EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Too cheap to be true

Detecting invalid values in product prices and index values

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Finding the needle in the haystack

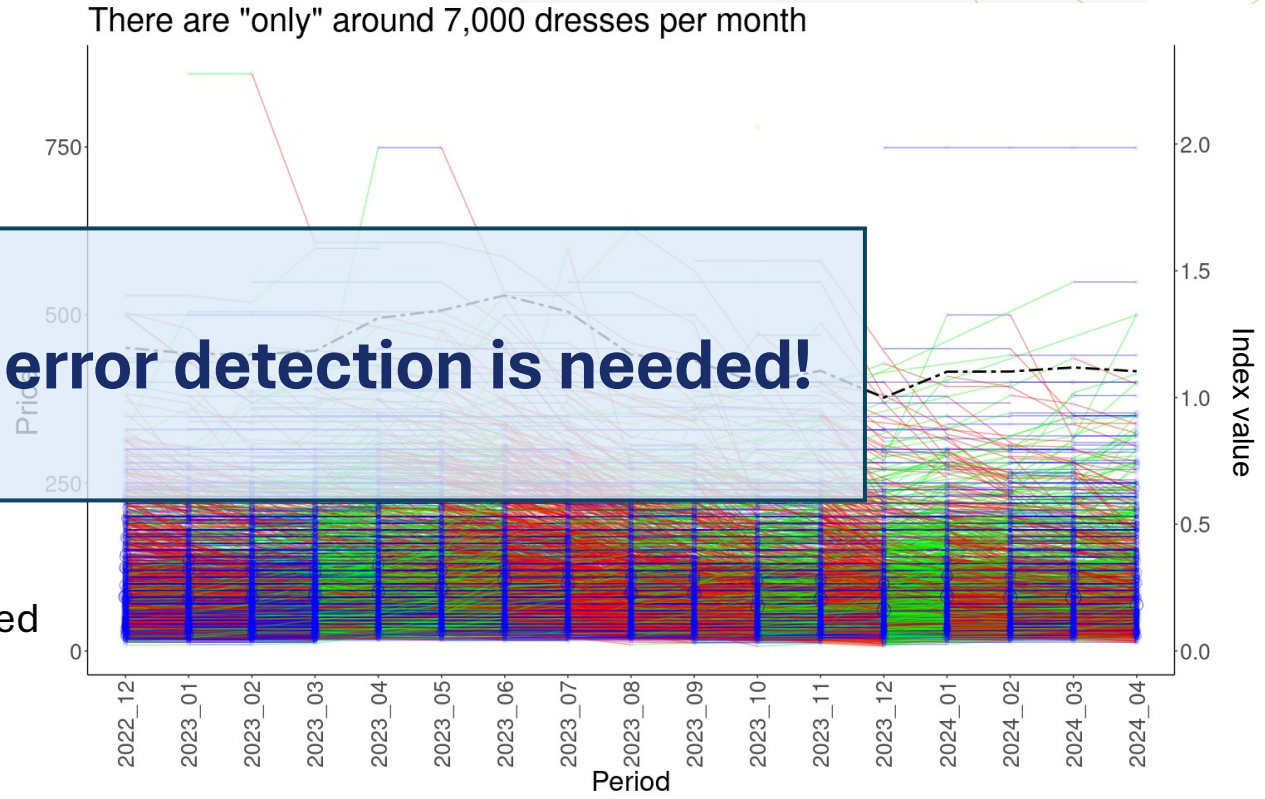
CPI/HICP are based on product information/prices collected

- manually in local stores (~ 20,000/month)
- by barcode scanners at the checkout (~ 670,000/month)
- online via webscraping (~ 140,000/month)

Mistakes in data collection are rare but not impossible

Assuming all outliers are indeed invalid values per se is not justified

A standardised framework for error detection is needed!



