Outlier detection for alternative data sources

Mario Spina

Meeting of the Group of Experts on Consumer Price Indices,
7-9 June 2023, Geneva, Switzerland

Introduction

- Background to data cleaning
  - Junk filters vs outlier detection
  - Main application & methods
- Results:
  - Second-hand cars
  - Rail fares
  - Discussion
- Future developments and conclusions

Background to data cleaning

- Introducing new, bigger data sources in CPI, bi-annual research
- Transforming rail fares and second-hand cars first
- New methods and techniques to ensure high-quality
- Adapting existing strategies to big data
- Data cleaning selects transactions used for index calculation

Junk filters vs outlier detection

Data cleaning consists of two underlying components:

Junk filter

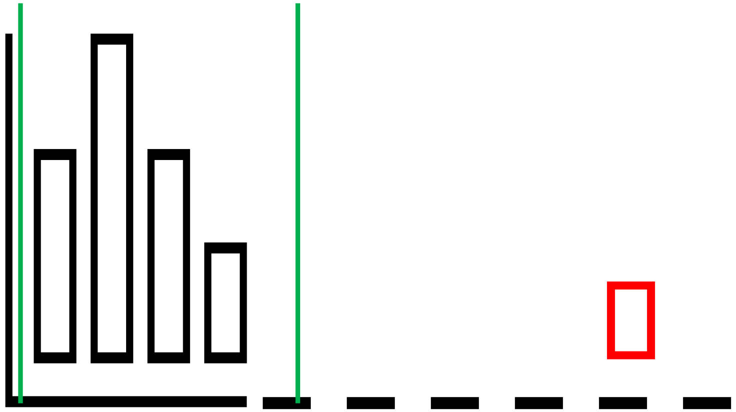
Determines observations out of scope by removing as example:

- ‘minibus’ from cars
- ‘underground’ fares from rail fares

More information on junk filters is available at this publication

Outlier detection

Identifies products with extreme and potentially erroneous prices or price movements



Main applications & methods

We investigated three applications of outlier detection:

- Global (transaction-level, global distribution)
- Observation-level (transaction-level, product distribution)
- Relative-based (unit value-level, global distribution)

Main applications & methods

Methods explored in the publication:

Method	Fences
User-defined fence	LF, UF: Manually selected
Tukey (interquartile)	LF: Q1 – k*(Q3-Q1) UF: Q3 + k*(Q3-Q1)
Kimber	LF: Q1 - k*(Q2-Q1) UF: Q3 + k*(Q3-Q2)
k-sigma	LF: mean – k*sd UF: mean + k*sd
Benchmark	No fences

- Note: Q1, Q2, Q3 are the first, second or third interquartile respectively
- mean and sd are mean value and standard deviation of a gaussian distribution

Case studies

- Explored a combination of applications and methods
- Second-hand petrol cars
  - Diesel cars in backup
- Rail fares

