From Reviews to Real Output

John Lourenze Poquiz[1, 2]

ESCoE Conference
19-21 May 2026

[1]Bennett School of Public Policy, University of Cambridge
[2]Economic Statistics Centre of Excellence

Accounting for quality change

$$Y_t^n \Rightarrow Y_t^r$$

- From nominal (Y_t^n) to real (Y_t^r), we adjust for prices.
- We also account for quality change.

- Account for changes in features
- better camera, faster software, USB Type C.

Accounting for quality change

$$Y_t^n \Rightarrow Y_t^r$$

- From nominal (Y_t^n) to real (Y_t^r), we adjust for prices.
- We also account for quality change.

$$\text{iPhone 16} \Rightarrow \text{iPhone 17}$$

- Account for changes in features
- better camera, faster software, USB Type C.

Why is this important?

$$Y_t^n = P_t \cdot Q_t$$

- Y_t^n**:** Output, P_t**:** Price, Q_t**:**

  The quantities compared over time must be those for homogeneous items and the resulting quantity changes for different goods and services must be weighted by their economic importance, as measured by their relative values in one or other, or both, periods. **For this reason volume is a more correct and appropriate term than quantity in order to emphasize that quantities must be adjusted to reflect changes in quality (2008 SNA par 15.13).**

Why is this important?

$$Y_t^n = P_t \cdot Q_t$$

- Y_t^n: Output, P_t: Price, Q_t: Volume

The quantities compared over time must be those for homogeneous items and the resulting quantity changes for different goods and services must be weighted by their economic importance, as measured by their relative values in one or other, or both, periods. **For this reason volume is a more correct and appropriate term than quantity in order to emphasize that quantities must be adjusted to reflect changes in quality (2008 SNA par 15.13).**

Quality Adjustment in the National Accounts

ICT Products

- Computers and mobile phones (Pakes 2003; Schwartz and Scafidi 2009)
- Scanner data (De Haan, Hendriks, and Scholz 2021)

Public Sector

- Health, education (Weale 2007; Gemmell, Nolan, and Scobie 2018)

Why are services different?

Services are the result of a production activity that **changes the conditions of the consuming units**, *or facilitates the exchange of products or financial assets (2008 SNA par 6.17 & 2025 SNA par 7.17).*

Types of change:

- Changes in the physical condition of persons
- Changes in the mental condition of persons

How do we capture:

- A restaurant hiring an award-winning chef.
- A hotel opening a heated indoor pool.
- Better staffing resulting to shorter queues in a fast food joint

What we do...

- We use **sentiment analysis** on consumer reviews to construct an annual index of perceived service quality.

  $\rightarrow$ The index captures implied improvements and deterioration in user experience.
- We integrate this sentiment-based quality index into an **output deflation**, treating sentiment as a proxy for unobserved quality change.
- We adjust real output (value added) of the **Accommodation and Food Services industries** (SIC I) for changes in service quality.

  $\rightarrow$ This aligns volume measures with the *SNA* concept of volume as reflecting both quantity and quality.

Measuring Quality

Lancaster (1966): Services as bundles

A service produces quality $z \geq 0$. Consumer θ derives utility:

$$u = q + \theta z$$

Participation requires $\theta z \geq p$. The *marginal consumer* has type:

$$\theta^* = \frac{p}{z}$$

The **hedonic price function** $P(z) = \theta^* z$ embeds the marginal consumer's valuation of quality.

Implication for measurement

If quality z changes, a conventional price index that ignores z conflates quality change with inflation — real output is mis-measured.

Sentiment as a Quality Proxy

Reviewers report experienced quality z with noise:

$$S_{jt} = z_{jt} + \varepsilon_{jt}, \qquad \mathbb{E}[\varepsilon_{jt}] = 0$$

The revenue-weighted index $\bar{S}_t = \sum_j \omega^r_{jt} S_{jt}$ proxies Δz.

Substituting into the quality-adjusted deflator:

$$P^*_L = \mathring{P}_L - \hat{\alpha} \cdot \Delta \ln \bar{S}_t$$

where $\hat{\alpha}$ scales sentiment changes into deflator units.

$\Delta \ln \bar{S}_t > 0$ (quality up) $\Rightarrow$ deflator falls $\Rightarrow$ **real output rises**

Measuring feelings

Definition

Sentiment analysis is a computational technique that identifies and extracts subjective information from text, determining whether the expressed opinion is positive, negative, or neutral.

Positive	Neutral	Negative
"Excellent service, highly recommend!"	"The restaurant opened at 6pm."	"Terrible food, never coming back."

- Widely used in customer reviews, social media, market research
- Enables quantitative analysis of qualitative data
- Can reveal trends in consumer opinions over time

Reviews as sentiment

Pros

- Readily available and direct representation of *perceived* service quality.
- Captures intangible aspects of products (e.g., cleanliness, convenience, experience, ambiance).

Cons

- Reviewers are not a random sample and may be biased towards people with strong feelings.

Is this okay?

We focus on *changes over time* rather than levels; if the bias is stable, it cancels out in the dynamics.

Reviews as sentiment

Pros

- Readily available and direct representation of *perceived* service quality.
- Captures intangible aspects of products (e.g., cleanliness, convenience, experience, ambiance).

Cons

- Reviewers are not a random sample and may be biased towards people with strong feelings.

Is this okay?

We focus on *changes over time* rather than levels; if the bias is stable, it cancels out in the dynamics.

