

A Service of



Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

von Auer, Ludwig; Weinand, Sebastian

Article — Published Version The Country-Product-Dummy Method With Product-Specific Spatial Price Variation

Review of Income and Wealth

Suggested Citation: von Auer, Ludwig; Weinand, Sebastian (2025) : The Country-Product-Dummy Method With Product-Specific Spatial Price Variation, Review of Income and Wealth, ISSN 1475-4991, Vol. 71, Iss. 1, https://doi.org/10.1111/roiw.70005

This Version is available at: https://hdl.handle.net/10419/319272

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.



http://creativecommons.org/licenses/by/4.0/

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



WWW.ECONSTOR.EU

The Country-Product-Dummy Method With Product-Specific Spatial Price Variation

Ludwig von Auer¹ | Sebastian Weinand² D

¹Universität Trier, Trier, Germany | ²Deutsche Bundesbank, Frankfurt am Main, Germany

Correspondence: Sebastian Weinand (sebastian.weinand@bundesbank.de)

Accepted: 3 February 2025

Keywords: CPD method | Gini-Eltetö-Köves-Szulc index | measurement bias | multilateral price index | regional price levels

ABSTRACT

The present paper shows that product-specific spatial price variation usually causes the Country-Product-Dummy (CPD) method to be biased. In cases where it is not, the CPD method is still inefficient and statistical inference is invalid. The paper develops a non-linear generalization of the CPD method. Even for product-specific spatial price variation, this NLCPD method remains unbiased and allows for inference. A comprehensive simulation reveals that the NLCPD method's root mean squared error is smaller than that of the Gini-Eltetö-Köves-Szulc (GEKS) index and the CPD method. Finally, this paper applies the NLCPD method to spatial price information derived from Germany's consumer price index micro data of 2019. Price levels of the 401 German districts are computed.

JEL Classification: C43, E31

1 | Introduction

Important areas of economic theory and economic policy utilize subnational indicators of regional economic activity. Well-known examples of these are regional real wages and output levels. However, the high demand for such indicators is not matched by the available supply. The reasons for this gap in economic statistics are not hard to find. The production of regional real indicators requires reliable information on regional price levels, while statistical offices' primary task is tracking intertemporal price level changes. The latter requires a very broad sample of different products. Thus, for pasta products, say, in different regions, prices of different types of pasta are recorded. By contrast, spatial price comparisons would benefit from a more selective sample in which the same type of pasta is recorded in all regions. However, it is laborious and costly to establish and maintain a sample that serves the needs of both intertemporal and spatial price comparisons. Therefore, only very few countries publish regional price levels (Weinand and Auer 2020, 416-418).

Matters are made worse by the methodological challenges of spatial price comparisons. While intertemporal price comparisons usually apply bilateral index theory, spatial price comparisons require a multilateral approach. A wide spectrum of multilateral methods are available and have been applied in subnational case studies of countries from all over the world (surveyed by Majumder and Ray 2020, 111-113 and Weinand and Auer 2020, 416-419). The choice between the various methods also depends on the available data set. Some subnational studies cover only parts of a country. Others cover the complete country, but the regions within this country are very large. Another distinguishing feature is the number and range of products for which prices are available. For example, housing costs are not always included. Usually, the data have been collected for other purposes. Micro price data are rarely available.

Unfortunately, large data gaps are the rule rather than the exception. This is true for subnational as well as international price level comparisons. For the latter case, Summers (1973)

Sebastian Weinand: Currently seconded to the European Commission (Eurostat).

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). Review of Income and Wealth published by John Wiley & Sons Ltd on behalf of International Association for Research in Income and Wealth.

introduced the Country-Product-Dummy (CPD) method. This method is equally applicable to subnational price level comparisons. The same is true for the analysis of the present study. Both, its theoretical findings and its simulation results apply to subnational as well as to international price level comparisons. It is primarily for illustrative purposes that this paper is framed as an analysis of subnational price level comparisons, where "regions" stands for districts or other spatial entities within a country. Accordingly, this study concludes with a subnational empirical application.

The CPD method is a linear regression approach that allows for statistical inference. This advantage rests on the implicit assumption that the products' spatial price variation is uniform. However, economic models (e.g., Tabuchi 2001, p. 105) as well as empirical studies (e.g., Weinand and Auer 2020, p. 430; Rokicki and Hewings 2019, p. 94; Aten 2017, 132-134) show that spatial price variation usually differs across products. The higher the level of aggregation and the more heterogeneous the included products (e.g., pasta vs. shoes), the less plausible the CPD method's assumption of a uniform spatial price variation.

Accordingly, in applied work the CPD method is primarily used for the computation of the regional price levels of products with a common consumption purpose (e.g., pasta products). At this initial stage of the aggregation process, often denoted as "below basic heading level", no expenditure information is available and the unweighted variant of the CPD method must be used. However, it cannot be ruled out that the included products' spatial price variation is heterogeneous (e.g., the regional pricing policy of large pasta brands deviates from the pricing policy of smaller regional competitors) which would constitute a violation of the CPD method's underlying assumption. A violation is even more likely when the price levels of the various basic headings are aggregated into the overall regional price levels (aggregation "above basic heading level"). Therefore, this second stage of aggregation usually applies an alternative method. As a consequence, the final result involves a mix of different methods.

The above considerations raise several fundamental questions. What are the statistical consequences if the CPD method is applied even though the spatial price variation is not uniform, but product-specific? Do the estimated regional price levels remain unbiased? Is inference still valid? If not, is there a practical way to check whether a set of products exhibits a uniform spatial price variation? Are there alternative estimation methods that remain unbiased even when the products' spatial price variation is not uniform, but product-specific?

The present paper answers all of these questions. With product-specific spatial price variation, the CPD method's statistical inference is invalid. The CPD estimates of the regional price levels remain unbiased as long as the set of price data is complete (a situation in which the CPD method is rarely used) or data gaps occur completely at random (a situation that is difficult to achieve in real-world price data samples). In other words, statistical offices normally process data sets with systematic data gaps and for such data sets the CPD estimates of the regional price levels are biased. As a solution to these problems, this paper introduces the *NLCPD method*, a non-linear generalization of the CPD method. Below basic heading level, the unweighted NLCPD method can be applied, while above basic heading level the weighted NLCPD method is preferable.

In multilateral comparisons, any direct comparison between two regions should give the same price levels as an indirect comparison of these two regions via a third one. In index number theory, this requirement is called transitivity (e.g., Rao and Banerjee 1986, p. 304). Both the CPD and the NLCPD method produce transitive regional price levels. In addition, these multilateral methods compute the general values of the individual products. However, only the NLCPD method also provides estimates of the spatial price variation of the individual products. These estimates indicate whether the assumption of a uniform spatial price variation would be justified. Even more important, the paper shows that the regional price levels estimated by the NLCPD method remain unbiased even when the price data exhibit product-specific spatial price variation and systematic data gaps exist. In addition, the variance of the estimators can be estimated, providing a basis for valid statistical inference. Even if the data set were complete or the data gaps were completely at random, the NLCPD method would still outperform the CPD method. Thus, the CPD method should be avoided unless all products included have exactly the same spatial price variation.

The rest of the paper is organized as follows. Section 2 provides an intuitive explanation for the source of the CPD method's bias. How the NLCPD method addresses this problem is explained in Section 3. A more formal treatment of the NLCPD method is presented in Section 4. Section 5 provides a comprehensive simulation that confirms and complements the theoretical predictions and makes a strong case for the use of the NLCPD method. Section 6 applies this method to a large data set of regional prices. Section 7 concludes.

2 | Problem

To illustrate the problems of the CPD method, we use a highly stylized example of a subnational price comparison. In empirical studies of subnational price comparisons, the prices of manufactured goods are found to be rather uniform across space, while the cost of housing varies considerably (e.g., Weinand and Auer 2020, 430-431, for Germany; Aten 2017, 130-131, for the United States). The prices of services take an intermediate position. Table 1 shows the same features. It lists the prices of three products, $\mathcal{N} = \{i : i = \text{goods}, \text{housing}, \text{services}\}$, in four different regions, $\mathcal{R} = \{r : r = A, B, C, D\}$. For simplicity, it is assumed that no expenditure information is available.

TABLE 1 | Prices of goods, housing, and services in four regions (grayprices may be missing).

	Α	В	С	D
1: Goods	2.9	3.0	3.0	2.9
2: Housing	3.5	5.6	6.7	10.1
3: Services	7.0	8.3	11.7	14.8

In Section 2.1, we describe the CPD method. In Section 2.2, we graphically illustrate the CPD regression with our highly stylized example of price data. Why the CPD regression may lead to biased estimates of the regional price levels is explained in Section 2.3. Further issues of the CPD method are sketched out in Section 2.4.

2.1 | CPD Method

The CPD method has been originally developed by Summers (1973) for international price comparisons. It can also be used in subnational comparisons to derive the general price levels of the regions within a country. The CPD regression assumes that the price of product *i* in region *r*, p_i^r , can be explained by the linear relationship

$$\ln p_i^r = \ln \pi_i + \ln P^r + e_i^r \tag{1}$$

where P^r is the price level of region r, π_i is the general value of product i, and $e_i^r \sim N(0, \sigma^2)$ is a normally distributed disturbance term with zero mean and variance σ^2 (see Summers 1973, p. 10).

