The use of scanner data to compile regional price indices in Germany with an outlook to income inequality

Alina Happ[1] and Chris-Gabriel Islam[2,*,3]

[1] affiliation: University of Bamberg; residence: Bamberg (Germany); e-mail: alina.happ@stud.uni-bamberg.de

[2] affiliation: Federal Statistical Office of Germany and University of Göttingen; residence: Mainz (Germany); e-mail: chris-gabriel.islam@destatis.de; *corresponding author

Abstract:
Statistics about regional price levels in Germany are rare due to the lack of suitable data in official statistics so far. Regional price indices, however, are of major importance in order to determine intra- and international wealth differences or to compare for regional economic power. Using newly acquired scanner data of the German Federal Statistical Office, this work examines how prices for food and beverages differ regionally in the year 2017 on NUTS 3 level by calculating a multilateral GEKS-Fisher-Index. Subsequently, we perform an exploratory multiple linear regression in order to find the most important drivers for high regional prices. We find that price levels rise with local average income and wage, with population growth and with the area of the respective NUTS 3 region. In contrast, price levels decrease with the amount of competition and are, ceteris paribus, lower in Western Germany. Lastly, we merge our data with income data of the Sample Survey of Income and Expenditure (EVS) and find that regional income inequality measured by the Gini coefficient decreases by 0.0009 when considering regional price disparities.

Keywords:
Transaction data; multilateral index number methods; price statistics; regional wealth differences

1. Introduction:

In 2019 the Federal Government of Germany started a commission for equal living conditions. Their recommendations include the investigation whether the statistical institutes are able to enhance their offer of regional price indices (FMI 2019). Until now, German price statistics comprises only intertemporal price comparisons on the NUTS 1 level on a regular basis. There are some studies that focus on the interregional price comparison, also on a deeper NUTS level, that all find a prevalence of regional disparities (Ströhl 1994; BMEATT 2003; Kawka 2009; Lippe and Breuer 2009). Nonetheless, until now no continuous statistics has been published. The main reason is missing data as price data is mostly collected manually so far and collecting additional data would be too costly (Auer 2012; Behrmann et al. 2009; Kawka 2009; Weinand 2020). Regional price indices, however, are of major importance in order to determine intra- and international wealth differences or to compare for regional economic power.

[3] The authors thank Timm Behrmann, Matthias Bieg, Daniel Gorba, Stephan Länge and Dr. Jürgen Schiener. The views expressed in this article are solely those of the authors and do not necessarily reflect those of the mentioned institutions.

This work builds a framework that allows to compile regional price indices for Germany on the NUTS 3 level and to analyse the most important drivers for regional disparities. As the Federal Statistical Office of Germany acquired scanner data in 2018, this task becomes possible.

2. Data:

As a main resource, we use weekly scanner data of one German food retail chain.[4] This data includes turnover, sales and Metadata for food, beverages and tobacco articles. Nonfood articles are being excluded from our calculations. The time range comprises all full weeks in 2017.[5] After cleaning the data and performing plausibility checks[6] we end up with data for 279 stores with a total of 81,923 observations on the article level. Each article has a unique Global Trade Item Number (GTIN) which is depicted by the barcode. For the interregional comparisons turnovers and sales are aggregated for all stores within a NUTS 3 area and for all full weeks in 2017, yielding 192 different unit value prices per article. As unit for comparison, we chose the NUTS 3 level as these areas are interrelated and a homogeneous price development can be expected within (FOBRP 1991; Linz and Dexheimer 2005). Note that the 279 stores are not distributed equally in Germany. The maximum of stores per NUTS 3 area is 6. In total, we end with 68,299 articles, as an article must be sold at least in two different NUTS 3 areas in order to have a price comparison.
Additional data on the NUTS 3 level for the following regression and income analysis is gathered by various sources. The specific source will be indicated in detail when the data is used.