Error c

Well establishe

Medcouple adju

Tmatrix-Check - Ergebnisse 2024_04 (Zeitreihen Basis: 202112, Aggvergleich Basis: 202312)

Nicht kontrollierbare | check_frame_vvm | check_frame_vvj | VVM_Aggvergleich | VVJ_Aggvergleich

10 Einträge anzeigen

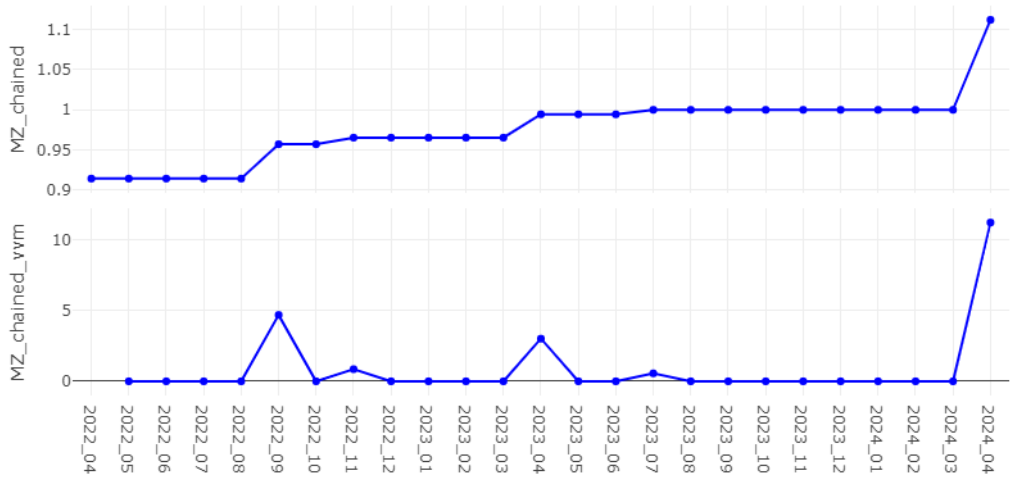
TEXTKURZ	AGGEBENE	AGG	CODE	CODEALT	MIN	Quant1
All	["7"			Al		
Teilkaskoversicherung 21.000-30.000	7	125410	105300	0920	-0.2	0
Privatzimmer im Inland	7	112010	097100	0839	-1.8	-0.1

TEXTKURZ	CODE	UCODE	VERARB_2024_03	VERARB_2024_04	VERARB_vvm_2024_04	MZ_chained_2024_03	MZ_chai
All			All	All	All	All	All
Teilkaskoversicherung 21.000-30.000	0920	0			FALSE	1	
Teilkaskoversicherung 21.000-30.000	0920	0			FALSE	1	
Teilkaskoversicherung 21.000-30.000	0920	0			FALSE	1	
Teilkaskoversicherung 21.000-30.000	0920	0			FALSE	1	

Zusammenfassung

parameter	min	MW	max	anz_changes	N
RPREIS	626.92	651.3136	722.35	4	25
MZ_chained	0.9142	0.976288	1.1123	5	25
VERARB				2	25

MZ | RPREIS



10 Einträge anzeigen | Suchen

period	VERARB	VERARB_verändert	RPREIS	RPREIS_vvm	MZ_chained	MZ_chained_vvm
2023_12		false	649.43	0	1	0
2024_01		false	649.43	0	1	0
2024_02		false	649.43	0	1	0
2024_03		false	649.43	0	1	0
2024_04		false	722.35	11.228	1.1123	11.23

Nächste Nachbarn für 0199 - Tagesmenü / Mittagsmenü; Jahr = 2024, Monat = 01, Basis = 2015

Target-Daten

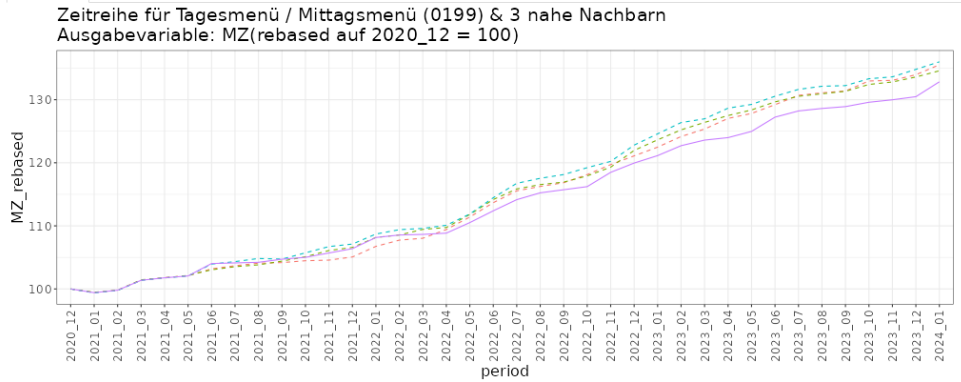
CODEALT	TEXTKURZ	MZGER	VVM	VVJ	EINFLVM	EINFLVJ
0199	Tagesmenü / Mittagsmenü	157.8	1.8	9.7	0.006	0.03