The term $(P^r/P^s - 1)$ indicates the percentage change in expenditures when the basket of products defined by \mathcal{N} is purchased in region *r* instead of region *s*. Equation (1) implies that the following relationship exists between the price level ratio P^r/P^s and the price ratios of the products, p_i^r/p_s^r (i = 1, ..., N):

$$\ln \frac{P^r}{P^s} = \mathbb{E}\left(\frac{1}{N}\sum_{i\in\mathcal{N}}\ln \frac{p^r_i}{p^s_i}\right) \tag{2}$$

where \mathbb{E} is the expectation operator.¹ The term in brackets corresponds to the logarithm of the Jevons (1865) index formula. For the plausibility of the CPD model (1), this is an appealing property.

To obtain the ordinary least squares estimates of the unknowns $\ln P^r$ and $\ln \pi_i$, the CPD model (1) must be transformed into a regression equation with a set of dummy variables that represent the regions and the products.² In Section 2.2, this regression is applied to the prices listed in Table 1.

The CPD regression approach, however, has a significantly understated drawback. It implicitly assumes that the products included in the comparison have the same (spatial) price variation. The prices in Table 1 violate this assumption. The cost of housing and the prices of services considerably vary across regions, while the prices of goods are all but constant. In Section 2.3, we demonstrate that, with product-specific price variation, the CPD regression produces biased estimates of the regional price levels (as formally shown in Online Appendix A.3), barring two cases that are rarely satisfied in real-world measurement problems. Even if those two exceptional cases applied, the CPD regression would still be inefficient and inference would become invalid (see Section 2.4 and Online Appendix A.4.2).

2.2 | Graphical Illustration

For the complete data set of Table 1, the CPD regression yields the three products' estimated logarithmic general values, $\ln \pi_i$, as well as the estimated logarithmic price levels, $\ln P^r$, of the four regions. Taking anti-logs gives the following regional price levels:

$$\hat{P}^{A} = 0.74, \ \hat{P}^{B} = 0.92, \ \hat{P}^{C} = 1.10, \ \hat{P}^{D} = 1.34$$
 (3)

where the price levels have been normalized such that $\hat{P}^{A} \cdot \hat{P}^{B} \cdot \hat{P}^{C} \cdot \hat{P}^{D} = 1.^{3}$

A graphical illustration of this CPD regression is provided in the upper left panel of Figure 1 (the other panels should be ignored for the moment). The panel shows on the vertical axis the observed values of the dependent variable, $\ln p_i^r$, and on



FIGURE 1 | CPD and NLCPD regressions for the price data of Table 1, respectively, either with complete price data (top panels) or with missing prices for "goods" (bottom panels).

the horizontal axis the unknown regional logarithmic price levels, $\ln P^r$. For each region *r*, three price observations exist. In the diagram, these three observations are depicted by a circle (goods), a square (services), and a triangle (housing). The three observations are positioned along a dashed vertical line. The position of that line is the outcome of the CPD regression. More specifically, the intersection of each line with the horizontal axis is the estimated value $\ln P^r$. Thus, the four intersection points are the logarithms of the price levels listed in (3).

Another element in the upper left panel are the colored straight lines. Each line relates to a different product *i* and, by definition, has slope one. The intersections of these lines with the vertical axis define the estimated values $\widehat{\ln \pi_i}$. Our normalization of the regional price levels in (3) implies that the logarithmic prices of product *i* observed in regions A to D, ln p_i^r , fluctuate around $\widehat{\ln \pi_i}$.

Changing the estimated value $\ln \pi_i$ causes a parallel vertical shift of the colored solid line relating to product *i*. Changing the estimated value $\ln P^r$ causes a horizontal shift of the dashed vertical line of region *r* and, therefore, of the three observations relating to that region. Both types of shifts would alter the vertical distance between the observations and their respective solid line. This vertical distance is the residual, \hat{e}_i^r . Graphically speaking, the CPD regression simultaneously shifts the solid lines and the dashed vertical lines (together with their three observations) such that the sum of the squared residuals is minimized. The upper left panel of Figure 1 depicts the solution to this minimization problem, resulting in the price levels listed in (3).

2.3 | Biased Price Levels

Next, we turn to the consequences of incomplete data. Let \mathbf{x}_i^r represent the regressor vector of product *i* in region *r*, that is, the values of the two sets of dummy variables. Irrespective of the set of missing price observations, the CPD regression assumes that the conditional expected value of the disturbance term is zero: $\mathbb{E}(e_i^r | \mathbf{x}_i^r) = 0$. However, the left panels of Figure 1 illustrate that missing prices usually lead to $\mathbb{E}(e_i^r | \mathbf{x}_i^r) \neq 0$ (the panels on the right-hand side of Figure 1 should be still ignored).

The upper left panel's two outer vertical dashed lines indicate the estimated logarithmic price levels of regions A and D, respectively. Clearly, region A is the cheapest region, while region D is the most expensive one. Now suppose that there is a systematic pattern of missing observations. An example is given in Table 1 when the two gray prices are missing. This scenario is depicted in the lower left panel of Figure 1. The product "goods" is observed in regions B and C, but missing in regions A and D. Thus, the red circles corresponding to the latter two regions need to be deleted. As a consequence, in region A the large positive disturbance in the upper left panel of Figure 1 vanishes, that is, $\mathbb{E}(e_i^A | \mathbf{x}_i^A) < 0$. To reduce the sum of squared residuals of region A's remaining two price observations, the CPD regression moves the vertical dashed line of region A to the left (see lower left panel of Figure 1). More generally, when a product with a low price variation is missing in the cheapest region, the CPD method's estimated price level of that region always decreases below the level with complete data-in other words, downward bias arises. Similarly, the missing observation in region D leads to $\mathbb{E}(e_i^{\rm D}|\mathbf{x}_i^{\rm D}) > 0$. The dashed vertical line of that region moves to the right, that is, the estimated price level of region D is upward biased (see lower left panel of Figure 1).

The corresponding price level estimates are

$$\hat{P}^{\rm A} = 0.64, \ \hat{P}^{\rm B} = 0.92, \ \hat{P}^{\rm C} = 1.09, \ \hat{P}^{\rm D} = 1.57$$

They can be compared to the price levels derived from complete price data. These price levels were reported in (3). The data gaps reduce the price level of region A by 14% while they raise the price level of region D by 17%. The price levels of regions B and C barely change. If the price observations missing in regions A and D were related to "housing" (the product with the largest price variation) instead of "goods" (the product with the lowest price variation), bias in the opposite direction would arise.

2.4 | Further Issues

The previous section showed that systematic gaps in the price data cause biased CPD estimates of the regional price levels (lower left panel of Figure 1). By contrast, if no prices were missing, the CPD regression would be unbiased (upper left panel of Figure 1). The same would be true if the prices were missing completely at random. However, even if these two exceptional cases applied, the CPD regression would be inefficient and inference would be invalid because the residuals would be both correlated and heteroskedastic.

This can be seen in the upper left panel of Figure 1. The correlation arises from the systematic relationship between the residuals and the general price levels of the regions. For example, there is a very strong negative correlation between the residuals \hat{e}_1^r (goods) and the estimated values of the general price levels, $\ln P^r$. This correlation is caused by the uniform prices of goods. Similarly, there is a strong positive correlation between the residuals \hat{e}_3^r (housing) and the estimated values of $\ln P^r$ because the differences in housing costs are more pronounced than the differences in the general price levels. Only the price variation of services is similar to that of the general price levels. As a consequence, the CPD regression's residuals related to services vary less than those related to goods and housing. Thus, heteroskedasticity arises.

The residuals' correlation and heteroskedasticity imply that the CPD regression is inefficient and that the estimation of the disturbances' standard deviation is biased. Therefore, inference is invalid. These conclusions are formally proven in Online Appendix A.4.2. Theoretically, the issue of invalid inference could be remedied along the lines proposed by Crompton (2000, p. 368), who advocates White's heteroskedasticity-robust specification of the variance matrix for the CPD regression. However, this remedy requires unbiased estimates of the price levels $\ln P^r$, that is, either complete price data (as in the upper left panel of Figure 1) or data gaps that arise completely at random. Real-world data rarely satisfy these requirements.

Therefore, a novel approach would be desirable that can handle missing observations regardless of their structure. The present paper introduces such an approach. It is a non-linear generalization of the CPD model (1) and can be applied in weighted or unweighted form.

3 | Solution

This section introduces the NLCPD method. In Section 3.1, we sketch out its basic concept and explain why this estimation approach requires an additional restriction. Section 3.2 applies the NLCPD method to the example of the previous section and discusses its advantages in comparison to the CPD method. The formal exposition of the NLCPD method is deferred to Section 4 and Online Appendix A.

3.1 | NLCPD Method

Again, we start with the case of complete data and product-specific price variation. In this case, the CPD regression is unbiased, but inefficient and inference is invalid. The three colored solid lines in the upper left panel of Figure 1 have a slope of one. The residuals could be markedly reduced if each solid line had its individual slope. This is accomplished when, instead of CPD model (1), the following relationship is estimated:

$$\ln p_i^r = \ln \pi_i + \delta_i \ln P^r + u_i^r \tag{4}$$

where $u_i^r \sim N(0, \sigma^2)$ is a normally distributed disturbance term with zero mean and variance σ^2 . The parameter δ_i determines the slope of product *i*'s colored line in the upper right panel of Figure 1. We denote the non-linear relationship (4) as the NLCPD regression model.