Approaches to Sentiment Analysis

Lexicon-based

- **Syuzhet / Bing**
- Word-level
- Ignores context
- Unbounded scale
- Fast, scalable
- Misses negation
- **Use:** trends, large samples

Transformer-based

- **BERT**
- Token-level
- Full bidirectional
- Probability [-1, +1]
- Highest accuracy
- Expensive, black-box
- **Use:** complex language
- Non-reproducible (stochastic)

Lexicon + shifters

- **Sentimentr**
- Sentence-level
- Context modifiers
- Normalized by word count
- Captures nuance
- More intensive
- **Use:** precision, quality
- Does not capture sarcasm

Example: *"The food was not bad but service was very slow."*

Approaches to Sentiment Analysis

Lexicon-based

- **Syuzhet / Bing**
- Word-level
- Ignores context
- Unbounded scale
- Fast, scalable
- Misses negation
- **Use:** trends, large samples

Transformer-based

- **BERT**
- Token-level
- Full bidirectional
- Probability [-1, +1]
- Highest accuracy
- Expensive, black-box
- **Use:** complex language
- Non-reproducible (stochastic)

Lexicon + shifters

- **Sentimentr**
- Sentence-level
- Context modifiers
- Normalized by word count
- Captures nuance
- More intensive
- **Use:** precision, quality
- Does not capture sarcasm

Example: *"The food was not bad but service was very slow."*

Approaches to Sentiment Analysis

Lexicon-based

- **Syuzhet / Bing**
- Word-level
- Ignores context
- Unbounded scale
- Fast, scalable
- Misses negation
- **Use:** trends, large samples

Transformer-based

- **BERT**
- Token-level
- Full bidirectional
- Probability [-1, +1]
- Highest accuracy
- Expensive, black-box
- **Use:** complex language
- Non-reproducible (stochastic)

Lexicon + shifters

- **Sentimentr**
- Sentence-level
- Context modifiers
- Normalized by word count
- Captures nuance
- More intensive
- **Use:** precision, quality
- Does not capture sarcasm

Example: *"The food was not bad but service was very slow."*

Same Rating, Different Stories

"Very aggressive nasty unhelpful people. Very aggressive nasty unhelpful people, only concerned about how many millions of pounds they will get each years. Definitely do not book any holiday with these rogue traders. Be warned."

Rating: 1/5
Sentiment: -2.41

"Very disappointed unfortunately. The lodge was in a lovely spot but was not well maintained inside and it was filthy. There was lots of hair in the bath plug hole, the bath and tiles were dirty, there was toothpaste all over the sink. In all very disappointing."

Rating: 1/5
Sentiment: -2.07

Where to Source Reviews?

Trustpilot

- Covers a broad range of service industries beyond hospitality.
- Easier access for large-scale text extraction.
- Include “verified” and “invited” reviews indicator
- Data is at the **enterprise level** *not* **establishment level**.

Note: I will be using other portals such as **Tripadvisor** in the next iteration.

Initial results

Can we trust Trustpilot reviews?

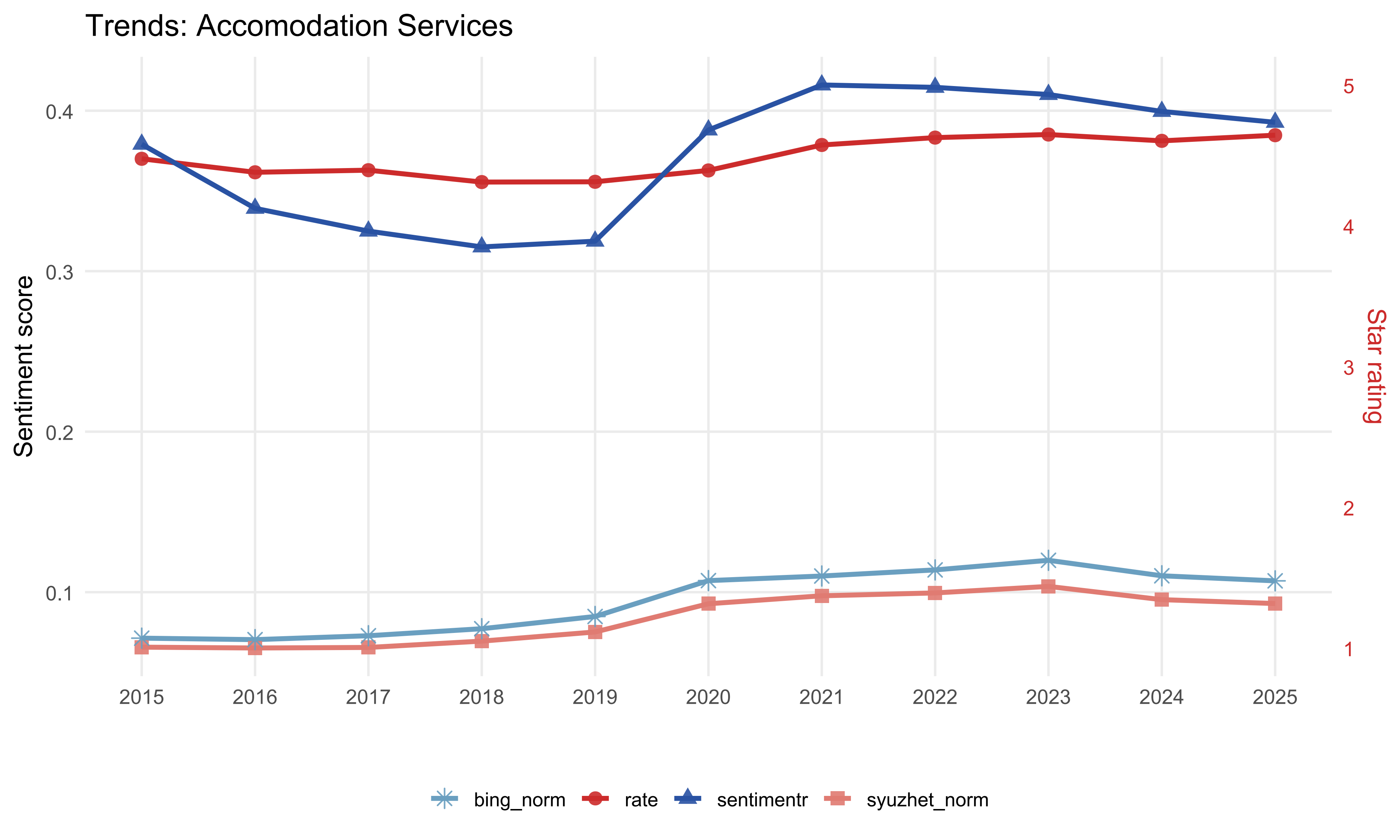
	N	Share (%)	Mean reviews
Panel A: Accommodation Services			
Invited	54,335	10.44	329.30
Not verified	183,803	35.31	173.24
Verified	282,330	54.25	1,509.79
Panel B: Food Services			
Invited	10,988	3.44	159.25
Not verified	119,291	37.33	76.22
Verified	189,287	59.23	1,051.59