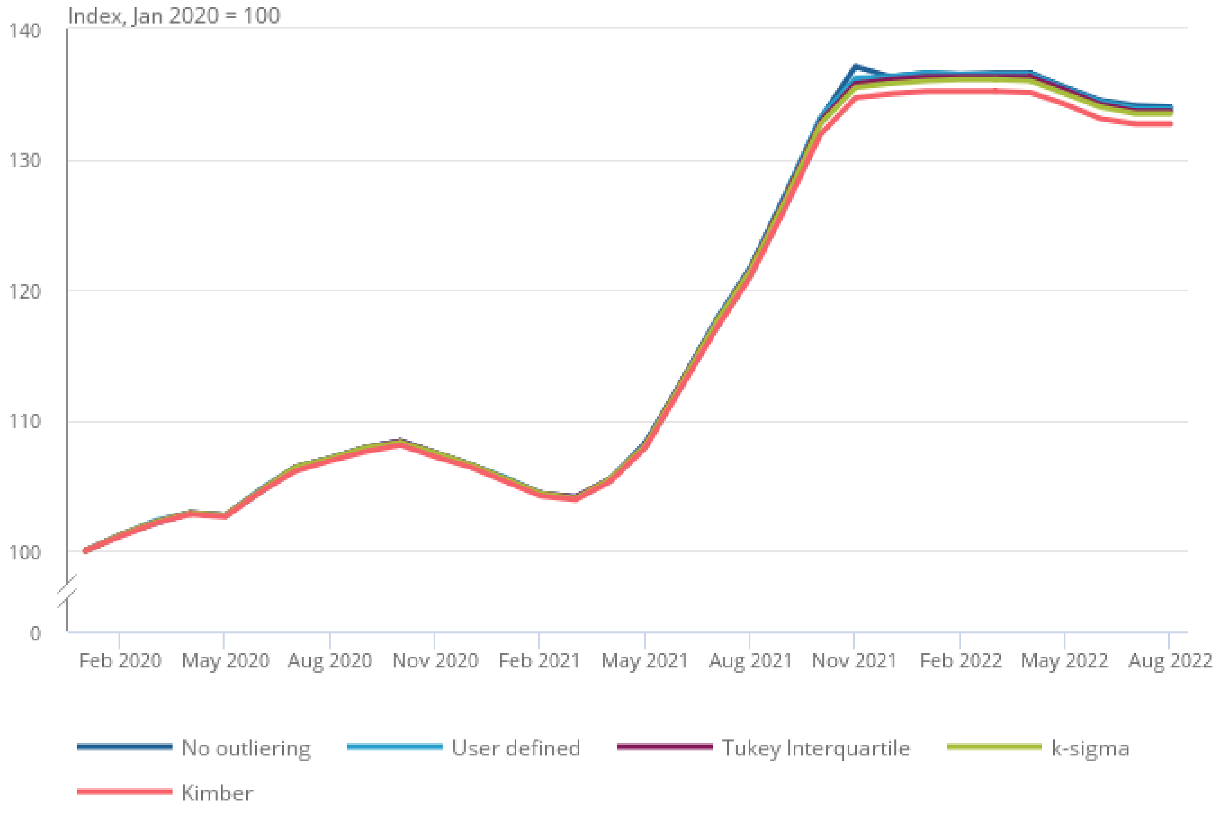
Results: second-hand cars

Methods of outlier detection explored with second-had cars

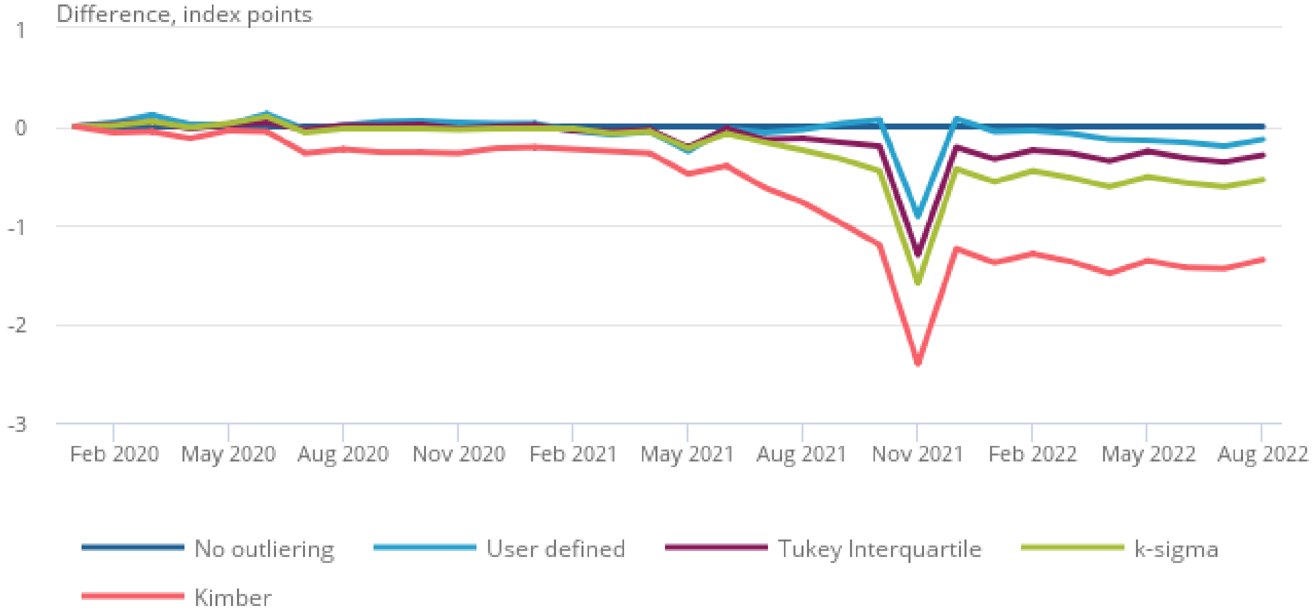
Approach	Method	Parameters	Flagged, petrol (%)	Flagged, diesel (%)
Benchmark	No outlier detection removal	N/A	0	0%
Global	User-defined	LF = 400, UF = 60000	0.91%	0.29%
Observation	Tukey (interquartile)	k = 3	0.15%	0.10%
Observation	Kimber	k = 3	1.18%	0.89%
Observation	k-sigma	k = 3	0.21%	0.16%
Relative	User-defined	LF= 1/3, UF = 3	0.03%	0.04%
Relative	Tukey (interquartile)	k = 3	1.56%	0.96%
Relative	Kimber	k = 3	5.04%	3.41%
Relative	k-sigma	k = 3	0.90%	0.67%