From an economic perspective, the parameters δ_i are the *elasticities* of prices p_i^r with respect to regional price levels P^r . As a consequence, a one percent increase in the regional price level P^r increases the price of product *i* in region *r* by δ_i percent. Products with price level elasticity $\delta_i > 1$ react more sensitively to changes in the regional price level P^r than the average of all products.⁴ In Table 1, this applies to the products "housing" and "services". They exhibit a stronger spatial price variation than the product "goods". Products with prices that are all but invariant with respect to the regional price levels have a slope parameter, δ_i , close to 0.

The NLCPD model (4) extends the CPD model (1) by the price level elasticities, δ_i . However, the NLCPD model (4) is not identified because $\delta_j \ln P^r = \tilde{\delta}_i \ln \tilde{P}^r$, with $\tilde{\delta}_i = (\delta_j / \lambda)$ and $\tilde{P}^r = P^{r\lambda}$. Thus, without any restriction on the δ_i -values, the regional price level ratios, $\tilde{P}^r / \tilde{P}^s = (P^r / P^s)^{\lambda}$, could be arbitrarily scaled up or down by the parameter λ . Consequently, an infinite number of solutions would exist.

To remove this ambiguity, the specification of the NLCPD model must be complemented by a restriction on the δ_i -values. This restriction arises quite naturally from an appealing property of the CPD method that the NLCPD method should also possess: The logarithm of the price level ratio of any pair of regions, P^r/P^s , is equal to the expected value of the logarithm of the Jevons index formula; see Equation (2). However, the price level ratios implied by the NLCPD model (4) are

$$\ln \frac{P^r}{P^s} = \mathbb{E}\left(\frac{(1/N)}{\sum_{i \in \mathcal{N}} (1/N)\delta_i} \sum_{i \in \mathcal{N}} \ln \frac{p_i^r}{p_i^s}\right)$$
(5)

Thus, the δ_i -values in Equation (5) must be such that $\sum_{i \in \mathcal{N}} (1/N)\delta_i = 1.^5$ Accordingly, the estimation of the NLCPD model (4) must be accompanied by the restriction

$$\sum_{i \in \mathcal{N}} (1/N)\hat{\delta}_i = 1 \tag{6}$$

where $\hat{\delta}_i$ is the estimator of δ_i .

3.2 | Graphical Illustration, Unbiased Price Levels, and Further Advantages

To obtain the non-linear least squares estimates of the unknowns $\ln P^r$, $\ln \pi_i$, and δ_i in the NLCPD model (4), this model is transformed into a regression equation with the same set of dummy variables as the regression equation corresponding to the CPD model (1). In other words, the NLCPD method requires no additional information.

For the complete price data listed in Table 1, the outcome of the NLCPD estimation is depicted in the upper right panel of Figure 1. The estimates of the slopes of the colored regression lines are $\hat{\delta}_1 = 0.00$, $\hat{\delta}_2 = 1.71$, and $\hat{\delta}_3 = 1.29$. They satisfy restriction (6). The estimated price levels are

$$\hat{P}^{A} = 0.74, \ \hat{P}^{B} = 0.92, \ \hat{P}^{C} = 1.09, \ \hat{P}^{D} = 1.35$$
 (7)

Because no price data are missing, they are very similar to those obtained from the CPD regression listed in (3).

The lower right panel of Figure 1 depicts the NLCPD estimation for the incomplete data set of Table 1 (the prices of "goods" are missing in regions A and D). In contrast to the CPD regression, these data gaps cause hardly any change in the estimated NLCPD price levels \hat{P}^A to \hat{P}^D . They are almost identical to those reported in (7). In other words, incomplete data no longer lead to estimation bias.

Another major advantage of the NLCPD regression is a better model fit. In the case of complete data (upper panels of Figure 1), the sum of squared residuals divided by the degrees of freedom falls from 0.055 (CPD regression) to 0.004 (NLCPD regression).⁶ Furthermore, in contrast to the CPD regression, the NLCPD method provides meaningful estimates of the standard errors of all estimated parameters (formally shown in Online Appendix A.4.1). Thus, the statistical significance of the coefficients $\ln P^r$, $\ln \hat{\pi}_i$, and $\hat{\delta}_i$ can be examined. When in the NLCPD regression at least one coefficient $\hat{\delta}_i$ significantly deviates from one, the CPD model is misspecified (see Section 4.3 on model specification).

4 | Method

The NLCPD model (4) is a generalization of the linear CPD model (1). The model function is non-linear in its parameters. Consequently, parameter estimates must be derived by

a non-linear regression approach. This approach is formally described in Sections 4.1 and 4.2 and compared to the CPD estimators. In contrast to the previous sections, we now allow for weighting information in the estimations. Section 4.3 expounds on the issues of weighting and inference. A short discussion of the limitations and possible generalizations of the NLCPD method is presented in Section 4.4.

4.1 | Model Identification

Let $\mathcal{R} = \{r : r = 1, 2, ..., R\}$ denote the set of regions. In subnational price comparisons "region" relates to some sort of "district", while in international price level comparisons "region" means "country". The set of products included in the price comparison is denoted by $\mathcal{N} = \{i : i = 1, 2, ..., N\}$. Furthermore, \mathcal{N}_r denotes the set of products for which a price is available in region *r*. Analogously, \mathcal{R}_i defines the set of regions in which product *i* is priced. The set's number of products is denoted by \mathcal{R}_i .

In both the CPD model (1) and the NLCPD model (4), perfect multicollinearity would arise. To avoid this problem, one of the $\widehat{\ln \pi_j}$ -values or $\widehat{\ln P^s}$ -values can be set equal to 0. Alternatively, the normalizations $\sum_{i \in \mathcal{N}} \widehat{\ln \pi_i} = 0$ or $\sum_{r \in \mathcal{R}} \widehat{\ln P^r} = 0$ can be applied. The $\widehat{\ln P^r}$ -values reported in (3) and (7) were based on the latter variant of normalization. Then, one of the values $\widehat{\ln P^r}$ is derived as a residual instead of being estimated by the regression. Any of the $\widehat{\ln P^r}$ -values can be selected for this purpose. In the following, we use $\widehat{\ln P^1}$, that is,

$$\widehat{\ln P^1} = -\sum_{s \in \mathcal{R} \setminus \{1\}} \widehat{\ln P^s}$$
(8)

Above basic heading level, expenditure shares are usually available. Let w_i^r denote the expenditure share of product *i* within region *r*. Thus, the average expenditure share of product *i* is

$$w_i = \frac{1}{R} \sum_{r \in \mathcal{R}} w_i^r \tag{9}$$

Note that $\sum_{i \in \mathcal{N}} w_i = 1.^7$

The CPD model (1) and the NLCPD model (4) had one principle in common: When no weights are available, the logarithm of the price level ratio of any pair of regions, P^r/P^s , is equal to the expected value of the logarithm of the Jevons index formula. For the NLCPD model, this principle resulted in restriction (6). The same principle can be adapted to *weighted* variants of the CPD and NLCPD models. Then, the logarithm of the price level ratio of any pair of regions should be equal to the expected value of the logarithm of a weighted Jevons index formula, where the weights are given by w_i .⁸ For the NLCPD method, this principle leads to a generalized version of restriction (6).

To derive this generalized restriction, we transform the weighted NLCPD model into the following relationship:⁹

$$\ln \frac{P^r}{P^s} = \mathbb{E}\left(\frac{1}{\sum_{i \in \mathcal{N}} w_i \delta_i} \sum_{i \in \mathcal{N}} w_i \ln \frac{p_i^r}{p_i^s}\right)$$
(10)

To turn the term in brackets into the logarithm of a weighted Jevons index formula (with weights w_i), we need the restriction $\sum_{i \in N} w_i \delta_i = 1$.

Accordingly, the NLCPD estimation must be accompanied by the restriction

$$\sum_{i \in \mathcal{N}} w_i \hat{\delta}_i = 1 \tag{11}$$

Substituting in Equations (10) and (11) the weights w_i by 1/N, gives Equations (5) and (6). In other words, restriction (6) is merely a special case of restriction (11).

The previous considerations imply that the weighted estimation of the NLCPD regression model (4) must be accompanied by restriction (11). Therefore, one $\hat{\delta}_i$ -value is residually computed. Any $\hat{\delta}_i$ -value can be used for this purpose. We choose $\hat{\delta}_1$, that is,

$$\widehat{\delta}_1 = \frac{1 - \sum_{i \in \mathcal{N} \setminus \{1\}} w_i \widehat{\delta}_i}{w_1} \tag{12}$$

4.2 | Estimators

The residuals \hat{u}_i^r of the NLCPD regression model (4) are defined by $\hat{u}_i^r = \ln p_i^r - \hat{\delta}_i \widehat{\ln P^r} - \widehat{\ln \pi}_i$. Accordingly, the weighted sum of squared residuals, $S_{\hat{u}_i^r \hat{u}_i^r}$, can be written as

$$S_{\hat{u}_{i}^{r}\hat{u}_{i}^{r}} = \sum_{r \in \mathcal{R}} \sum_{i \in \mathcal{N}_{r}} w_{i} \left(\ln p_{i}^{r} - \widehat{\delta_{i}} \widehat{\ln P^{r}} - \widehat{\ln \pi_{i}} \right)^{2}$$
$$= \sum_{i \in \mathcal{N}} \sum_{r \in \mathcal{R}_{i}} w_{i} \left(\ln p_{i}^{r} - \widehat{\delta_{i}} \widehat{\ln P^{r}} - \widehat{\ln \pi_{i}} \right)^{2}$$
(13)

where $\hat{\delta}_1$ and $\widehat{\ln P^1}$ are defined by normalization (8) and restriction (12). The NLCPD-estimators can be derived by minimizing $S_{\hat{u}_i^r \hat{u}_i^r}$ with respect to $\widehat{\ln \pi_i}$ (i = 1, ..., N), $\hat{\delta}_i$ (i = 2, ..., N), and $\widehat{\ln P^r}$ (r = 2, ..., R). In the following we state and discuss these estimators (they are derived in Online Appendix A.1). The values of $\widehat{\ln P^1}$ and $\hat{\delta}_1$ are residually derived from Equations (8) and (12).