Data: By reviews

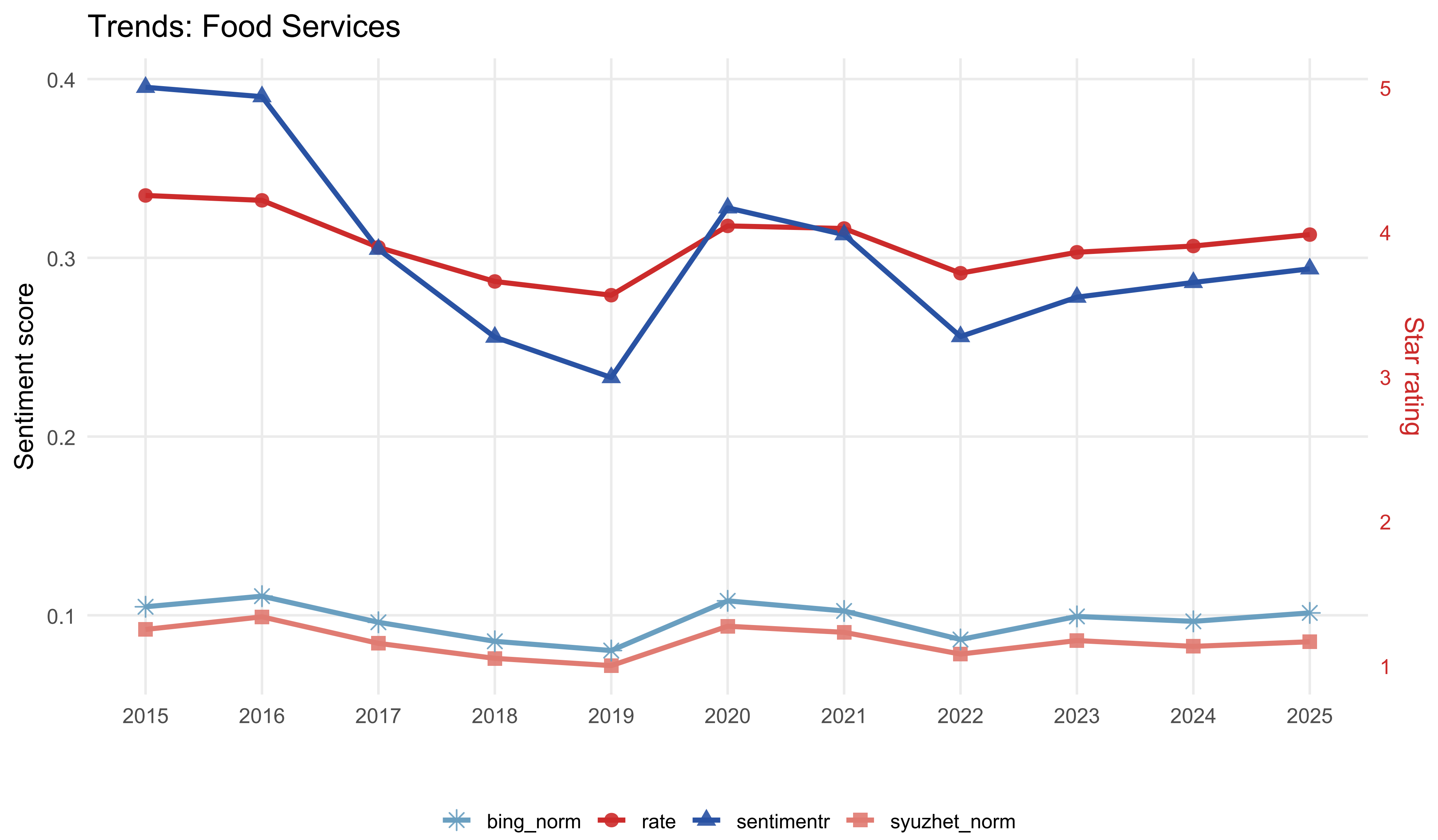
	N	Mean	St. Dev.	Min	Max
Panel A: Accommodation Services					
Rating	520,464	4.37	1.23	1	5
SentimentR	520,468	0.35	0.40	−2.76	3.05
Syuzhet	520,468	2.30	2.34	−22.80	50.45
Bing	520,468	2.37	3.05	−32	50
Panel B: Food Services					
Rating	319,565	3.92	1.63	1	5
SentimentR	319,566	0.30	0.45	−2.78	3.61
Syuzhet	319,566	1.78	2.05	−17.85	44.80
Bing	319,566	1.68	2.68	−27	49

Data: By firm

	N	Mean	St. Dev.	Min	Max
Panel A: Accommodation Services					
Rating	1,073	3.88	1.40	1.00	5.00
SentimentR	1,073	0.22	0.30	−1.27	1.48
Syuzhet	1,073	3.23	2.62	−5.55	16.68
Bing	1,073	3.18	3.69	−20.00	21.50
Panel B: Food Services					
Rating	1,577	3.66	1.29	1.00	5.00
SentimentR	1,578	0.20	0.31	−1.61	1.19
Syuzhet	1,578	2.42	2.03	−8.95	16.65
Bing	1,578	2.20	2.73	−17.00	20.00