Results: second-hand petrol cars

Global and observation-based methods

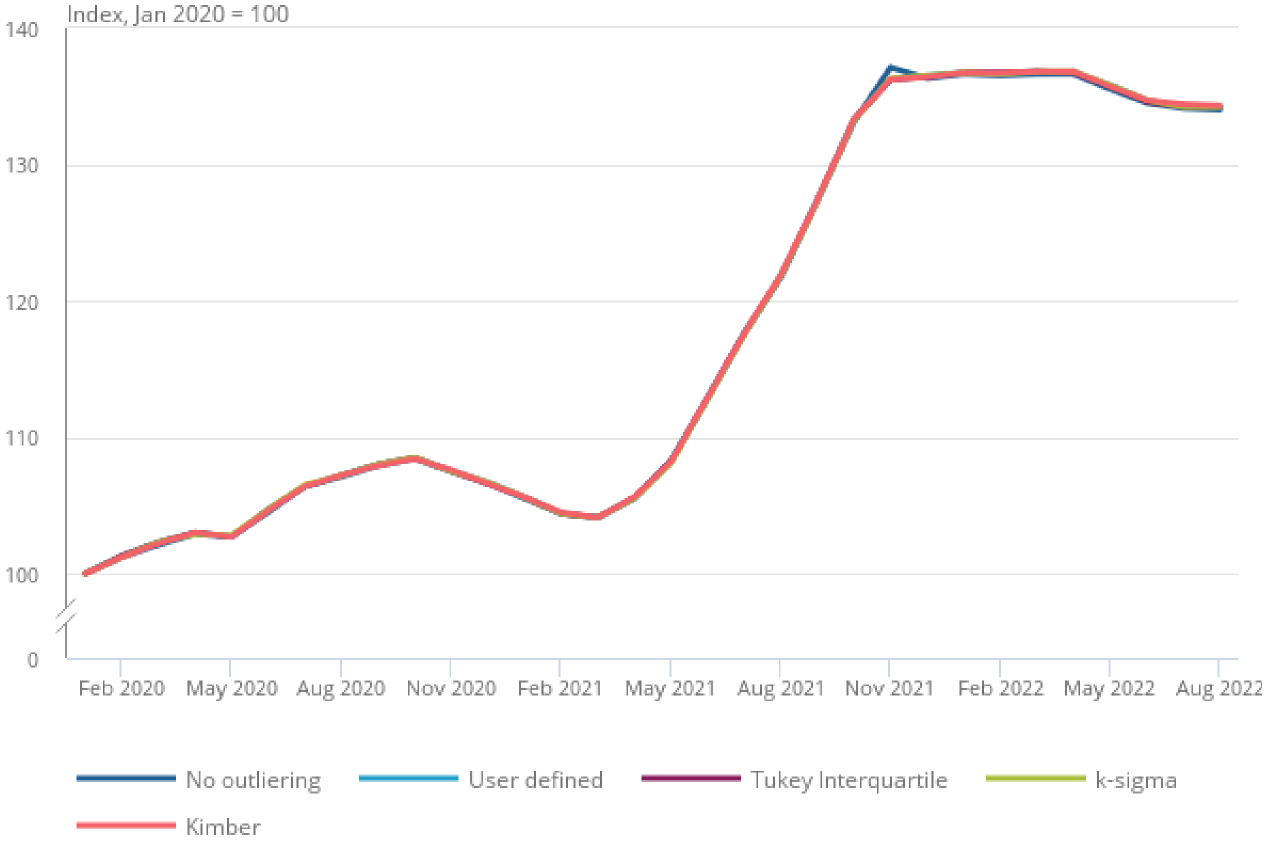


- Observation-based methods biased

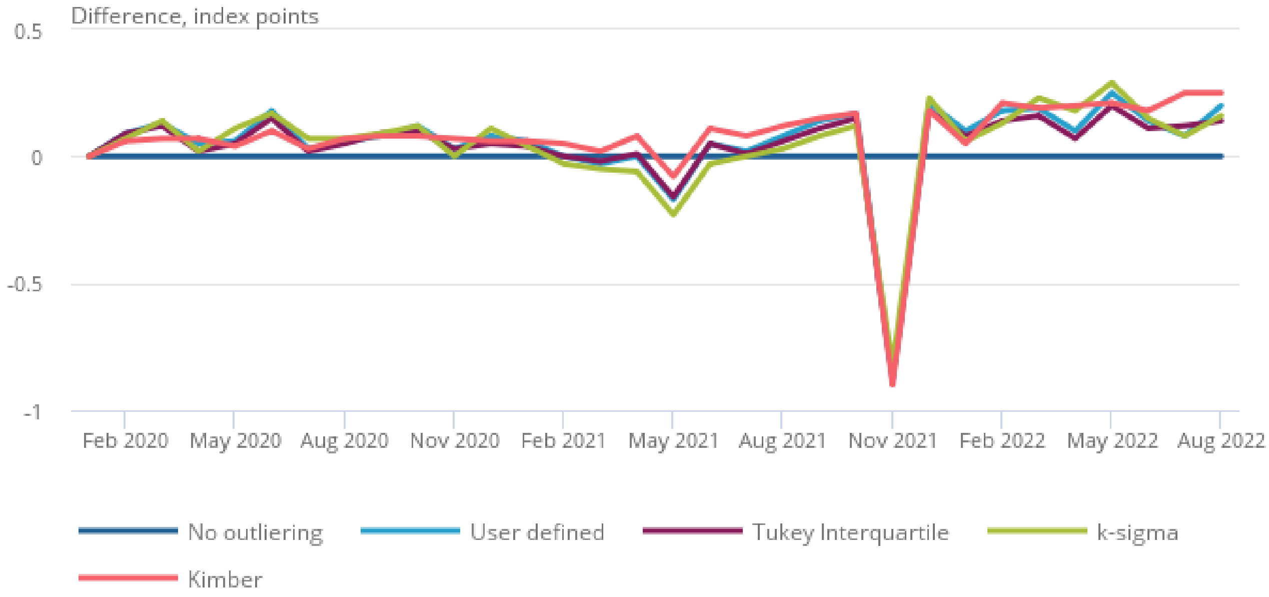


Results: second-hand petrol cars

Relative-based methods



- Methods behave similarly