The NLCPD estimator of the logarithmic general value of product i, $\ln \pi_i$, is

$$\widehat{\ln \pi_i} = \frac{1}{R_i} \sum_{r \in \mathcal{R}_i} \left(\ln p_i^r - \widehat{\delta_i} \widehat{\ln P^r} \right)$$
(14)

In this NLCPD estimator, each region receives the same weight, $1/R_i$, because in the weighted sum (13) the weights w_i are uniform across regions. Setting $\hat{\delta}_i = 1$ for all products $i \in \mathcal{N}_r$, the NLCPD estimator (14) simplifies to the corresponding CPD estimator:

$$\widehat{\ln \pi_i'} = \frac{1}{R_i} \sum_{r \in \mathcal{R}_i} \left(\ln p_i^r - \widehat{\ln P^{r'}} \right)$$
(15)

where $\ln P'$ is the CPD estimator of the regional price levels as defined in Equation (17), below. For a product *i* that is priced in all regions ($R_i = R$), both the NLCPD estimator (14) and the CPD estimator (15) simplify to the CPD formula in Diewert (2004, p. 7):

$$\widehat{\ln \pi_i} = \widehat{\ln \pi_i'} = \frac{1}{R} \sum_{r \in \mathcal{R}} \ln p_i^r$$

The NLCPD estimator of the logarithmic price level of region r, $\ln P^r$, is

$$\widehat{\ln P^{r}} = \frac{\sum_{i \in \mathcal{N}_{r}} w_{i} \widehat{\delta}_{i} \left(\ln p_{i}^{r} - \widehat{\ln \pi_{i}} \right)}{\sum_{i \in \mathcal{N}_{r}} w_{i} \left(\widehat{\delta}_{i} \right)^{2}}$$
(16)

The numerator is the covariation (across products) of $\left(\ln p_i^r - \widehat{\ln \pi_i}\right)$ and the elasticity $\widehat{\delta_i}$. The denominator is the variation (across products) of $\widehat{\delta_i}$. The same formula would be applied in a weighted least squares regression where the dependent variable $\left(\ln p_i^r - \widehat{\ln \pi_i}\right)$ is a linear function of the independent variable $\widehat{\delta_i}$. A negative value, $\widehat{\ln P^r}$, indicates a relatively cheap region. It arises when the numerator is negative, that is, when in region *r* prices, $\ln p_i^r$, below the general value, $\widehat{\ln \pi_i}$, dominate in the sense that they are either more frequent and/or more often arise for products with a large elasticity, $\widehat{\delta_i}$. In expensive regions $\left(\widehat{\ln P^r} > 0\right)$, prices above the general level dominate.

Setting $\hat{\delta}_i = 1$ for all products $i \in \mathcal{N}_r$, the NLCPD estimator (16) simplifies to the corresponding CPD estimator:

$$\widehat{\ln P^{r'}} = \frac{\sum_{i \in \mathcal{N}_r} w_i \left(\ln p_i^r - \widehat{\ln \pi_i^r} \right)}{\sum_{i \in \mathcal{N}_r} w_i}$$
(17)

When in region *r* no data gaps occur, we have $\sum_{i \in N_r} w_i = 1$ and the resulting estimator (17) simplifies to the well-known CPD formula (e.g., Rao 2005, p. 577; Rao and Hajargasht 2016, p. 417):

$$\widehat{\ln P^{r'}} = \sum_{i \in \mathcal{N}_r} w_i \left(\ln p_i^r - \widehat{\ln \pi_i'} \right)$$

The parameter δ_i represents the price elasticity of product *i* with respect to the general price level, ln P^r . The parameter's estimator is

$$\widehat{\delta}_{i} = \frac{\sum_{r \in \mathcal{R}_{i}} \left(\ln p_{i}^{r} - \widehat{\ln \pi_{i}} \right) \widehat{\ln P^{r}}}{\sum_{r \in \mathcal{R}_{i}} \left(\widehat{\ln P^{r}} \right)^{2}}$$
(18)

The numerator is the covariation (across regions) of the logarithmic regional price levels, $\ln P^r$, and $\left(\ln p_i^r - \ln \pi_i\right)$. The denominator is the variation (across regions) of the logarithmic regional price levels. Therefore, the estimator (18) can be viewed as the ordinary least square estimator of the slope parameter of a simple linear model where $\left(\ln p_i^r - \ln \pi_i\right)$ is regressed on $\ln P^r$. The covariation represented by the numerator is usually positive. The larger this covariation, the stronger the elasticity of the prices $\ln p_i^r$ with respect to the price levels $\ln P^r$. If some product *i* has a uniform price, then $\ln \pi_i = \ln p_i^r$ and, therefore, the fraction becomes 0.

For the derivation of the non-linear least squares formulas (14), (16), and (18), the expenditure share weights, w_i , were used. They are uniform across regions and add up to unity. This is in line with the weighting information usually available for subnational

price comparisons. In other contexts, however, one may want to apply NLCPD estimators with expenditure share weights that vary across regions (w_i^r instead of w_i) or/and one may want to apply weights that do not necessarily reflect expenditure shares. This general case is considered in Online Appendix A.1.¹⁰

Below basic heading level, no weighting information exists. Even in such cases, the products' elasticities, δ_i , may be heterogeneous. For these cases, the unweighted NLCPD method can be used. Its estimators can be derived by minimizing the unweighted sum of squared residuals under restriction (6). The resulting estimators of $\ln \pi_i$ and δ_i are identical to the weighted NLCPD-estimators (14) and (18). The $\ln P^r$ -estimator (16) simplifies to

$$\widehat{\ln P^{r}} = \frac{\sum_{i \in \mathcal{N}_{r}} \widehat{\delta}_{i} \Big(\ln p_{i}^{r} - \widehat{\ln \pi_{i}} \Big)}{\sum_{i \in \mathcal{N}_{r}} \Big(\widehat{\delta}_{i} \Big)^{2}}$$

The non-linear least squares formulas (14), (16), and (18) do not provide explicit solutions for the coefficients $\ln \pi_i$, $\ln P^r$, and $\hat{\delta}_i$. Instead, an iterative optimization routine with appropriate start values is necessary. For $\ln \pi_i$ and $\ln P^r$, such start values can be obtained from the CPD method. Their insertion into formula (18) yields start values for $\hat{\delta}_i$. Further details on this strategy can be found in Online Appendix B, which also discusses two alternative strategies along with the applied optimization approach.

4.3 | Weighting, Inference, and Model Specification

For the weighted least squares approach to yield consistent estimators, Gorajek (2022, 86-87) stresses that the weights should be explicitly incorporated in the econometric model.¹¹ Accordingly, the following discussion considers the weights w_i as part of the error terms of the CPD and NLCPD models.

The weights w_i were defined in the previous sections as the products' average expenditure shares. For the weighted least squares approach in Equation (13), this is not necessarily the most appropriate form of weighting. Generally, the weight of an observation can represent its economic importance (e.g., Diewert 2005, 562-563; Rao 2005, 574-575) and/or it can reflect the econometric reliability of the observation's information for estimating the regional price levels. A natural measure of an observation's economic importance is the product's expenditure share, while the observation's reliability of information is inversely related to the variance of the disturbance term u_i^r . Thus, the economic and the econometric motivation for weighting may lead to different sets of weights. This complicates the estimation of standard errors for the CPD and NLCPD methods.

In the NLCPD model (4), the disturbance term, u_i^r , can be homoskedastic or heteroskedastic. The latter case implies that the reliability of the observations' information is not uniform. However, Clements and Izan (1987) argue that a product's expenditure share usually is a reasonable approximation to the product's reliability of information. More specifically, they assume that the variance of the disturbance term, u_i^r , is given by σ^2/w_i where σ^2 is a constant and $\sum_{i \in \mathcal{N}} w_i = 1$. Weighting each observation by the square root of the product's expenditure share, $\sqrt{w_i}$, yields a homoskedastic weighted disturbance term, $\sqrt{w_i}u_i$. If, at the same time, product *i*'s expenditure share w_i is considered an appropriate measure of its economic importance, no contradiction arises between the economic and the econometric motivation for weighting.

This coincidence simplifies the derivation of the NLCPD estimators' standard errors. In non-linear regression models, approximations of these standard errors can be computed from the Jacobian matrix evaluated at final parameter estimates. These approximations are derived in Online Appendix A.4.1. In Online Appendix A.4.2 it is shown that the estimates of the CPD estimators' standard errors are biased.

