- The seasonality studies reinforce the hypothesis that most outliers come from dump prices.

Office for National Statistics

Future development

- ONS plan to introduce grocery scanner data in 2025 according to our programme of transformation across UK consumer price statistics.

- ONS might explore outlier detection at transaction price level, which might allow to remove only the outlier transaction(s) instead of all transactions in a month.

Conclusions

- The Outlier detection for grocery scanner data in consumer price statistics was presented, discussing the impact of outlier detection methodologies on grocery scanner data.
- Several outlier detection strategies were discussed, and the chosen method combines price relative and price-quantity relative filters.
- The impact of the method on indices depends on the level of aggregation, ranging from 0.2 to 6 index points.
- The indices show a seasonal pattern due to the removal of dump prices.

Thanks for your attention!