Results: rail fares

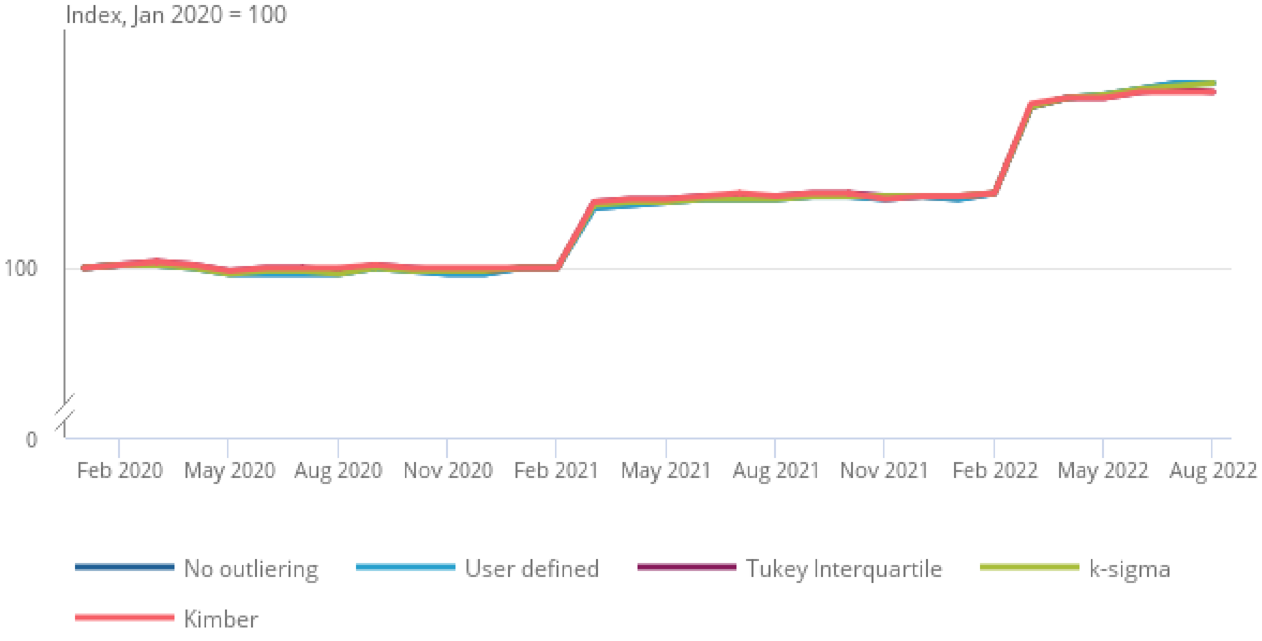
Methods of outlier detection explored with rail fares

- Negligible impact of global outlier detection
- Observation-base strategy not applicable due to bimodal distributions

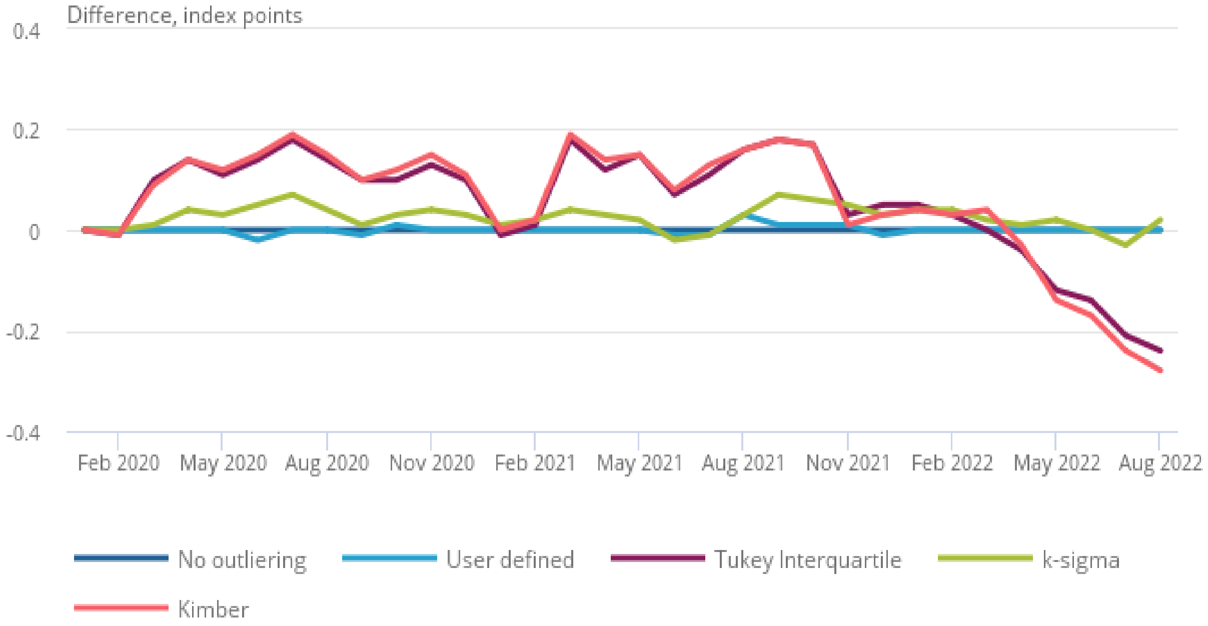
Type	Method	Parameters	Flagged	Percent
Benchmark	No outlier detection	N/A	0	0%
Relative	User-defined	LF = 1/3, UF = 3	132,796	0.02%
Relative	Kimber	k = 3	182,006,519	29.91%
Relative	k-sigma	k = 3	5,751,068	0.95%
Relative	Tukey (interquartile)	k = 3	145,194,524	23.85%