Clements and Izan (1987) argue that the weight w_i correctly addresses product *i*'s economic importance and that the weighted disturbance term $\sqrt{w_i}u_i^r$ is homoskedastic because the weight w_i is negatively related to the variance of the disturbance term u_i^r . Clements et al. (2006) give two justifications for this negative relationship. First, statistical offices spend more effort on the collection of correct prices when the products are of greater relevance to the budget. Second, by definition, the true price level is closer to the prices of the products with larger budget shares.

However, this justification is not always backed by empirical evidence (Diewert 1995, p. 20). For example, when all observations can be considered as equally reliable, the unweighted disturbance term, u_i^r , is homoskedastic and the weighted disturbance term, $\sqrt{w_i}u_i^r$, is heteroskedastic. Rao (2004, 17-18) and Hajargasht and Rao (2010, S44-S46) describe how this should be accounted for when, in a CPD regression, the standard errors of the estimated parameters are to be computed.

The upper left panel of Figure 1 revealed that product-specific elasticities δ_i result in a heteroskedastic disturbance term. In this case, the NLCPD regression model is preferable. If its unweighted disturbance term, u_i^r , is homoskedastic, the weighted disturbance term, $\sqrt{w_i}u_i^r$, is heteroskedastic and the standard errors of the estimated parameters must be computed using a formula that resembles the CPD formula stated in Hajargasht and Rao (2010, S45). If both, the unweighted disturbance term, $\sqrt{w_i}u_i^r$, are heteroskedastic, an even more general formula is required (see Online Appendix A.4.1).

Unbiased estimates of the standard errors are a prerequisite for meaningful *t*- and *F*-tests. Since the CPD model is nested in the NLCPD model, a two-sided *F*-test of the null hypothesis $H_0: \delta_1 = \delta_2 = \ldots = \delta_N = 1$ directly indicates whether the CPD model is misspecified. The examination could be refined by individual *t*-tests of the null hypotheses $H_0: \delta_i = 1$ for all $i \in \mathcal{N}$.

4.4 | Limitations and Generalizations

The only drawback of the NLCPD method as compared to the CPD method is its non-linear specification. As a consequence, iterative estimation procedures are required. When regional price levels are similar and a product has only very few observations, the iterative estimation of the product's elasticity might not converge. To avoid such problems, one may treat such a product in the same way it would have been treated in a CPD regression. That is, instead of estimating the product's elasticity, one can impose the restriction $\delta_i = 1$. Recall that the CPD method imposes this restriction on *all* products. Furthermore, for products that are known to be sold everywhere at the same price, one can impose the restriction $\delta_i = 0$. Such a restricted NLCPD regression would clearly outperform the CPD regression.

The CPD regression assumes that each price can be explained by the linear relationship $\ln p_i^r = \ln \pi_i + \ln P^r + e_i^r$, while the NLCPD method amends the logarithmic price level $\ln P^r$ in this relationship by the price level elasticity δ_i . Is it possible to make an additional extension of the CPD relationship? More specifically, can we complement the logarithmic general value $\ln \pi_i$ by the parameter μ^r ? This parameter would represent the elasticity of the price of product *i* in region *r* with respect to changes of this product's general value π_i . This elasticity would be region-specific (just as δ_i is product-specific). Notwithstanding that the region-specific general value elasticity, μ^r , appears less relevant than the product-specific price level elasticity, δ_i , one can show that the additional inclusion of μ^r would lead to a singular regressor matrix.

Cuthbert and Cuthbert (1988, p. 55) point out that the price of some product *i* tends to be negatively related to its degree of representativeness in region *r*. For the NLCPD model (4) to correctly capture the impact of representativeness on product prices, it would be necessary that a product's degree of representativeness is correlated with the regional price level, P^r . However, such a correlation is quite unlikely. Therefore, the issue of representativeness requires a separate treatment. Cuthbert and Cuthbert (1988) propose to extend the CPD model (1) by *N* dummy variables, each relating to a different product. The dummy variable of product *i*, say, takes on the value 1 only if the observation relates to product *i* and, at the same time, this product is *not* representative in the region to which the observation relates. For the NLCPD model (4), the same extension is conceivable.

5 | Simulation

Imposing the restriction $\delta_i = 1$ for all products *i* in the NLCPD model (4) yields the CPD model (1). However, the restriction is quite unrealistic as the elasticities δ_i can be expected to vary across basic headings and sometimes even within basic headings. Hence, the NLCPD method should theoretically provide more accurate price level estimates than the CPD method. To examine this hypothesis in a statistical context, we perform a Monte Carlo simulation, which compares the performances of the CPD and NLCPD methods. A popular alternative to the CPD method is the GEKS approach, named after its authors Gini (1924, 1931), Eltetö and Köves (1964), and Szulc (1964). Therefore, we also include the GEKS approach in our simulation study.

Section 5.1 provides a few remarks on the GEKS approach. The simulation setting is described in Section 5.2, while the results are provided in Section 5.3.

5.1 | GEKS Approach

The GEKS approach is used in the International Comparison Program (World Bank 2020) and in various national studies (surveyed in Majumder and Ray 2020, 105-109; Weinand and Auer 2020, 416-418). Most commonly, the GEKS approach derives the ln P^r -estimates from geometric averages of chained Fisher indices (e.g., Rao and Hajargasht 2016, p. 416). We refer to this variant as the GEKS index in the rest of the paper. The GEKS-Törnqvist index uses the Törnqvist index instead of the Fisher index (e.g., Caves et al. 1982, Selvanathan and Rao 1994). If no weighting information is available, the GEKS-Jevons index can be applied (e.g., World Bank 2013, 100-105). It replaces the Fisher index by the Jevons index.

In the literature, it is well known that the unweighted CPD method and the GEKS-Jevons index provide identical results when the data set is complete (e.g., World Bank 2013, p. 108). Weinand and Auer (2019, 35-37) show that the weighted CPD method and the GEKS-Törnqvist index coincide for the expenditure share weights w_i . Even when data gaps are present, there is a close relationship between the two methods (Weinand 2022). Consequently, we suspect that any issue of one approach is likely to also apply to the other one.

5.2 | Setting

In the simulation, we consider four different scenarios. They differ with respect to the number and structure of missing observations and with respect to the variance of the δ_i -values. The first of the four scenarios is the most artificial one, while the fourth scenario is the most realistic one. The other two scenarios allow us to identify the separate effects of missing observations and varying δ_i -values. In all scenarios, we consider N = 15 products or basic headings available in R = 20 regions.

- Scenario 1: We assume that the price data are complete, that is, there is exactly one price per product and region. This gives NR = 300 observations. The true δ_i -parameters are 1 for all *N* products. Note that this corresponds to the CPD model (1).
- Scenario 2: We still assume that the price data are complete. Now, however, the true δ_i -parameters are allowed to differ from 1.
- Scenario 3: We assume that every third price is missing. This gives a total of 200 remaining observations. The missing prices are chosen completely at random. All other parameters are the same as in the second scenario.
- Scenario 4: We keep the setting of the third scenario but introduce the missing prices in a systematic manner: The larger the δ_i , the smaller the probability that prices for product *i* are missing.

For each scenario, we perform the following steps. First, we generate artificial price data by inserting randomly sampled values of $\ln P^r$, $\ln \pi_i$, δ_i , w_i , and u_i^r into the data generating process defined in Equation (4). The sampling of these values is described in

Online Appendix C. Second, we order the regions according to their true price levels $\ln P^r$ and then label the regions by their rank. In other words, region r = 1 always denotes the cheapest region and region r = 20 the most expensive one. Similarly, we arrange the products according to their δ_i -parameter. Thus, product i = 1 always exhibits the lowest elasticity. Third, we apply the (weighted) NLCPD method, the (weighted) CPD method, and the GEKS index to the price data generated during the first step.¹² For the starting values of the NLCPD method, we use the CPD method's estimates for $\ln P^r$ and $\ln \pi_i$ as starting values. These values are also used to calculate the starting values of all δ_i using formula (18).

We repeat these three steps L = 2000 times (with iterations l = 1, 2, ..., L) and obtain for each region, r, 2000 vectors of $\ln P^{r'}$ -values for the NLCPD method, 2000 vectors of $\ln P^{r''}$ -values for the CPD method, and 2000 vectors of $\ln P^{r''}$ -values for the GEKS index. Afterwards, we compare the performance of the three methods. To this end, we use the NLCPD results of the L iterations to compute for each region r the absolute value of the bias, $|\text{Bias}(\widehat{\ln P^r})|$, and also the root mean squared error, $\text{RMSE}(\widehat{\ln P^r})$. Then, we take the average of these numbers across all regions:

$$\operatorname{Bias}\left(\widehat{\ln P}\right) = \frac{1}{R} \sum_{r \in R} \left| \operatorname{Bias}\left(\widehat{\ln P^{r}}\right) \right|$$
$$= \frac{1}{R} \sum_{r \in R} \left| \frac{1}{L} \sum_{l=1}^{L} \left(\widehat{\ln P^{r}_{l}} - \ln P^{r}_{l} \right) \right|$$
(19a)

$$RMSE(\widehat{\ln P}) = \frac{1}{R} \sum_{r \in \mathcal{R}} RMSE\left(\widehat{\ln P^{r}}\right)$$
$$= \frac{1}{R} \sum_{r \in \mathcal{R}} \sqrt{\frac{1}{L} \sum_{l=1}^{L} \left(\widehat{\ln P^{r}_{l}} - \ln P^{r}_{l}\right)^{2}}$$
(19b)

where $\widehat{\ln P_l'}$ denotes the *estimated* parameter of region *r*'s price level obtained in iteration *l* by the NLCPD method, while $\ln P_l'$ is the corresponding *true* parameter. Following the same strategy, we derive for the CPD method the values of $\operatorname{Bias}\left(\widehat{\ln P'}\right)$ and $\operatorname{RMSE}\left(\widehat{\ln P'}\right)$ and for the GEKS index the values of $\operatorname{Bias}\left(\widehat{\ln P''}\right)$ and $\operatorname{RMSE}\left(\widehat{\ln P''}\right)$.