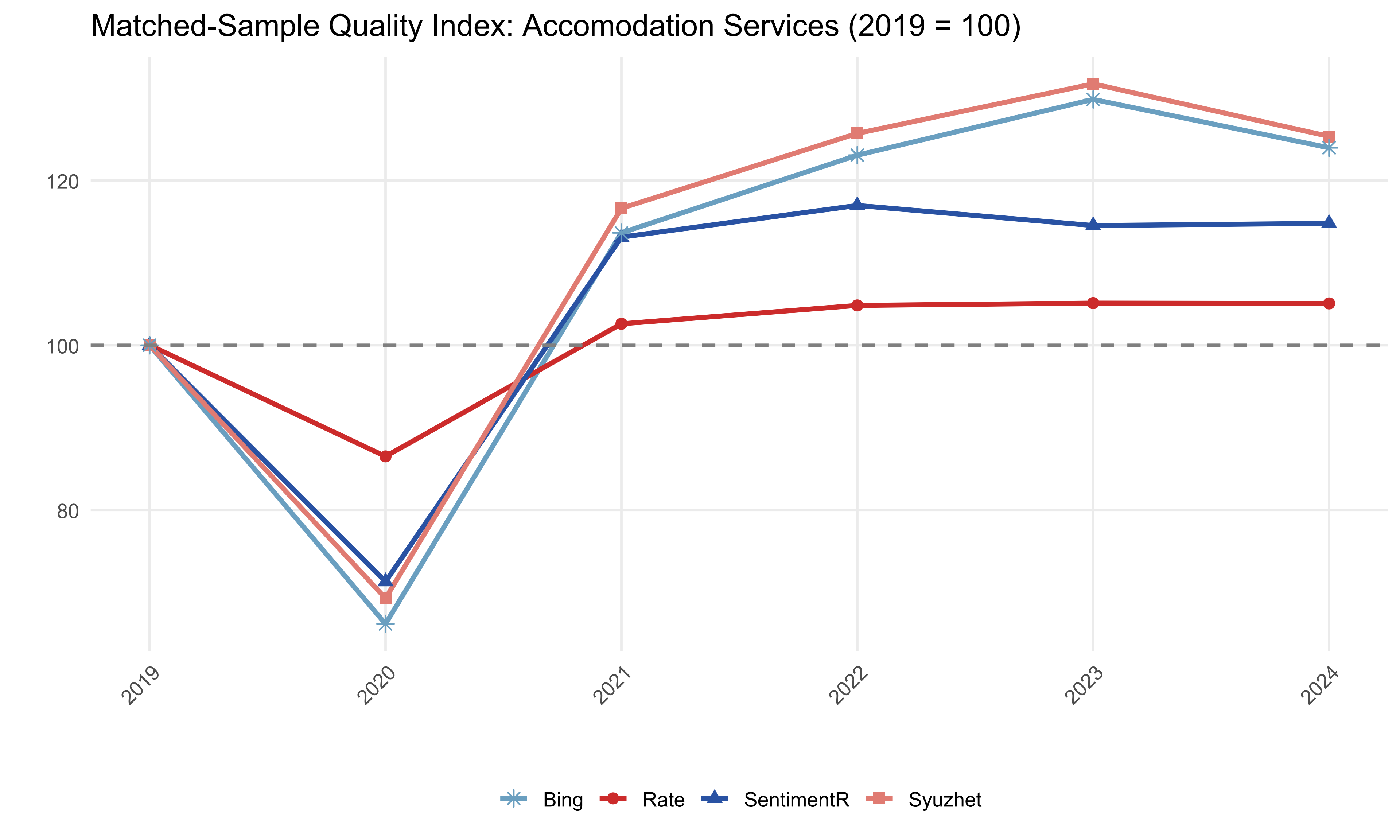
Trends: Accomodation Services
Sentiment score
Star rating
0.4
0.3
0.2
0.1
5
4
3
2
1
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
bing_norm
rate
sentimentr
syuzhet_norm

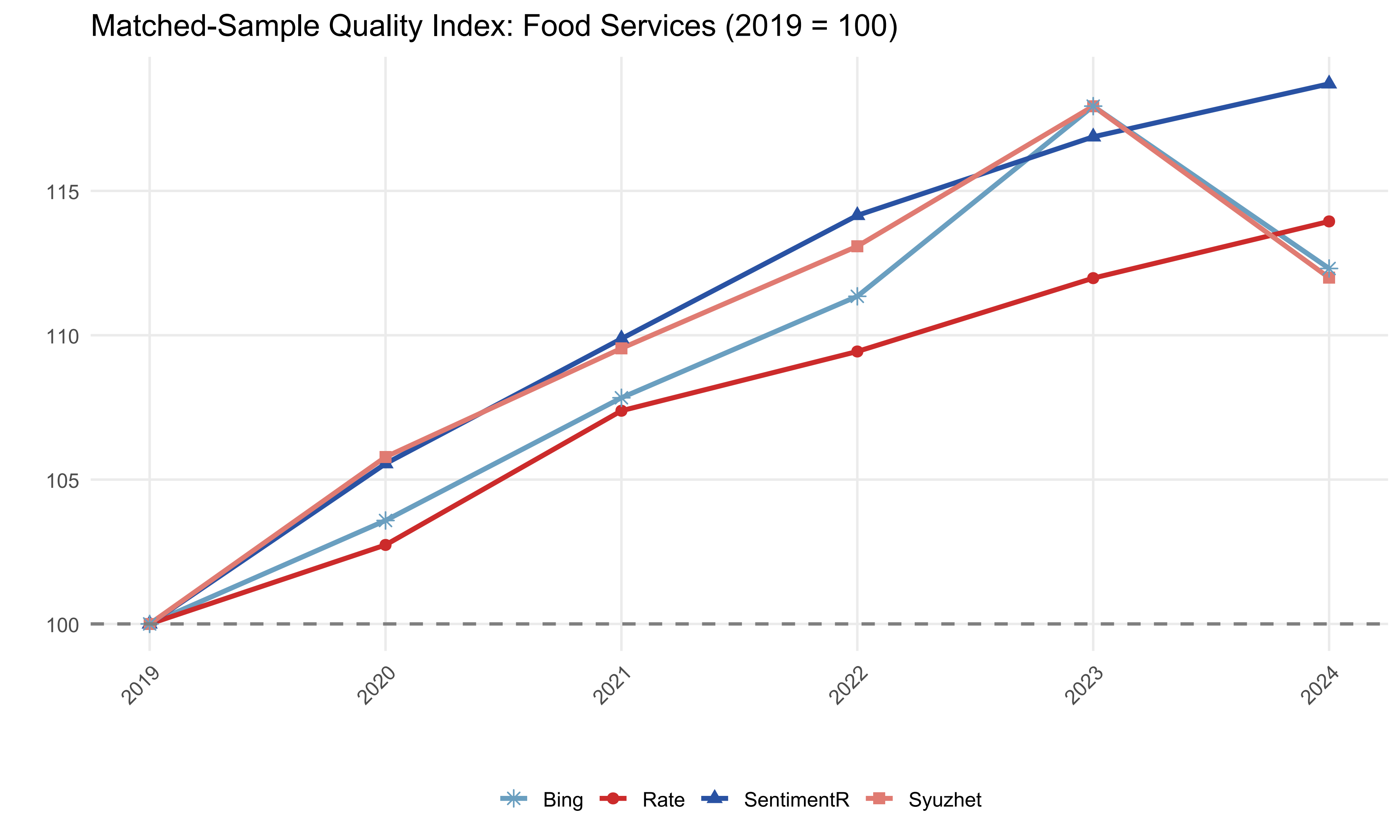


Trends: Food Services
Sentiment score
Star rating
0.4
0.3
0.2
0.1
5
4
3
2
1
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
bing_norm
rate
sentimentr
syuzhet_norm