Results: rail fares

Relative-based methods



- Difference affected by narrow distribution of relatives



Results: discussion

We prefer relative-based outlier detection with a user-defined lower fence of one third and upper fence of 3

- Corrects potentially erroneous spikes
- Very mild change otherwise
- Removes minimal data
  - Reduces risk no-price-change bias
- Reduces outdated fences risk
- Avoids risk of poor fit
- Consistent across categories
  - Bespoke k parameter

Future developments & Conclusions

- Monitoring outliers and indices to avoid bias
- Account for genuine large relatives
- Exploring outlier detection on grocery scanner data
  - Investigating other methodologies

Future developments & Conclusions

- Presented Outlier detection for rail fares and second-hand cars dynamic price data

- Discussed potential strategies

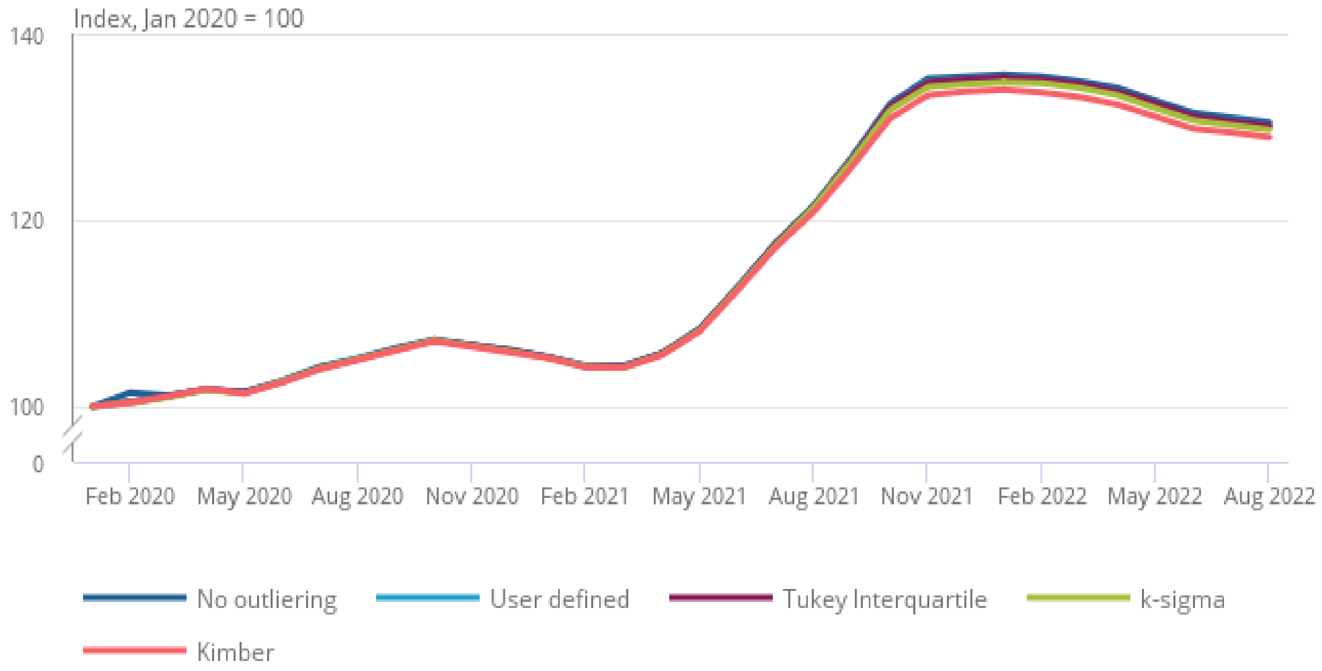
- Relative-based outlier detection
  - Mild impact on indices
  - 0.25 and 0.03 index points for used cars and rail fares