For Scenarios 1–3, we expect all methods to produce unbiased estimates for $\ln P^r$. However, when data gaps are introduced in a systematic manner, as in Scenario 4, $\ln P^r$ -estimates of the CPD method are expected to be biased (see Section 2). The $\ln P^r$ -estimates of the GEKS index are derived from geometric averages of chained Fisher indices. Since systematic data gaps can lead to biased Fisher indices, the GEKS index is also susceptible to bias.

Although the degrees of freedom in the NLCPD method are lower than in the CPD method, we expect that the NLCPD model's higher flexibility results in higher accuracy. Consequently, the RMSE should be lower for the NLCPD method in all simulation scenarios. The only exception should be the first scenario where the true δ_i -values are equal to 1, as implicitly assumed in the CPD method.

5.3 | Discussion of Results

Table 2 shows the simulation results for the mean absolute bias and the mean RMSE of the $\ln P^r$ -estimates.¹³ Regional price level estimates seem to be unbiased for all three methods if price data are complete or if gaps occur completely at random (Scenarios 1–3). The mean absolute bias over all regions is all but zero. However, if data gaps occur systematically (Scenario 4), the $\ln P^r$ -estimates of the CPD method are—in absolute terms—biased by more than 1% and those of the GEKS index by more than 2% on average. By contrast, the NLCPD method's estimates are still unbiased.

In general, a lower RMSE indicates higher accuracy. Since regional price levels are measured on the logarithmic scale, even small differences in the RMSE significantly impact accuracy. In Scenarios 2–4, the computed mean RMSE of $\ln P^r$ -estimates is lower for the NLCPD method than for the CPD method and the GEKS index (see Table 2). If the price data are complete, the difference in the mean RMSE is relatively small. With missing prices, however, this difference noticeably increases. In Scenario 1, the RMSE of the NLCPD method is (almost) as

small as that of the CPD method and the GEKS index. In other words, when the true δ_i -values are equal to 1, the efficiency loss of the NLCPD method is negligible.

The NLCPD method's better performance is not only valid on average, but can be observed for each region and each scenario. This is shown in Figure 2. Its structure is similar to Table 2 but it depicts the bias and RMSE for each region r. The regions are listed on the horizontal axis. They are ordered with respect to their true price level.

The top row of Figure 2 reveals that in all regions the NLCPD method, the CPD method, and the GEKS index are unbiased as long as the data are complete or missing completely at random (Scenarios 1–3), but that the CPD method and the GEKS index are biased when the data gaps are systematic (see the red and green dots in Scenario 4). More specifically, the more a region's true price level deviates from the average price level of all regions, the larger the bias will be. As predicted in Section 2, in the cheap regions, downward bias arises, while the expensive regions exhibit upward bias. Consequently, the CPD method overestimates the price level spread between the most expensive region and the cheapest region. The bias is even more pronounced for the GEKS index. Recall that in Scenario 4, the number of data gaps is negatively correlated with the product's true elasticity, δ_i . Switching to a positive correlation, one would observe the

TABLE 2 | Mean absolute bias and mean RMSE of the NLCPD estimates, $\widehat{\ln P^r}$, the CPD estimates, $\widehat{\ln P^{r'}}$, and the GEKS estimates, $\widehat{\ln P^{r''}}$.

	Scenario 1		Scenario 2		Scenario 3		Scen	Scenario 4	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	
NLCPD	0.000	0.010	0.000	0.008	0.000	0.011	0.001	0.010	
CPD	0.000	0.010	0.000	0.010	0.000	0.020	0.013	0.025	
GEKS	0.000	0.010	0.000	0.010	0.000	0.020	0.024	0.033	



FIGURE 2 | Bias and RMSE of the NLCPD estimates, $\widehat{\ln P^r}$, the CPD estimates, $\widehat{\ln P^{r'}}$, and the GEKS estimates, $\widehat{\ln P^{r''}}$, for the four simulation scenarios.

opposite effects, that is, cheap regions appear too expensive, expensive regions appear too cheap and, therefore, the regional price level spread is underestimated. The NLCPD method avoids all these problems. In Scenario 4, the blue dots remain close to the horizontal baseline.

The NLCPD method also outperforms the CPD method and the GEKS index with respect to the RMSE. This is shown in the bottom row of Figure 2. The blue dots are closer to the baseline than the red and green dots. As long as the data are complete (Scenario 2), the advantage of the NLCPD method does not depend on a region's true price level. However, when data gaps occur (Scenarios 3 and 4), the accuracy problems of the CPD method and the GEKS index become more pronounced. The *u*-shape of the red and green dots implies that the largest inaccuracies arise for the cheapest and the most expensive regions.

6 | Empirical Application

In the following, we apply the NLCPD method to regional price levels above the basic heading level, compiled from German official consumer price index (CPI) micro data. This is of particular interest because price level elasticities, δ_i , can be expected to vary between basic headings (e.g., rents vs. manufactured goods), while the CPD method assumes that these elasticities are uniform. Therefore, we also compare the results of the NLCPD method to those we would obtain from the CPD method. In addition, we report the results of the GEKS index. The estimated price levels are transformed into a regional price index for Germany.¹⁴

6.1 | Price Data and Aggregation Approach

We have the privilege to work with German CPI micro data from May 2019. These data were provided to us by the Research Data Center of the Federal Statistical Office and Statistical Offices of the Länder. In total, the data contain more than 400,000 price observations for goods, services, and rents, which were collected in the 401 districts of Germany (henceforth, we speak of regions). Because the prices of few items are collected in all regions, the micro price data exhibit gaps.

The observations of the German CPI are classified into 12 divisions (see Table 3) and further into 783 basic headings. This classification follows the United Nations' Classification of Individual Consumption by Purpose (COICOP). In the German CPI, the expenditure weights of the basic headings are uniform across regions.

Due to methodological reasons, 70 basic headings with centrally collected prices cannot be exploited in a regional analysis.¹⁵ They represent a combined expenditure weight of 13.44%. Thirty-six other basic headings with a combined weight of 1.45% were too fragmentary to convey useful information for the subnational price comparison.¹⁶ As can be seen from Table 3, the largest problems are in division "09: Recreation and culture", where 2.66 percentage points of the 4.75% reported can be attributed solely to the basic heading of package holidays. By contrast, the divisions 01 to 03 (food, beverages, and clothing) are almost fully covered in the regional price comparison.

This leaves us with 677 basic headings for which the price information can be included in the regional price comparison. For each of these basic headings, we assume that the spatial price variation of the items within a basic heading is identical. Thus, the set of regional price levels of a given basic heading can be estimated with the CPD method. Since the expenditure weights of the individual items are not known, a weighted estimation is not feasible. Principally, we apply the CPD method to each basic heading (except for rents). Apart from a few improvements, the data preparation and aggregation have been very similar to the process documented in Weinand and Auer (2020).

TABLE 3 | Number of basic headings included in the price level estimation ("#BH") and their expenditure weights in the German CPI (as a percentage, base year 2015). Usable and unusable weights add up to 100%.

			Expenditure weight	
ID	Division	#BH	Usable	Unusable
01	Food and non-alcoholic beverages	172	9.69	0.00
02	Alcoholic beverages, tobacco, and narcotics	18	3.78	0.00
03	Clothing and footwear	62	4.45	0.08
04	Housing, water, electricity, gas, and other fuels	38	29.95	2.52
05	Furnishings, household equipment, and maintenance	93	4.50	0.50
06	Health	31	3.92	0.69
07	Transport	53	11.29	1.62
08	Communication	1	0.05	2.62
09	Recreation and culture	100	6.58	4.75
10	Education	7	0.90	0.00
11	Restaurants and hotels	36	3.60	1.07
12	Miscellaneous goods and services	66	6.39	1.03
		677	85.11	14.89

There are almost 300 basic headings that also contain prices related to the outlet type "internet and mail-order business". These prices are constant across regions. Their combined expenditure weight is 2.96%. Furthermore, the prices of 56 other basic headings (weight 10.18%) are uniform across Germany (e.g., cigarettes). We combine all prices that are constant across regions in two separate price level vectors. Together, they account for 13.14% of the total expenditure weight.

In the German CPI, five basic headings represent rents (weight 19.63%). The rent data are collected by the Federal Statistical Office. The sample includes the qualitative features of the flats. Therefore, we do not use a CPD regression, but estimate the regional rent levels by means of a hedonic regression that takes into account the individual characteristics of each flat. The details of this procedure are documented in Weinand and Auer (2020, 423-424, see second aggregation stage). As a result, the five basic headings are aggregated into one basic heading. However, this basic heading covers mainly existing tenancies. Therefore, we add another basic heading featuring the rent levels of new contracts. These rent levels were provided to us by the Federal Office for Building and Regional Planning (BBSR) for the second quarter of 2019.