3. Methodology:

In order to compare different regions, we need to calculate price indices. All price indices developed so far tend to either follow more the principle of a pure price comparison, that does only account for changes in prices and not in quantities, or the principle of representativity, that accounts for sudden changes in the basket of goods. In sum, there is no one best index for all situations. As scanner data contain very exact information about the quantities sold in every point in time, one should consider using this data and trading some purity of price comparison for more representativity. While points in time follow a cardinal scale and therefore bilateral price indices can be applied for an intertemporal comparison, this does not hold for interregional comparisons as there is no natural scale between regions. So that we are able to calculate regional comparisons, a multilateral index is necessary (Auer 2012; Balk 1996; Balk 2008; Eurostat 2012; Lippe 2007; Neubauer 1996). The latter also allows for a better comparison between regions with a disjoint basket of goods.[7] This is due to the transitive nature of multilateral indices. In the end, we chose the GEKS-F index for our analysis. This is an index developed in parts by Gini, Éltetö, Köves and Szulc that uses a Fisher index for the respective bilateral comparisons.[8] We chose this index as it is one of the most researched indices and because Eurostat and the OECD have been using it for the calculation of purchasing power parities since 1990 (Eurostat 2012). The GEKS-F for a comparison between region 1 and *r* while having *R* regions in total is calculated as follows:

[4] This retailer wishes to remain anonymous, hence we do not mention any name here.
[5] I.e. the two weeks that are partly in 2016 and 2017 or 2017 and 2018 are excluded.
[6] E.g. removing articles with negative turnover due to returning items, removing articles with prices less than 0.1 € or greater than 1,000 €, removing a branch as it was closed in the end of 2017.
[7] Consider e.g. region A and B with goods that are only sold in one of the regions but also in another region C. While bilateral indices allow only to compare A and B using the intersection of their goods sold, multilateral indices allow for indirect comparisons using C.
[8] One could also use a Törnqvist instead, which is known as GEKS-T.

$$GEKS_1^r = \prod_{z=1}^{R} \left(\frac{F_1^z}{F_r^z}\right)^{\frac{1}{R}}$$

with *F* as the bilateral Fisher index between the respective regions.

4. Results:

4.1 Regional price disparities in Germany

The left-hand side of Figure 1 shows price indices per NUTS 3 area computed with GEKS-F using scanner data from 2017. In some areas no index could be calculated as there is no store located from our data providing retailer. The most expensive area is Düsseldorf city with a value of 101.6, meaning that food, tobacco and beverages are 1.6 % more expensive than the German average. Bamberg city is the cheapest area with a value of 98.4. The standard deviation is rather low with 0.587. More variation and clear patterns can be seen when considering different subcategories of food, beverages and tobacco. The most extreme values of the GEKS-F are achieved when considering only articles that are served by an assistant, e.g. meat or cheese, which is depicted on the right-hand side of Figure *1*. Bamberg city holds the minimum with 93.0, while Krefeld city holds the maximum with 104.1. The standard deviation is 1.25. We deduce that wages play an important role for the price setting. Subsequently (not depicted), we find that fruits, vegetables and meat products are most expensive in the south of Germany while being cheaper in Eastern Germany. The latter holds for milk products and bread as well. In contrast, edible fats are more expensive in Eastern Germany while being cheaper in the north-west. We find that our results are robust as we calculate similar values by using data from 2018 and by using another multilateral method instead of GEKS-F, namely Geary-Khamis[9]. Some of the previous studies in the literature only include articles that are sold in every region, i.e. that use only a joint basket of goods. If we restrict our sample to these 12,602 articles, we find fewer disparities. This might give reasons that previous studies could have underestimated disparities.

[9] See Chessa et al. (2017) for more details.