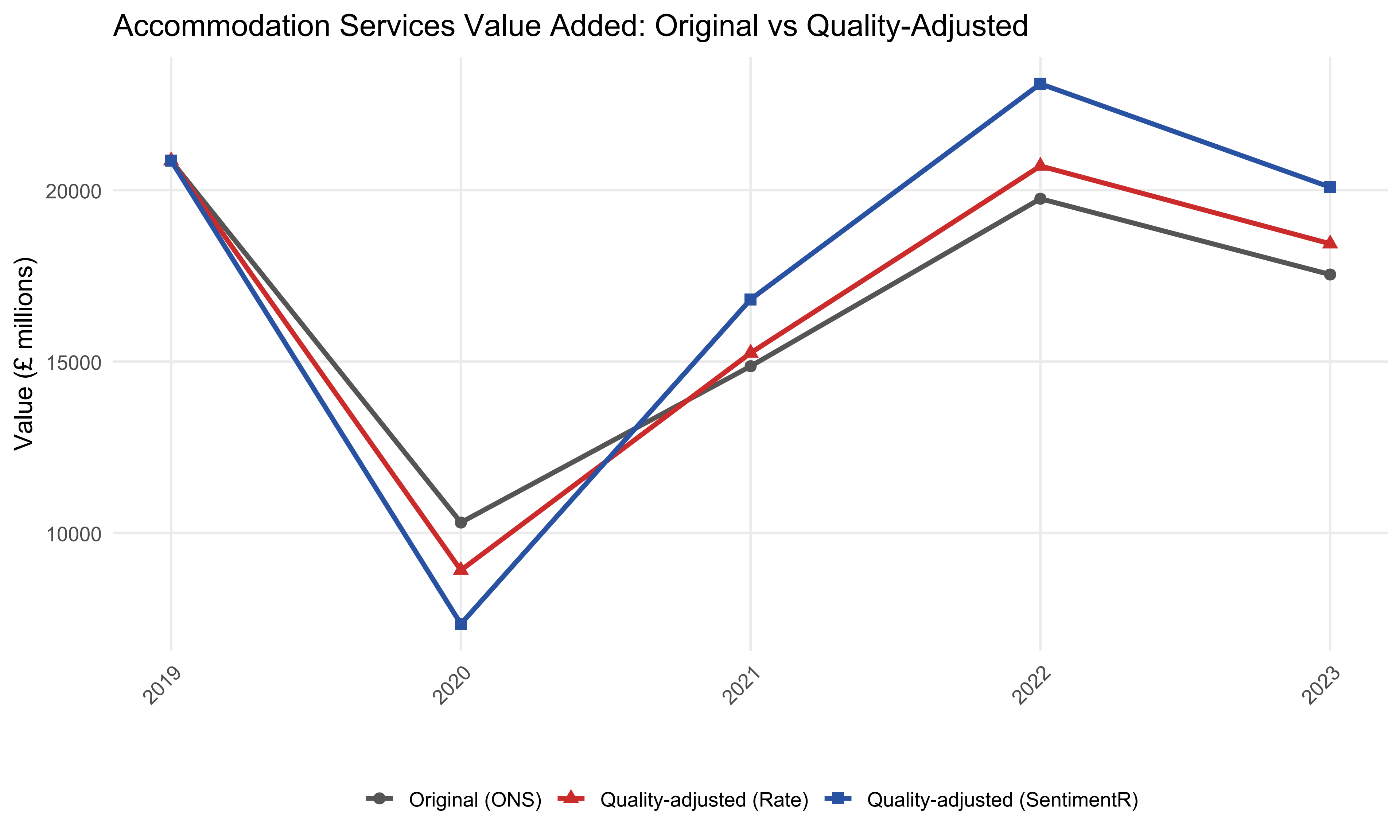
Quality Adjustment

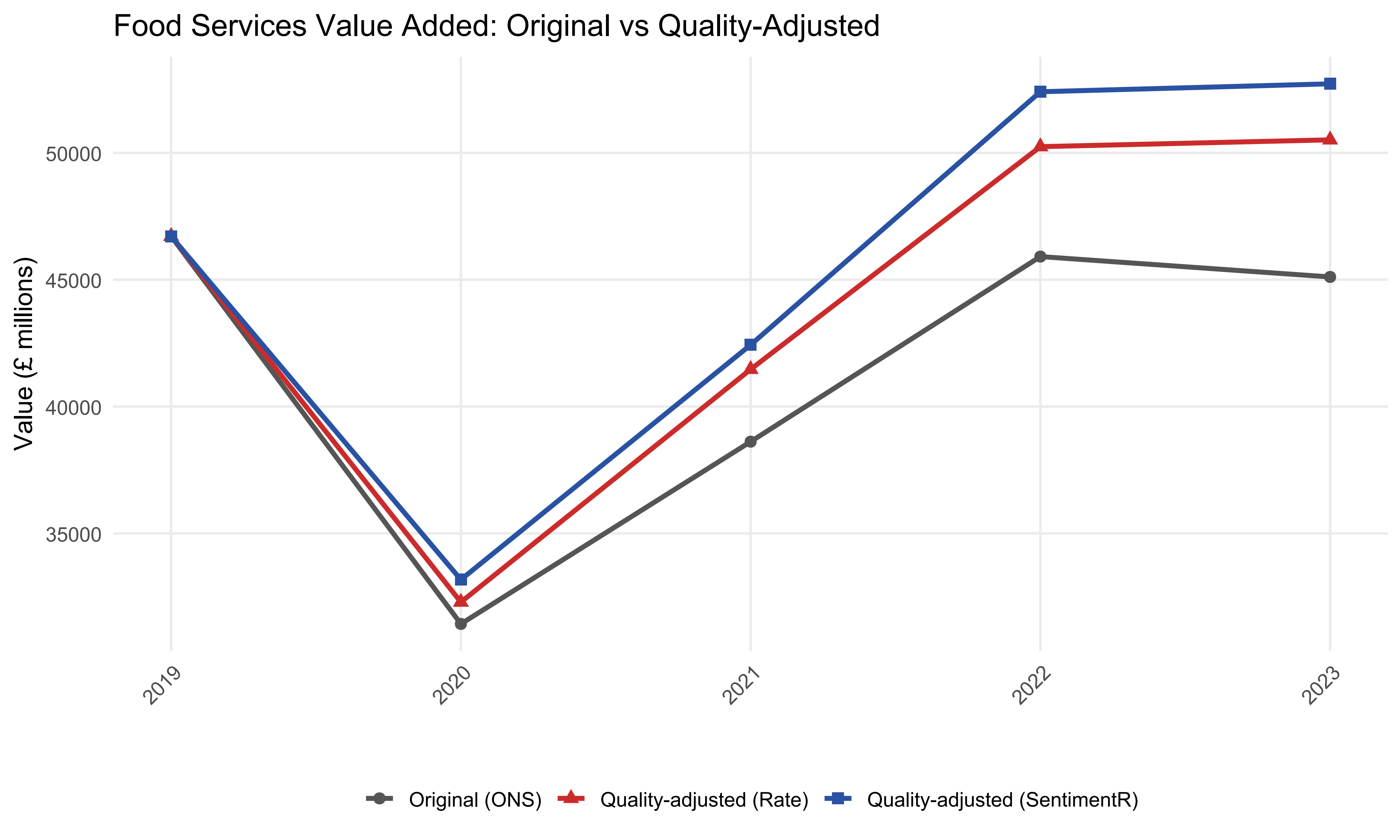
From Reviews to Quality Index — Four Steps
① Average
Mean score per establishment per year
Rest. A
mean=4.2
Rest. B
mean=3.0
Rest. C
mean=4.6
Rest. D
mean=2.4
Smooths individual reviewer noise
within each establishment
② Match
Only establishments present
in both adjacent years
2022
A: 4.0
B: 3.2
C: 4.5
D: 2.8
2023
A: 4.3
B: 3.5
D: 3.1
C: new entrant, excluded
Standard matched-model method
used by statistical agencies
③ Chain
Stack year-on-year changes into
a single index (2019 = 100)
100
119
Normalised by base-year SD