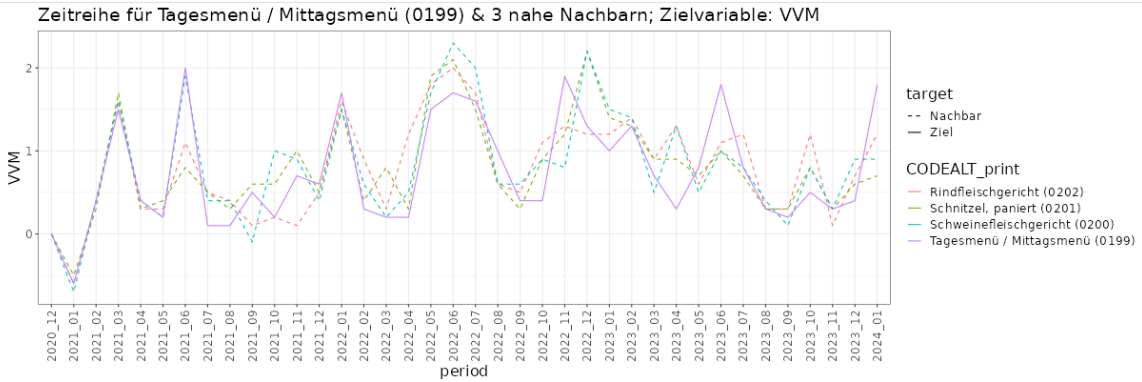
Gefundene Nachbarn

CODEALT	TEXTKURZ	MZGER	VVM	VVJ	EINFLVM	EINFLVJ	near_nbg	dist
0200	Schweinefleischgericht	162.6	0.9	9.1	0.002	0.019	true	2.41867733001709
0201	Schnitzel, paniert	161.8	0.7	8.9	0.003	0.042	true	2.574878454208374
0202	Rindfleischgericht	162.1	1.2	10.6	0.005	0.047	true	2.632489442825317
0204	Mehlspeise, warm	164	0.2	9.6	0	0.018	false	2.736786603927612

Charts

Near Nbg | All Nbg





093500	0199	Tagesmenü / Mittagsmenü	3	1.8	0.2	1.2	1	0.6	0.7	1.2	1	1.2
073700	0513	Innerstädtischer Verkehr, Jahreskarte	4	0.1	0	0	1	0	0	0	1	

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Erro
mod

Arima Ausreisser-Check

Jahr: 2024

Monat: 04

Arima Ergebnis-Übersicht für 2024_04

20 Einträge anzeigen

Suchen

CODE	CODEALT	TEXTKURZ	MZ_orig	VVM	VVJ	arima_pred	lo_95	lo_80	hi_80	hi_95	mean_residuals	flag_arima	flag_arima_smoothing	over_max_param_95