The prices of fuels collected by the Federal Statistical Office represent four different basic headings. We replace them with two basic headings computed from a full sample that was collected by the German Market Transparency Unit for Fuels in May 2019.¹⁷

In total, our compilation procedures yield 618 price level vectors, one for each basic heading. They cover 85.11% of the total expenditure weight. The remaining 14.89% of total expenditure weight is proportionally assigned to these 618 basic headings. This set of weights and price level vectors forms the database for the NLCPD and CPD methods as well as for the GEKS index. The weighted NLCPD and CPD estimations are conducted as described in Section 4. The empirical results not only provide us with a reliable regional price index for Germany but also allow us to verify the theoretical predictions made in the previous sections.

6.2 | Discussion of Empirical Results

The NLCPD and CPD methods as well as the GEKS index provide estimates of the overall logarithmic price levels of the 401 German regions, $\ln P^r$, $\ln P^{r'}$, and $\ln P^{r''}$ (r = 1, ..., 401), respectively. The NLCPD and CPD methods also estimate the basic headings' general values, $\ln \pi_i$ and $\ln \pi'_i$ (i = 1, ..., 618), respectively. Only the NLCPD method additionally provides estimates of the basic headings' elasticities, $\hat{\delta}_i$ (i = 1, ..., 618).

Except for very few outliers, the NLCPD method's estimates $\hat{\delta}_i$ appear highly plausible. For the two basic headings with

constant regional price levels, the NLCPD method yields an estimated elasticity of $\hat{\delta}_i = 0$. For rents (existing tenancies) and for new lease rents we get $\hat{\delta}_i = 3.23$ and $\hat{\delta}_i = 4.82$, respectively. On average, the $\hat{\delta}_i$ -values of goods are the smallest ones. The $\hat{\delta}_i$ -values of rents are among the largest ones, while most of the $\hat{\delta}_i$ -values of services take a middle position. The results clearly confirm that the elasticities vary between the basic headings. The *F*-test (1% significance level) between the CPD and NLCPD models rejects the null hypothesis that all price level elasticities are equal to one. Thus, the implicit working hypothesis of the CPD method is falsified by our results.

The logarithmic price level estimates of the CPD and NLCPD methods are found to be highly correlated (Pearson correlation: 0.97).¹⁸ The estimated logarithmic price levels obtained from the NLCPD method, $\ln P^r$, range between -0.09 and 0.22, while those of the CPD method, $\ln P^{r'}$, exhibit a much larger spread ranging from -0.17 to 0.31. This empirical finding is perfectly in line with the theoretical predictions made in Section 2. There, it was argued that a negative correlation between a product's number of data gaps and its elasticity δ_i results in an upward biased estimate of the spread of the estimated regional price levels. In the present case, the Spearman correlation of the number of data gaps and the NLCPD's estimates $\hat{\delta}_i$ is -0.13. Consequently, the CPD method produces biased price level estimates. The bias inherent in the GEKS index is even more pronounced. The logarithmic price levels, $\ln P^{r''}$, range between -0.17 and 0.33.

In order to transform the logarithmic price level estimates into a regional price index, they are expressed in relation to their respective population-weighted averages. For the NLCPD method, the transformation is

$$P^r = 100 \cdot \exp\left(\widehat{\ln P^r} - \ln P^{\mathrm{Ger}}\right)$$

where $\ln P^{\text{Ger}} = \sum_{r=1}^{401} g^r \ln P^r$ and g^r is the population share of region *r*. The same transformation is applied to the CPD price level estimates $\ln P^{r'}$ and the GEKS price level estimates $\ln P^{r''}$. Summary statistics of the resulting price index numbers are reported in Table 4.

As can be seen from Table 4, the price level of the cheapest region is 10.8% below the population-weighted average when the NLCPD method is applied. The most expensive region exceeds that average by 21.8%. The spread between the most expensive and the cheapest region is (121.8 - 89.2)/89.2 = 36.5%. These numbers are more pronounced for the CPD method, resulting in a regional price spread of 61.9%. The GEKS index generates an even larger spread: 64.5%. For all three methods, the unweighted mean is below the population-weighted mean, indicating that a region's price level tends to increase with its population.

TABLE 4 Price index numbers in relation to their population-weighted average (= 100).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd.
NLCPD	89.2	94.3	96.9	98.0	100.4	121.8	5.2
CPD	82.2	92.5	95.9	97.5	101.2	133.1	7.6
GEKS	82.0	91.4	95.1	97.1	101.0	134.9	8.3





The spatial pattern of the price index numbers of the 401 German regions is depicted in Figure 3. As expected, the price level dispersion estimated by the CPD method is much larger than that estimated by the NLCPD method. The seven biggest cities in Germany all exhibit price index numbers above the population-weighted average. The NLCPD method ranks Munich as the most expensive region. Its price level is 21.8% above the population-weighted average. The numbers for Stuttgart and Frankfurt are 14.7%, Hamburg 12.1%, Cologne 9.2%, Dusseldorf 7.1%, and Berlin 5.6%. In the CPD method and the GEKS index, the same ranking of the seven cities arises.

7 | Concluding Remarks

Spatial price comparisons often suffer from incomplete price data. To deal with such situations, Summers (1973) introduced the CPD method. This regression approach provides estimates of the regional price levels along with their standard errors.

The present paper focused primarily on situations in which data gaps occur and the spatial price variation of the individual products is different. It was shown that in such situations the CPD estimates of the standard errors are biased. Even worse, the estimates of the regional price levels are biased, unless the data gaps are completely at random.

As a solution, this paper introduced the NLCPD method, a non-linear generalization of the CPD method. To properly address product-specific spatial price variation, the NLCPD method adds parameters to the CPD model. The new parameters represent the price level elasticities of the individual products. The estimates of these new parameters indicate whether the implicit CPD assumption of uniform elasticities would have been reasonable.

In a simulation, the superiority of the NLCPD method over the CPD method was shown. Furthermore, the GEKS index performed even worse than the CPD method.

In the theoretical parts as well as in the simulation, it was irrelevant whether the spatial price comparison is subnational or international. Thus, the issues of the CPD method and the GEKS index are likely to be relevant also in international price comparisons such as the International Comparison Program.

Finally, in a price level comparison of the 401 German districts, the practical applicability of the NLCPD method was demonstrated. In the available data set, the data gaps were negatively correlated with the products' price level elasticities. Therefore, the CPD method and the GEKS index overestimated the spread in the price levels of the German districts.

The CPD method's drawbacks in the presence of product-specific elasticities are of relevance not only for spatial but also for intertemporal price comparisons. Here, the dummy variables of the regions in the CPD regression model are replaced with dummy variables for the time of price collection. Therefore, this regression is denoted as the Time-Product-Dummy (TPD) method (e.g., de Haan et al. 2021). The TPD method provides estimates of the price level change over time. While the prices of some products decline or remain constant over time, other prices increase, and some products exhibit seasonal patterns in their price movements. Thus, it is unlikely that the variation of prices over time is uniform across products. Consequently, the TPD method has the same statistical issues as the CPD method. By contrast, the non-linear TPD (NLTPD) method estimates for each individual product the elasticity of its prices with respect to the general price level. High elasticities indicate inflation-sensitive products. Ongoing research is examining the NLTPD method in this intertemporal context (Auer and Weinand 2024).

Acknowledgments

We are indebted to the Research Data Center of the Federal Statistical Office and Statistical Offices of the Länder for granting us access to the consumer price index micro data of May 2019. We also wish to express our gratitude to Alexander Schürt and Rolf Müller from the Federal Office for Building and Regional Planning (BBSR) for providing us with the results of their rent data sample from 2019. We gratefully acknowledge comments from participants of a research seminar at Deutsche Bundesbank, of the conference "Statistische Woche 2022" in Münster, and of the workshop "Messung der Preise" in Berlin. The paper has greatly benefited from the generous advice we received from Gholamreza Hajargasht and Mark Trede as well as from the editor and one anonymous referee. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Deutsche Bundesbank, the Eurosystem, or the European Commission. Open Access funding enabled and organized by Projekt DEAL.

Endnotes

¹ The derivation of Equation (2) is described in Footnote 9.