so all 4 measures are comparable
④ Adjust
Multiply ONS real output
by the quality index
Y* = Y_ONS × I_t / 100
Adj.
ONS
If quality ↑ → deflator ↓
→ real output higher than ONS
Star ratings
SentimentR
Syuzhet
Bing
Four independent measures — consistent results validate the approach

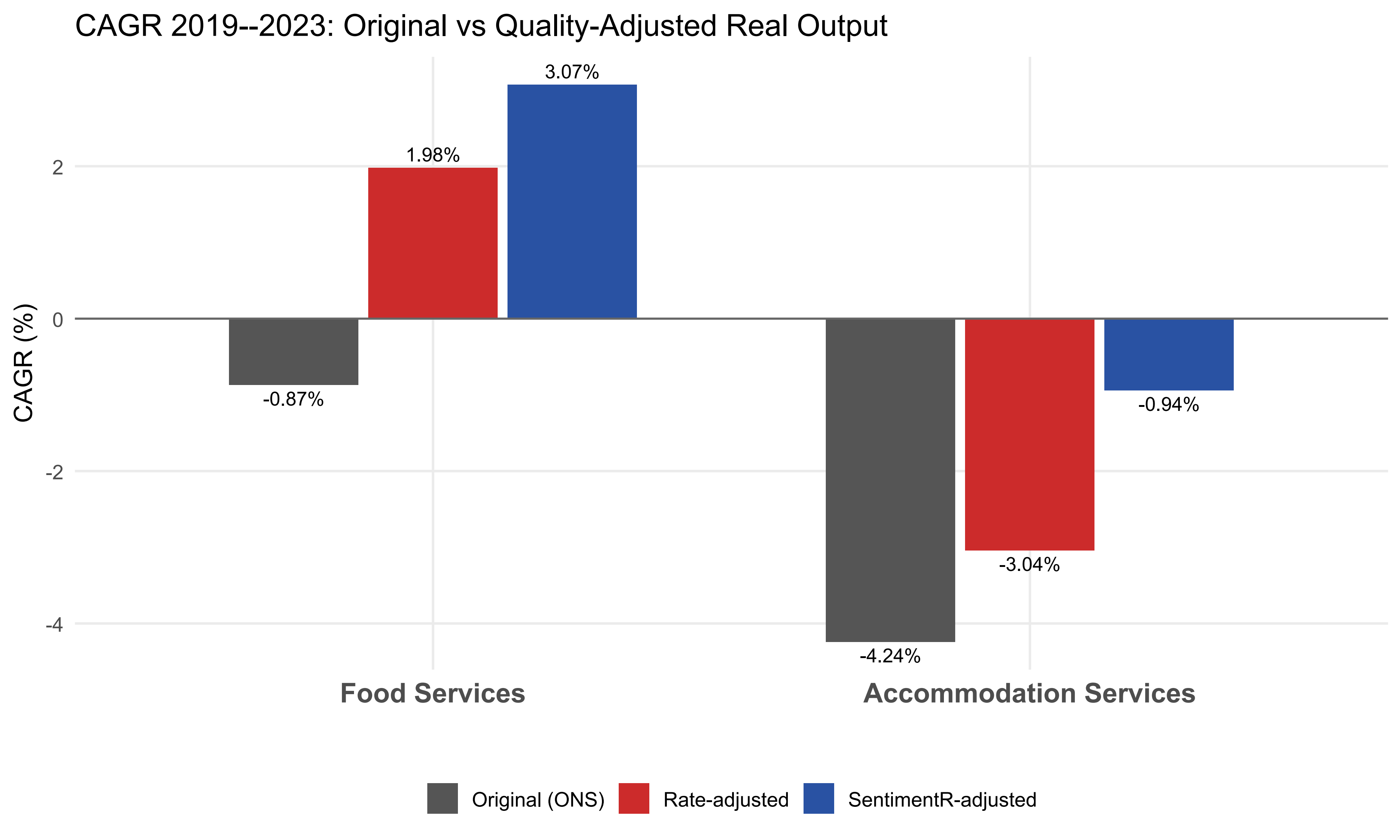

Matched-Sample Quality Index: Accomodation Services (2019 = 100)
120
100
80
2019
2020
2021
2022
2023
2024
Bing
Rate
SentimentR
Syuzhet


Matched-Sample Quality Index: Food Services (2019 = 100)
115
110
105
100
2019
2020
2021
2022
2023
2024
Bing
Rate
SentimentR
Syuzhet

Does this impact the growth story?