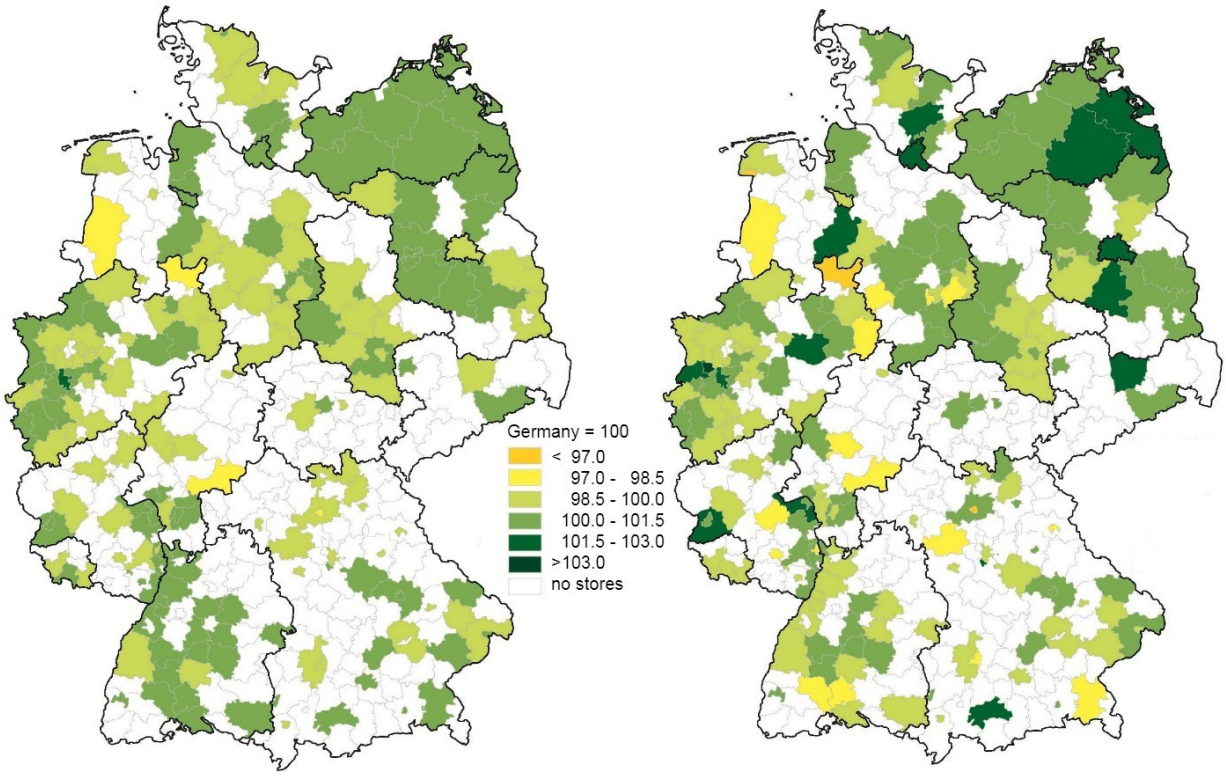

Figure 1: Price indices calculated with GEKS-F on NUTS 3 level in Germany in 2017 for food, beverages and tobacco (left-hand side) and for articles served by an assistant (right-hand side).

4.2 Drivers for regional price disparities

In order to analyse the most important drivers for regional price disparities, we collect possible variables based on the literature of pricing. As a first step, we remove variables with only a low correlation to the dependent variable ($r < 0.05$). Then we perform a principal component analysis using the Varimax rotation. We pick only the interpretable factors, remove variables with high correlations to other independent variables and perform a backward regression.[10] The results of this exploratory analysis are depicted in Table 1. All drivers show a positive sign with the exception of the competition. A positive coefficient for Eastern Germany seems surprising and contrary to the literature about regional differences. However, it can be explained as the regression already controls for other important differences between Western and Eastern Germany, namely yearly income, wages and population growth. The low Adjusted R^2 in comparison to other models in the literature (e.g. Kawka 2009) emphasises the exploratory character of this regression. The high value of the constant indicates the problem of having only data from one retailer.