12 Einträge anzeigen

Suchen

JAHR	MONAT	MZ_orig	MZ	VVM	VVJ	arima_pred

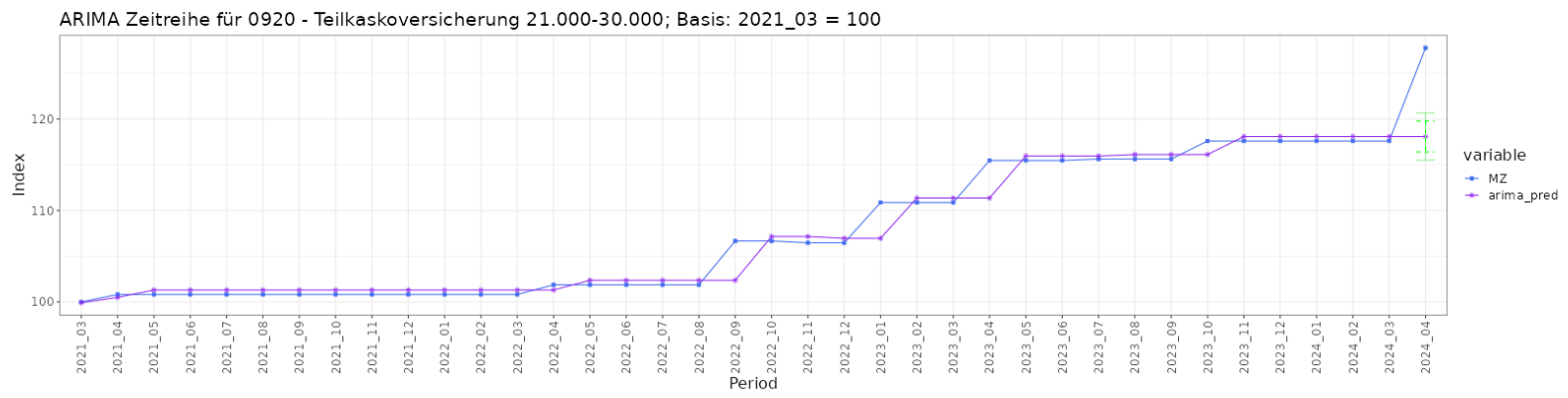
12 Einträge anzeigen

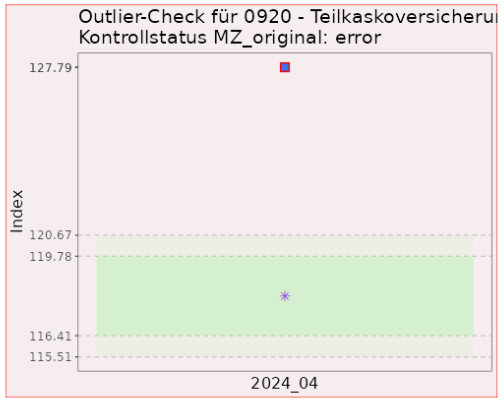
Suchen

JAHR	MONAT	MZ_orig	MZ	VVM	VVJ	arima_pred
["2024"]	All	All	All	All	All	All
2024	01	117.6	117.6	0	6.1	118.09
2024	02	117.6	117.6	0	6.1	118.09
2024	03	117.6	117.6	0	6.1	118.09
2024	04	127.79	127.79	8.7	10.7	118.09

1 bis 4 von 4 Einträgen (gefiltert von 38 Einträgen)

Zurück 1 Nächste





INSTITUTO NACIONAL DE ESTATÍSTICA | STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Summary and outlook

False positive over false negative

Increasing evaluation speed is easier than increasing precision

Evaluation on monthly change rate is the most useful approach in production

Nearest neighbours search well suited for regular but not strict periodical cyclic patterns (travel expenditures around Easter)

Time series approach well suited for seasonal products (clothing) and strict periodical cycles (taxes)

Omit periods where COVID had a large impact on the market

Constant work in progress

Feedback from users is highly appreciated

Bug fixing

Parameter selection for Nearest Neighbours and ARIMA

New features (mail notification system, ...)

The conference is partly financed by the European Union

References & R-Packages

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., & Li, S. (2023). *FNN: Fast nearest neighbor search algorithms and applications. R package version 1.1.3.2.* https://CRAN.R-project.org/package=FNN.

Brys, G., Hubert, M., & Struyf, A. (2004). A Robust Measure of Skewness. *Journal of Computational and Graphical Statistics, 13*(4), 996–1017. https://doi.org/10.1198/106186004X12632

Chang W, Borges Ribeiro B (2021). *shinydashboard: Create Dashboards with 'Shiny'.* R package version 0.7.2. https://CRAN.R-project.org/package=shinydashboard

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2024). *shiny: Web Application Framework for R*. R package version 1.8.1.1. https://github.com/rstudio/shiny, https://shiny.posit.co/

Dang, T. T., Ngan, H. Y., & Liu, W. (2015, July). Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In *2015 IEEE International Conference on Digital Signal Processing (DSP)* (pp. 507-510). IEEE. https://doi.org/10.1109/ICDSP.2015.7251924

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186-5201. https://doi.org/10.1016/j.csda.2007.11.008

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

Isbister, T. (2015). Anomaly detection on social media using ARIMA models.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Statistics Austria (2022). *Standard-Dokumentation zum Verbraucherpreisindex und Harmonisierten Verbraucherpreisindex.* https://www.statistik.at/fileadmin/shared/QM/Standarddokumentationen/VW/std_v_vpi_hvpi.pdf

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019, July). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2828-2837). https://doi.org/10.1145/3292500.3330672

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Too cheap to be true

Detecting invalid values in product prices and index values

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union