Accommodation Services Value Added: Original vs Quality-Adjusted
Value (£ millions)
20000
15000
10000
2019
2020
2021
2022
2023
Original (ONS)
Quality-adjusted (Rate)
Quality-adjusted (SentimentR)


Food Services Value Added: Original vs Quality-Adjusted
Value (£ millions)
50000
45000
40000
35000
2019
2020
2021
2022
2023
Original (ONS)
Quality-adjusted (Rate)
Quality-adjusted (SentimentR)


CAGR 2019--2023: Original vs Quality-Adjusted Real Output
CAGR (%)
2
0
-2
-4
-0.87%
1.98%
3.07%
-4.24%
-3.04%
-0.94%
Food Services
Accommodation Services
Original (ONS)
Rate-adjusted
SentimentR-adjusted

Key Takeaways

1. Conventional deflators overstate inflation and understate real output in service sectors by masking quality change with price change
2. Quality adjustment **materially changes** the picture, revealing fast post-COVID growth.
3. Online reviews offer a **scalable, reproducible** option for quality adjustment.

What's next?

Priory

1. Weighting establishments with revenue data (using Orbis).
2. Alternative portals: Tripadvisor, Google Reviews.
    - Weighted by traffic/reviews
3. Price effects (?)
4. Alternative NLP approaches for validation (BERT, sentiment.ai).

If able...

1. Use verified reviews only & adjust for fake reviews (if necessary).
2. Comparison with other countries: US, Canada, Australia.

Possible Extensions

Other Industries

Retail ★★

Transportation ★★★★

Communication ★★

Banking & insurance ★★

Real estate & renting ★

Recreation ★★★★

Personal services ★★★★★

If we include
Accommodation and Food Services
these industries accounts for
25% of UK GDP in 2023

Thank you

This super early stage so all comments are welcome!

jlp87@cam.ac.uk

Step 1: Constructing the Quality Index

Establishment-year average (establishment e, year t):

$$\bar{s}_{e,t} = \frac{1}{n_{e,t}} \sum_{i=1}^{n_{e,t}} s_{e,t,i}$$

Matched sample: only establishments in *both* adjacent years with $n \geq 10$:

$$\overline{\Delta S}_t = \frac{\sum_{e \in \mathcal{M}_t} w_{e,t} \left(\bar{s}_{e,t} - \bar{s}_{e,t-1}\right)}{\sum_{e \in \mathcal{M}_t} w_{e,t}}, \qquad w_{e,t} = \frac{n_{e,t-1} + n_{e,t}}{2}$$

Chained index (base 2019 = 100), normalised by base-year SD σ_0:

$$I_t = 100 \times \left(1 + \frac{\sum_{\tau=2020}^{t} \overline{\Delta S}_\tau}{\sigma_0}\right)$$

Note: Computed separately for each sentiment measure (SentimentR, Syuzhet, Bing) and star ratings.

Step 2: Quality-Adjusting Real Output

Adjust the price deflator to remove the quality component:

$$P_t^* = \frac{P_t}{I_t/100}$$

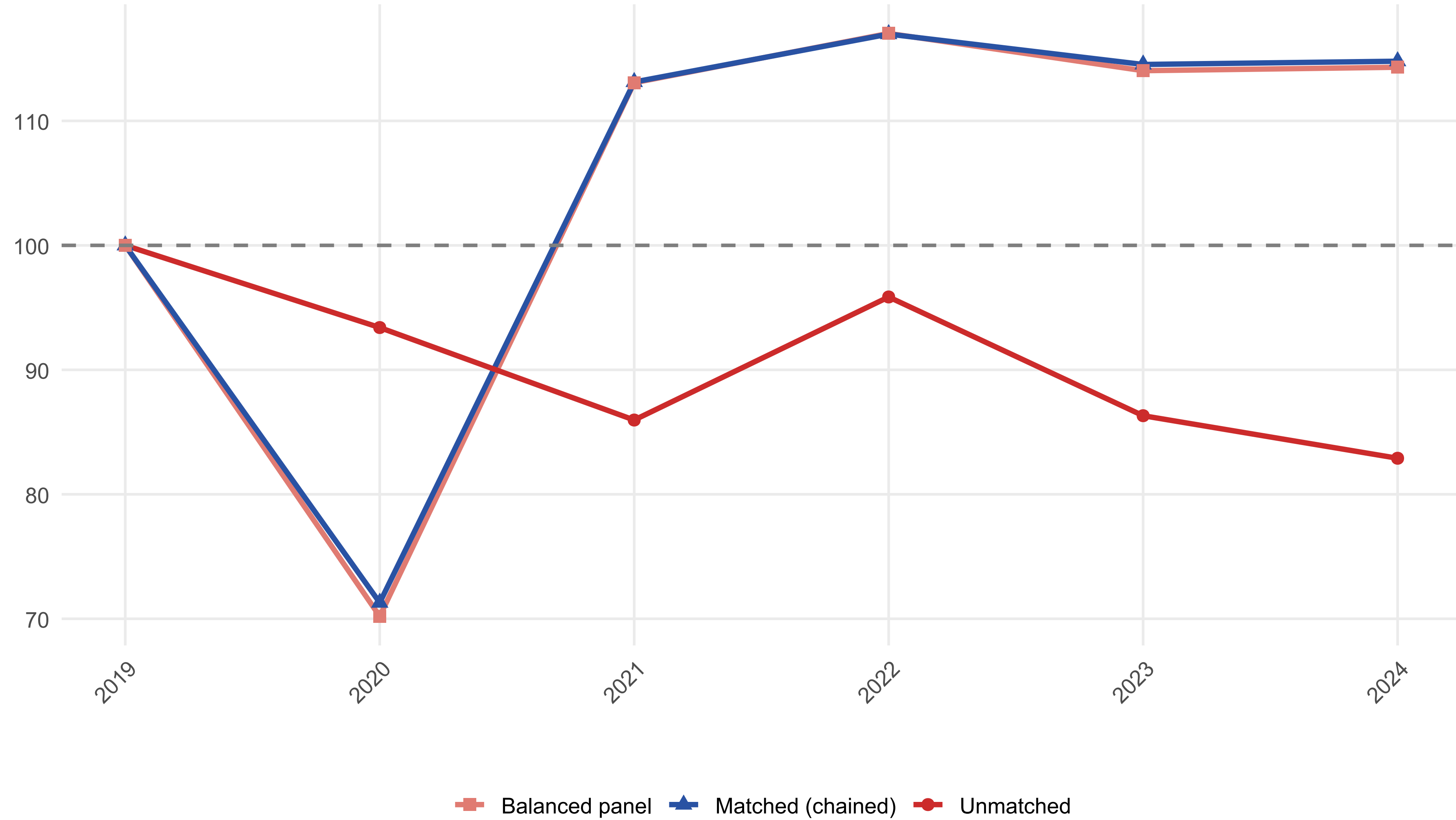
If quality improved ($I_t > 100$), $P_t^* < P_t$: part of what looked like inflation was actually better service.