- ² When expenditure shares or other indicators of the products' importance are available, it is recommended that weighted least squares be used instead (e.g., Clements and Izan 1981, 745-746; Selvanathan and Rao 1992, 338-339; Diewert 2005, 562-563; Rao 2005, 574-575; Hajargasht and Rao 2010, p. S39).
- ³ Dividing the price levels in expression (3) by 0.74 renormalizes them such that the price levels of regions B, C, and D are measured relative to region A's price level $\hat{P}^{A} = 1$. Of course, the price level ratios (e.g., \hat{P}^{B}/\hat{P}^{C}) would not be affected by such a renormalization.
- ⁴ We deliberately speak of "price level elasticity" instead of "price elasticity" because the name of an elasticity usually relates to the variable that *causes* the change in the other variable. In our context, a change in the regional price level causes a change in the product price.
- ⁵ The derivation of Equation (5) is described in Footnote 9. The CPD model (1) implicitly assumes that all δ_i -values are equal to one. Then, Equation (5) simplifies to Equation (2).
- ⁶ When all $δ_i$ -values are close to 1, the NLCPD method "waists" information for estimating these parameters. Then, the sum of squared residuals divided by the degrees of freedom may be larger for the NLCPD method than for the CPD method.
- 7 The weights, w_i , are "democratic" in the sense that regions with large total expenditures are not more influential than regions with small total expenditures.
- ⁸ If only two regions are involved (R = 2) in the weighted NLCPD price level comparison, the weights w_i are defined as $w_i = (w_i^r + w_i^s)/2$. Then, the weighted Jevons index becomes the Törnqvist index (Diewert 1995, 11-12; 2005, 564-565). If instead of an arithmetic mean of the expenditure shares w_i^r and w_i^s , a geometric mean were used, the index formula would turn into the Walsh-Vartia index. With a logarithmic mean, the Sato-Vartia index would result.
- ⁹ The parameters $\ln P^r$, $\ln \pi_i$, and δ_i in Equation (4) are unknown constants. Furthermore, the model assumes that $\mathbb{E}(u_i^r) = 0$. Thus, to derive

Equation (10), we take expectations on both sides of Equation (4) and multiply by w_i . Next, we solve for $\delta_i \ln P^r$ and sum over all products in \mathcal{N} . We factor out $\ln P^r$ and divide by $\sum_{i \in \mathcal{N}} w_i \delta_i$. Afterwards, we derive the same equation for region *s*. Subtracting the latter equation from the former, yields Equation (10). Equations (2) and (5) are special cases of Equation (10). Replacing in Equation (10) the weights w_i by 1/N gives Equation (5). The additional assumption $\delta_i = 1$ leads to Equation (2).

¹⁰ To express restriction (11) in terms of region-specific expenditure weights, w_i^r , one merely has to substitute w_i with the right-hand side of Equation (9).

- ¹¹ In the words of Gorajek (2022, p. 86), "... the expenditure shares in the weights are endogenous; they are functions of prices and thus the error terms in the price equation." We agree with this statement but add that, in practice, the expenditure shares and, thus, their underlying prices often refer to an earlier period than the one under consideration.
- ¹² Conducting the simulation without expenditure share weights, w_i , and applying the unweighted NLCPD and CPD methods as well as the GEKS-Jevons index does not change any of the results presented in the next section.
- ¹³ In Online Appendix C, mean absolute bias and mean RMSE are also reported for the estimates of $\ln \pi_i$ and δ_i , respectively.
- ¹⁴ The price index numbers for the German districts are available in the online supplemental material.
- ¹⁵ For example, prices of package holidays are collected from a big sample (e.g., Egner 2019, p. 97). However, this sample of prices is already aggregated by the Federal Statistical Office into a single index number in the micro data set.
- ¹⁶ For example, the priced items of the basic heading "gloves" were not identical and, therefore, not comparable.
- ¹⁷The data were downloaded from https://creativecommons. tankerkoenig.de/ where historical fuel prices are provided on a daily basis.
- ¹⁸ This correlation is 0.96 for the $\ln \pi_i$ -estimates of the two methods. For the CPD method, the $\ln \pi_i$ -estimates range from -1.33 to 0.64, while this range is -1.33 to 0.90 for the NLCPD method.

References

Aten, B. H. 2017. "Regional Price Parities and Real Regional Income for the United States." *Social Indicators Research* 131, no. 1: 123–143.

Auer, L. v., and S. Weinand. 2024. "Multilateral Approaches in Inflation Measurement: Why Does the TPD Method Fail Us and Can We Do Something About It? 18th Meeting of the Ottawa Group." https://stats.unece. org/ottawagroup/meeting/18.

Caves, D. W., L. R. Christensen, and W. E. Diewert. 1982. "Multilateral Comparisons of Output, Input, and Productivity Using Superlative Index Numbers." *Economic Journal* 92, no. 365: 73–86.

Clements, K. W., and H. Y. Izan. 1981. "A Note on Estimating Divisia Index Numbers." *International Economic Review* 22, no. 3: 745–747.

Clements, K. W., and H. Y. Izan. 1987. "The Measurement of Inflation: A Stochastic Approach." *Journal of Business & Economic Statistics* 5, no. 3: 339–350.

Clements, K. W., H. Y. Izan, and E. A. Selvanathan. 2006. "Stochastic Index Numbers: A Review." *International Statistical Review* 74, no. 2: 235–270.

Crompton, P. 2000. "Extending the Stochastic Approach to Index Numbers." *Applied Economics Letters* 7, no. 6: 367–371.

Cuthbert, J. R., and M. Cuthbert. 1988. "On Aggregation Methods of Purchasing Power Parities." In *OECD Economics Department Working Paper* 56. OECD Publishing. de Haan, J., R. Hendriks, and M. Scholz. 2021. "Price Measurement Using Scanner Data: Time-Product Dummy Versus Time Dummy Hedonic Indexes." *Review of Income and Wealth* 67, no. 2: 394–417.

Diewert, W. E. 1995. Axiomatic and Economic Approaches to Elementary Price Indexes. Working Paper 5104. National Bureau of Economic Research.

Diewert, W. E. 2004. *On the Stochastic Approach to Linking the Regions in the ICP*. Discussion Paper 04/16. University of British Columbia.

Diewert, W. E. 2005. "Weighted Country Product Dummy Variable Regressions and Index Number Formulae." *Review of Income and Wealth* 51, no. 4: 561–570.

Egner, U. 2019. "Verbraucherpreisstatistik auf neuer Basis 2015." In *Wirtschaft und Statistik*, vol. 5, 86–106. Statistisches Bundesamt.

Eltetö, Ö., and P. Köves. 1964. "On a Problem of Index Number Computation Relating to International Comparison." *Statisztikai Szemle* 42: 507–518.

Gini, C. 1924. "Quelques Considerations au Sujet de la Construction des Nombres Indices des Prix et des Questions Analogues." *Mentron* 4, no. 1: 3–162.

Gini, C. 1931. "On the Circular Test of Index Numbers." *International Statistical Review* 9, no. 2: 3–25.

Gorajek, A. 2022. "Generalizing the Stochastic Approach to Price Indexes." *Review of Income and Wealth* 70, no. 1: 80–101.

Hajargasht, G., and D. S. P. Rao. 2010. "Stochastic Approach to Index Numbers for Multilateral Price Comparisons and their Standard Errors." *Review of Income and Wealth* 56, no. s1: S32–S58.

Jevons, W. S. 1865. "On the Variation of Prices and the Value of the Currency since 1782." *Journal of the Statistical Society of London* 28, no. 2: 294–320.

Majumder, A., and R. Ray. 2020. "National and Subnational Purchasing Power Parity: A Review." *Decision* 47, no. 2: 103–124.

Rao, D. S. P. 2004. The Country-Product-Dummy Method: A Stochastic Approach to the Computation of Purchasing Power Parities in the ICP. Working Paper 03/2004. Centre for Efficiency and Productivity Analysis.

Rao, D. S. P. 2005. "On the Equivalence of Weighted Country-Product-Dummy (CPD) Method and the Rao-System for Multilateral Price Comparisons." *Review of Income and Wealth* 51, no. 4: 571–580.

Rao, D. S. P., and K. S. Banerjee. 1986. "A Multilateral Index Number System Based on the Factorial Approach." *Statistische Hefte* 27, no. 1: 297–313.

Rao, D. S. P., and G. Hajargasht. 2016. "Stochastic Approach to Computation of Purchasing Power Parities in the International Comparison Program (ICP)." *Journal of Econometrics* 191, no. 2: 414–425.

Rokicki, B., and G. J. D. Hewings. 2019. "Regional Price Deflators in Poland: Evidence from NUTS-2 and NUTS-3 Regions." *Spatial Economic Analysis* 14, no. 1: 88–105.

Selvanathan, E. A., and D. S. P. Rao. 1992. "An Econometric Approach to the Construction of Generalized Theil-Tornqvist Indices for Multilateral Comparisons." *Journal of Econometrics* 54, no. 1: 335–346.

Selvanathan, E. A., and D. S. P. Rao. 1994. *Index Numbers: A Stochastic Approach*, 47–73. Palgrave Macmillan.

Summers, R. 1973. "International Price Comparisons Based upon Incomplete Data." *Review of Income and Wealth* 19, no. 1: 1–16.

Szulc, B. J. 1964. "Indices for Multiregional Comparisons." *Przeglad Statystyczny* 3: 239–254.

Tabuchi, T. 2001. "On Interregional Price Differentials." *Japanese Economic Review* 52, no. 1: 104–115.

Weinand, S. 2022. "Measuring Spatial Price Differentials at the Basic Heading Level: A Comparison of Stochastic Index Number Methods." *AStA Advances in Statistical Analysis* 106, no. 1: 117–143.

Weinand, S., and L. v. Auer. 2019. *Anatomy of Regional Price Differentials: Evidence from Micro Price Data*. Discussion Paper 2019/04. Deutsche Bundesbank.

Weinand, S., and L. v. Auer. 2020. "Anatomy of Regional Price Differentials: Evidence from Micro-Price Data." *Spatial Economic Analysis* 15, no. 4: 413–440.

World Bank. 2013. Measuring the Real Size of the World Economy: The Framework, Methodology, and Results of the International Comparison Program. World Bank.

World Bank. 2020. Purchasing Power Parities and the Real Size of World Economies: Results From the 2017 International Comparison Program. World Bank.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.