Table 1: Multiple linear regression on GEKS-F

Independent variables	Coefficients	*t*-values
Constant	99.860***	2314.55
Population growth	0.147***	3.56
Yearly nominal household income per capita	0.239***	5.35

[10] Overall, we excluded the following variables from the regression: Tourism (measured via the number of stays in accommodations), the number of employed persons and of inhabitants in general, the density of competition per capita and per square kilometre, the population density, a dummy for Northern Germany and a dummy for rural areas.

Wages	0.167**	2.87
Competition	−0.135*	2.50
Area	0.175**	2.82
Eastern Germany	0.765***	5.61
Controls		
Average store space of retailer	0.056	1.50
Number of stores of retailer	0.116**	2.86
Observations	192	
Adjusted R^2	0.306	
F-test	11.52	

Notes: Results from a multiple linear regression are shown. The dependent variable is the GEKS-F index for each of the 192 NUTS 3 areas considering the articles sold by the data providing retailer. All variables except the dummy for Eastern Germany and the constant are presented as *z*-scores. Population growth is measured as the difference of inhabitants between 2017 and 2007. This data as well as the information of the area is collected from Eurostat. Yearly household income per capita is collected from the EVS. Wages is a factor built using average wages for women, men and people with completed vocational training in the whole NUTS 3 area. This data is collected from the Federal Employment Agency. Competition is measured by the number of stores in the NUTS 3 area that do not belong to our data-providing retailer. This data is collected from Nielsen.
Coefficients marked with *, ** and *** are significant at the 0.05, 0.01 and 0.001 probability level (two-sided).

4.3 Impact on income inequality

With the aim of achieving equal living conditions on a regional level, usually the nominal income is considered (FMI 2019). Data about real income is rare, but it considers regional prices and is more suitable for calculating purchasing power parities (Auer 2012). In a first step, we use data from the Sample Survey of Income and Expenditure (EVS) provided by the Federal and State Statistical Offices to get the average nominal household income per capita per NUTS 3 level[11] in order to calculate a quasi-real income per household:

$$\text{quasi-real income} = \text{nominal income} - (\text{nominal income} \cdot 0.0966 \cdot \text{GEKS-F}/100),$$

with 0.0966 as the estimated portion of income that an average German citizen spends for food and beverages (FSO 2018). In a second step, we calculate the Gini coefficient as well as the coefficient of variation for the 192 NUTS 3 areas in our data considering the nominal income and the quasi-real income. We find that the Gini coefficient lowers from 0.05912 to 0.05903 when using the quasi-real income and the variation coefficient lowers from 0.10443 to 0.10426. Thus, we find a very small tendency that income differs less between regions when controlling for price levels of food and beverages.

5. Discussion and Conclusion:

We showed that scanner data can be used to compute regional price indices. Even within one retailer pricing differs between regions. Using GEKS-F we depicted different regional price patterns depending on the group of articles or the level of service considered.
By executing an exploratory multiple linear regression, we found the most important drivers for high regional prices, i.e. price levels rise with population growth, local average income and wage and the area of the respective NUTS 3 region. On the contrary, price levels decrease with the amount of competition. This is in line with the literature about pricing. Additionally, prices are, ceteris paribus, lower in Western Germany. For policy makers that are interested in equalising the differences between Western and Eastern Germany a takeaway could be to focus less on daily consumer products but more on other parts of the consumer baskets, e.g. costs of housing.

[11] Note that the data is not equivalence-weighted.

Note that the literature about regional prices, to the best of our knowledge, mostly depends on small manually gathered samples while this work is the first that only uses automatically collected scanner data of one whole retail chain while focusing on foods, beverages and tobacco. It therefore resembles a tremendously larger sample size and allows for a more robust regression analysis. An additional advantage is the usage of up-to-date quantity information, which is not possible with a static sample. However, a major caveat is the presence of only one retailer. Results may alter if more retailers are included, although this is not to be expected as retailers usually align their prices to other retailers nearby. Moreover, the retailer used in this study is not present in every region in Germany with only a few stores in Eastern Germany. This might not change current results, but due to this fact a regression at the aggregated Eastern-Western level instead of the NUTS 3 level could suffer from small sample bias and was therefore not calculated.
All in all, this research is a framework for further studies using more retailers or other groups of articles. It may even serve as a framework for a continuous statistics production for interregional comparisons. An application of this work in social sciences could be the analysis of regional income or wealth inequality while controlling for different regional price levels. Lastly, knowing drivers of regional price differences can lead to better policy making.