- Future application to new data sources

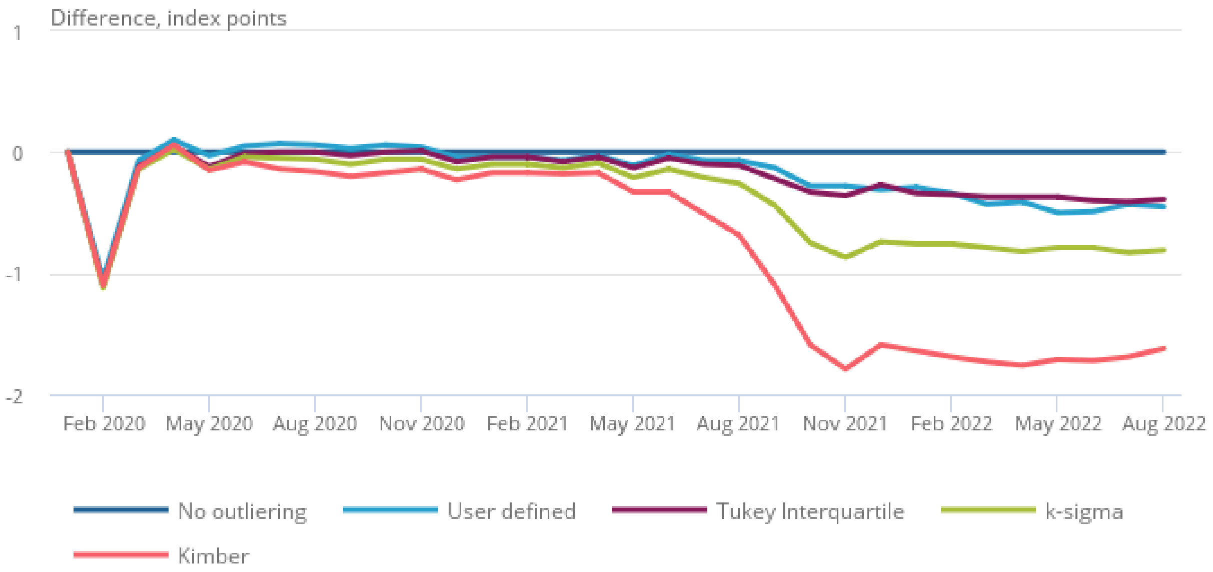
Thanks for your attention!