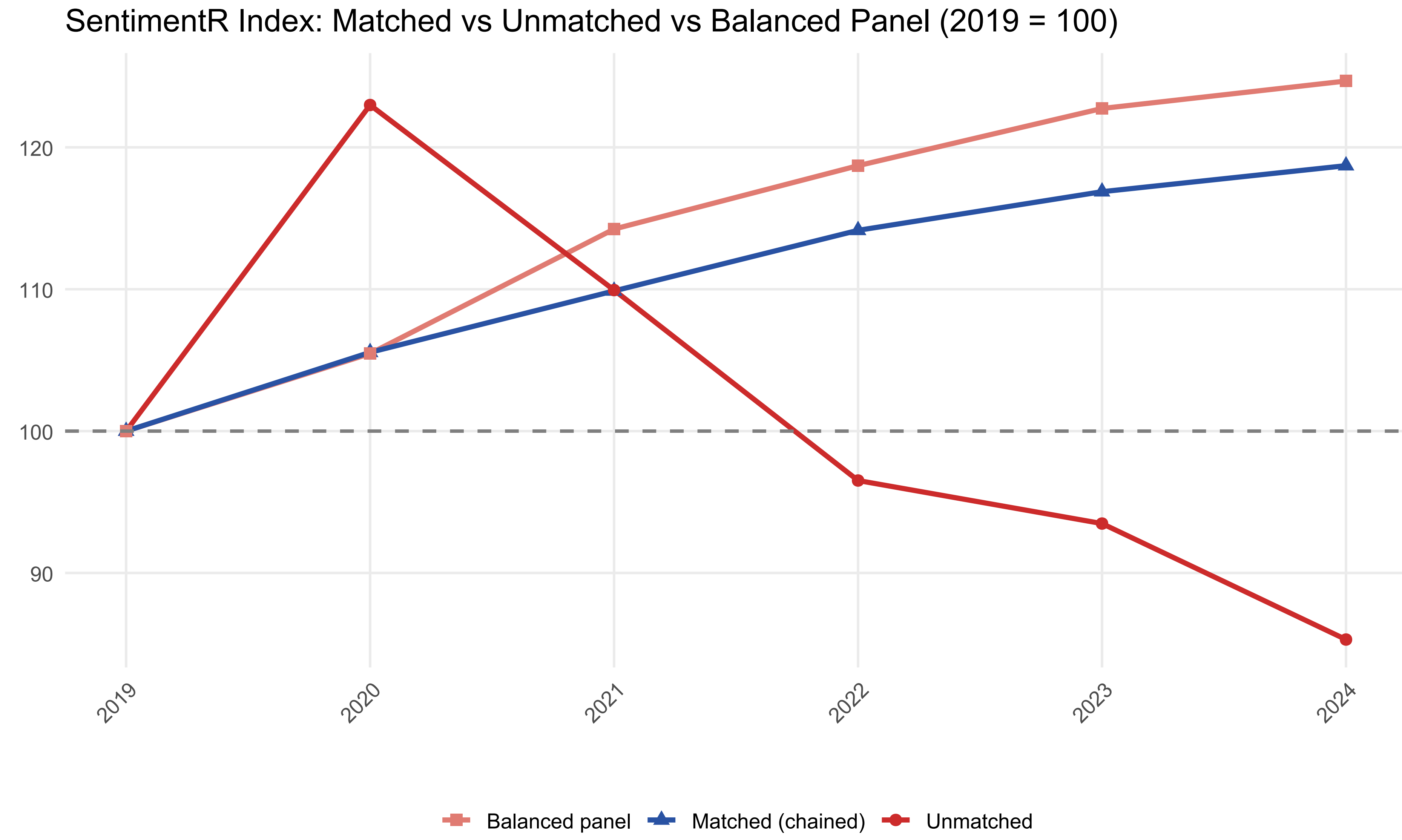
Quality-adjusted real output via standard deflation:

$$\tilde{Y}_t^r = \frac{Y_t^{\mathrm{CP}}}{P_t^*} \cdot P_{2019}^* \;=\; Y_t^{\mathrm{CVM}} \times \frac{I_t}{100}$$

- $I_t = 100$ for $t < 2019$: **no adjustment** — series identical to ONS
- $I_t > 100$: quality improved $\Rightarrow$ $\tilde{Y}_t^r$ exceeds ONS real output
- $I_t < 100$: quality declined $\Rightarrow$ $\tilde{Y}_t^r$ falls below ONS real output

SentimentR Index: Matched vs Unmatched vs Balanced Panel (2019 = 100)





SentimentR Index: Matched vs Unmatched vs Balanced Panel (2019 = 100)
120
110
100
90
2019
2020
2021
2022
2023
2024
Balanced panel
Matched (chained)
Unmatched