References:

Auer, Ludwig von (2012). „Räumliche Preisvergleiche: Aggregationskonzepte und Forschungsperspektiven“. In: *AStA Wirtschafts- und Sozialstatistisches Archiv*6.1-2, pp. 27–56.

Balk, Bert M. (1996). “A Comparison of Ten Methods for Multilateral International Price and Volume Comparison”. In: *Journal of Official Statistics* 12.2, pp. 199–222.

— (2008). *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*. Cambridge University Press.

Bavarian ministry of economic affairs, transport and technology (2003). *Die reale Kaufkraft in Bayern 2002. Zwischenörtliche Preis- und Einkommensunterschiede*. München.

Behrmann, Timm, Simon Deml, and Stefan Linz (2009). “Verwendung von Einzeldaten aus der Verbraucherpreisstatistik für regionale Preisvergleiche”. In: *RatSWD Research Note* 36.

Chessa, Antonio G., Johan Verburg, and Leon Willenborg (2017). “A Comparison of Price Index Methods for Scanner Data”. In: *Ottawa Group, 15th meeting*. Germany.

Eurostat (2012). *Eurostat-OECD Methodological Manual on Purchasing Power Parities*. Luxemburg: Publications Office of the European Union.

Federal Ministry of the Interior, Building and Community, Federal Ministry for Food and Agriculture and Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (2019). *Unser Plan für Deutschland. Gleichwertige Lebensverhältnisse überall*.

Federal Office for Building and Regional Planning (1991). *Raumentwicklung in den alten und neuen Bundesländern*. Bonn.

Federal Statistical Office of Germany (2018). “Wirtschaftsrechnungen. Einkommen, Einnahmen und Ausgaben privater Haushalte 2017”. In: *Laufende Wirtschaftsrechnungen* 15 (1).

Kawka, Rupert (2009). *Regionaler Preisindex*. Berichte 30. Bonn: Self-publisher of the Federal Institute for Research on Building, Urban Affairs, Spatial Development in the Federal Office for Building, and Regional Planning.

Linz, Stefan and Verena Dexheimer (2005). “Weiterentwicklung der Stichprobe der Verbraucherpreisstatistik”. In: *Wirtschaft und Statistik* 6, pp. 582–586.

Lippe, Peter von der (2007). “Bemerkungen zur Studie "Die reale Kaufkraft in Bayern 2002" zur Vorlage beim Bundesverfassungsgericht. Statistisch-methodische Analyse einer Studie vom Bayerischen Staatsministerium für Wirtschaft, Verkehr und

Technologie". In: *University of Duisburg-Essen, Department of Economics, Discussion Paper Series* 163.

Lippe, Peter von der and Claus Christian Breuer (2009). "Regionale Kaufkraftvergleiche in Deutschland: Bedarf, Methoden und Machbarkeit". In: *AStA Wirtschafts- und Sozialstatistisches Archiv* 3.1, pp. 25–40.

Neubauer, Werner (1996). *Preisstatistik*. Vahlens Handbücher der Wirtschafts- und Sozialwissenschaften. München: Vahlen.

Ströhl, Gerd (1994). "Zwischenörtlicher Vergleich des Verbraucherpreisniveaus in 50 Städten". In: *Wirtschaft und Statistik* 6, pp. 415–434.

Weinand, Sebastian (2020). "Measuring spatial price differentials: A comparison of stochastic index number methods". In: *Deutsche Bundesbank Discussion Paper* 12.