Results: second-hand diesel cars

Global and observation-based methods

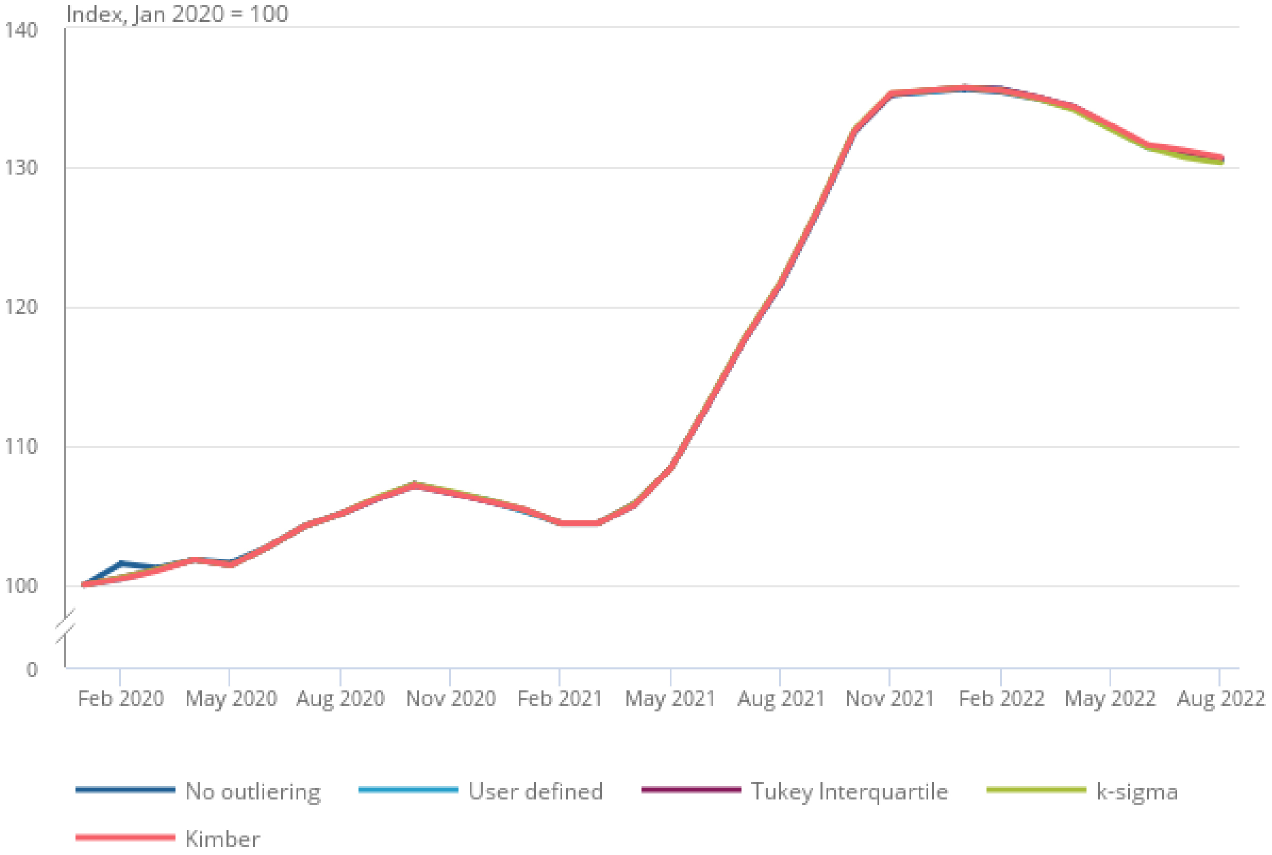


- Observation-based methods biased

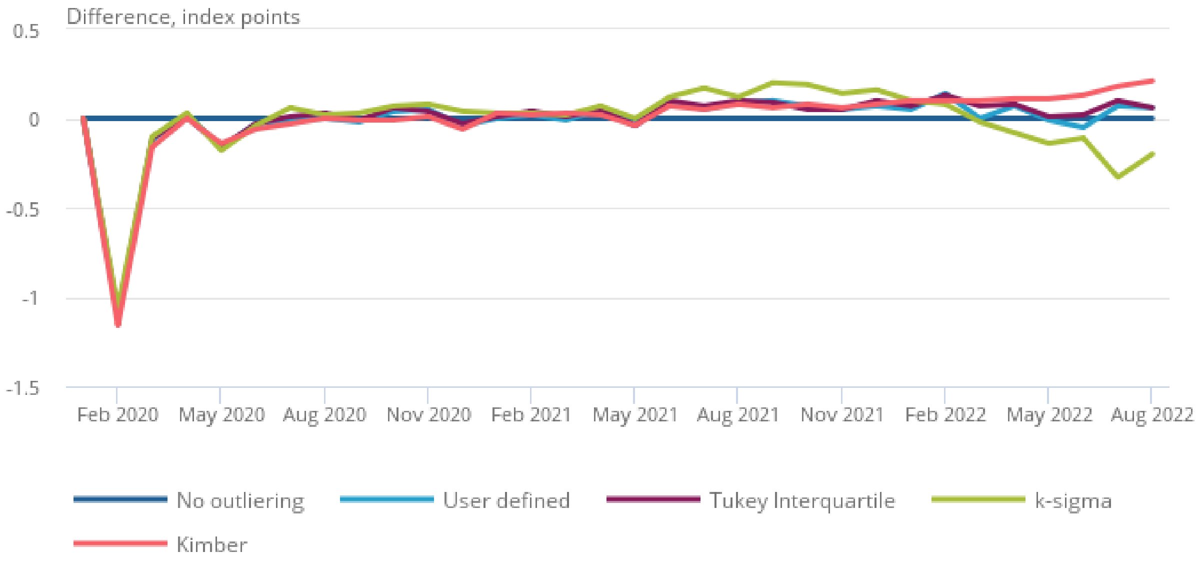


Results: second-hand diesel cars

Relative-based methods



- Methods behave similarly


Difference, index points
Feb 2020
May 2020
Aug 2020
Nov 2020
Feb 2021
May 2021
Aug 2021
Nov 2021
Feb 2022
May 2022
Aug 2022
No outliering
User defined
Tukey Interquartile
k-sigma
Kimber